

# WHYACT: Identifying Action Reasons in Lifestyle Vlogs

Oana Ignat and Santiago Castro and Hanwen Miao and Weiji Li and Rada Mihalcea

Computer Science and Engineering

University of Michigan

{oignat, sacastro, hwmiao, weijili, mihalcea}@umich.edu

## Abstract

We aim to automatically identify human action reasons in online videos. We focus on the widespread genre of lifestyle vlogs, in which people perform actions while verbally describing them. We introduce and make publicly available the WHYACT dataset, consisting of 1,077 visual actions manually annotated with their reasons. We describe a multimodal model that leverages visual and textual information to automatically infer the reasons corresponding to an action presented in the video.

## 1 Introduction

Significant research effort has been recently devoted to the task of action recognition (Carreira and Zisserman, 2017; Shou et al., 2017; Tran et al., 2018; Chao et al., 2018; Girdhar et al., 2019; Feichtenhofer et al., 2019). Action recognition works well when applied to well defined/constrained scenarios, such as people following scripts and instructions (Sigurdsson et al., 2016; Miech et al., 2019; Tang et al., 2019), performing sports (Soomro et al., 2012; Karpathy et al., 2014) or cooking (Rohrbach et al., 2012; Damen et al., 2018, 2020; Zhou et al., 2018). At the same time however, action recognition is limited and error-prone once the application space is opened to everyday life. This indicates that current action recognition systems rely mostly on pattern memorization and do not effectively understand the action, which makes them fragile and unable to adapt to new settings (Sigurdsson et al., 2017; Kong and Fu, 2018). Research on how to improve action recognition in videos (Sigurdsson et al., 2017) shows that recognition systems for actions with known intent have a significant increase in performance, as knowing the reason for performing an action is an important step for understanding that action (Tosi, 1991; Gilovich et al., 2002).

In contrast to action recognition, action causal reasoning research is just emerging in computational applications (Vondrick et al., 2016; Yeo et al.,

2018; Zhang et al., 2021; Fang et al., 2020). Causal reasoning has direct applications on many real-life settings, for instance to understand the consequences of events (e.g., if “there is clutter,” “cleaning” is required), or to enable social reasoning (e.g., when “guests are expected,” “cleaning” may be needed – see Figure 1). Most of the work to date on causal systems has relied on the use of semantic parsers to identify reasons (He et al., 2017), however this approach does not work well on more realistic every-day settings. As an example, consider the statement “This is a mess and my friends are coming over. I need to start cleaning.” Current causal systems are unable to identify “this is a mess” and “friends are coming over” as reasons, and are thus failing to use them as context for understanding the action of “cleaning.”

In this paper, we propose the task of multimodal action reason identification in everyday life scenarios. We collect a dataset of lifestyle vlogs from YouTube that reflect daily scenarios and are currently very challenging for systems to solve. Vloggers freely express themselves while performing most common everyday activities such as cleaning, eating, cooking, writing and others. Lifestyle vlogs present a person’s everyday routine: the vlogger visually records the activities they perform during a normal day and verbally express their intentions and feelings about those activities. Because of these characteristics, lifestyle vlogs are a rich data source for an in depth study of human actions and the reasons behind them.

The paper makes four main contributions. First, we formalize the new task of multimodal action reason identification in online vlogs. Second, we introduce a new dataset, WHYACT, consisting of 1,077 (action, context, reasons) tuples manually labeled in online vlogs, covering 24 actions and their reasons drawn from ConceptNet as well as crowdsourcing contributions. Third, we propose several models to solve the task of human action reason



Figure 1: Overview of our task: automatic identification of action reasons in online videos. The reasons for *cleaning* change based on the visual and textual (video transcript) context. The videos are selected from YouTube, and the actions together with their reasons are obtained from the ConceptNet (Speer et al., 2017) knowledge graph which we supplement with crowdsourced reasons. The figure shows two examples from our WHYACT dataset.

identification, consisting of single-modalities models based on the visual content and vlog transcripts, as well as a multimodal model using a fill-in-the-blanks strategy. Finally, we also present an analysis of our new dataset, which leads to rich avenues for future work for improving the tasks of reason identification and ultimately action recognition in online videos.

## 2 Related Work

There are three areas of research related to our work: identifying action motivation, commonsense knowledge acquisition, and web supervision.

**Identifying Action Motivation.** The research most closely related to our paper is the work that introduced the task of predicting motivations of actions by leveraging text (Vondrick et al., 2016). Their method was applied to images from the COCO dataset (Lin et al., 2014), while ours is focused on videos from YouTube. Other work on human action causality in the visual domain (Yeo et al., 2018; Zhang et al., 2021) relies on object detection and automatic image captioning as a way to represent videos and analyze visual causal relations. Research has also been carried out on detecting the intentions of human actions (Pezzelle et al., 2020); the task definition differs from ours, however, as their goal is to automatically choose the correct action for a given image and intention. Other related work includes (Synakowski et al., 2020), a vision-based classification model between intentional and non-intentional actions and Intentionomy (Jia et al.,

2021), a dataset on human intent behind images on Instagram.

**Commonsense Knowledge Acquisition.** Research on commonsense knowledge often relies on textual knowledge bases such as ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019), COMET-ATOMIC 2020 (Hwang et al., 2021), and more recently GLUCOSE (Mostafazadeh et al., 2020).

Recently, several of these textual knowledge bases have also been used for visual applications, to create more complex multimodal datasets and models (Park et al., 2020; Fang et al., 2020; Song et al., 2021). VisualCOMET (Park et al., 2020) is a dataset for visual commonsense reasoning tasks to predict events that might have happened before a given event, events that might happen next, as well as people intents at a given point in time. Their dataset is built on top of VCR (Zellers et al., 2019), which consists of images of multiple people and activities. Video2Commonsense (Fang et al., 2020) uses ATOMIC to extract from an input video a list of intentions that are provided as input to a system that generates video captions, as well as three types of commonsense descriptions (intention, effect, attribute). KVL-BERT (Song et al., 2021) proposes a knowledge enhanced cross-modal BERT model by introducing entities extracted from ConceptNet (Speer et al., 2017) into the input sentences, followed by testing their visual question answering model on the VCR benchmark (Zellers et al., 2019). Unlike previous work that broadly addresses com-

nonsense relations, we focus on the extraction and analysis of action reasons, which allows us to gain deeper insights for this relation type.

**Webly-Supervised Learning.** The space of current commonsense inference systems is often limited to one dataset at a time, e.g., COCO (Lin et al., 2014), VCR (Zellers et al., 2019), MSR-VTT (Xu et al., 2016). In our work, we ask commonsense questions in the context of rich, unlimited, constantly evolving online videos from YouTube.

Previous work has leveraged webly-labeled data for the purpose of identifying commonsense knowledge. One of the most extensive efforts is NELL (Never Ending Language Learner) (Mitchell et al., 2015), a system that learns everyday knowledge by crawling the web, reading documents and analysing their linguistic patterns. A closely related effort is NEIL (Never Ending Image Learner), which learns commonsense knowledge from images on the web (Chen et al., 2013). Large scale video datasets (Miech et al., 2019) on instructional videos and lifestyle vlogs (Fouhey et al., 2018; Ignat et al., 2019) are other examples of web supervision. The latter are similar to our work as they analyse online vlogs, but unlike our work, their focus is on action detection and not on the reasons behind actions.

### 3 Data Collection and Annotation

In order to develop and test models for recognizing reasons for human actions in videos, we need a manually annotated dataset. This section describes the WHYACT dataset of action reasons.

#### 3.1 Data Collection

We start by compiling a set of lifestyle videos from YouTube, consisting of people performing their daily routine activities, such as cleaning, cooking, studying, relaxing, and others. We build a data gathering pipeline to automatically extract and filter videos and their transcripts.

We select five YouTube channels and download all the videos and their transcripts. The channels are selected to have good quality videos with automatically generated transcripts containing detailed verbal descriptions of the actions depicted. An analysis of the videos indicates that both the textual and visual information are rich sources for describing not only the actions, but why the actions in the videos are undertaken (action reasons). We present qualitative and quantitative analyses of our data in section 6.

|                                       |       |
|---------------------------------------|-------|
| Initial                               | 9,759 |
| Actions with reasons in ConceptNet    | 139   |
| Actions with at least 3 reasons in CN | 102   |
| Actions with at least 25 video-clips  | 25    |

Table 1: Statistics for number of collected actions at each stage of data filtering.

We also collect a set of human actions and their reasons from ConceptNet (Speer et al., 2017). Actions include verbs such as: *clean*, *write*, *eat*, and other verbs describing everyday activities. The actions are selected based on how many reasons are provided in ConceptNet and how likely they are to appear in our collected videos. For example, the action of *cleaning* is likely to appear in the vlog data, while the action of *yawning* is not.

#### 3.2 Data Pre-processing

After collecting the videos, actions and their corresponding reasons, the following data pre-processing steps are applied.

**Action and Reason Filtering.** From ConceptNet, we select actions that contain at least three reasons. The reasons in ConceptNet are marked by the “motivated by” relation. We further filter out those actions that appear less than 25 times in our video dataset, in order to assure that each action has a significant number of instances.

We find that the reasons from ConceptNet are often very similar to each other, and thus easy to confound. For example, the reasons for the action *clean* are: “dirty”, “remove dirt”, “don’t like dirtiness”, “there dust”, “dirtiness unpleasant”, “dirt can make ill”, “things cleaner”, “messy”, “company was coming”. To address this issue, we apply agglomerative clustering (Murtagh and Legendre, 2014) to group similar actions together. For instance, for the action *clean*, the following clusters are produced: [“dirty”, “remove dirt”, “there dust”, “things cleaner”], [“don’t like dirtiness”, “dirtiness unpleasant”, “dirt can make ill”], [“messy”], [“company was coming”]. Next, we manually select the most representative and clear reason from each cluster. We also correct any spelling mistakes and rename the reasons that are either too general or unclear (e.g., we rename “messy” to “declutter”). Finally, after the clustering and processing steps, we filter out all the actions that contain less than three reasons.

We show the statistics before and after the additive filtering steps in Table 1.

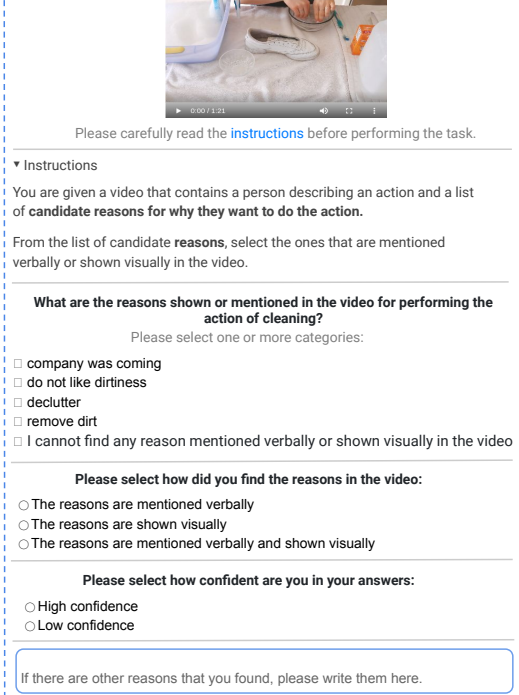
**Transcript Filtering.** We want transcripts that reflect the reasons for performing one or more actions shown in the video. However, the majority of the transcripts contain mainly verbal descriptions of the action, which are not always helpful in determining their reason. We therefore implement a method to select candidate transcript sequences that contain at least one causal relation related to the actions shown in the video.

We start by automatically splitting the transcripts into sentences using spaCy (Honnibal et al., 2020). Next, we select the sentences with at least one action from the final list of actions we collected from ConceptNet (see the previous section). For each selected sentence, we also collect its context consisting of the sentences before and after. We do this in order to increase the search space for the reasons for the actions mentioned in the selected sentences.

We want to keep the sentences that contain action reasons. We tried multiple methods to automatically determine the sentences more likely to include causal relations using Semantic Role Labeling (SRL) (Ouchi et al., 2018), Open Information Extraction (OpenIE) (Angeli et al., 2015) and searching for causal markers. We found that SRL and OpenIE do not work well on our data, likely due to the fact that the transcripts are more noisy than the datasets these models were trained on. Most of the language in the transcripts does not follow simple patterns such as “I clean because it is dirty.” Instead, the language consists of natural everyday speech such as “Look at how dirty this is, I think I should clean it.”

We find that a strategy sufficient for our purposes is to search for causal markers such as “because”, “since”, “so that is why”, “thus”, “therefore” in the sentence and the context, and constrain the distance between the actions and the markers to be less than 15 words – a threshold identified on development data. We thus keep all the transcript sentences and their context that contain at least one action and a causal marker within a distance of less than the threshold of 15 words.

**Video Filtering.** As transcripts are temporally aligned with videos, we can obtain meaningful video clips related to the narration. We extract video clips corresponding to the sentences selected



Please carefully read the [instructions](#) before performing the task.

▼ Instructions

You are given a video that contains a person describing an action and a list of candidate reasons for why they want to do the action.

From the list of candidate reasons, select the ones that are mentioned verbally or shown visually in the video.

**What are the reasons shown or mentioned in the video for performing the action of cleaning?**

Please select one or more categories:

- ☐ company was coming
- ☐ do not like dirtiness
- ☐ declutter
- ☐ remove dirt
- ☐ I cannot find any reason mentioned verbally or shown visually in the video

**Please select how did you find the reasons in the video:**

- ☐ The reasons are mentioned verbally
- ☐ The reasons are shown visually
- ☐ The reasons are mentioned verbally and shown visually

**Please select how confident are you in your answers:**

- ☐ High confidence
- ☐ Low confidence

If there are other reasons that you found, please write them here.

Figure 2: Instructions for the annotators.

from transcripts (described in the section above).

We want video clips that show why the actions are being performed. Although there can be many actions along with reasons in the transcript, if they are not depicted in the video, we cannot leverage the video information in our task. Videos with low movement tend to show people sitting in front of the camera, describing their routine, but not performing the action they are talking about. We therefore remove clips that do not contain enough movement. We sample one out of every one hundred frames of the clip, and compute the 2D correlation coefficient between these sampled frames. If the median of the obtained values is greater than a certain threshold (0.8, selected on the development data), we filter out the clip. We also remove video-clips that are shorter than 10 seconds and longer than 3 minutes.

### 3.3 Data Annotation

The resulting (video clip, action, reasons) tuples are annotated with the help of Amazon Mechanical Turk (AMT) workers. They are asked to identify: (1) what are the reasons shown or mentioned in the video clip for performing a given action; (2) how are the reasons identified in the video: are they mentioned verbally, shown visually, or both; (3) whether there are other reasons other than the ones provided; (4) how confident the annotator is



|                  |         |
|------------------|---------|
| Video-clips      | 1,077   |
| Video hours      | 107.3   |
| Transcript words | 109,711 |
| Actions          | 24      |
| Reasons          | 166     |

Table 2: Data statistics.

|             | Test | Development |
|-------------|------|-------------|
| Actions     | 24   | 24          |
| Reasons     | 166  | 166         |
| Video-clips | 853  | 224         |

Table 3: Statistics for the experimental data split. The methods we run are unsupervised with fine-tuning on development set.

in their response. The guidelines and interface for annotations are shown in Figure 2. In addition to the guidelines, we also provide the annotators with a series of examples of completed assignments with explanations for why the answers were selected. We present them in the supplemental material in Figure 6.

We add new action reasons from the ones added by the annotators if they repeat at least three times in the collected answers and are not similar to the ones already existing.

Each assignment is completed by three different master annotators. We compute the agreement between the annotators using Fleiss Kappa (Fleiss, 1971) and we obtain 0.6, which indicates a moderate agreement. Because the annotators can select multiple reasons, the agreement is computed per reason and then averaged.

We also analyse how confident the workers are in their answers: for each video, we take the confidence selected by the majority of workers: out of 1,077 videos, in 890 videos the majority of workers are highly confident.

Table 2 shows statistics for our final dataset of video-clips and actions annotated with their reasons. Figure 1 shows a sample video and transcript, with annotations. Additional examples of annotated actions and their reasons can be seen in the supplemental material in Figure 8.

## 4 Identifying Causal Relations in Vlogs

Given a video, an action, and a list of candidate action reasons, our goal is to determine the reasons mentioned or shown in the video. We develop a multimodal model that leverages both visual and

textual information, and we compare its performance with several single-modality baselines.

The models we develop are unsupervised in that we are not learning any task-specific information from a training dataset. We use a validation set only to tune the hyper-parameters of the models.

### 4.1 Data Processing and Representation

**Textual Representations.** To represent the textual data – transcripts and candidate reasons – we use sentence embeddings computed using the pre-trained model Sentence-BERT (Reimers and Gurevych, 2019).

**Video Representations.** In order to tie together the causal relations, both the textual, and the visual information, we represent the video as a bag of object labels and a collection of video captions. For object detection we use Detectron2 (Wu et al., 2019), a state-of-the-art object detection algorithm.

We generate automatic captions for the videos using a state-of-the-art dense captioning model (Iashin and Rahtu, 2020). The input to the model are visual features extracted from I3D model pre-trained on Kinetics (Carreira and Zisserman, 2017), audio features extracted with VGGish model (Hershey et al., 2017) pre-trained on YouTube-8M (Abu-El-Haija et al., 2016) and caption tokens using GloVe (Pennington et al., 2014).

### 4.2 Baselines

Using the representations described in Section 4.1, we implement several textual and visual models.

#### 4.2.1 Textual Similarity

Given an action, a video transcript associated with the action, and a list of the candidate action reasons, we compute the cosine similarity between the textual representations of the transcript and all the candidate reasons. We predict as correct those reasons that have a cosine similarity with the transcript greater than a threshold of 0.1. The threshold is fine-tuned on development data.

Because the transcript might contain information that may be unrelated to the action described or its reasons, we also develop a second version of this baseline. When computing the similarity, instead of using the whole transcript, we select only the part of the transcript that is in the vicinity of the causal markers (before and after a fixed number words, fine-tuned on development data).

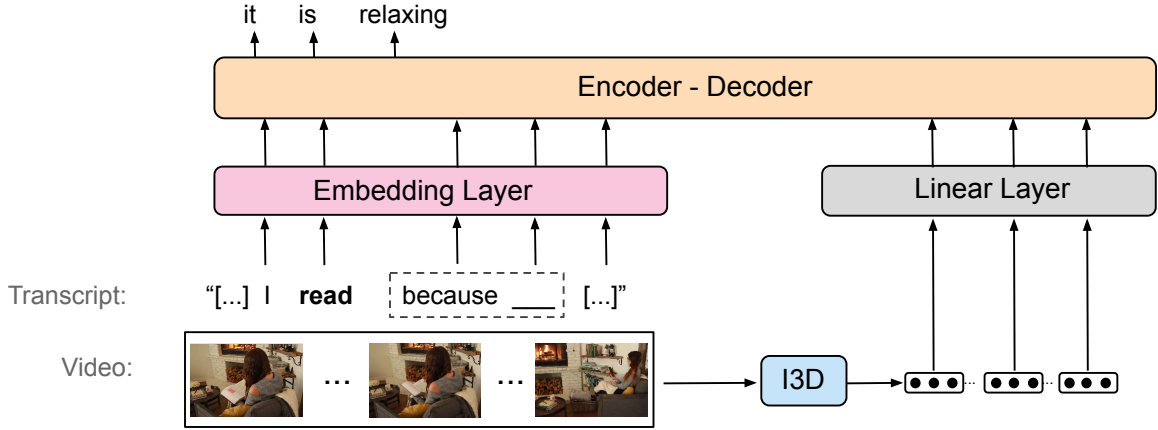


Figure 3: Overview architecture of our Multimodal Fill-in-the-blanks model. The span of text “because \_\_\_\_” is introduced in the video transcript, after the appearance of the action. This forces the T5 model to generate the words missing in the blanks. We then compute the probability of each potential reason and take as positive those that pass a threshold.

#### 4.2.2 Natural Language Inference (NLI)

We use a pre-trained NLI model (Yin et al., 2019) as a zero-shot sequence classifier. The NLI model is pre-trained on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018), a collection of sentence pairs annotated with textual entailment information.

The method works by posing the sequence to be classified as the NLI premise and constructing a hypothesis from each candidate label: given the transcript as a premise and the list of reasons as the hypotheses, each reason will receive a score that reflects the probability of entailment. For example, if we want to evaluate whether the label “declutter” is a reason for the action “cleaning”, we construct the hypothesis “The reason for cleaning is declutter.”

We use a threshold of 0.8 fine-tuned on the development data to filter the reasons that have a high entailment score with the transcript.

**Bag of Objects.** We replace the transcript in the premise with a list of object labels detected from the video. The objects are detected using the Detectron2 model (Wu et al., 2019) on each video frame, at 1fps. We select only the objects that pass a confidence score of 0.7.

**Automatic Video Captioning.** We replace the transcript in the premise with a list of video captions detected using the Bi-modal Transformer for Dense Video Captioning model (Iashin and Rahtu, 2020). The video captioning model generates captions for several time slots. We further filter the generated captions to remove redundant captions: if a time slot is heavily overlapped or even covered

by another time slot, we only keep the caption of the longer time slot. We find that captions of longer time slots are also more informative and accurate compared to captions of shorter time slots.

#### 4.3 Multimodal Model

To leverage information from both the visual and linguistic modalities, we propose a new model that recasts our task as a Cloze task, and attempts to identify the action reasons by performing a fill-in-the-blanks prediction, similarly to Castro et al. (2021) that proposes to fill blanks corresponding to noun phrases in descriptions based on video clips content. Specifically, after each action mention for which we want to identify the reason, we add the text “because \_\_\_\_”. For instance, “I clean the windows” is replaced by “I clean the windows because \_\_\_\_”. We train a language model to compute the likelihood of filling in the blank with each of the candidate reasons. For this purpose, we use T5 (Raffel et al., 2020), an encoder-decoder transformer (Vaswani et al., 2017) pre-trained model, to fill in blanks with text.

To incorporate the visual data, we first obtain Kinetics-pre-trained I3D (Carreira and Zisserman, 2017) RGB features at 25fps (the average pooling layer). We input the features to the T5 encoder after the transcript text tokens. The text input is passed through an embedding layer (as in T5), while the video features are passed through a linear layer. Since T5 was not trained with this kind of input, we fine-tune it on unlabeled data from the same source, without including data that contains the causal marker “because”. Note this also helps the

model specialize on filling-in-the-blank with reasons. Finally, we fine-tune the model on the development data. We obtain the reasons for an action by computing the likelihood of the potential ones and taking the ones that pass a threshold selected based on the development data. The model architecture is shown in Figure 3.

We also use our fill-in-the-blanks model in a single modality mode, where we apply it only on the transcript.

## 5 Evaluation

We consider as gold standard the labels selected by the majority of workers (at least two out of three workers).

For our experiments, we split the data across video-clips: 20% development and 80% test (see Table 3 for a breakdown of actions, reasons and video-clips in each set). We evaluate our systems as follows. For each action and corresponding video-clip, we compute the Accuracy, Precision, Recall and F1 scores between the gold standard and predicted labels. We then compute the average of the scores across actions. Because the annotated data is unbalanced (in average, 2 out of 6 candidate reasons per instance are selected as gold standard), the most representative metric is F1 score. The average results are shown in Table 4. The results also vary by action: the F1 scores for each action, of the best performing method, are shown in the supplemental material in Figure 12.

Experiments on WHYACT reveal that both textual and visual modalities contribute to solving the task. The results demonstrate that the task is challenging and there is room for improvement for future work models.

Selecting the most frequent reason for each action on test data achieves on average an F1 of 40.64, with a wide variation ranging from a very low F1 for the action “writing” (7.66 F1) to a high F1 for the action “cleaning” (55.42 F1). Note however that the “most frequent reason” model makes use of data distributions that our models do not use (because our models are not trained). Furthermore, we believe that it is expected that for certain actions the distribution of reasons is unbalanced, as in everyday life there are action reasons much more common than others (e.g. for “cleaning”, “remove dirt” is a more common/frequent reason than “company was coming”).

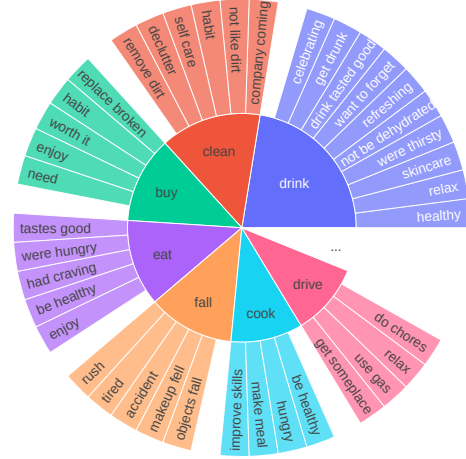


Figure 4: Distribution of the first seven actions, in alphabetical order, and their reasons, in our dataset. The rest of the actions and their reasons are shown in the appendix, in Figure 7.

## 6 Data Analysis

We perform an analysis of the actions, reasons and video-clips in the WHYACT dataset. The distribution of actions and their reasons are shown in Figure 4. The supplemental material includes additional analyses: the distribution of actions and their number of reasons (Figure 11) and videos (Figure 10) and the distribution of actions and their worker agreement scores (Figure 9).

We also explore the content of the videos by analysing their transcripts. In particular, we look at the actions and their direct objects. For example, the action clean is depicted in various ways in the videos: “clean shower”, “clean body”, “clean makeup”, “clean dishes”. The action diversity assures that the task is challenging and complex, trying to cover the full spectrum of everyday activities. In Figure 5 we show what kind of actions are depicted in the videos: we extract all the verbs and their most five most frequent direct objects using spaCy (Honnibal et al., 2020) and then we cluster them by verb and plot them using t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008).

Finally, we analyse what kind of information is required for detecting the action reasons: what is verbally described, visually shown in the video or the combination of visual and verbal cues. For this, we analyse the worker’s justifications for selecting the action reasons: if the reasons were verbally mentioned in the video, visually shown or both. For each video, we take the justification selected by the majority of workers. We find that the rea-

| Method                   | Input                                | Accuracy     | Precision    | Recall       | F1           |
|--------------------------|--------------------------------------|--------------|--------------|--------------|--------------|
| BASELINES                |                                      |              |              |              |              |
| Cosine similarity        | Transcript                           | 57.70        | 31.39        | 55.94        | 37.64        |
|                          | Causal relations from transcript     | 50.85        | 30.40        | 68.91        | 39.73        |
| SINGLE MODALITY MODELS   |                                      |              |              |              |              |
| Natural                  | Transcript                           | <b>68.41</b> | <b>41.90</b> | 48.01        | 40.78        |
| Language Inference       | Video object labels                  | 54.49        | 31.70        | 59.93        | 36.79        |
|                          | Video dense captions                 | 49.18        | 29.54        | 68.47        | 37.40        |
|                          | Video object labels & dense captions | 36.93        | 27.34        | 87.97        | 39.11        |
| Fill-in-the-blanks       | Transcript                           | 44.04        | 30.70        | 87.10        | <b>43.59</b> |
| MULTIMODAL NEURAL MODELS |                                      |              |              |              |              |
| Fill-in-the-blanks       | Video & Transcript                   | 32.6         | 27.56        | <b>94.76</b> | 41.11        |

Table 4: Results from our models on test data.

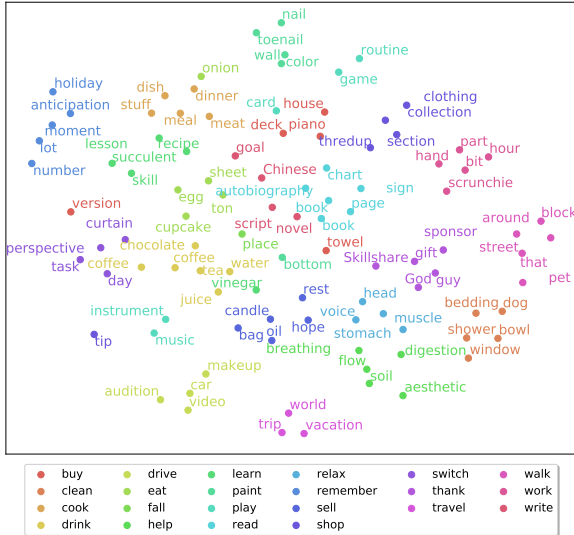


Figure 5: The t-SNE representation of the five most frequent direct objects for each action/verb in our dataset. Each color represents a different action.

sons for the actions can be inferred only by relying on the narration for less than half of the videos (496 / 1,077). For the remaining videos, the annotators answered that they relied on either the visual information (in 55 videos) or on both visual and audio information (in 423 videos). The remaining 103 videos do not have a clear agreement among annotators on the modality used to indicate the action reasons. We believe that this imbalanced split might be a reason for why the multimodal model does not perform as well as the text model. For future work, we want to collect more visual data

that contains action reasons.

**Impact of reason specificity on model performance.** The reasons in WHYACT vary from specific (e.g., for the verb “fall”, possible reasons are: “tripped”, “ladder broke”, “rush”, “makeup fell”) to general (e.g., for the verb “play”, possible reasons are: “relax”, “entertain yourself”, “play an instrument”). We believe that a model can benefit from learning both general and specific reasons. From general reasons such as “relax”, a model can learn to extrapolate, generalize, and adapt to other actions for which those reasons might apply (e.g., “relax” can also be a reason for actions like “drink” or “read”) and use these general reasons to learn commonalities between these actions. On the other hand, from a specific reason like “ladder broke”, the model can learn very concise even if limited information, which applies to very specific actions.

**Data Annotation Challenges.** During the data annotation process, the workers had the choice to write comments about the task. From these comments we found that some difficulties with data annotation had to do with actions expressed through verbs that have multiple meanings and are sometimes used as figures of speech. For instance, the verb “jump” was often labeled by workers as “jumping means starting” or “jumping is a figure of speech here.” Because the majority of videos containing the verb “jump” are labeled like this, we decided to remove this verb from our initial list of 25 actions. Another verb that is used (only a



few times) with multiple meanings is “fall” and some of the comments received from the workers are: “she mentions the season fall, not the action of falling,” “falling is falling into place,” “falling off the wagon, figure of speech.” These examples confirm how rich and complex the collected data is and how current state-of-the-art parsers are not sufficient to correctly process it.

## 7 Conclusion

In this paper, we addressed the task of detecting human action reasons in online videos. We explored the genre of lifestyle vlogs, and constructed WHY-ACT – a new dataset of 1,077 video-clips, actions and their reasons. We described and evaluated several textual and visual baselines and introduced a multimodal model that leverages both visual and textual information.

We built WHYACT and action reason detection models to address two problems important for the advance of action recognition systems: adaptability to changing visual and textual context, and processing the richness of unscripted natural language. In future work, we plan to experiment with our action reason detection models in action recognition systems to improve their performance.

The dataset and the code introduced in this paper are publicly available at [https://github.com/MichiganNLP/vlog\\_action\\_reason](https://github.com/MichiganNLP/vlog_action_reason).

## Ethics and Broad Impact Statement

Our dataset contains public YouTube vlogs, in which vloggers choose to share episodes of their daily life routine. They share not only how they perform certain actions, but also their opinions and feelings about different subjects. We use the videos to detect actions and their reasons, without relying on any information about the identity of the person such as gender, age or location.

The data can be used to better understand people’s lives, by looking at their daily routine and why they choose to perform certain actions. The data contains videos of men and women and sometimes children. The routine videos present mostly ideal routines and are not comprehensive of all people’s daily lives. Most of the people represented in the videos are middle class Americans.

In our data release, we only provide the YouTube urls of the videos, so the creator of the videos can always have the option to remove them. YouTube videos are a frequent source of data in research

papers (Miech et al., 2019; Fouhey et al., 2018; Abu-El-Haija et al., 2016), and we followed the typical process used by all this previous work of compiling the data through the official YouTube API and only sharing the urls of the videos. We have the rights to use our dataset in the way we are using it, and we bear responsibility in case of a violation of rights or terms of service.

## Acknowledgements

We thank Pingxuan Huang for his help in improving the annotation user interface. This research was partially supported by a grant from the Automotive Research Center (ARC) at the University of Michigan.

## References

- Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, A. Natsev, G. Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *ArXiv*, abs/1609.08675.
- Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. [Leveraging linguistic structure for open domain information extraction](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.
- Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Nan Liu, Jonathan Stroud, and Rada Mihalcea. 2021. [Fill-in-the-blank as a challenging video understanding evaluation framework](#).
- Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and R. Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1130–1139.
- Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: Extracting visual knowledge from web data. *2013 IEEE International Conference on Computer Vision*.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. Rescaling egocentric vision. *CoRR*, abs/2006.13256.

- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*.
- Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. [Video2Commonsense: Generating commonsense descriptions to enrich video captioning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860, Online. Association for Computational Linguistics.
- Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.
- J. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. 2018. From lifestyle vlogs to everyday interactions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4991–5000.
- Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.
- Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–253.
- Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. [Deep semantic role labeling: What works and what’s next](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.
- Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. [Cnn architectures for large-scale audio classification](#). In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. [\(comet-\) atomic 2020: On symbolic and neural commonsense knowledge graphs](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.
- Vladimir Iashin and Esa Rahtu. 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*.
- Oana Ignat, Laura Burdick, Jia Deng, and Rada Mihalcea. 2019. [Identifying visible actions in lifestyle vlogs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6406–6417, Florence, Italy. Association for Computational Linguistics.
- Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. 2021. Intentionomy: a dataset and study towards human intent understanding. In *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- A. Karpathy, G. Toderici, Sanketh Shetty, Thomas Leung, R. Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732.
- Yu Kong and Yun Fu. 2018. Human action recognition and prediction: A survey. *ArXiv*, abs/1806.11230.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Antoine Miech, D. Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, I. Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.
- Tom Michael Mitchell, William W. Cohen, Estevam R. Hruschka, Partha P. Talukdar, Bo Yang, J. Betteridge, Andrew Carlson, B. D. Mishra, Matt Gardner, Bryan Kiesel, J. Krishnamurthy, N. Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil Antonios Platanios, Alan Ritter, M. Samadi, Burr Settles, R. C. Wang, D. Wijaya, A. Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2015. Never-ending learning. *Communications of the ACM*, 61:103 – 115.
- Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. [GLUCOSE: Generalized and Contextualized story explanations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

- F. Murtagh and P. Legendre. 2014. Ward’s hierarchical agglomerative clustering method: Which algorithms implement ward’s criterion? *Journal of Classification*, 31:274–295.
- Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. [A span selection model for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642, Brussels, Belgium. Association for Computational Linguistics.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Motlaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision – ECCV 2020*, pages 508–524, Cham. Springer International Publishing.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. [Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2751–2767, Online. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Marcus Rohrbach, S. Amin, M. Andriluka, and B. Schiele. 2012. A database for fine grained activity detection of cooking activities. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. [Atomic: An atlas of machine commonsense for if-then reasoning](#). In *AAAI*, pages 3027–3035.
- Zheng Shou, J. Chan, Alireza Zareian, K. Miyazawa, and S. Chang. 2017. Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1417–1426.
- Gunnar A. Sigurdsson, Olga Russakovsky, and A. Gupta. 2017. What actions are needed for understanding human actions in videos? *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2156–2165.
- Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pages 510–526, Cham. Springer International Publishing.
- Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. [Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning](#). *Knowledge-Based Systems*, 230:107408.
- K. Soomro, A. Zamir, and M. Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *AAAI*, pages 4444–4451.
- Stuart Synakowski, Q. Feng, and A. Martínez. 2020. Adding knowledge to unsupervised algorithms for the recognition of intent. *International Journal of Computer Vision*, pages 1–18.
- Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, L. Zhao, Jiwen Lu, and J. Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1207–1216.
- Henry L. Tosi. 1991. A theory of goal setting and task performance. *Academy of Management Review*, 16:480–483.
- Du Tran, Heng Wang, L. Torresani, Jamie Ray, Y. LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Carl Vondrick, Deniz Oktay, H. Pirsiavash, and A. Torralba. 2016. Predicting motivations of actions by

- leveraging text. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2997–3005.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. <https://github.com/facebookresearch/detectron2>.
- J. Xu, Tao Mei, Ting Yao, and Y. Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.
- Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo, and Seung-won Hwang. 2018. [Visual choice of plausible alternatives: An evaluation of image-based commonsense causal reasoning](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hongming Zhang, Yintong Huo, Xinran Zhao, Yangqiu Song, and Dan Roth. 2021. Learning contextual causality between daily events from time-consecutive images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1752–1755.
- Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. [Towards automatic learning of procedures from web instructional videos](#). In *AAAI*, pages 7590–7598.



## A Appendix

### ▼ Instructions

You are given a video that contains a person describing an action and a list of **candidate reasons for why they want to do the action**.

From the list of candidate **reasons**, select the ones that are mentioned verbally or shown visually in the video.

Please see three examples below:

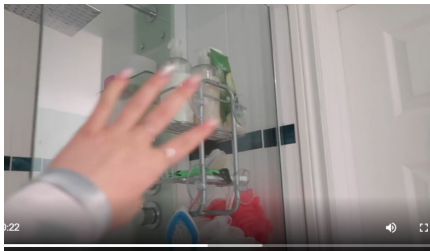
#### 1. Action reasons are mentioned verbally, and shown visually in the video



##### Answers:

- ☒ remove dirt (because it shown and metioned in the video)
- ☒ don't like dirtiness (because it is mentioned in the video)
- ☐ declutter
- ☐ company is coming
- ☐ feel productive

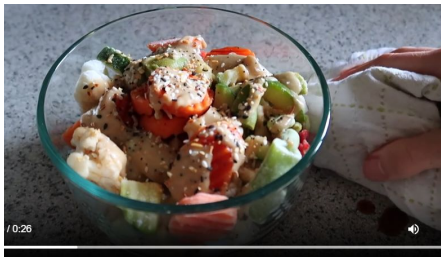
#### 2. Action reasons are mentioned verbally, but not shown visually in the video



##### Answers:

- ☒ remove dirt (because it is mentioned in the video)
- ☐ don't like dirtiness
- ☐ declutter
- ☐ company is coming
- ☐ feel productive

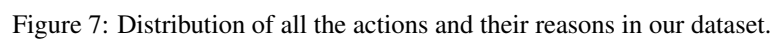
#### 3. Action reasons are shown visually, but not mentioned verbally in the video



##### Answers:

- ☒ remove dirt (because it shown in the video)
- ☐ don't like dirtiness
- ☐ declutter
- ☐ company is coming
- ☐ feel productive

Figure 6: Instructions and examples of completed assignments with explanations for why the answers were selected.



- ☒ tastes good
- ☐ were hungry
- ☐ had craving
- ☒ be healthy
- ☒ enjoy

- x clean walls
- ✓ DIY craft project
- ✓ express yourself
- ✓ enhance appearance
- ✓ feel creative
- x change colors in home

Figure 8: Other examples of actions and their annotated action reasons in our dataset.

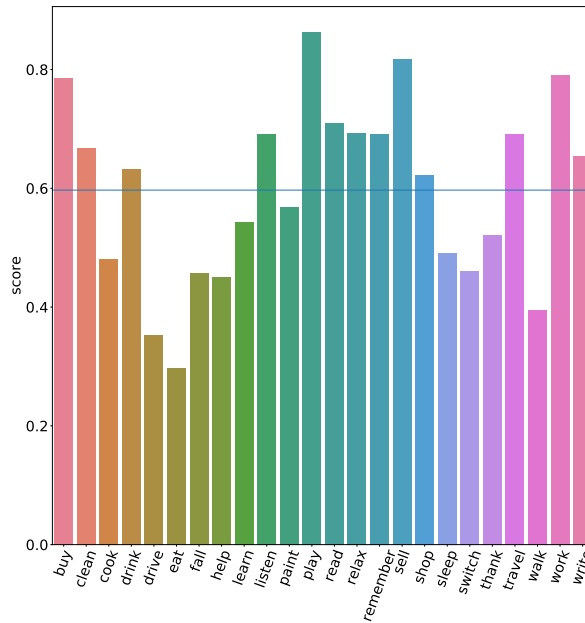


Figure 9: Distribution of all the actions and their worker agreement score: Fleiss kappa score (Fleiss, 1971).

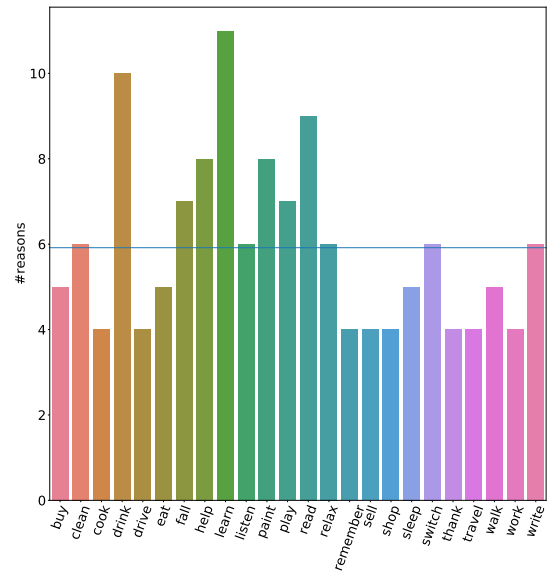


Figure 11: Distribution of all the actions and their number of reasons.

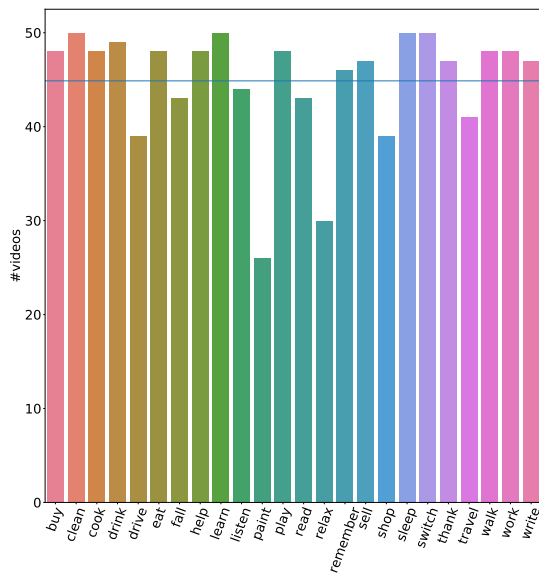


Figure 10: Distribution of all the actions and their number of videos.

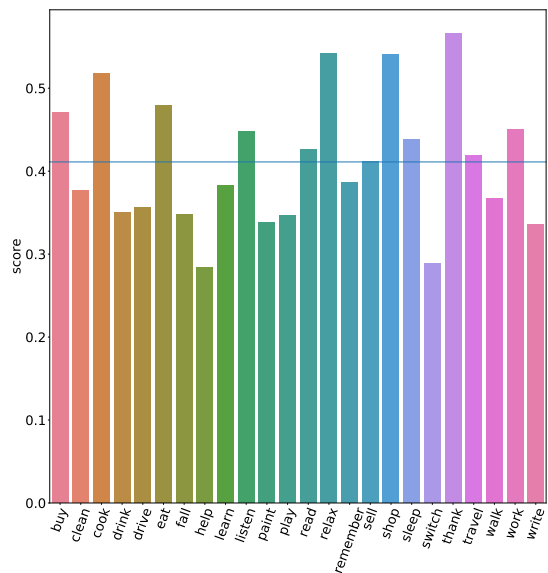


Figure 12: Distribution of all the actions and their F1 score obtained with the highest performing model (Fill-in-the-blanks with Text).