## A Comparison with Existing Datasets

There are publicly available datasets related to understanding instructional videos:

- AllRecipes (Kiddon et al., 2015) (AR). The authors collected 2,456 recipes from AllRecipes weibsite[8]. The sentences in the dataset are mostly simple imperative English describing concise steps to make a given dish, where the first word is usually the verb describing the action. The ingredient list information is also available. In contrast, our task seeks to extract procedural information from more noisy, oral and erroneous languages in real life video context.

- YouCook2[9] (Zhou et al., 2018b) (YC2). The procedure steps for each video are annotated with temporal boundaries in the video and described by human-written imperative English sentences. However, this dataset does not contain more fine-grained annotations in a structured form.

- HowTo100M[10](Miech et al., 2019). This is a large scale how-to videos dataset, searched on YouTube using the task taxonomy on WikiHow[11] as a source. However, it does not contain any annotations although the domain is more general.

- CrossTask[12](Zhukov et al., 2019) (CT). Based on HowTo100M, this dataset is used for weakly supervised learning with 18 tasks fully labeled and 65 related tasks unlabeled. Although the dataset is annotated in a structured way by separating verbs and objects, the label space is closed with predefined sets of verbs and objects. The dataset also does not allow multiple verbs or objects to be extracted for a single segment.

- COIN[13] (Tang et al., 2019). This contains instructional (how-to) videos, in a closed taxonomy of tasks and steps. The authors annotated time spans of steps in a video with pre-defined steps, however the biggest drawback is that it is unstructured and closed domain.

- How2[14] (Sanabria et al., 2018). This dataset annotates ground truth transcript text to help abstractive summarization, a very different task than ours of structured data extraction.

- HAKE[15] (Li et al., 2019). Human Activity Knowledge Engine (HAKE) is a large-scale knowledge base of human activities, built upon existing activity datasets, and supplies human instance action labels and corresponding body part level atomic action labels. However, HAKE uses closed activity and part state classes. It also does not contain videos of activities accompanied with narrative transcripts.

- TACOS[16] (Regneri et al., 2013). This dataset considers the problem of grounding sentences describing actions in visual information extracted from videos in kitchen settings. The dataset contains expert annotations of low level activity tags, with a total of 60 different activity labels with numerous associated objects, and sequences of NL sentences describing actions in the kitchen videos. This dataset also does not support open extraction and the videos are provided using human annotated caption sentences, rather than transcript texts with noise.

## B Neural Selection Model

Figure 5 presents the overall detailed structure of the neural selection model for combining utterance and video information for key clip selection.

**Sentence token encoding** Each input clip is accompanied with a sentence $S = \{t_1, \ldots, t_k\}$ which has $k$ tokens. We use a pre-trained BERT (Devlin et al., 2018) model as the encoder and extract the sentence representation $\mathbf{s}$.

**Video frame features** For each clip we uniformly sample $T = 10$ frames and use an ImageNet-pretrained (Deng et al., 2009) ResNet50 (He et al., 2016) to extract the feature vector of each frame as $X = \{x_1, \cdots, x_T\}$.

[8] https://www.allrecipes.com/
[9] http://youcook2.eecs.umich.edu/
[10] https://www.di.ens.fr/willow/research/howto100m/
[11] https://www.wikihow.com
[12] https://github.com/DmZhukov/CrossTask
[13] https://coin-dataset.github.io/
[14] https://github.com/srvk/how2-dataset
[15] http://hake-mvig.cn
[16] http://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos

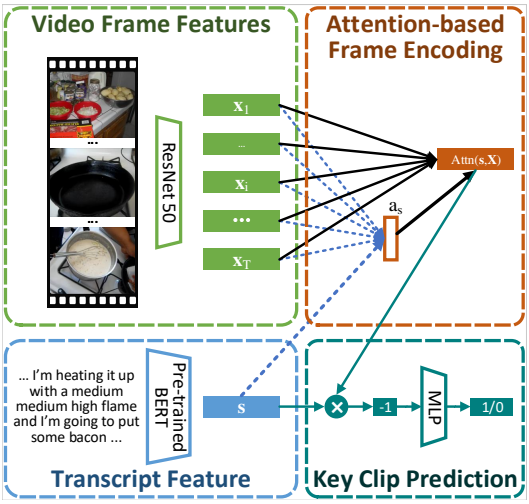


Figure 5: Neural key clip selection model.

**Attention-based frame encoding** To model the interaction between the encoded sentence and the feature of each frame, we adopt an attention-based method. We first calculate the attention weight $\mathbf{a}_s$ by a tensor product of sentence feature $\mathbf{s}$ with each video frame $\mathbf{x}_i$ followed by a softmax layer. Then we perform a weighted sum on all frame features to get $\text{Attn}(s, \mathbf{X})$.

**Visual-utterance fusion** Finally, we fuse the extracted transcript features $s$ with the attended video features $\text{Attn}(s, \mathbf{X})$ by a tensor product and flatten it into a vector. Then we use a non-linear activation layer to map these features into a real number, which represents the probability of the clip being a key clip.

**Experiment details** In the presented experiments, we use a pre-trained BERT (Devlin et al., 2018) model[17] to extract the continuous representation of each sentence. During fine-tuning, the model is optimized by Adam optimizer (Kingma and Ba, 2014) with the starting learning rate of $1e-4$. The model is trained in a supervised fashion with a separate key clip/sentence classification dataset that is *not* related to YouCook2. This auxiliary dataset will also be publicly released. All of them are general domain instructional videos harvested from from YouTube. Human annotators labeled whether it is a key clip when given a video clip-sentence pair. In the end, we have 1,034 videos (40,146 pairs) for training the classification model. We split the dataset into two subset as 772 videos (28,519 pairs) and 312 videos (11,627 pairs) for training and validation (hyper-parameter tuning) respectively. The testing set is our proposed dataset with key clips and sentences annotated (see §3), containing 356 videos and 15,523 pairs. The testing set used is the same as all other compared methods.

[17] https://github.com/hanxiao/bert-as-service

## C SRL Argument Filtering

The argument types that we deem to not contribute as the procedural knowledge for completing the task and filter out include: ARG0 (usually refers to the subject, usually a person), AM-MOD (modal verb), AM-CAU (cause), AM-NEG (negation marker), AM-DIS (discourse marker), AM-REC (reciprocal), AM-PNC/PRP (purpose), AM-EXT (extent), and R-ARG* (in-sentence references).

## D Fuzzy Matching and Partial Fuzzy Matching

**Fuzzy matching** Denote the Levenshtein distance between string $a$ and string $b$ as $d(a, b)$. We then define a normalized pairwise score between 0 to 1 as $s(a, b) = d(a, b)/max\{|a|, |b|\}$ Given a set of $n$ predicted phrases $X = \{x_1, ..., x_n\}$ and a set of $m$ ground truth phrases $G = \{g_1, ..., g_m\}$, we can find a set of $min(n, m)$ string pairs between predicted $X$ and ground truth $G$, as $M = \{(x_i, g_j)\}$ that maximizes the sum of scores $\sum_{(x_i,g_j)\in M} s(x_i, g_j)$. This assignment problem can be solved efficiently with Kuhn-Munkres (Munkres, 1957) algorithm[18]. Since this fuzzy pairwise score is normalized, it can be regarded as a soft version for calculating $TP = max \sum_{(x_i,g_j)\in M} s(x_i, g_j)$.

**Partial fuzzy matching** The only difference from *"fuzzy"* matching is that the scoring function now follows the "best partial" heuristic that assuming the shorter string $a$ is length $|a|$, and the longer string $b$ is length $|b|$, we now calculate the score between shorter string and the best *"fuzzy"* matching length-$|a|$ substring.

$$s(a, b) = max\{d(a, t)\}/|a|, t \in \text{substring of } b, \\ |t| = |a|, |a| < |b|$$

Both fuzzy metric implementations are based on FuzzyWuzzy[19].

[18] http://software.clapper.org/munkres/

[19] https://github.com/seatgeek/fuzzywuzzy

## E Example Extractions

In this section, we showcase some example extractions from our overall best-performing baseline model "SRL w/ heur." in Table 6 (next page). We show both the annotated structured extractions as well as the model output, for a recipe of type *"pizza marghetta"*.[20] Note that the table includes only transcript sentences that are annotated as key steps. We can see that some sentences do no have extraction output from the model, while others tend to be over-extracted as long spans of text or incorrect due to the noisy nature of the transcript, degrading the extraction quality. Verbs are also relatively better extracted than arguments for the proposed model. From the extraction examples, we can also see that the model sometimes omits important action verbs in the extraction, and extract pronouns like "it", "this", "here", etc. as arguments. This suggests that the utterance-only model cannot handle *coreference* and *ellipsis* scenarios very well, which is one of the key difficulties of the proposed task (discussed in Section 3.2). This also partly implies the need for utilizing the visual information to extract actions and arguments that are not included in transcript, as well as visual co-reference and language grounding to help resolve what objects the pronouns in the extracted arguments are referring to.

## F Reproducibility Details

All the experiments in this paper are performed on a workstation with one 8-core Intel i7 processor, 32GB RAM and one NVIDIA Tesla K80 GPU.

The SRL-based models' runtime are bounded by the inference speed of the pretrained model, and take 30 minutes on average to complete the evaluation. The code is based on the semantic role labeling pretrained model from https://github.com/allenai/allennlp-hub/blob/master/allennlp_hub/pretrained/allennlp_pretrained.py.

Kiddon et al. (2015) models are trained from scratch with default hyperparameter settings on our dataset using only CPU, and take 2 hours on average to complete training and evaluation. We directly use the released code (https://github.com/uwnlp/recipe-interpretation) to train the unsupervised model on our dataset for evaluation.

The neural selection model (Section 4.2) takes 6 hours on average to train on both transcript and video data. The number of parameters for the neural model is the addition of those from BERT-base model, ResNet-50 model, an attention module and a MLP layer. More details are described in Appendix B.

The action and object detection model (Section 5.2) pretrained on EpicKitchen dataset is retrieved from https://github.com/epic-kitchens/action-models. In our use case, we only need to perform the inference steps on our dataset, which costs 3 hours on average for all the video clips in our proposed dataset.

[20] https://www.youtube.com/watch?v=FHvZgt3ExDI

Table 6: Example extractions compared with gold annotations for a recipe of type “pizza marghetta”. Only transcript sentences that are annotated as key steps are included. Some long and incorrect extractions are quoted with omission.

| Transcript Sentence | Procedure Summary | Gold Verbs | Gold Arguments | Extracted Verbs | Extracted Arguments |
|---|---|---|---|---|---|
| so we ’ve placed the dough directly into the caputo flour that we import from italy. | place dough in caputo flour | place | dough, caputo flour | place | so, the dough, directly, into the caputo flour that we import from italy |
| and then we give it a flip as i ’ve read in some, some manuals for italian pizza that are neapolitan style. | flip dough | flip | dough | | |
| but that ’s what we do anyway, we sprinkle the surface and then real quick. | sprinkle the surface | sprinkle | surface | sprinkle | the surface |
| we just give a squish with our palm and make it flat in the center. | squish dough with palm; flatten center | squish; flatten | dough, with palm; center of dough | make | just, it flat in the center |
| we dimple the rest of the pizza, moving the pizza around definitely handmade, definitely handmade, not trying your best not to disturb the edge ’cause that ’s where you ’re going to get a lot of natural bubbles from the fermentation. | dimple the rest of pizza; move the pizza around | dimple; move | pizza; pizza | move | the pizza, around, “definitely handmade, definitely handmade” |
| once you ’ve dimpled, you ’re going to do a quick stretch, while you ’re rotating. | stretch dough;rotate dough | stretch; rotate | dough; dough | rotate | you |
| and then we ’re going to put it, pick it up and put it on the backs of our hands and just let it kind of hang down on the backs of our knuckles. | pick up dough and let it hang on backs of knuckles | pick up; hang on | dough; dough, backs of knuckles | put; pick; put; let | it; it; it, on the backs of our hands; just, it kind of hang down on the backs of our knuckles |
| do you want a quarter to two and a half ounces of sauce, and i ’m going to put the cheese on. | put cheese on | put on | cheese | put | the cheese, on |
| olive oil actors come out of the oven, and we put on here. | put on fresh basil | put on | fresh basil | put | here |
| so that ’s all we do to the pizza and then we go into the oven just going to place it directly on the stone and give it a shake. | place pizza in oven, give it a shake | place; shake | pizza, oven; pizza | place | it, directly on the stone |
| so will have to rotate it every thirty to forty five seconds, your home oven will take you about ten minutes. | rotate pizza every 30-45 seconds | rotate | pizza, 30-45 seconds | rotate, take | it, every thirty to forty five seconds; will, you, about ten minutes |
| we ’re going to put a little bit of extra virgin olive oil directly on. | put extra vigrin olive oil on | put on | extra virgin olive oil | put | a little bit of extra virgin olive oil, directly, on |
| yeah , over the whole pizza and then in america we cut it, but i ’m told in, uh, in italy you cut it yourself. | cut the pizza | cut | pizza | cut, cut | in america, it, in; in italy, it, yourself |
| and we start to knead, so we need about though for this fifteen twenty minutes until we have a very nice texture. | knead dough for 15-20 minutes | knead | dough, 15-20 minutes | | |
| you make a nice bowl with cover, the door to avoid crusting an you let it rise at room temperature for three four hour. | make bowl with cover, rise dough at room temperature for 3-4 hours | make; rise | bowl, with cover; dough, room temperature, 3-4 hours | make; let | a nice bowl, with cover; it rise at room temperature for three four hour |
| one sourdough, as almost double is volume we make six small balls. | make six small balls | make | six small balls | make | one sourdough, as almost double is volume, six small balls |
| so we make nice round shape now. | make a round shape | make | round shape | make | so, nice round shape, now |

| | | | | | |
|---|---|---|---|---|---|
| we cover again with film to avoid the door crafting an we elect, leave at room times or for a couple of hour or if you like you can place in a warm place for forty five minutes around pizza dough is ready. | cover with film | cover | film | cover; leave; place | again, with film, “to avoid the door ... pizza dough is ready”; at room times or for a couple of hour; in a warm place, for forty five minutes, around |
| and we start making flat with our hands. | make the dough flat with hands | make flat | dough, with hands | make | flat, with our hands |
| now we had a tomato and we spread all over our pizza dough. | spread tomato over pizza dough | spread | tomato, pizza dough | spread | all, over our pizza dough |
| we had the mozzarella cheese shredded little basil pizza is ready to be back. | add mozzarella cheese, shredded basil | add | mozzarella cheese, shredded basil | shred | pizza |
| now we should make pizza at around seven hundred fifty degrees. | make pizza | make | pizza, 750 degrees | make | now, should, pizza, at around seven hundred fifty degrees |
| fahrenheit four five, seven finish our pizza with basil and fresh olive oil. | finish pizza with basil and olive oil | finish | pizza, with basil, olive oil | | |
| OK , well, to make the margarita pizza we ’re going to start off with by stretching the dough and we do make gardell everyday fresh in our kitchen after that we ’re going to add some shredded mozzarella cheese. | stretch the dough, add shredded mozzarella cheese | stretch; add | dough; shredded mozzarella cheese | make; stretch; make; add | “the margarita pizza we ’re going to start off with”, by stretching the dough; the dough; gardell, everyday, fresh, in our kitchen, after that; some shredded mozzarella cheese |
| then our version of a margarita pizza has four dollops of tomato sauce, along with some fresh, chopped tomato, then it goes into the wood burning oven, they certainly, we ’ve been cooking with for fifteen years. | put chopped tomato and tomato sauce into oven | put | chopped tomato and tomato sauce, into oven | oven; cook | wood burning; with, for fifteen years |
| after the margarita pizza comes out of the oven, we finish it with some fresh, grated parmesan cheese. | finish with more cheese, fresh basil | finish | pizza, with more cheese, fresh basil | | |
| here you can find how i do it out on the website, and then i ’ve also got my basic classic tomato sauce plus that i use for just about all my pizzas, an i ’m going to go ahead and lay that down in a, in a nice coat and i like you know, probably frankly i probably like a medium amount of sauce people, people have said little bit less a little bit more i kind of like it a little bit right in the middle use, the back of the spoon there, yeah, you spoon it out with the front, then just kind of use the back spread it out. | lay down tomato sauce | lay down | tomato sauce | lay; spoon; spread | that, down, “in a , in a nice coat”; it, with the front; just, kind of, it |
| and then i like to use our leave about half an inch all the way around the pizza. | leave half an inch around pizza | leave | half an inch leaf, around pizza | | |
| i ’m going to go ahead and lay this down. | lay down cheese | lay down | cheese | lay | this, down |
| and then you can lay some basil down now. | lay basil down | lay down | basil | lay | and, then, can, some basil, down, now |

| | | | | | |
|---|---|---|---|---|---|
| i like to press down a little bit to make sure that it stays down, stays down as it cooks if it stays down it. | press basil down | press down | basil | press; make | down, a little bit, “to make sure ... it stays down it”; “sure ... it stays down it” |
| let ’s go ahead and pop this in the oven, so in the interest of full disclosure. | pop pizza in oven | pop | pizza, oven | let; pop | “’s go ahead and ... full disclosure”; this, in the oven, so in the interest of full disclosure |
| i actually kept it on the pizza stone. | keep it on pizza stone | keep | pizza, pizza stone | keep | actually, it, on the pizza stone |