Detours for Navigating Instructional Videos

Supplementary Material

1. Supplementary video

We attach a supplementary video containing details about high-level idea, overview of the problem, detour dataset visualizations and some result visualizations.

2. Dataset visualization

We attach visualization pages that show the outputs from LLM for (a) summary generation and (b) weakly-supervised detour training data. Both these visualizations contain samples that are automatically generated using narrations with LLAMA 2. Most of the training samples are valid detour prompts with correct time windows; with some noise due to imperfections in narrations as a supervision and due to errors in LLAMA 2 generations. Finally, we also have a visualization that shows the samples from (c) manually collected testing data. These visualizations show the good quality manual annotations that we have collected.

3. Detour dataset generation details

In this section, we present additional details about the detour dataset generation (Sec. 3.2 in the main paper) process. We discuss the input prompt used to generate weakly-supervised detour annotations in $\mathcal{D}_D^{tr}$ and resulting sample outputs (including failure cases). Next, we detail the manually annotated dataset $\mathcal{D}_D^{te}$ collection details and sample visualizations.

3.1. Weakly-supervised training set.

The first step involves generating summaries from given narrations N_i. The narrations are obtained using ASR from narrated videos and we use the sentencified version, provided in [32]. We use the narrations along with the timestamp and provide the following prompt to LLAMA 2 [68]:

> *System:* Help summarize the steps of this recipe whose narrations with timestamps are given. Timestamp is given in HH:MM:ss.
> *User:* Given the narrations of a video, tell the recipe being made in this video and list down the steps and start and end timestamps in the video. Answer in this format: 'Recipe: Name of the recipe and brief detail Step 1: [HH:MM:ss - HH:MM:ss] description of the step Step 2: [HH:MM:ss - HH:MM:ss] description of the step and so on'. Here are narrations with timestamps in HH:MM:ss format: <insert>

where <insert> is replaced by narrations with timestamps. Some sample outputs are shown in Fig. 1 along with a row of failure cases (bottom). We create an automated parser that extracts steps as a tuple of timestamps and text description. Many of the videos do not contain meaningful narrations or have no narrations, and hence the outputs from these prompts do not fit into the desired output format. They are rejected automatically by the parser. There are rare instances where even though the narrations are meaningful, the outputs are incorrect, e.g. garbage output or no output (bottom right, Fig. 1). Finally, there is a small fraction of outputs ($\sim$3%) where the timestamps are incorrect or missing altogether (bottom left, Fig. 1). To mitigate this, we make sure steps' coverage is at least 80% of the duration of the video. This process results in a high-quality text summary dataset. The overall process results in a summary dataset of 187K samples. Please also see the attached summary visualization page. It contains parsed summaries and we can observe the good quality summary generations using LLAMA 2 with only few failure cases.

Finally, we input two similar summaries into LLAMA 2 [68] and generate detour instances. The process to filter similar summaries is detailed in Sec. 3.2 (main paper). For every pair of similar summaries, we use the following prompt:

> *System:* Help understand why a user would pause watching one video and take a detour to another cooking video.
> *User:* There are two cooking videos A and B. The steps of the recipe along with timestamps in HH:MM:ss format is given. Suppose a person is watching video A, can you tell me what the user would prompt to take a detour and watch video B? The answer can be some extra/missing ingredients, tools or procedural step. Some examples of such queries can be 'How to do this step without adding yeast?', 'Can I add chilli powder here?', 'Can I do this step without blender?', 'Can you give a video that shows other way to roll a sushi?' and so on. Also, tell the time when the user would stop watching Video A and the time range in Video B and answers the user query. Answer in this format: 'Detour time in Video A: HH:MM:ss, Detour time window in Video B: [HH:MM:ss - HH:MM:ss], Detour text prompt: One sentence question a user would prompt to take a detour'. Here are the reciped: Video A: <insert> and Video B: <insert>

where we insert source and detour video narrations in <insert>, respectively. Same as above, we create an automated parser to convert the text outputs into dataset tuples. A small fraction of outputs cannot be parsed by the automated parser due to incorrect output format by LLAMA 2. We ignore these instances since they are small in number in comparison to the successful parse. Fig. 2 shows some output samples along with failure cases. Please also see the attached visualization page for training data samples that contains automatically annotated valid detour annotations and some failure cases. We manually verify a subset of the generations and observe good quality. Furthermore, we observe a strong correlation coefficient of > 0.85 between validation set (created automatically using narrations) and the manually collected test set across all training runs for both detour video retrieval and detour window localization task.

This automatically generated data is used for training only—never for ground truth evaluation of any model.

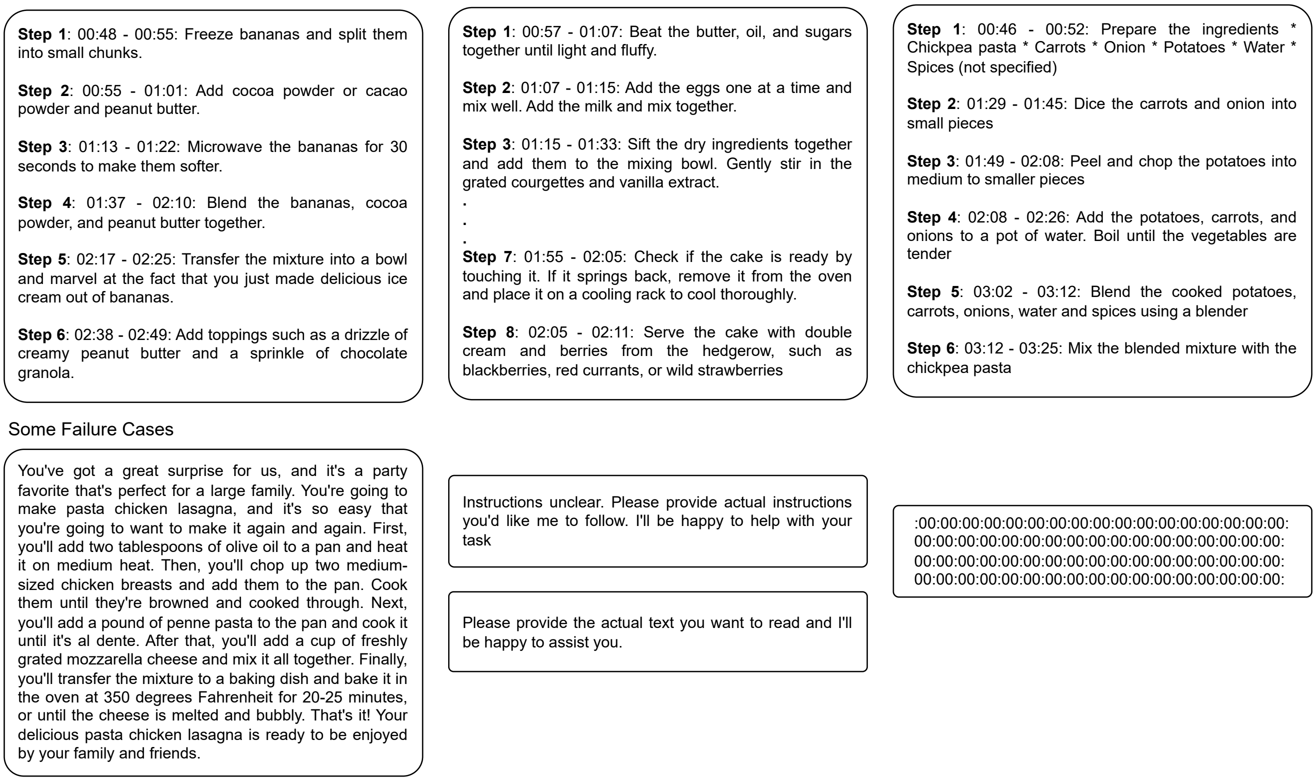

Figure 1. **Weakly-supervised summaries generated using narrations with LLAMA 2 [68]**. While majority of the outputs contains step details and timestamps in the desired format, a few outputs are incorrect (bottom).

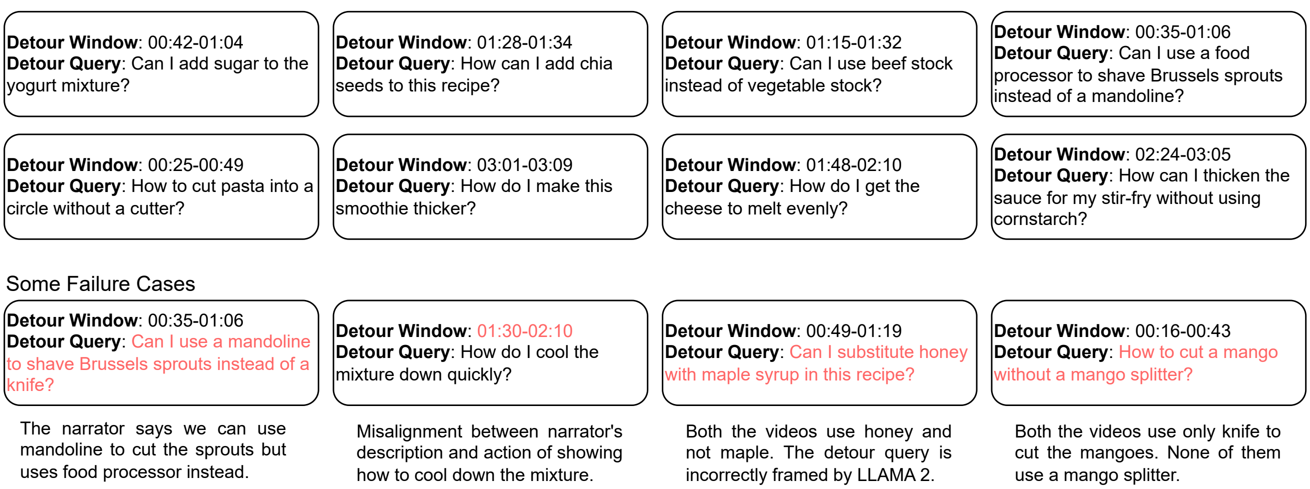

Figure 2. **Weakly-supervised detour annotation sample for training and validation.** It also contains a row of failure cases with reasons. Please also see the attached visualization for more visualizations.

3.2. Manually collected testing set

We hire 24 professional annotators for manually generating video detour instances. All of them are trained for a few days on what constitutes a detour (along with examples), how to mark the time instances t_a and detour window T_d and what types of samples to reject. The training was followed by a pilot collection to evaluate their understanding. Finally, they annotate using a designed interface shown in Fig. 3. We randomly sample a subset and manually verify the annotations for quality control.

The resulting dataset consists of 3.9K source-detour video pairs resulting in $16,207$ samples. Due to our annotator trainings and quality control, the resulting dataset is of high-quality. Fig. 4

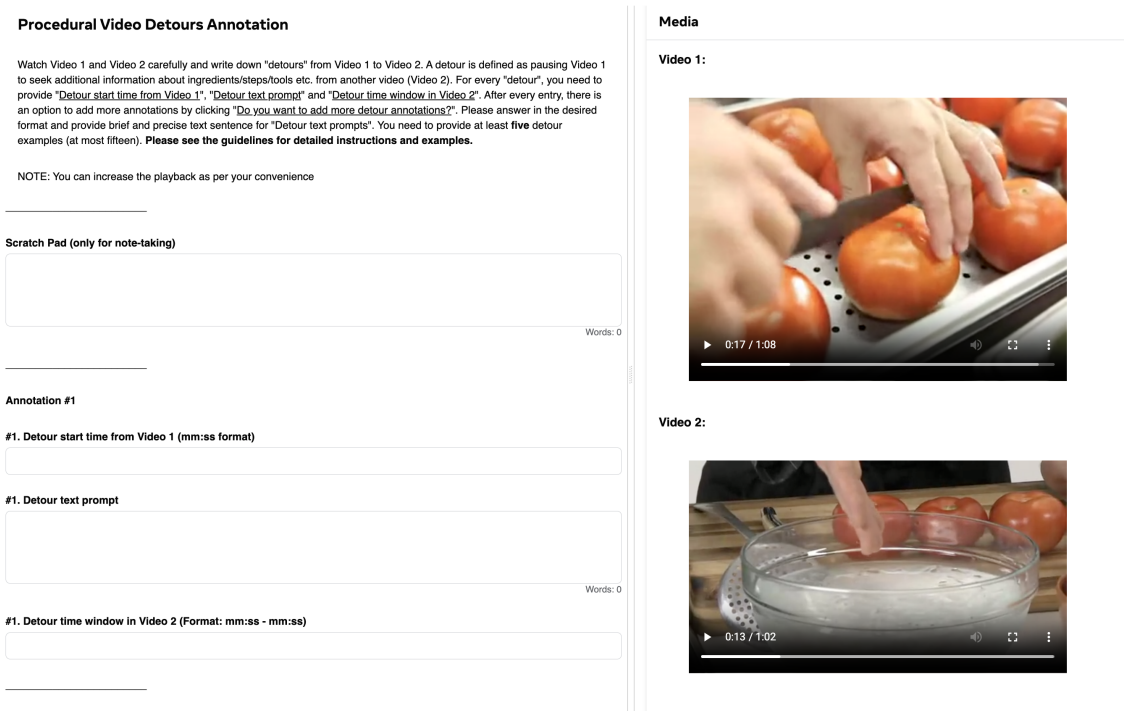

Figure 3. **Annotation interface for manual test dataset collection**. The interface reiterates important details in addition to a separate document (top). There is a scratch for the annotators to take notes while watching the two videos on the right. Finally, each instance of annotation contains a detour start time from the source video, a detour text prompt and finally a detour time window in the target video. The interface supports up to 15 annotations but only three is required.

Detour Window: 01:10-01:29 **Detour Query**: Can I use fresh coconut water?	**Detour Window**: 00:56-01:06 **Detour Query**: How do I do this without a microwave safe tray?	**Detour Window**: 02:24-03:08 **Detour Query**: Show me how to finish the cake differently	**Detour Window**: 02:05-02:12 **Detour Query**: Can I use a food processor instead of an immersion blender?
Detour Window: 00:49-00:55 **Detour Query**: Should i cover the turkey burger with foil while cooking?	**Detour Window**: 00:07-01:43 **Detour Query**: Show me the process of rinsing the rice in detail	**Detour Window**: 00:13-00:15 **Detour Query**: Can I skip slicing the jalapeno peppers?	**Detour Window**: 02:01-02:03 **Detour Query**: Would it be fine to have crushed walnuts as an alternative for raisins and nuts?
Detour Window: 02:51-02:54 **Detour Query**: Can I add cheese cubes?	**Detour Window**: 01:57-02:41 **Detour Query**: How can I stuff the jalapenos?	**Detour Window**: 00:37-00:57 **Detour Query**: Is there other way to make milk froth?	**Detour Window**: 00:21-00:24 **Detour Query**: I do not want it to be spicy, can I skip adding chili?

Figure 4. **Manually collected detour annotation for testing.** Please also see the attached visualization for more samples with videos that showcases the good quality annotations that we collected.

shows representative examples. Please also see the attached visualization of test data samples that shows the high-quality samples. These manually created detours are used for evaluation.

4. Experimental results expanded

We expand on to the results in Sec. 4 (main paper) and show the generalizability of our method (Sec. 4.1) and performance at different input combinations (Sec. 4.2).

4.1. Generalizability to *novel* tasks

As discussed in Sec. 3.4, we have two splits of the test data — *common* tasks containing video pairs from most frequent recipes of HowTo100M [55], and *novel* tasks consisting of video pairs from least frequent recipes of the dataset. We do not include any video pairs from *novel* tasks in the training set to evaluate the generalizability of our method.

Results. Tab. 1 contains the performance split for each testing subset for both detour video retrieval and detour window localization tasks. In the main paper, we showed only the performance on

Method	Common Task MedR ↓	Novel Task MedR ↓	MedR ↓
Text-only	508	524	512
CLIP [63]	348	310	342
CLIP-Hitchhiker [7]	339	305	336
InternVideo [72]	315	296	313
DistantSup. [45]	320	350	329
MLLM [80]	127	155	139
CoVR [69]	464	485	473
Ours	**29**	**35**	**30**
Ours w/o hard-negatives	49	63	55
Ours w/ parser	76	88	81

Method	Common Task Mean R@1	Novel Task Mean R@1	Mean R@1
Text-only	4.0	4.5	4.2
2D-TAN [86]	8.9	8.2	8.6
VSLNet [84]	9.2	9.8	9.4
UMT [47]	9.6	9.3	9.4
DistantSup. [45]	8.8	7.9	8.3
MLLM [80]	9.7	10.8	10.2
STALE [58]	9.7	9.5	9.6
Ours	**13.3**	**12.3**	**12.8**
Ours w/ parser	12.0	11.3	11.6

Table 1. Results for detour video retrieval (left) and detour window localization (right) tasks on common task and novel task splits. Our method outperforms all prior methods and baselines by a significant margin even on novel tasks.

Method	V_s	$\mathcal{Q}$	R@5	R@10	R@50	MedR ↓
CLIP [63]	✓		9.6	13.2	26.9	314
		✓	11.2	16.4	32.0	191
	✓	✓	7.9	11.8	25.2	342
CLIP-Hitch. [7]	✓	—	—	—	—	—
		✓	11.3	17.7	33.2	186
	✓	✓	8.4	12.3	25.6	336
InternVideo [72]	✓		11.2	17.0	31.8	150
		✓	13.1	19.2	37.2	138
	✓	✓	9.7	13.2	27.2	313
DistantSup. [45]	✓		4.9	10.2	15.9	384
		✓	8.0	12.0	25.4	370
	✓	✓	8.4	12.6	25.1	329
MLLM [80]	✓		11.3	17.8	32.3	189
		✓	11.4	16.8	31.4	158
	✓	✓	5.9	10.5	32.1	139
CoVR [69]	✓		4.9	10.1	15.9	388
		✓	4.1	10.0	15.6	401
	✓	✓	4.3	9.2	15.3	473
Ours	✓		6.1	11.0	32.3	128
		✓	6.1	10.8	32.6	116
	✓	✓	**17.6**	**27.8**	**62.4**	**30**

Method	V_s	$\mathcal{Q}$	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	Mean R@1
2D-TAN [86]	✓		8.9	3.2	0.9	5.5
		✓	10.0	3.8	1.2	8.0
	✓	✓	10.3	4.2	1.5	8.6
VSLNet [84]	✓		9.2	3.1	1.1	6.1
		✓	10.9	4.0	1.5	8.5
	✓	✓	11.8	5.8	1.7	9.4
UMT [47]	✓		9.7	3.5	1.2	6.5
		✓	11.2	5.4	1.7	8.7
	✓	✓	12.0	6.1	1.6	9.4
DistantSup. [45]	✓		9.8	3.7	1.2	7.6
		✓	10.0	3.8	1.2	7.9
	✓	✓	10.6	4.0	1.5	8.3
MLLM [80]	✓		12.2	6.0	1.5	9.1
		✓	12.3	6.2	1.7	9.7
	✓	✓	12.7	6.5	1.8	10.2
STALE [58]	✓		10.0	3.8	1.2	6.9
		✓	11.6	5.5	1.5	8.8
	✓	✓	12.1	6.1	1.7	9.6
Ours	✓		12.0	5.9	1.5	8.9
		✓	14.6	6.9	2.2	11.2
	✓	✓	**16.7**	**7.7**	**2.8**	**12.8**

Table 2. Comparison of our method with prior methods at different input combinations for detour video retrieval (left) and detour window localization (right) on all metrics. Our method outperforms all the prior works for all input combinations.

overall dataset for conciseness. We see that our method achieves significant gains over the baselines for both the tasks. Moreover, the performance drop in novel tasks is minimal compared to the gain. This result shows that the learned model is able to generalize to detour in newer recipes *without* being explicitly trained on them. We attribute this effect to the strong interconnected nature of demonstrations in instructional videos, cooking in particular. Some recipes like *making pancake* and *making crepe* will only differ at some steps and detour learned with *making pancake* should transfer to *making crepe*.

4.2. Results at different input combinations

Tab. 2 contains an expanded version of Tab. 2 from the main paper for all metrics. We showed performance only on one metric for brevity. We see that for both the tasks, the previous source video context and the query context is useful for the model. It is also interesting to note that for state-of-the-art methods InternVideo [72] and CLIP [63], combining source video context directly with query features degrades the performance. This underscores the need for a smarter method to fuse the two contexts, as we show in our method.