Detours for Navigating Instructional Videos

Kumar Ashutosh[1,2], Zihui Xue[1,2], Tushar Nagarajan[2], Kristen Grauman[1,2]
[1]UT Austin, [2]FAIR, Meta

Abstract

We introduce the video detours *problem for navigating instructional videos. Given a source video and a natural language query asking to alter the how-to video's current path of execution in a certain way, the goal is to find a related "detour video" that satisfies the requested alteration. To address this challenge, we propose VidDetours, a novel video-language approach that learns to retrieve the targeted temporal segments from a large repository of how-to's using video-and-text conditioned queries. Furthermore, we devise a language-based pipeline that exploits how-to video narration text to create weakly supervised training data. We demonstrate our idea applied to the domain of how-to cooking videos, where a user can detour from their current recipe to find steps with alternate ingredients, tools, and techniques. Validating on a ground truth annotated dataset of 16K samples, we show our model's significant improvements over best available methods for video retrieval and question answering, with recall rates exceeding the state of the art by 35%.*

1. Introduction

Instructional or "how-to" videos are a compelling medium for people to share and learn new skills. From everyday home fix-it projects, cooking, sports, to aspirational goals like playing piano beautifully, there are so many things that people of all ages and backgrounds want to learn or do a bit better. Indeed, online how-to's are among the top few dominating categories of all content on YouTube, alongside entertainment and music. Advances in computer vision for keystep recognition [6, 21, 45, 55, 56, 90, 91], procedural task understanding [11, 12, 89], and video summarization [5, 60] have the potential to make such content more searchable and accessible.

However, while today's how-to content is a vast resource, it is nonetheless disconnected. Human learners access how-to's in doses of one video at a time, studying the advice and visual demonstrations of one expert at a time.

Website: https://vision.cs.utexas.edu/projects/detours

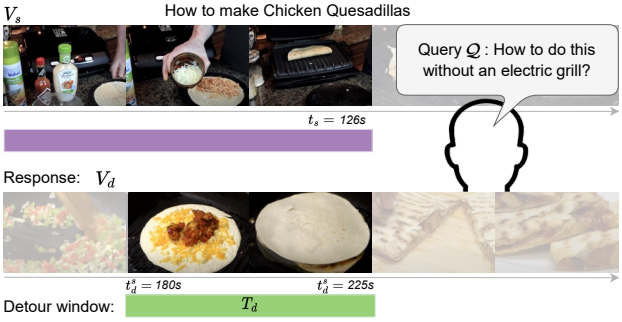

Figure 1. **An example video detour.** In the *Chicken Quesadillas* recipe, the source video V_s (top) shows the use of an electric grill at time instant t_s. A user watching this video does not have a grill and asks a query $\mathcal{Q}$ *"how to do this without an electric grill?"*. In response, the system identifies a detour video V_d and timepoint T_d showing a similar recipe but using a heating pan instead of a grill.

While there are thousands and thousands of videos addressing, for example, *"how to repair a bike" or "how to make a samosa"*, any given video offers only a single execution of a task using a fixed set of ingredients, tools, approach, and assuming a certain skill level. When those criteria do not align, users face a dilemma whether to improvise, risking "breaking" the final output, or to find and watch another video hoping it better matches their constraints. Manually synthesizing the information across videos is time consuming if not prohibitively taxing.

What if the wealth of knowledge in online instructional videos was not an array of isolated lessons, but instead an interconnected network of information? What would it take to transform a pile of videos into a how-to knowledge base?

Towards this vision, we explore how to intelligently navigate between related how-to videos, conditioned on a natural language query. Suppose a user watching a given video discovers they do not have the desired ingredients, tools, or skill-level. They may ask, *"can I do this step without a wrench?"* or *"I am on a diet, can I skip adding cheese here?"* or *"how could I prepare the mix from scratch instead of using a pre-made one?"* or *"is there a simpler way to do the corners?"* and so on. Conditioned on the content watched so far in the *source video*, the goal is to identify a *detour video*—and a temporal segment within it—that

would allow the user to continue their task with the adjustment specified by their language query, then return to the original source video and complete execution. See Figure 1.

At the core this requires new technical advances in multimodal video understanding. Standard text-to-video retrieval models [17, 22, 50, 55, 56, 75, 88] are insufficient because the query text alone may not reveal enough details about the task (e.g., *"can I do this step without an electric grill?"*) Similarly, existing video localization [47, 81, 84, 86] and question answering [37, 38, 41, 64, 78, 79] methods do not consider the viewing history of the source video, which is essential to properly identify a detour. For example, answering *"how to do this step without a wrench?"* requires the model to understand which steps are already done and which part of the target detour video shows the same effect *without* using a wrench.

We introduce VidDetours: a video-language model that benchmarks this new problem. Our approach formulates the video navigation task in two parts: (a) retrieval of the detour video (b) temporal localization of the relevant portion of the detour video—both conditioned on the source video and text query. Building on ideas from the video retrieval [50, 56, 75, 92] and localization [58, 84, 86] literature, we develop an architecture and training objective that accounts for all the essential components of this task: gauging the relatedness of any two instructional videos; capturing the *interchangeability* of their component steps; and interactively indexing into the alternatives with language.

Since there are no existing datasets labeled for video detours, we devise a framework leveraging large language models (LLMs) to generate weakly-supervised training data. Using HowTo100M [55], a large-scale instructional video dataset of in-the-wild how-to's accompanied by the transcribed speech of the narrator, we automatically generate plausible user queries at targeted timepoints in training source videos together with their detour counterparts in closely related videos. This procedure makes it possible to obtain ample effective training data without manual annotations. To rigorously evaluate the video detours, we introduce a gold-standard test set comprised of manually labeled data from 4K full-length videos and 16K human-generated questions.

In extensive experiments, we validate our model and illustrate the promise of the novel task. VidDetours strongly outperforms state-of-the-art video-language and video retrieval methods. We will release our training and test annotations to establish a formal benchmark for navigating instructional videos.

In short, ours is the first work to investigate personalized query-based navigation of instructional videos. Our main contributions are the innovative task definition, our video-language model to address it, and the high quality eval set and benchmark. These results help pave the way towards an interconnected how-to video knowledge base that would transcend the expertise of any one teacher, weaving together the myriad of steps, tips, and strategies available in existing large-scale video content.

2. Related Work

Learning from instructional videos. Several recent video datasets like HowTo100M [55], COIN [67], CrossTask [93] are based on instructional videos and have enabled research in procedure planning [11, 12, 89], task graph learning [6, 21, 31, 90], and alignment detection [5, 32]. The availability of large-scale instructional videos on the internet also facilitates video representation learning for action recognition [23, 29, 40, 43, 73], action anticipation [1, 24, 26, 28, 52], and object detection [3, 10]. All the prior work either focuses on short-term representations [14, 15, 50, 56, 75] or video-level understanding [6, 21]. To our knowledge, we are the first to establish a means to navigate across instructional videos, which is essential for holistic task understanding and optimal skill learning.

Vision and language learning. Videos often also contain text—whether converted from narrations through automatic speech recognition (ASR) [55, 67, 93] or manually annotated [30]. Using both text and video for representation learning [4, 14, 15, 44, 56, 75] helps in multi-modal tasks like retrieval [17, 22, 50, 75, 88], localization [47, 64, 75, 76, 81, 83, 84, 93], captioning [34, 42, 61, 71, 82], question answering [37, 38, 41, 78, 79], and episodic memory queries [30, 44, 64]. Most of these tasks focus on images or clip-level understanding, typically a few seconds long. Recent work [4, 7, 25, 30, 54, 62, 64] extends this further for video-level understanding spanning minutes. None of the existing methods or benchmarks answer text queries by navigating *between* long videos, as we propose.

Interactive retrieval. Dialog-based retrieval has been studied for fashion image retrieval [16, 33, 74] and conversation-based e-commerce shopping [66, 87], where a user wants a specific product and gives feedback on successive retrievals. In Visual Dialog [13, 18, 19, 35, 57], an agent is given an image and its caption and has to answer questions about the image e.g. *"what color is the mug?"* Recent work in composed image/video retrieval [8, 9, 36, 48, 69, 70, 77] uses an image/clip with a modification text to retrieve an improved version, e.g. a fountain image with text *"at night"* retrieves a clip of the fountain at night. Similarly, StepDiff [59] generates the difference between two clips in instructional videos.

All the previous work focuses on improving video or image retrieval through dialog, and the inputs are image or short-duration video. In contrast, we focus on action demonstrations where the prompt can be about ingredients, tools, or even step executions, e.g., *"how to prepare the mixture instead of using a pre-made mix?"* which is cru-

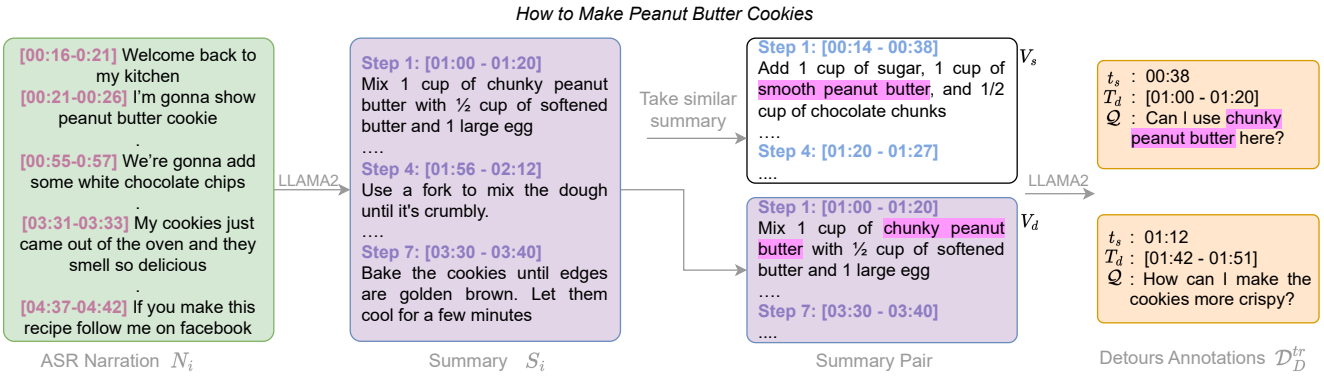

Figure 2. **Overview of the detours dataset ($\mathcal{D}_D^{tr}$) curation**. Given unlabeled instructional videos for training (we use HowTo100M [55]), we first input their narrations with timestamps to a language model (LLAMA2 [68]) to obtain summaries of their steps. Next, we automatically select pairs of similar summaries along with their timestamps and use a language model to generate weakly-supervised detours annotation tuples $(V_s, t_s, \mathcal{Q}, V_d, T_d)$. As an example, the source video here uses smooth peanut butter. A possible detour question is *"can I use chunky peanut butter here?"* and the window at T_d in the detour video (top right, orange) shows the use of crunchy peanut butter.

cial for a holistic task-level understanding of many actions and dependencies amongst them. Furthermore, our setup considers full instructional videos that typically span several minutes, as opposed to static images and short clips.

3. Approach

In this section, we first define our detour task formulation (Sec. 3.1). Next, we detail the dataset collection process (Sec. 3.2) and our model architecture (Sec. 3.3). Finally, we discuss implementation and training details (Sec. 3.4).

3.1. Video detour task formulation

We define a *video detour* as a mapping from a source video V_s at timestamp t_s to a response segment $T_d = (t_d^s, t_d^e)$ in a detour video V_d, based on a query text $\mathcal{Q}$. This is illustrated in Fig. 1, where after watching the source video for some time (purple bar, top panel), a user issues a query *"how to do this without an electric grill?"*, for which the response is a segment in a different video showing the step in a pan instead of a grill (green bar, bottom panel). By construction, V_s and V_d are related demonstrations from the same high-level task, that differ slightly in their demonstration (e.g., two videos demonstrating how to make chicken quesadillas).

Formally, we cast this as a video segment retrieval task conditioned on *both a source video and a query text*. The goal is to find functions $\mathcal{F}_R$ and $\mathcal{F}_L$ such that

$$V_d = \underset{V_i \in \mathcal{D}}{\operatorname{argmax}}\, \mathcal{F}_R(V_i | V_s[1:t_s], \mathcal{Q}) \quad (1)$$

$$T_d = \underset{T_i}{\operatorname{argmax}}\, \mathcal{F}_L(T_i | V_s[1:t_s], \mathcal{Q}, V_d) \quad (2)$$

where $V_s[1:t_s]$ refers to a video watched from the beginning, until time t_s, and T_i refers to a temporal window (start and end time) in video V_d.

Here, $\mathcal{F}_R$ is a *retrieval* mapping that finds the correct full instructional video, typically minutes long, given the source video segment and the text query, while $\mathcal{F}_L$ is the *localization* function that finds the start and end time in the detour video.

3.2. Detour dataset generation

Our goal is to learn retrieval and localization functions to find the correct detour segment given a source video and a query, for which we require a training dataset with tuples of the form $(V_s, t_s, \mathcal{Q}, V_d, T_d)$. The detour queries can focus on any aspect of the demonstration of the recipe: ingredients (e.g., *"can I add eggs here?"*), tools, (e.g., *"how to do this without an blender"*) or steps (e.g., *"how do I serve this on a plate?"*). Existing procedural video datasets only offer a subset of the required information [2, 55, 67, 93]: they provide narrations or keystep labels, but are missing interrelations between different procedural demonstrations and thus cannot be used for detours training directly. Moreover, collecting detour annotations at a large-scale may be impractical due to the amount of time required for annotators to watch and parse long detour videos.

To address this, we propose an approach to automatically create a training dataset $\mathcal{D}_D^{tr}$ for our task using how-to video narrations and language models. Subsequently, for rigorous testing, we also manually collect ground truth test data $\mathcal{D}_D^{te}$ from human annotators.

Weakly-supervised training set $\mathcal{D}_D^{tr}$. We start with a dataset of unlabeled instructional videos: $\mathcal{D} = \{(V_i, N_i)\}_{i=1}^{|\mathcal{D}|}$ containing *narrations* N_i in addition to the videos V_i. A narration is the spoken component of the how-to video, where the expert describes their actions ("now we mix for 3 minutes") and gives other commentary ("oh it looks great!"). We concentrate on the broad

domain of cooking due to its prominence in instructional video datasets (∼370K videos in HowTo100M [55]), well-structured recipes, and strong interconnection between different instances.

We use the narrations to generate labels for our training dataset $\mathcal{D}_D^{tr}$. Despite being noisy, narrations have been used successively for weakly-supervised training labels for video-language pretraining [50, 55, 56, 75] and keystep recognition [6, 45, 90]. Here we explore their utility for mining candidate detour pairs. We do this in two stages. See Figure 2. First, we generate text summaries for the key steps in each video. Specifically, we prompt LLAMA 2 [68], a recent open source large language model, to summarize the instructional videos using the timestamped narrations N_i. The prompt is of the form "Given the following narrations from a video, what recipe is this, and summarize each step along with its timestamps..." The exact text prompt is given in the Supp. along with example outputs. We obtain the summary of video V_i as a tuple of the step start time, end time, and text description. This process yields an intermediate summary dataset; see Figure 2 (second panel).

Next, we generate detour queries and time windows given a pair of summaries. For this, we identify video pairs that share an activity (and therefore have similar summaries), as unrelated pairs are unlikely to yield a meaningful detour query $\mathcal{Q}$. Specifically, we sort summary pairs by cosine similarity of their MPNet [65] sentence embeddings, discarding dissimilar pairs (score < 0.75). With a video pair in hand, we design a prompt for the LLM to generate detour queries and time windows $(t_s, \mathcal{Q}, T_d)$ of roughly the form "Given video summaries with timestamps, suppose a person is watching video A, identify a text prompt that a user might issue to take a detour and watch video B, along with the detour timestamps? Some examples of detours ...". See Supp. for full prompt details. Note that a source video can have multiple valid queries and matching detour videos, which our generation strategy allows.

The entire process yields a pair of videos and the detour annotations $(V_s, t_s, \mathcal{Q}, V_d, T_d)$ (Figure 2, right panel). More examples are in Supp. While $\mathcal{D}_D^{tr}$ will naturally have some noise due to language model errors and misaligned or non-visual narrations [5, 32], we find them quite reasonable (85% satisfactory) from manual inspection of a subset of the data, and our dataset creation strategy diminishes the noise. Most importantly, they are effective for training a detour model, as our experiments testing on manually labeled data will show.

Manually collected testing set $\mathcal{D}_D^{te}$. While the weakly-supervised data is sufficient for training, for reliable evaluation of our trained models and baselines, we manually collect ground truth test data. Similar to $\mathcal{D}_D^{tr}$, we identify a pair of similar videos, and ask the professional annotators to watch the videos completely, and then annotate

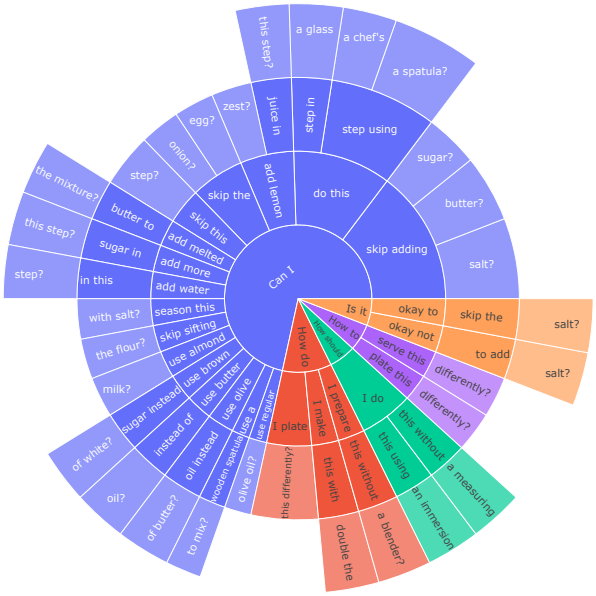

Figure 3. **Visualization of most frequent bigrams of the queries in the manually annotated test set**. We see that most of the queries have little or no context about the current recipe and the step being executed, e.g. *"how do I plate this differently?"* or *"can I do this step using a spatula?"*—emphasizing the need for source video context, as we explore in the proposed model.

$(t_s, \mathcal{Q}, T_d)$. Annotators are allowed to reject pairs for which detours cannot be constructed (e.g., if they are too dissimilar). Since there can be multiple detours possible for a given pair of videos, we ask the annotators to identify at least three detours. We ensure that the videos in the train and test set are disjoint. The annotation process results in a high quality, benchmark test set for our new task.

Figure 3 shows the diversity in queries that annotators provide. We see a variety of questions arise about substitutions of ingredients, tools, and steps. Further, we see contextual queries, e.g. *"can I skip adding butter?"* that do not reveal much information about the recipe and the current step, underscoring the need to reference the source video when localizing a detour.

3.3. Detour retrieval and localization modules

Next, we describe our training framework to learn the mapping functions $\mathcal{F}_R$ and $\mathcal{F}_L$ using our generated detour dataset. Our objective is to design a multimodal (video and text) architecture that fuses the source video context with the language query enabling $\mathcal{F}_R$ and $\mathcal{F}_L$ to utilize both the viewing history and the query. As we will see in the experiments, this idea is more effective than late fusion of video and query features [7, 63, 72]. For this, we leverage the reasoning capabilities of large language models (LLM). In short, we encode both videos and the detour query as a sequence of tokens to pass to the LLM, which aggregates mul-

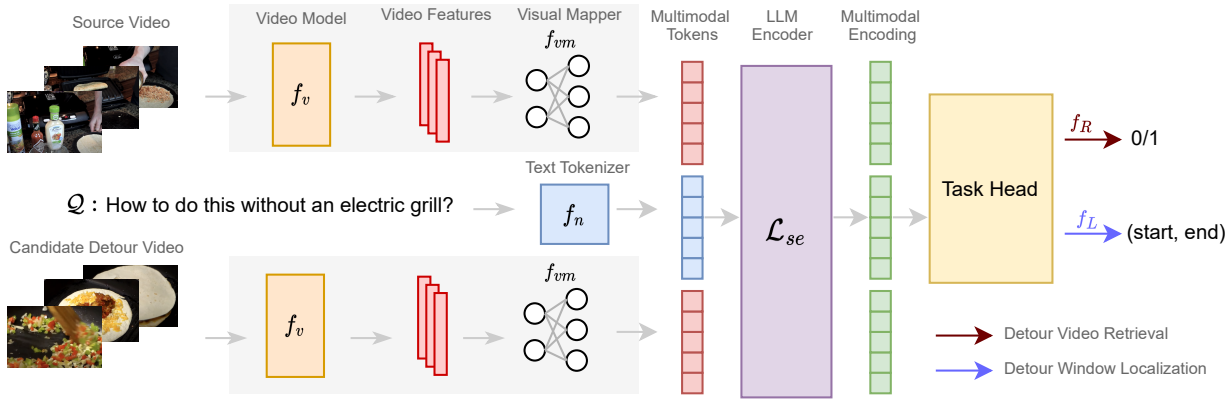

Figure 4. **Overview of the proposed approach**. The source video and the candidate detour video are converted to visual tokens by passing through a video encoder f_v, followed by a visual mapper f_{vm}. We obtain a similar text token using standard tokenizer f_n. The processed tokens are then passed through a multimodal sequence encoder $\mathcal{L}_{se}$ to obtain output features. Finally, we have specific task heads for detour video retrieval and detour window localization.

timodal information from the inputs in its output encoding. These encodings are finally used to retrieve detour videos and detour segments using task specific heads. Our overall framework is shown in Fig. 4 and each component is described in detail below.

Tokenizing videos and detour queries. We begin by encoding the videos as a sequence of tokens compatible with the LLM. We encode each video as spatio-temporal features using a video encoder f_v (e.g., InternVideo [72]), extracted at one feature per second [6, 75] Next, we use a low-parameter visual mapper f_{vm} [27, 39, 46, 51, 85] to convert the video features into visual tokens $\mathbf{v} = f_{vm}(f_v(V))$ that are in the same embedding space as the text tokens, making them compatible with language encoders. Finally, we encode the detour query into text tokens using a standard text tokenizer (e.g. [20, 68]) $\mathbf{n}_{\mathcal{Q}} = f_n(\mathcal{Q})$. We now have both visual and query tokens in the same space to feed into our language models.

LLMs as multimodal sequence encoders. Next, we use a LLAMA2 model as our multimodal sequence encoder ($\mathcal{L}_{se}$). LLMs are an ideal choice here, given our goal of encoding *dialogue-driven detours*—the idea that a user is watching the source video and pauses it to ask a query, followed by a response from the LLM. Specifically, we append source video tokens (until time t_s) with the query tokens $\mathbf{v}_s[1:t_s] \mid \mathbf{n}_{\mathcal{Q}}$, followed by candidate response video tokens. The multimodal encoder captures the cross-modal interactions between the source video, query context and the candidate video, resulting in an encoded output for the candidate video $O_i = \mathcal{L}_{se}\left(\mathbf{v}_s[1:t_s] \mid \mathbf{n}_{\mathcal{Q}} \mid \mathbf{v}_i\right)$.

Detour retrieval and localization heads. Finally, we use the updated multimodal encodings O_i to score detour video candidates and segments using two task heads f_R and f_L. f_R is a classifier that scores the relevancy of candidate video V_i given the viewing context and the user query, while f_L is a classifier that identifies the highest score time segment T_i inside the correct detour video V_d, given O_d. These two heads are trained separately for each task.

For our video retrieval network $\mathcal{F}_R$, we minimize the binary cross entropy loss f_{BCE} between the prediction and the ground truth label:

$$\min_{V_i \in \mathcal{D}} f_{BCE}(\mathcal{F}_R(V_i|V_s[1:t_s], \mathcal{Q}), \mathbb{1}(V_i, V_d)).$$

We assign a positive label to the correct detour video V_d and negative to other videos sampled from the dataset, i.e. $V_i \neq V_d$. In particular, for every correct training instance, we randomly sample an incorrect video from which to curate a negative sample—either from the same task (hard negatives) or other tasks.

The localization network $\mathcal{F}_L$ training objective is:

$$\min_{T_i=[t_i^s,t_i^e]} \frac{1}{2}\left[f_{CE}(t_i^s, t_d^s) + f_{CE}(t_i^e, t_d^e)\right],$$

where f_{CE} is the cross-entropy loss. Similar to the video-language grounding model VSLNet [84], this objective minimizes the error in the distribution of the start and end times across the video. At inference, we find the candidate video V_d and time duration T_d that maximizes the scores from $\mathcal{F}_R$ and $\mathcal{F}_L$, respectively (Sec. 3.1). Note that there may be multiple plausible detour videos for a given source and query (e.g., multiple videos can use a *heating pan* instead of an *electric grill*); our scoring-based approach ensures other related (and valid) pairings can also score highly.

3.4. Implementation details

Dataset and statistics. Our training and test sets are both derived from HowTo100M [55] (average length 6.5

Method	R@5	R@10	R@50	MedR ↓
Text-only	3.9	8.7	14.0	512
CLIP [63]	7.9	11.8	25.2	342
CLIP-Hitchhiker [7]	8.4	12.3	25.6	336
InternVideo [72]	9.7	13.2	27.2	313
Distant Supervision [45]	8.4	12.6	25.1	329
Multi-modal LLM [80]	5.9	10.5	32.1	139
CoVR [69]	4.3	9.2	15.3	473
Ours	**17.6**	**27.8**	**62.4**	**30**
Ours w/o hard-negatives	16.5	24.9	56.3	55
Ours w/ parser	13.9	21.6	50.0	81

Method	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	Mean R@1
Text-only	5.2	2.7	0.6	4.2
2D-TAN [86]	10.3	4.2	1.5	8.6
VSLNet [84]	11.8	5.8	1.7	9.4
UMT [47]	12.0	6.1	1.6	9.4
Distant Supervision [45]	10.6	4.0	1.5	8.3
Multi-modal LLM [80]	12.7	6.5	1.8	10.2
STALE [58]	12.1	6.1	1.7	9.6
Ours	**16.7**	**7.7**	**2.8**	**12.8**
Ours w/ parser	13.4	7.0	2.5	11.6

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

mins). We consider cooking tasks (i.e., recipes) containing $370K$ videos. We use the weakly-supervised automatically curated dataset $\mathcal{D}_D^{tr}$ for training and validation. Following the steps described in Sec. 3.2, we obtain $586,603$ training and $18,308$ validation detour annotation tuples $(V_s, t_s, \mathcal{Q}, V_d, T_d)$.

The manual annotation results in a large-scale test dataset $\mathcal{D}_D^{te}$ containing $16,207$ detour instances $(V_s, t_s, \mathcal{Q}, V_d, T_d)$. The test set is based on $3,873$ unique videos across $1,080$ recipes/tasks, e.g. *"how to make chicken quesadillas"* is one task and has multiple video instances. We curate the test set so that there are 834 *common* tasks $(14,450/16,207$ annotations) with the training data $\mathcal{D}_D^{tr}$ having videos from those tasks. There are 246 additional *novel* recipes $(1,757/16,307$ annotations) that do not appear in the training set. For retrieval evaluation, the detour candidates are all videos in the dataset, i.e. $3,873$ candidates per detour annotation. No video exists in both the training and testing split.

Network architecture. We use InternVideo [72] as the video feature extractor f_v. The features are extracted at one feature per second, following [32, 75]. f_v is frozen and f_{vm} is a trainable linear layer, inspired by [46]. $\mathcal{L}_{se}$ is a LLAMA2-13B-chat [68] language model. We try both variants of keeping $\mathcal{L}_{se}$ frozen and trainable; performance is better if $\mathcal{L}_{se}$ is trainable. Lastly, f_n is the LLAMA-2 tokenizer. The retrieval head f_R is a transformer classifier and we take the CLS token of the output followed by a linear layer to output the score. Finally, the localization head f_L is the VSLNet [84] architecture. We remove the tokenizer from VSLNet since the text input is already processed.

Training parameters. We train both networks on 8 nodes with 8 NVIDIA A100 GPUs for 5 epochs. The training time is 8 hours. We use AdamW [49] optimizer with learning rate 3×10^{-5} and batch size of 16 per device. The transformer classifier f_R uses an input dimension of 4096 (consistent with the LLAMA2 output dimension), 4 heads, 4 layers and 1024 dimensional feed-forward network. All other parameters are defaults of the respective models.

4. Experiments

We show the results for the detour video retrieval (Sec. 4.1 and detour localization (Sec. 4.2) subtasks.

4.1. Detour video retrieval

First, we benchmark models on finding the correct detour video given the source video and the query, from amongst all videos in the test set (3,873 videos).

Baselines. We adapt state-of-the-art video retrieval methods for detour retrieval. All the baselines embed text and video in a shared space using an encoder ϕ. We find the detour video that has the most similar embedding to a reference $V_d = \text{argmax}_{V_i \in \mathcal{D}} \langle \phi(V_i), \psi(V_s, \mathcal{Q}) \rangle$, where $\psi(V_s, \mathcal{Q})$ computes the reference embedding from the source and query videos, and $\langle\rangle$ is cosine-similarity. We evaluate three variants corresponding to different inputs:

- *with* $V_s, \mathcal{Q}$: where $\psi(V_s, \mathcal{Q}) = 1/2[\phi(V_s) + \phi(\mathcal{Q})]$.
- *with* V_s: where $\psi(V_s, \mathcal{Q}) = \phi(V_s)$.
- *with* $\mathcal{Q}$: where $\psi(V_s, \mathcal{Q}) = \phi(\mathcal{Q})$.

These variants test whether individual embeddings (or simple combinations of them) are sufficient for detour retrieval.

- **Text-only** computes the similarity between summary and query embeddings, ignoring visual cues. This baseline evaluates the impact of text-bias in the automatically curated dataset.
- **CLIP [63], InternVideo [72]** are state-of-the-art vision-language models used extensively for multi-modal tasks. Since they take short video clips as input, the video representation is the average of all short-term features.
- **CLIP-Hitchhiker [7]** is similar to CLIP [63] but uses a weighted average of frame features, instead of uniformly averaging for $\phi(V_i)$. The weights are the similarity score between the frame and query features (eq. 1 in [7]). Note that the *source-only* variant is not evaluated as query features are needed to compute the weights.
- **Distant Supervision [45]** uses WikiHow as an external knowledge base to map steps in the video with keysteps. We replace the narrations with this keystep assignment for the detour dataset generation.
- **Multi-modal LLM [80]** can be used to generate dense

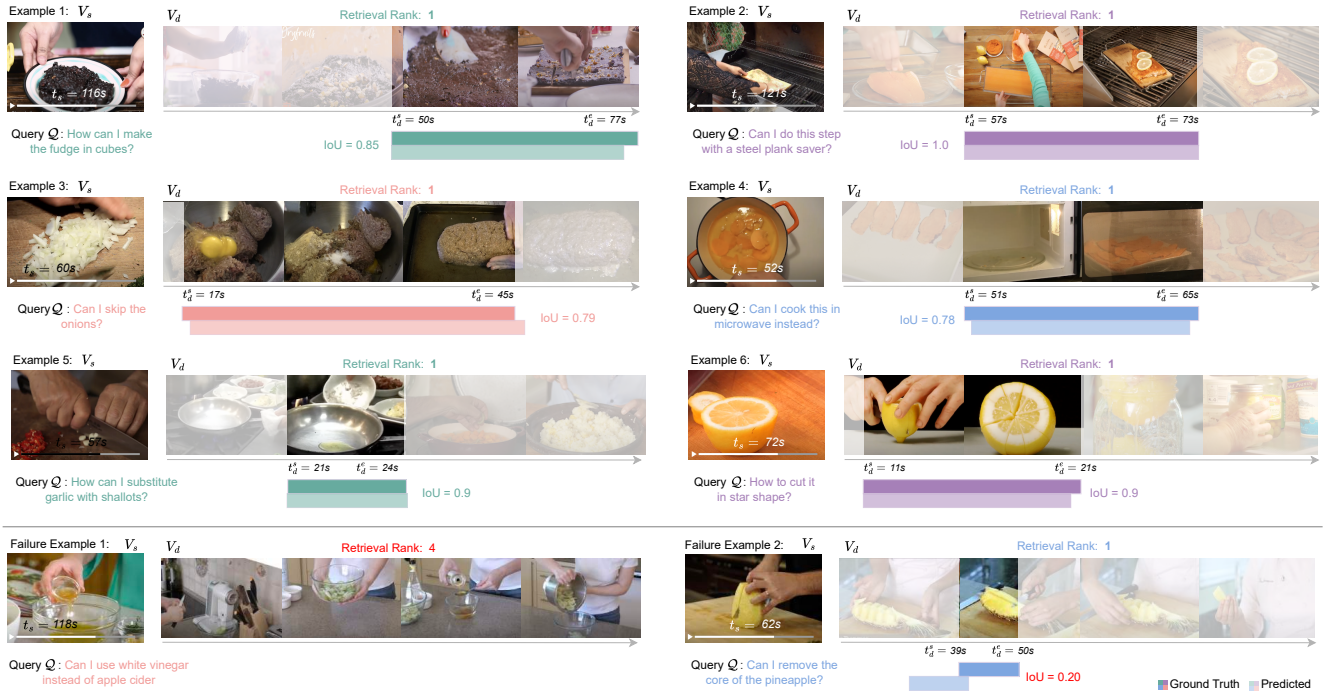

Figure 5. **Visualization of our model's predictions** for the detour video retrieval and window localization tasks, including failure cases (last row). For all the successful examples, our method ranks the detour video at top-1. Furthermore, in the detour window localization, our method is able to predict the detour window with a high overlap (> 0.7 IoU). Our model is able to correctly use source video context and nuanced queries like *"how to cut it in star shape?"* (Example 6) and *"can I skip the onions?"* (Example 3). Best viewed with zoom.

captions to replace the narrations. We use Vid2Seq [80] to first densely annotate the videos and convert the task to text-only video detours using those captions.

- **CoVR [69]** is a state-of-the-art method for composed video retrieval. Following CoVR, we sample a frame from the source video and use the user query as the modification text to obtain detour retrieval.

Note that we train [45, 69, 80] on the same detour dataset as our model, whereas [7, 63, 72] are used in zero-shot setting for its good retrieval capabilities in instructional videos.

Ablations. Similar to how we evaluate the baselines with different inputs, we compare the performance of our method (i.e. "with $V_s, \mathcal{Q}$") with ablations—"with V_s" and "with $\mathcal{Q}$". Next, recall that we use a language model to generate timestamps for both the summaries and the detour timestamps (see Fig. 2 and Sec. 3.2), which can be noisy. Instead, for this ablation, we parse the summaries and queries and assign timestamps based on the similarity score between the narrations and the summaries. The language model does not handle timestamps for this "Ours w/ parser" baseline. Finally, we evaluate the retrieval performance without sampling negatives from the same task, i.e. no hard negatives in "Ours w/o hard negatives" baseline.

Metrics. Following standard retrieval evaluation [53, 55, 56, 75], we report recall@k for $k \in [5, 10, 50]$ and the median rank. Recall@k ranges in $[0, 1]$, higher the better. The median rank ranges between 1 and the size of the retrieval candidate set $(3,873)$, the lower the better.

Results. Table 1 (left) reports the results with both inputs, i.e. "with $V_s, \mathcal{Q}$" for all baselines on all metrics. We also report medR for different inputs in Table 2 (left). Our method outperforms all the baselines by a significant margin. Our performance gain is 7.9% w.r.t. InternVideo [72] at R@5 and the gap further increases to **35.2%** for R@50. Across all the different input combinations, InternVideo [72] "with $\mathcal{Q}$" is the second best and attains a medR of 138 — much lower than our medR of 30 (Tab. 2).

We attribute our large gain to three crucial factors: (a) thanks to its tokens and long sequence length in the LLM, our model is capable of long video understanding, unlike CoVR [69], which samples a few frames from the video; (b) appropriately using the source video context V_d, unlike CLIP [63], CLIP-Hitchhiker [7], and InternVideo [72], since our model is more than a standard text-to-video retrieval model; and (c) conditioning the retrieval on the query text. Furthermore, we find that using narrations for generating a weakly-supervised training set is better than using keysteps [45] or captions [80], which miss details typically present in narrations.

Fig. 5 shows example detours inferred by our model. In all the successful cases, our retrieval model ranks the correct detour video at top-1. It is evident that detour video re-

Method	V_s	$\mathcal{Q}$	MedR ↓
CLIP [63]	✓		314
		✓	191
	✓	✓	342
CLIP-Hitch. [7]	✓		—
		✓	186
	✓	✓	336
InternVideo [72]	✓		150
		✓	138
	✓	✓	313
DistantSup. [45]	✓		384
		✓	370
	✓	✓	329
MLLM [80]	✓		189
		✓	158
	✓	✓	139
CoVR [69]	✓		388
		✓	401
	✓	✓	473
Ours	✓		128
		✓	116
	✓	✓	**30**

Method	V_s	$\mathcal{Q}$	R@1
2D-TAN [86]	✓		5.5
		✓	8.0
	✓	✓	8.6
VSLNet [84]	✓		6.1
		✓	8.5
	✓	✓	9.4
UMT [47]	✓		6.5
		✓	8.7
	✓	✓	9.4
DistantSup. [45]	✓		7.6
		✓	7.9
	✓	✓	8.3
MLLM [80]	✓		9.1
		✓	9.7
	✓	✓	10.2
STALE [58]	✓		6.9
		✓	8.8
	✓	✓	9.6
Ours	✓		8.9
		✓	11.2
	✓	✓	**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). Our method outperforms all the prior works for all input combinations. See Supp. for all metrics.

trieval is more challenging than conventional text-to-video-retrieval because the detour query can have missing context, e.g. *"how do I add garnish after pouring?"* does not reveal the task in the video; the previous video context is needed for the correct retrieval. The failure cases (bottom row) show example errors in retrieval rank (left) and localization (right). The performance trend is similar for *common* and *novel* tasks of the dataset (see Supp.), reinforcing that training on a sufficiently large dataset with many recipes enables good generalization.

4.2. Detour window localization

Next, we show results on detours window localization where the task is to determine the correct window given the query, the source video, and the ground-truth detour video.

Baselines. While there are no existing methods that use previous minutes-long viewing history along with the query to perform temporal localization, we use state-of-the-art localization methods as baselines and also strengthen them by providing source video context, as described below.

- **Text-only:** Similar to the detours video retrieval, we have a text-only baseline to evaluate the text-bias in the automatically curated train set.
- **2D-TAN [86], VSLNet [84], UMT [47]:** All these prior models aim to localize text in videos. To apply them for detour window localization, we train them to accept video context as input, namely using a visual mapper f_{vm} similar to our approach—thus providing a late fusion of video V_s context. These visual tokens are prepended to the text token, same as our token sequence (see Fig. 4). This enhancement enables us to evaluate "with V_s" and "with V_s, $\mathcal{Q}$" in addition to the standard "with $\mathcal{Q}$"
- **STALE [58]:** This is *zero-shot* temporal detection method uses vision-language prompting. Same as above, we evaluate this baseline at three combination on inputs.

The baselines in [47, 84, 86] are trained on our same detour dataset for localization, whereas [58] is zero-shot.

Metrics. Following [30, 84, 86], we report recall@1 for IoU thresholds in $[0.3, 0.5, 0.7]$. We also report the recall@1 at the average IoU. All the recall metrics range in $[0, 1]$, higher the better.

Results. Tab. 1 shows the results with greatest input context ("with V_s, $\mathcal{Q}$"), and Tab. 2 (right) compares against different input combinations. We outperform all the baselines and ablations by a clear margin. Our mean recall@1 is 3.2% higher than the second best performing method, STALE [58]. The same trend is true for all IoU thresholds, with higher thresholds having lower recall, as expected. Again, the LLM-based parser is better for obtaining the timestamps, and using narrations is better than an external knowledge base or captions.

Our gains can be attributed to our model design involving early fusion of previous viewing context and the query features. The detour queries have less context than in a typical text localization task. For example, *"can I skip adding salt here?"* requires a model to first understand the step being done in the source video, followed by interpreting the query to localize a *similar* step without salt. Existing methods are incapable of capturing this multimodal dependency, even if we strengthen them with late fusion. We also see that our mean recall@1 performance is better than all other methods at all input combinations, including ablations.

Fig. 5 also shows example detour localizations. We see that our method is able to use the source video context and the user query to localize the detour window. For example, when a user asks *"can I do this step with a steel plank saver"*, our model correctly localizes the use of a steel plank saver to put the salmon of the grill in the target video.

5. Conclusion

We propose a novel task of finding detours for navigating instructional videos. Building on video and language modeling, we develop a weakly-supervised training dataset and a novel method to train a detours network. Our results show how existing methods are insufficient to address this problem. The dataset will be released to the community to support research in navigating instructional videos.

Acknowledgements: UT Austin is supported in part by the IFML NSF AI Institute. KG is paid as a research scientist by Meta. We thank Suyog Jain, Austin Miller, Honey Manglani, and Robert Kuo for help with the data collection.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 2

[2] Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. Ht-step: Aligning instructional articles with how-to videos. In *NeurIPS*, 2023. 3

[3] Elad Amrani, Rami Ben-Ari, Inbar Shapira, Tal Hakim, and Alex Bronstein. Self-supervised object detection and retrieval using unlabeled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 954–955, 2020. 2

[4] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078, 2023. 2

[5] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. What you say is what you show: Visual narration detection in instructional videos. *arXiv preprint arXiv:2301.02307*, 2023. 1, 2, 4

[6] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 4, 5

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 2, 4, 6, 7, 8

[8] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474, 2022. 2

[9] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv preprint arXiv:2303.15247*, 2023. 2

[10] Gedas Bertasius and Lorenzo Torresani. Cobe: Contextualized object embeddings from narrated instructional video. *Advances in Neural Information Processing Systems*, 33: 15133–15145, 2020. 2

[11] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 1, 2

[12] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 334–350. Springer, 2020. 1, 2

[13] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2017. 2

[14] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 2

[15] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023. 2

[16] Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. *arXiv preprint arXiv:2211.07394*, 2022. 2

[17] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 2

[18] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 2

[19] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[21] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 319–335. Springer, 2022. 1, 2

[22] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2

[23] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[24] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2

[25] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2782–2795, 2013. 2

[26] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2

[27] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 5

[28] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 2

[29] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 2

[30] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 8

[31] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024. 2

[32] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2, 4, 6

[33] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5047–5056, 2019. 2

[34] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2

[35] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[36] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012. 2

[37] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 2

[38] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2

[39] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 5

[40] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022. 2

[41] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[43] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2

[44] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 2

[45] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 1, 4, 6, 7, 8

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 5, 6

[47] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 2, 6, 8

[48] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 2

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[50] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2, 4

[51] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 5

[52] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting@ ego4d challenge 2022. *arXiv preprint arXiv:2207.12080*, 2022. 2

[53] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations, 2023. 7

[54] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2

[55] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1, 2, 3, 4, 5, 7

[56] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 1, 2, 4, 7

[57] Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. Improving generative visual dialog by answering diverse questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 2

[58] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. *arXiv e-prints*, pages arXiv–2207, 2022. 2, 6, 8

[59] Tushar Nagarajan and Lorenzo Torresani. Step differences in instructional video. In *CVPR*, 2024. 2

[60] Medhini Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1

[61] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. *arXiv preprint arXiv:2207.09666*, 2022. 2

[62] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 612–619, 2014. 2

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4, 6, 7, 8

[64] Santhosh Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *CVPR*, 2023. 2

[65] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 4

[66] Yueming Sun and Yi Zhang. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 235–244, 2018. 2

[67] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2, 3

[68] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 3, 4, 5, 6

[69] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. *arXiv preprint arXiv:2308.14746*, 2023. 2, 6, 7, 8

[70] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 2

[71] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[72] 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, 2022. 4, 5, 6, 7, 8

[73] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[74] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference*

on computer vision and pattern recognition, pages 11307–11317, 2021. 2

[75] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2, 4, 5, 6, 7

[76] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13623–13633, 2023. 2

[77] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 776–791. Springer, 2016. 2

[78] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 2

[79] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022. 2

[80] 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 *CVPR*, 2023. 6, 7, 8

[81] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 2

[82] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 2

[83] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *arXiv preprint arXiv:2202.07925*, 2022. 2

[84] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020. 2, 5, 6, 8

[85] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 5

[86] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020. 2, 6, 8

[87] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186, 2018. 2

[88] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. *arXiv preprint arXiv:2205.00823*, 2022. 2

[89] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. *arXiv preprint arXiv:2303.17839*, 2023. 1, 2

[90] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10727–10738, 2023. 1, 2, 4

[91] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[92] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 2

[93] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 2, 3