Relational Space-Time Query in Long-Form Videos

Xitong Yang[1] Fu-Jen Chu[1] Matt Feiszli[1] Raghav Goyal[1,2] Lorenzo Torresani[1] Du Tran[1]
[1] Meta AI [2] University of British Columbia

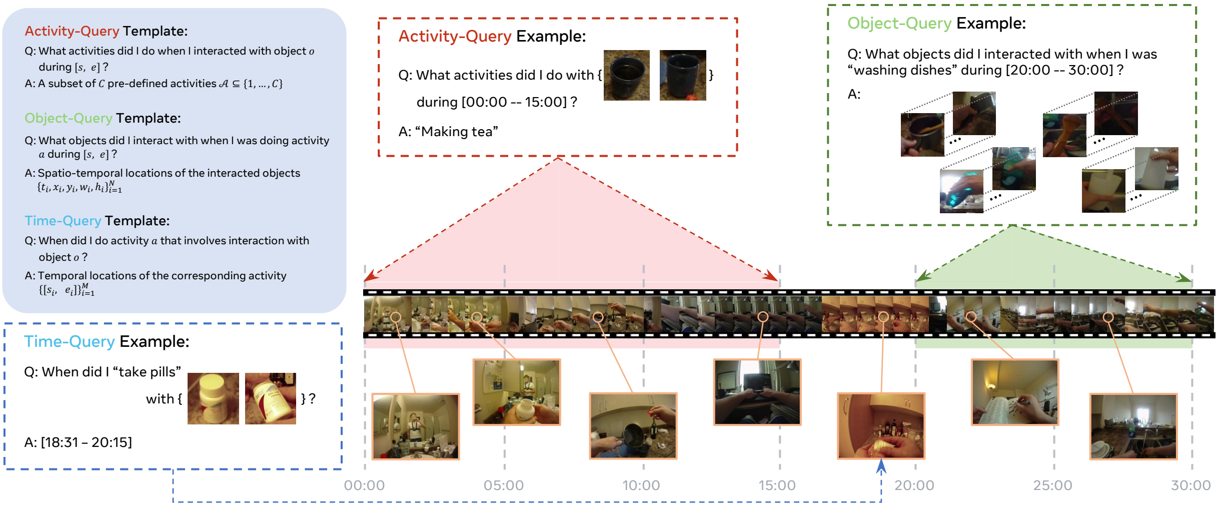
Activity-Query Template:
Q: What activities did I do when I interacted with object o during $[s,\ e]$?
A: A subset of C pre-defined activities $\mathcal{A} \subseteq \{1, \dots, C\}$
Object-Query Template:
Q: What objects did I interact with when I was doing activity a during $[s,\ e]$?
A: Spatio-temporal locations of the interacted objects $\{t_i, x_i, y_i, w_i, h_i\}_{i=1}^{N}$
Time-Query Template:
Q: When did I do activity a that involves interaction with object o ?
A: Temporal locations of the corresponding activity $\{[s_i,\ e_i]\}_{i=1}^{M}$
Time-Query Example:
Q: When did I "take pills" with { } ?
A: [18:31 – 20:15]
Activity-Query Example:
Q: What activities did I do with { } during [00:00 -- 15:00] ?
A: "Making tea"
Object-Query Example:
Q: What objects did I interacted with when I was "washing dishes" during [20:00 -- 30:00] ?
A:
00:00
05:00
10:00
15:00
20:00
25:00
30:00

Figure 1. Illustration of the three types of queries in our Relational Space-Time Query (ReST) framework. Given a long video spanning up to 30 minutes, a set of queries are provided to assess a model's ability to understand activities, objects, and their interactions in the video. All queries and answers are generated in the form of pre-defined templates (top-left) to avoid the ambiguity and bias introduced by language input / output. Note that ReST is a holistic framework that supports constructing queries with different levels of complexity beyond the three basic types described in this paper.

Abstract

Egocentric videos are often available in the form of uninterrupted, uncurated long videos capturing the camera wearers' daily life activities.Understanding these videos requires models to be able to reason about activities, objects, and their interactions. However, current video benchmarks study these problems independently and under short, curated clips. In contrast, real-world applications, e.g. *AR assistants, require bundling these problems for both model development and evaluation. In this paper, we propose to study these problems in a joint framework for long video understanding. Our contributions are three-fold. First, we propose an integrated framework, namely* **Re**lational **S**pace-**T**ime *Query (ReST), for evaluating video understanding models via templated spatiotemporal queries. Second, we introduce two new benchmarks, ReST-ADL and ReST-Ego4D* [1]*, which augment the existing egocentric video datasets with abundant query annotations generated by the ReST framework. Finally, we present a set of baselines and in-depth analysis on the two benchmarks and provide insights about the query tasks. We view our integrated framework and benchmarks as a step towards comprehensive, multi-step reasoning in long videos, and believe it will facilitate the development of next generations of video understanding models.*

1. Introduction

Thanks to the advances of modern, massive parallel hardware, *e.g.* GPUs, and the availability of large datasets, significant progress has been made in the last few years with large language models (*e.g.*, GPT-3 [6], BERT [11]) and image / video generative models (*e.g.*, DALLE [40], Imagen [41], Make-A-Video [46]). Meanwhile, current video understanding models mostly focus on processing short video clips [12, 13, 52] and solving basic perception tasks such as action recognition [16, 28, 31, 44, 48] and detection [7, 19]. One may ask the questions for video understanding research: "How far are current models progressing to a human-level performance on video understanding?", or

[1] The latest version of our benchmark and models will be available here.

"What is blocking us from building models that can understand complex relationships in long videos?"

Of course, there exists multiple blockers in practice such as GPU memory limitation and inefficient hardware support for processing long videos. Yet the first and most important reason is always the lack of the *right research problem* and the *right benchmark*. One drawback of current video understanding benchmarks [19, 28] is that they handle analysis of activities, objects and their interactions in a separate manner. However, understanding long-form videos usually requires a unified analysis of these factors because activities manifesting within these uncurated videos are primarily in the form of human-object interaction, especially for egocentric recordings of a camera wearer's daily lives [17]. In recent years, video-QA [18, 33, 49, 66] and video captioning [10, 50] have been proposed as alternative tasks for video understanding. These tasks require models to understand both visual and text modalities and perform cross-modal reasoning. On one hand, such vision-language based tasks have the benefit of bypassing the pre-defined taxonomy and closed-world assumptions by leveraging language as input and/or output. On the other hand, using language for vision tasks, either in the form of input query or output prediction, brings additional ambiguity in text generation and requires use of uninterpretable evaluation metrics (*e.g.*, BLEU, ROUGE). Language priors also introduce bias to the task as observed in prior work that the language-only model achieves comparable results with the VQA ones [4, 18].

In this paper, we present a holistic framework, **Re**lational **S**pace-**T**ime Query (ReST), for evaluating video understanding models via templated spatiotemporal queries. By combining analysis of activities, objects, and their interactions, our ReST framework captures much of the rich expressivity of natural language query while avoiding the ambiguity and bias introduced by the language input / output. Figure 1 illustrates an example of our ReST framework. Given a long video spanning up to 30 minutes, we evaluate a video understanding model by asking various queries about the activities and human-object interactions occurred in the video. Unlike VQA that relies on language-based questions and answers, all of our queries and answers are constructed in the form of pre-defined templates with visual or categorical input. Such a design helps the evaluation remain pure vision-focused and enjoy well-defined, well-posed evaluation metrics. Queries constructed in ReST can cover various questions with different levels of complexity. As shown in the examples in Figure 1, the questions can be: "what activities did I do with the coffee mug?", "what objects did I interact with when I was washing dishes?" or "when did I take the pills stored in this specific bottle?". We note that these questions are templated and only the time in square brackets, the activities in the double quotes, and the image crops in the curvy brackets are allowed to vary to form different questions. In order to perform well on these query tasks, a model needs not only be able to process the long video efficiently, but is also required to understand the temporal dependencies of activities and the fine-grained human-object interactions across the video.

We summarize our contributions as follows. We present **Re**lational **S**pace-**T**ime Query (ReST), a holistic framework and benchmarks for long video understanding with detailed-reasoning tasks. We provide annotations for our ReST queries on two existing benchmarks, ADL [37] and Ego4D [17]. We conduct a set of comprehensive baselines and analysis for ReST to gain insights into the tasks. We find that even with the initial set of the three basic tasks, current state-of-the-art models are not meeting desired performance, which indicates the need for further research and opportunities in this field.

2. Related Work

Our work is related to a broad range of prior studies on video understanding, especially in the areas of action recognition, detection, and question answering. This section provides an overview of well-known benchmarks and models that have been developed for these tasks.

Action recognition and detection. Action recognition has been one of the fundamental and fast growing research areas in video understanding. Since the introduction of 3D Convolutional Neural Networks (CNNs) [26, 51] to video classification, various CNN-based models have been proposed to learn better spatiotemporal representations [8, 13, 52, 59, 63]. Two-stream architectures [8, 14, 45] and motion extraction modules [32, 53, 65] are also introduced to better model motion information from raw video frames. Recently, Transformer-based models [3, 5, 12, 34] have shown promising results. Building upon the frame/clip-based features extracted from the classification backbones, action detection can be achieved by temporal localization models [9, 35, 68, 71] or spatiotemporal detection models [27, 64, 72].

While significant progress has been made, it is noteworthy that most of the action classification models are designed for conventional video benchmarks which comprise short videos spanning from a few seconds to three minutes, such as HMDB51 [31], UCF-101 [48], Kinetics [28], Something-something [16], and Charades [44]. Even for action detection or the recent works on "long-form video modeling" [23, 42, 55–57, 63], which involve datasets with untrimmed videos and longer durations (*e.g.*, THUMOS [24], ActivityNet [7], AVA [19], LVU [56]), they are still processing videos with duration shorter than 5 minutes and limited to basic classification or detection tasks.

Visual reasoning. Recent years have witnessed growing interest in visual reasoning tasks, most of which involve video-language understanding, such as image captioning [36], question answering (QA) [1, 67] and language grounding / query [17, 69]. Specifically, in the video domain, MovieQA [49] is one of the earliest works that explore video understanding via QA problems. After that, various QA benchmarks have been proposed by collecting human-generated questions on either video clips extracted from TV series (*e.g*., TVQA [33]), GIFs (*e.g*., TGIF-QA [25] or internet videos (*e.g*., ActivityNet-QA [66], Next-QA [58], Just-Ask [61]). There are also datasets that automatically generate question-answer pairs from descriptions (*e.g*., MSRVTT-QA [60], MSVD-QA [60]) or scene graphs (*e.g*., AGQA [18]). Language grounding / query is a related task that involves generating the (spatio-)temporal location of a language input, instead of the language answer. Several popular benchmarks, such as Charades-STA [15] and VidSTG [70], have been proposed in prior work.

In the recently published dataset, Ego4D [17], the authors propose three episodic memory tasks. (1) Visual Query (VQ): locating the most recent spatio-temporal tube corresponding to a query object; (2) Moment Query (MQ): locating the temporal segments corresponding to a moment category; (3) Natural Language Query (NLQ): locating the temporal segments corresponding to a language question. The three tasks can be cast to object retrieval and tracking, activity detection and language grounding, respectively. A QA dataset QAEgo4D [4] is later built upon the NLQ subset by collecting additional natural language answers.

Unlike all the these prior work, our ReST benchmarks emphasize the *joint analysis* of objects, activities, and human-object interactions in long-form, egocentric videos. In addition, ReST is designed to be entirely focused on vision, both in terms of the input and output of the query. We will provide additional elaboration on the distinctions between ReST and Ego4D episodic memory tasks in Sec. 3.

3. Relational Space-Time Query for Long Video Understanding

The key idea of Relational Space-Time Query (ReST) is to provide a unified framework for the analysis of activities, human-object interactions and eventually long video understanding. Drawing inspiration from recent advances in visual reasoning [2, 17, 49], we formulate the problem as a set of query tasks [2] that require the model to predict structured answers to input queries. In particular, we describe a basic *event* in an egocentric video using the following general form:

“I engaged in $\langle$*activity*$\rangle$ while interacting with $\langle$*object*$\rangle$ during $\langle$*time period*$\rangle$.”

This form naturally integrates the occurrence of activities and human-object interactions while also grounding them to specific time periods. In our framework, we denote an activity c as a category label defined by a close-set taxonomy, and a time period t as a temporal segment with a starting time s and an ending time e. An object o is represented as image crops of the object in selected frames. This is in fact more precise and concise than using natural language to refer to a specific object instance, as discussed in [17, 67].

We construct three basic types of queries by asking about one of the three key properties in the query: activity, object, or time. In other words, to answer the query, a model is required to understand two of the key properties *simultaneously*. This is in contrast with most existing tasks that study different vision problems in isolation. We believe that this multi-modal, compositional nature of our queries is the key to providing a unified framework for measuring holistic video understanding. We describe the three basic query types in details below. Note that the template descriptions only convey the motivation and semantic significance of the query, while the query input consists solely of activity labels, object crops, and time windows.

Activity-query: “Template: What activities did I perform with object o during time t ?” An example query is shown in the red box in Fig. 1. The answer to this type of query is a subset $\mathcal{C}$ of the activities from a pre-defined taxonomy $\{1 \dots C\}$, where C is the total number of activity categories. Note that $\mathcal{C}$ may contain multiple activities or be an empty set, indicating no interaction with the object o. Activity-query differs from conventional activity recognition and detection tasks by incorporating a condition on the interaction between the camera wearer and the queried object. This design not only increases the task's difficulty but also helps reduce ambiguity in the questions, especially when applied to long videos with numerous irrelevant activities. For instance, it is more meaningful to ask “What activity did I perform yesterday *while using this mug*?” than to simply ask “What activity did I perform yesterday?”

Object-query: “Template: What objects did I interact with when engaging in activity c during time t ?” An example query is shown in the green box in Fig. 1. The answer to this type of question is a set of “active” objects (*i.e*., objects that the camera wearer is interacting with) represented as bounding boxes with associated spatiotemporal locations $\{t_i, x_i, y_i, h_i, w_i\}_{i=1}^{N}$, where N is the total number of ground truth boxes. While object-query shares a similar prediction format with traditional object detection and visual query tasks, the objects of interest are fundamentally different. Object detection involves localizing objects from a pre-defined set of categories, while visual query [17] aims to localize the most recent track of the query object, without considering whether it was interacted with or not. In contrast, object-query requires identifying the active objects that are involved in the query activity, without making any assumptions about their semantic category.

[2] We use the term “query” instead of “question-answering” to make it explicit that our framework is not dependent on language input or output.

Time-query: "Template: When did I perform activity c with object o ?" An example query is shown in the blue box in Fig. 1. The answer to this query type is a set of time windows with starting and ending timestamps $\{s_i, e_i\}_{i=1}^{M}$, where M is the number of ground truth segments. Compared to traditional temporal localization tasks, time-query presents a more challenging but fine-grained problem since it involves understanding the joint occurrence of the query activity and the interaction with the querying object.

Although this paper focuses on the three basic types of queries, it is important to note that our ReST framework is holistic and supports queries that require varying levels of comprehension of a long video. It is possible to construct more complex queries by combining the three basic ones or incorporating additional information such as object states, ownership, and environment, into the event description. We leave the option of introducing more diverse query types to further expand the scope of our framework in the future work. Nevertheless, we have already observed that even with the three basic tasks, the complexity of the problem is increasing significantly due to various design choices for different perception components (as discussed in Sec. 5). We see our integrated framework and benchmarks as a crucial step towards achieving comprehensive, multi-step reasoning in long videos, and we believe that it will facilitate the development of next-generation video understanding models.

4. ReST Benchmarks

We introduce two new benchmarks for evaluating the Relational Space-Time query (ReST) tasks, based on the publicly available datasets ADL [37] and Ego4D [17]. Both datasets consist of egocentric videos with long durations ranging from 10 to 60 minutes, and are characterized by intensive human-object interaction. In particular, the ReST-ADL benchmark is smaller in scale but provides denser and cleaner annotations, making it more suitable for task analysis (as discussed in Sec. 6). The ReST-Ego4D benchmark, on the other hand, is larger in scale and more challenging with videos recorded from diverse scenarios. In this section, we introduce the query generation process in details, report basic statistics for both two benchmarks, and discuss the evaluation metrics for the three types of queries.

4.1. Benchmark Generation

Human-object annotation. ReST tasks are centered around activities and human-object interaction in long-form videos. To achieve this, we consolidate and augment the original annotation in ADL [37] and Ego4D [17] to include the following information: (1) activity annotation: labeling each video with temporal segments indicating the occurrence of pre-defined activities; (2) object annotation: labeling each frame with bounding boxes locating common objects and tagging each of them with a unique id indicating different object instances; (3) interaction annotation: labeling each object bounding box with the status of being interacted with or not. Annotations are collected in 1 FPS on ADL and 2 FPS on Ego4D.

Query generation. The query generation process is fully automatic given the densely collected human-object annotations. For each video, we first generate a set of candidate windows, which can be considered as independent "episodic memories" from which the three types of queries are generated. A valid candidate window should not contain truncated activity segments. We define three window sizes to evaluate different complexity levels in terms of memory durations – short (around 5 minutes) / medium (around 15 minutes) / long (around 30 minutes). To avoid highly overlapping windows, we randomly select query windows from the these candidates such that their temporal intersection-over-union (IoU) is less than 0.9. For each query window, we collect activity and object information $(S_m, l_m, \mathcal{A}_m, \mathcal{I}_m)$, where $m \in \{1, ..., M\}$ denotes the index of activity segments occurred in the current window. $S_m = [s_m, e_m]$ indicates the segment location, $l_m \in [1, C]$ is the activity label, $\mathcal{A}_m$ and $\mathcal{I}^m$ denote the "active" and "inactive" objects that are present during the activity, respectively.

To generate activity-query, we randomly sample an "active" object $o \in \{\mathcal{A}_1 \cup \mathcal{A}_2 \cup ...\mathcal{A}_M\}$ as the query, and collect the corresponding activities as the answers $\{l_m \mid o \in \mathcal{A}_m\}$. To generate object-query, we sample an activity category $c \in \{l_m\}_{m=1}^{M}$ as the query and collect the bounding boxes of the corresponding "active" objects as the answer $\{\mathcal{A}_m \mid l_m = c\}$. Similarly for time-query, the query is a pair of "active" object and activity category (o, c), and the answer is the temporal location of the corresponding activity segments $\{S_m \mid o \in \mathcal{A}_m, l_m = c\}$. In addition to the positive queries described above, we also generate negative queries where either the query activity or the object interaction is absent from the query window. In such cases, the answers are represented as empty sets and the model should "reject" those queries to prevent false alarms.

Comparison with Ego4D queries. ReST provides a unifying view of VQ and MQ – rather than locating the presence of a query object or activity independently, we take into account the relationships between these closely related problem. Although NLQ also supports such capacity in general, ReST differentiates itself by employing a substantially different query generation process, *i.e.*, operating on densely collected, low-level human-object annotations. The advantages are as follow: (a) Finer-grained control. First, it enables re-balancing or creating new query sets with different emphasis. For instance, we can balance the distribu-

Benchmarks	Vision focused	Avg. length (s)	Total hours	#Queries (K)
MovieQA [49]		202.7	381.2	6.5
TGIF-QA [25]		3	59.8	165
Act.-QA [66]		180	290	58
AGQA [18]		30	80	1920
QAEgo4D [4]		495	182	14.5
Just-Ask [61]		12.1	233K	83K
Next-QA [58]		44	66.5	52
ReST-ADL	✓	1631	9	185.7
ReST-Ego4D	✓	1104	92	303.3

Table 1. Compare ReST-ADL and ReST-Ego4D with common video question answering (QA) benchmarks.

tion of positive and negative queries, or re-sample the questions and answers to reduce intrinsic bias of the dataset. Second, it allows for evaluation under specific conditions, such as different window sizes, query types, and even difficulty and visibility levels (detailed in the supplementary material). (b) Improved data efficiency. Collecting human annotated question-answer pairs is labor-intensive – NLQ only involves sparse annotation with 52.2 queries per hour on average. In contrast, low-level annotations can be efficiently collected or even generated by models, leading to 20.6(3.3)K queries per hour in ReST-ADL(Ego4D) with dense coverage of events. In addition, ReST is the first work that explicitly takes human-object interaction into account when constructing queries. Our framework is also vision-centric, providing a more precise and concise method to refer to object instances than NLQ / QAEgo4D [4] and avoiding potential bias from language models

4.2. Dataset Statistics

ReST-ADL ADL [37] is one of the earliest datasets for detecting activities of daily living in first-person camera views. The dataset involves 9 hours of videos amassed from 20 people performing non-scripted activities in 20 different homes. We re-define 24 activity categories that are more common and clear in real-world applications and manually clean up the annotation of instance id and "active" objects.

We generate over 185K queries in total for ReST-ADL, with 1,020 unique object instances. Fig. 2 highlights some basic statistics of the generated queries. The activity-query and time-query are more frequent due to the large number of interacted objects for query generation. Query windows with a medium duration (around 15 minutes) are the most frequent because they involve more diverse activities and interactions than short windows and some of the 20 videos are not long enough to generate long query windows. We also observe that the distribution of an activity served as the query activity shows a natural long-tail distribution since some activities occur more often and have longer duration than the others. Since ADL is relatively small scale, we divide the videos into five splits and report results on 5-fold cross validation unless otherwise stated.

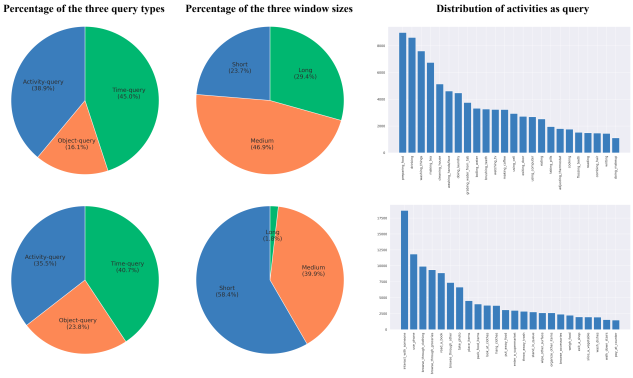

Figure 2. Dataset statistics of ReST-ADL (top) and ReST-Ego4D (bottom). (Left) The percentage of activity-query, object-query and time-query. (Center) The percentage of queries from different window sizes. (Right) Distribution of an activity category served as the query activity. The top-24 most frequent activities are selected for the visualization of ReST-Ego4D.

ReST-Ego4D Ego4D [17] is a recently-published dataset offering a huge amount of egocentric video recordings in diverse scenarios. In particular, we extend a subset of the dataset that was originally proposed for the Moment Query (MQ) challenge. This subset provides temporal segment annotation for a set of pre-defined "moments", and we additionally collect the object and interaction annotations described in Sec. 4.1 for each of the moment segments. ReST-Ego4D consists of 301 video recordings with 92 hours in total. We generate over 303K queries and the dataset statistics is shown in Fig. 2. Note that the duration of moment segments in ReST-Ego4D is much shorter than those in ReST-ADL as reported in [17], therefore identifying and locating the query activities is very challenging even for the query windows with a small size (around 5 minutes). We randomly split the 301 videos into 70%, 10% and 20% for training, validation, and testing, respectively.

4.3. Evaluation metrics

Prior works on question answering mostly use accuracy as the evaluation metric since they either involve multi-choice questions [49, 58] or only short and simple answers [2, 61]. When evaluation for more complex and practical answers, metrics from machine translation (*e.g.*, BLEU, ROUGE) are adopted in QAEgo4D [4]. Inspired by the work on information retrieval [30, 47] and language grounding [62, 69], we evaluate the effectiveness of vision models using the standard **Recall@kx**. The metric measures the percentage of ground truth labels identified in top-

kx predictions, where x stands for the number of ground truth labels in the answer. For object-query and time-query that involve (spatio-)temporal localization, an IoU threshold is used to identify the correct detection. We also measure the **Rejection recall** for negative queries, which measures the percentage of negative queries that are successfully rejected by the models.

5. Experiments: Baselines for ReST Tasks

Solving ReST queries requires understanding of both fine-grained human-object interaction and long-range temporal dependencies in videos. In this section, we present a simple, modularized system that generates probabilistic answers to the queries by consolidating predictions from state-of-the-art perception modules. We then report baseline results for the three query tasks on our benchmark.

5.1. Probability-based Modularized System

Given a query, our modularized system first estimates the likelihood of a time step being relevant to the query task within the query window, which we refer to as the "relevance likelihood". A time step is regarded as "relevant" if there exists an "active" object matches the query object or the ongoing activity matches the query activity. Query rejection is performed if no relevant time step is identified. Then, the system collects information within the relevant time steps and generates probabilistic answers to the query.

Activity-query. Given a query object o and a query window $[s, e]$, the goal is to either reject the query or generate a probability estimation of the C activity classes as the answer. We first estimate the relevance likelihood $\mathcal{L}_1(t)$ for each time step $t \in [s, e]$, defined by the probability of finding an "active" object that matches the query object:

$$\mathcal{L}_1(t) = \max_{i=1,\dots,N_t} P_{match}(b_i,\ o) \times P_{inter}(b_i), \tag{1}$$

where b_i denotes the i-th detected bounding box and N_t is the number of detected boxes at time t. $P_{match}(b_i, o)$, $P_{inter}(b_i)$ denote the probability of matching a detected box b_i with o and the probability of b_i being interacted with, respectively. We will elaborate on how to obtain these two probabilities later. A query is rejected if $\mathcal{L}_1(t) < \sigma_1$ at all time steps. If a query is not rejected, we re-weight the activity recognition prediction $P_{act}(t)$ with the relevance likelihood and generate the answer $A_{\text{act}} \in \mathbb{R}^C$ by performing a max-pooling for each activity class c along all time steps:

$$A^c_{\text{act}} = \max_{t=s,\dots e} \mathcal{L}_1(t) \times P^c_{act}(t). \tag{2}$$

Object-query. Given a query activity c and a query window $[s, e]$, the goal is to either reject the query or generate a list of spatio-temporal bounding boxes with confidence scores. We define the relevance likelihood as the probability of the occurrence of the query activity: $\mathcal{L}_2(t) = P^c_{act}(t)$. A query is rejected if $\mathcal{L}_2(t) < \sigma_2$ for all time steps. For each detected box $b_{t,i}$ at time t, its confidence score is defined as the probability of the object being interacted with, weighted by the "relevance likelihood" of the time step:

$$w_{t,i} = \mathcal{L}_2(t) \times P_{inter}(b_{t,i}). \tag{3}$$

The answer to the query is the collection of all detected boxes within the query window:

$$A_{\text{obj}} = \{b_{t,i}, w_{t,i}\},\ t \in [s, e],\ i = 1, ..., N_t \tag{4}$$

Time-query. Given a query instance o and a query activity c, the goal to either reject the query or generate a list of temporal segments with confidence scores. We obtain the "relevance likelihood" by estimating the joint likelihood: $\mathcal{L}_3(t) = \mathcal{L}_1(t) \times \mathcal{L}_2(t)$. A query is rejected if $\mathcal{L}_3(t) < \sigma_3$ for all time steps. To generate the answer, we follow the standard post-processing strategies in weakly-supervised action detection [20, 39], which generate activity proposals by applying different thresholds to the activity recognition scores at each time step. We score each of these activity proposals $\hat{S}_i = [s_i, e_i]$ by combining its outer-inner-contrastive (OIC) score [43] over P^c_{act} and the maximum object matching score within the proposal:

$$v_i = \text{OIC}(\hat{S}_i) \times \max_{t\in[s_i,e_i]} \mathcal{L}_1(t). \tag{5}$$

The answer is obtained after applying temporal non-maximum suppressed (NMS) to the proposals:

$$A_{\text{time}} = \left\{\hat{S}_i,\ v_i\right\},\ i = 1, ..., K, \tag{6}$$

where K is the total number of predicted activity segments.

Perception modules. We finetune the recent action recognition models [12, 34] on the training split of our dataset and use the model to predict the activity recognition results $P_{act}(t)$ in Eq. (2). Off-the-shelf object detectors [21,29,54] are adopted to detect potential objects in each frame and generate the corresponding objectness scores $P_{obj}(b_i)$. We then crop the image with the predicted bounding boxes and extract feature embeddings using a pre-trained ResNet-50 model [22]. For brevity, we extend the use of the symbol o_i to also represent the embedding of the bounding box b_i.

The matching probability between a detected box and a query object $P_{match}(b_i, o)$ is defined as the joint probability of the box successfully detecting an object and this object being the same as the query one: $P_{match}(b_i, o) = P_{obj}(b_i) \times P_{same}(o_i, o)$. P_{same} is obtained by computing the cosine similarity between the two embeddings followed by probabilistic calibration [38]. Similarly, we define the interaction probability of a detect box $P_{inter}(b_i)$ as the joint

	Activity-query			Object-query			Time-query		
	R@1x	R@3x	Rej.	R@1x	R@3x	Rej.	R@1x	R@3x	Rej.
Short	48.1 (±7.0)	69.2 (±8.0)	68.9 (±3.1)	9.6 (±1.3)	19.4 (±3.3)	84.5 (±2.1)	31.3 (±9.5)	32.6 (±9.8)	86.8 (±3.0)
Medium	50.7 (±13.1)	72.6 (±8.4)	63.3 (±5.9)	10.0 (±1.7)	20.23 (±3.7)	79.3 (±3.6)	31.8 (±9.4)	33.8 (±10.2)	84.3 (±4.4)
Long	46.3 (±7.6)	70.9 (±4.8)	67.0 (±8.8)	10.0 (±1.9)	21.0 (±4.3)	68.6 (±7.8)	30.0 (±6.3)	31.9 (±7.6)	85.8 (±5.1)

Table 2. Baseline results with the explicit system on ReST-ADL. The results are obtained with 5-fold cross validation and reported in the format *mean* (±*std*). Rej. stands for rejection recall. We report results with IoU=0.3 for object-query and time-query in the table and the complete results are provided in the supplementary material.

probability of the box detecting an object and this object being interacted with: $P_{inter}(b_i) = P_{obj}(b_i) \times P_{active}(o_i)$. We train a binary classifier using the annotated objects in the training split to predict $P_{active}(o_i)$. More implementation details and ablations of the individual perception modules are provided in the supplementary material.

Discussion. Despite the simplicity and interpretability of our modularized system, we have identified certain limitations to the approach. (1) The perception modules are not optimized for the final ReST tasks, which may hinder the performance when scaling to larger amount of data. However, adapting end-to-end models to the tasks is non-trivial – there is no existing QA or video models that support detection of human-object interaction or reasoning on multiple perceptions (*e.g.*, predicting video activity given object crops and time windows). We provide preliminary studies on ReST-ADL using a modified version of TubeDETR [62] and Object-Transformer [56] in the supplementary material and leave more dedicated model designs to future work. (2) Our system requires computing and storing the results of all the perception modules beforehand, which implies an unlimited computation and storage footprint for solving the tasks. Such an assumption does not hold for most real-world applications and our system is designed for better analyzing the new tasks and evaluating different vision models. Our system also requires different inference strategies to generate answers with different structures.

5.2. Baseline Results

We present baseline results with our explicit system in Tab. 2 and 3 for ReST-ADL and ReST-Ego4D, respectively. We first observe that our ReST tasks are very challenging, especially for object-query and time-query. Even equipped with state-of-the-art perception models and unlimited computation and memory resource, the explicit system still suffers from identifying and localizing the ongoing activities and the objects being interacted. ReST-ADL and ReST-Ego4D benchmarks share the same trend of difficulties across three query types. The problem becomes more severe when processing videos in ReST-Ego4D with more diverse scenarios and more severe camera motion.

	Activity-query		Object-query		Time-query	
	R@1x	Rej.	R@3x	Rej.	R@3x	Rej.
Short	30.1	61.9	1.4	85.5	11.1	74.3
Medium	33.9	64.7	1.9	89.3	17.8	69.0
Long	22.6	76.1	1.4	90.3	21.8	76.6

Table 3. Baseline results on the test split of ReST-Ego4D. Our ReST tasks are highly challenging, especially for object-query and time-query that require precise detection of interacted objects.

5.3. Qualitative results

We present visualization results of our modularized system in Fig. 3. We also plot the "relevance likelihood" to illustrate how the query object and activity are identified within the query window. For activity-query (Q1, Q2), our system successfully detects the time steps where the query objects are interacted with and predicts the correct activity label at the corresponding time steps, therefore the correct answers are generated accordingly. For object-query (Q3), we can see that it is a very challenging task and the top-scoring predictions are usually dominated by the incorrect object detection results. We only recall one ground truth bounding box (the tap) in this example. For time-query (Q4), we observe that the model successfully detects the "washing hands" segment that involves interaction with the query object (hand soap), while suppressing the prediction on the other segment with the same activity label. However, the model still generates incorrect answers if the activity recognition results are incorrect (*e.g.* Q5).

6. Sensitivity Analysis

We conduct an "oracle" analysis to better understand the contribution of each sub-problem (*i.e.*, perception component) to the overall ReST task performance. In this setup, we replace model predictions with ground truth labels for all perception modules, and then systematically corrupt one module (*e.g.* by injecting noise at different levels) to gauge its sensitivity and importance to task performance. This

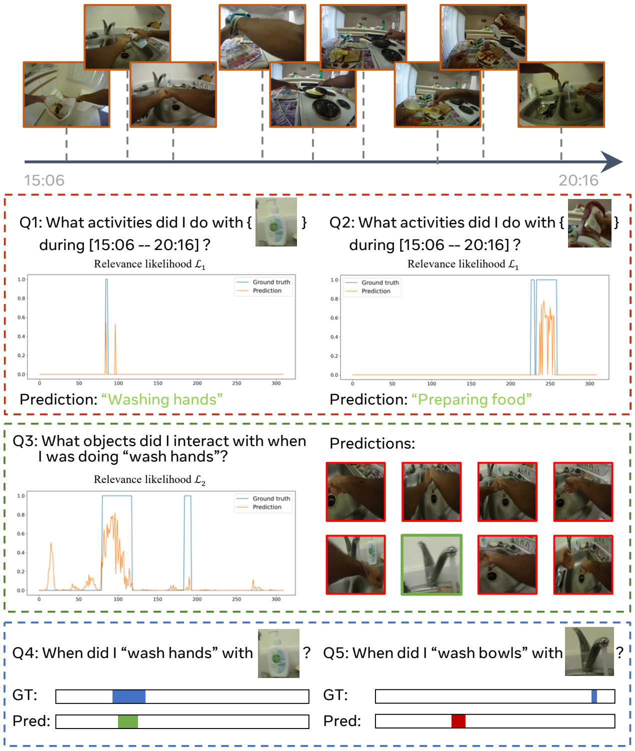

Figure 3. Visualization of the model predictions on ReST-ADL generated by the modularized system. The first row shows video snapshots within the query window. The following rows show the queries and predictions for activity-query, object-query, and time-query, respectively. We visualize the top-1x predictions for all the queries and use the green color to indicate correct predictions.

analysis helps inform future work on identifying which submodule may provide the greatest opportunities for performance improvements in the ReST task.

In particular, we conduct *label corruption* experiments by replacing the ground truth of one module with noisy labels. The noise we add to each module is described as follows. **Activity module:** assign a random activity label (including background) to a time step with probability ρ. **Detection module:** randomly remove a ground truth box with probability ρ (missed detection of an object). **Embedding module:** assign a random instance id to a ground truth box (one-hot embedding of the instance id is used for object matching in the experiments). **Interaction module:** assign a random interaction label to a ground truth box with probability ρ.

We note that by varying ρ we can control the level of corruption of the selected module. Experiment results are shown in Fig. 4. It is obvious that the performance on positive queries drops significantly as the noise level of the activity module increases. This indicates the key role of the activity module to the ReST tasks since all three types of query require recognizing and localizing activities to generate the answers. We also observe that the detection module and the interaction module have larger impacts to object-query, which requires predicting the bounding boxes of all "active" objects within the query activity. The interaction module has particularly large impact to the performance on negative queries because query rejection is based on the "relevance likelihood" computed with the prediction of the interaction module (Eqn. 1). The probabilistic modeling approach in our explicit system provides decent robustness to random noise when the overall noise level is low.

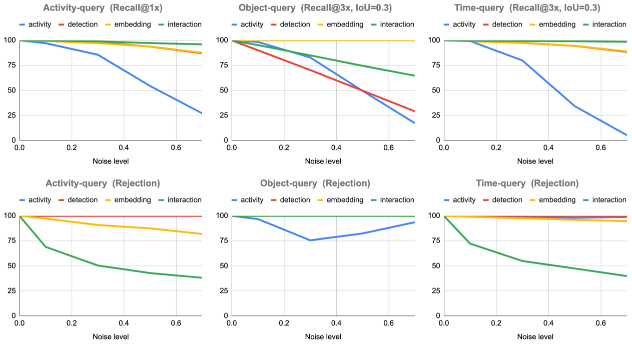

Figure 4. Label corruption experiments on split-1 of ReST-ADL. We observe that the activity module is critical for all types of queries, while the object detection and interaction modules have larger impacts to object-query and negative queries, respectively.

7. Conclusion and Future Work

We introduced Relational Space-Time query (ReST), a holistic framework that jointly studies the activities, objects, and their interactions in long videos with templated spatiotemporal queries. ReST is set up on long videos, *e.g.*, 5-to-30-minute long, and is designed to captures the rich expressivity of natural language query while avoiding the ambiguity and bias introduced by the language modality. We further introduced two new benchmarks, ReST-ADL and ReST-Ego4D, which are built upon the publicly available egocentric video datasets with long duration and intensive human-object interactions. Finally, we developed a probability-based explicit system and conducted a set of experiments to provide in-depth analysis about the ReST tasks. We believe our integrated framework and benchmarks will facilitate future research on long video understanding and inspire the development of next generations of video models.

Acknowledgement

The authors would like to thank Weiyao Wang for providing the open-world object detectors, Jing Huang for sharing the object embedding models, and Jiabo Hu for helping with data annotation.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3, 5

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *CVPR*, pages 6836–6846, 2021. 2

[4] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 2, 3, 5

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 2

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020. 1

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[9] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. 2

[10] David Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies.*, 2011. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, June 2019. 1

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 1, 2, 6

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1, 2

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 2

[15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 3

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *arXiv:1706.04261*, 2017. 1, 2

[17] 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, 3, 4, 5

[18] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2, 3, 5

[19] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. *CoRR*, abs/1705.08421, 2017. 1, 2

[20] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13925–13935, 2022. 6

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[23] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 2

[24] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thu-

mos challenge on action recognition for videos “in the wild”. *CVIU*, 155:1–23, 2017. 2

[25] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 3, 5

[26] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *TPAMI*, 2013. 2

[27] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *CVPR*, pages 4405–4413, 2017. 2

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1, 2

[29] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 2022. 6

[30] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. In *ACM Computing Surveys (CSUR)*, 2000. 5

[31] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 1, 2

[32] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *ECCV*, pages 345–362, 2020. 2

[33] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2, 3

[34] 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, 6

[35] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 3

[37] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 2, 4, 5

[38] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 6

[39] Sanqing Qu, Guang Chen, Zhijun Li, Lijun Zhang, Fan Lu, and Alois Knoll. Acm-net: Action context modeling network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2104.02967*, 2021. 6

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1

[42] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *ECCV*, pages 154–171, 2020. 2

[43] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018. 6

[44] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 1, 2

[45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS*, 27, 2014. 2

[46] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 1

[47] Amit Singhal. Modern information retrieval: A brief overview. In *IEEE Data Eng. Bull*, 2001. 5

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 1, 2

[49] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2, 3, 5

[50] Atousa Torabi, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Using descriptive video services to create a large data source for video annotation research. *CoRR*, abs/1503.01070, 2015. 2

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 1, 2

[53] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *CVPR*, pages 352–361, 2020. 2

[54] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. *CVPR*, 2022. 6

[55] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[56] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, June 2021. 2, 7

[57] 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 (CVPR)*, pages 13587–13597, June 2022. 2

[58] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 3, 5

[59] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. In *ECCV*, 2018. 2

[60] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 3

[61] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021. 3, 5

[62] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 5, 7

[63] Xitong Yang, Haoqi Fan, Lorenzo Torresani, Larry S. Davis, and Heng Wang. Beyond short clips: End-to-end video-level learning with collaborative memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7567–7576, June 2021. 2

[64] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[65] Xitong Yang, Xiaodong Yang, Sifei Liu, Deqing Sun, Larry Davis, and Jan Kautz. Hierarchical contrastive motion learning for video action recognition. *BMVC*, 2021. 2

[66] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 2, 3, 5

[67] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 3

[68] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510. Springer, 2022. 2

[69] 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*, volume 34, pages 12870–12877, 2020. 3, 5

[70] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, pages 10668–10677, 2020. 3

[71] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *CVPR*, pages 13658–13667, 2021. 2

[72] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *CVPR*, pages 13598–13607, 2022. 2