VIOLIN: A Large-Scale Dataset for Video-and-Language Inference

Jingzhou Liu[1]* Wenhu Chen[2]* Yu Cheng[3] Zhe Gan[3] Licheng Yu[3]
Yiming Yang[1] Jingjing Liu[3]
[1]Carnegie Mellon University [2]University of California, Santa Barbara
[3]Microsoft Dynamics 365 AI Research
{liujingzhou,yiming}@cs.cmu.edu, wenhuchen@ucsb.edu
{yu.cheng,zhe.gan,licheng.yu,jingjl}@microsoft.com

Abstract

We introduce a new task, Video-and-Language Inference, for joint multimodal understanding of video and text. Given a video clip with aligned subtitles as premise, paired with a natural language hypothesis based on the video content, a model needs to infer whether the hypothesis is entailed or contradicted by the given video clip. A new large-scale dataset, named VIOLIN *(VIdeO-and-Language INference), is introduced for this task, which consists of 95,322 video-hypothesis pairs from 15,887 video clips, spanning over 582 hours of video. These video clips contain rich content with diverse temporal dynamics, event shifts, and people interactions, collected from two sources: (i) popular TV shows, and (ii) movie clips from YouTube channels. In order to address our new multimodal inference task, a model is required to possess sophisticated reasoning skills, from surface-level grounding (e.g., identifying objects and characters in the video) to in-depth commonsense reasoning (e.g., inferring causal relations of events in the video). We present a detailed analysis of the dataset and an extensive evaluation over many strong baselines, providing valuable insights on the challenges of this new task.*

1. Introduction

Joint vision-and-language understanding sits at the nexus of computer vision and natural language processing (NLP), and has attracted rapidly growing attention from both communities. Popular tasks include visual question answering [4, 20], referring expression comprehension [69, 68], visual dialog [12], visual reasoning [27, 52, 25], visual commonsense reasoning [72], NLVR^2 [52], and visual entailment [61]. The emergence of these diverse Vision+Language tasks, benchmarked over large-scale human annotated datasets [39, 34], has driven tremendous progress in joint multimodal embedding learning [53, 42, 10, 51]. However, most of these datasets and models were centered on static images, leaving the joint modeling of video and its aligned textual information (*e.g.*, video-and-language understanding) a relatively under-explored territory.

Video Question Answering (Video QA) is one of the most popular tasks in current studies for video-and-language understanding. Video QA model aims to answer a natural language question given a video clip. Existing Video QA datasets include MovieFIB [44], MovieQA [54], TGIF-QA [26], PororoQA [32], and TVQA [35, 36]. While these datasets have covered a rich pool of video content (*e.g.*, cartoons, short GIFs and TV shows), they are limited to QA task only. On the other hand, in NLP field, one important benchmark for natural language understanding is natural language inference (NLI) [5, 60], where a model is presented with a pair of sentences (premise and hypothesis), and judges the relationship between the pair (*e.g.*, *Contradiction*, *Neutral*, and *Entailment*).

Inspired by NLI, we present a novel task, Video-and-Language Inference, to foster deeper investigations in video-and-language understanding. Specifically, given a video clip with aligned subtitles as premise, and a natural language statement as a hypothesis describing the video content, a model is expected to infer whether the statement is entailed or contradicted by the given video clip. This new task is easy to evaluate, since only binary classification is measured; but also challenging to solve, as a thorough interpretation of both visual and textual clues is required in order to achieve in-depth understanding and inference for a complex video scenario.

We introduce a large-scale dataset for this new task, **VI**de**O**-and-**L**anguage **IN**ference (VIOLIN)[2], built upon natural video content with rich temporal dynamics and social interactions. Video clips are collected from diverse sources to cover realistic visual scenes, and statements are

*This work was done while the authors were interns at Microsoft.

[2]Project page: https://github.com/jimmy646/violin.

00:00:03,576 --> 00:00:05,697
Gavin Mitchell's office.
Rachel Green's office.

00:00:05,870 --> 00:00:07,409
Give me that phone.

00:00:08,873 --> 00:00:12,293
Hello, this is Rachel Green.
How can I help you?

00:00:12,460 --> 00:00:17,629
Uh-huh. Okay, then.
I'll pass you back to your son.

00:00:18,800 --> 00:00:21,639
Hey, Mom. No, that's just my secretary.

(positive) The woman becomes upset when the man answers the phone because he pretends it is his own office.
(negative) The woman becomes upset when the man answers the phone because she is expecting a phone call from her mom.

Inferring reasons

(positive) The woman realizes it is the man's mother who is calling and she passes the phone back to the man.

(negative) The man realizes it is the woman's mother who is calling and he passes the phone back to the woman.

Identifying characters

(positive) The phone rings, a man picks it up, and a woman slams her hand on the desk and demands the man give her the phone.

(negative) The two people that the man in the glasses is talking to need to be briefed on something.

Global video understanding

Figure 1. An example from the VIOLIN dataset. The first two rows show a video clip with its aligned subtitles. The third row contains three pairs of positive/negative statements. The task is to independently decide whether each statement is supported or contradicted given the subtitled video. The first two negative statements are written by modifying part of the positive statements (marked in red), and the third is curated by adversarial matching (Sec. 3.1). The text box below each pair of statements indicates the reasoning skill required to infer the verdict of each statement.

collected from crowdsource workers via Amazon Mechanical Turk (AMT)[3], who watched the videos accompanied by subtitles (dialogue, scene description, etc). Our goal is to provide a dataset that can test a model's cross-modality reasoning skills over both video and textual signals. To this end, we require AMT workers to write statements based on joint understanding of both video and subtitles, which not only describe explicit information in the video (*e.g.*, objects, locations, characters, social activity), but also reveal in-depth comprehension of complex plots (*e.g.*, interpreting human emotions and relations, understanding the events, inferring causal relations of events throughout the video). This distinguishes our collected statements from the straightforward captions in video/image captioning dataset [39, 33, 59], which are dominated by explicit factual descriptions without deeper inference.

Writing negative statements for an inference task is challenging [5, 72]. To gather high-quality negative statements without artificial cues or biased priors, we employed two strategies in the data collection: (i) requiring annotators to write negative statements by changing just a few words or phrases in a positive statement, to ensure that the style and length of the statement remain unchanged; (ii) performing adversarial matching [72]: for each video, select challenging and confusing statements from the statement pool of other videos as the negative ones. The first strategy ensures the collected statements can test a model's in-depth inference ability, since only a small fraction of a positive statement is modified, which requires the model to distinguish highly similar statements with different meanings. The second strategy focuses more on testing a model's global understanding of the video, to distinguish statements with high-level scene difference between videos. When combined together, these two strategies produce a dataset with minimal visual or textual bias. Through this effort, we collected 95,322 video-statement pairs, containing 15,887 video clips spanning over 582 hours. Each video is paired with 6 statements and is 35.2 seconds long on average.

The main contributions of this paper are three-fold. (i) We propose a new task, Video-and-Language Inference, which requires a model to draw inference on whether a written statement entails or contradicts a given video clip. (ii) We introduce a new dataset VIOLIN for this task, providing a reliable benchmark for measuring joint video-and-language understanding models. (iii) We provide a detailed analysis of the VIOLIN dataset with evaluation over strong baselines, and suggest future directions for this new task.

2. Related Work

Natural Language Inference (NLI) Understanding entailment and contradiction relations between sentences (*i.e.*, Natural Language Inference) is fundamental to natural language understanding. Several large-scale datasets have been developed as NLI benchmarks, such as SNLI [5] and MultiNLI [60]. NLI is also included in the GLUE benchmark for evaluating general language understanding [57]. Recent introduction of large-scale pre-trained language models, such as BERT [14], XLNet [63], and RoBERTa [41], has propelled significant progress in NLI. Multi-task learning and adversarial training [40, 73] also prove to be helpful in improving model performance.

Inspired by NLI, we propose the task of Video-and-Language Inference to evaluate a system's multimodal reasoning ability. However, different from NLI, our task is more challenging in the sense that both video and text (subtitles) are provided; thus, a thorough joint understanding of both modalities is required for inference.

[3] https://www.mturk.com/

source	# episodes	# clips	avg clip len	avg pos. statement len	avg neg. statement len	avg subtitle len
Friends	234	2,676	32.89s	17.94	17.85	72.80
Desperate Housewives	180	3,466	32.56s	17.79	17.81	69.19
How I Met Your Mother	207	1,944	31.64s	18.08	18.06	76.78
Modern Family	210	1,917	32.04s	18.52	18.20	98.50
MovieClips	5,885	5,885	40.00s	17.79	17.81	69.20
All	6,716	15,887	35.20s	18.10	18.04	76.40

Table 1. Statistics of different video sources used to create our dataset.

Visual Entailment Visual Entailment (VE) [61] is a recently proposed task that extends NLI to the visual domain. In this task, a natural image premise and a natural language hypothesis are given, and the goal is to judge whether the textual hypothesis can be confirmed based on the visual content in the image. Three labels are assigned: *Entailment, Neutral*, and *Contradiction*. The dataset is created based on Flickr30k image captions [66] and SNLI [5]. Similarly, NLVR^2 [52] is proposed to investigate the grounding relationship between given images and a natural language description.

Our proposed task is different from VE in the following aspects. (i) VE considers images as input, while our task focuses on videos instead. Compared with static images, videos contain complex temporal dynamics, making the video-and-language inference task more challenging as the model needs to understand the relationship between different visual scenes to draw inference. (ii) Our proposed task requires deeper visual understanding. Images in the VE task are mostly natural images, while the videos in VIOLIN were collected from popular TV shows and movie clips, which contain rich social interactions and diverse scenes. This requires a model to not only understand explicit visual cues, but also infer in-depth rationale behind the scene. (iii) Our task requires more sophisticated language understanding. VE is a combination of Flickr30k [66] and SNLI [5], with no crowdsouring involved. The hypotheses in VE task are composed of captions only, containing factual descriptions that can be explicitly derived from the visual content in the image. On the other hand, VIOLIN mainly consists of implicit statements that cannot be solved without in-depth understanding of the video and text, designed specifically to evaluate a model's multimodal reasoning skills.

Video-and-Language Research With the emergence of large-scale video datasets [6, 1, 29, 11, 58], several video-and-language tasks have been proposed, such as video captioning [21, 56, 62, 18, 33, 16, 47, 59], localizing video segments from natural language queries [19, 3, 8, 37], video reasoning [65], and video question answering [54, 35]. Video captioning is a conditional text generation task, while the other three belong to video-and-language understanding. In particular, MovieQA [54], TGIF-QA [26] and TVQA [35, 36], which contain real-world videos and human-generated questions, are recently proposed for video question answering.

Our VIOLIN dataset also uses TV shows as one of the video sources, similar to TVQA [35]. The main differences are summarized as: (i) Our dataset contains richer video content, including 5,885 movie clips in additional to TV shows used in TVQA. (ii) Our dataset requires more sophisticated reasoning skills from a model, such as inferring reasons and interpreting human emotions, while most QA pairs in TVQA are focused on identifying explicit information.

Visual Question Answering Our proposed task is also related to Visual Question Answering (VQA) [4, 20]. The CLEVR dataset [27] serves as a popular synthetic diagnosis dataset that tests a model's compositional reasoning skills. Recently, GQA [25] was introduced to benchmark real-world visual reasoning, and VCR [72] for visual commonsense reasoning.

Many neural network models have been proposed for these tasks, such as more advanced attention mechanisms [64, 43, 70], better multimodal fusion methods [15, 71, 31, 30], the use of multi-step reasoning [24, 17, 7], the incorporation of relations [49, 38, 45], and neural module networks for compositional reasoning [2, 28, 23, 9]. Our proposed task can provide a new perspective for benchmarking these models.

3. Video-and-Language Inference Dataset

In our VIOLIN dataset for video-and-language inference, the input is a video clip V consisting of a sequence of video frames $\{v_i\}_{i=1}^T$, paired with its aligned text $S = \{s_i, t_i^{(0)}, t_i^{(1)}\}_{i=1}^n$ (s_i is the subtitle within time span $(t_i^{(0)} \rightarrow t_i^{(1)})$ in the video) and a natural language statement H as the hypothesis aiming to describe the video clip. For every (V, S, H) triplet, a system needs to perform binary classification: $f(V, S, H) \rightarrow \{0, 1\}$, deciding whether the statement H is entailed (label 1) from or contradicts (label 0) the given video clip. In order to increase the coverage and versatility, we collect the videos from diverse sources, including 4 popular TV shows of different genres and YouTube movie clips from thousands of movies. To ensure high video quality, we also provide carefully-designed protocols to guide crowdsource workers to select representative video segments for which to write positive/negative statements. The procedure of dataset collection is detailed in Sec. 3.1, and Sec. 3.2 provides a thorough analysis on the dataset.

Dataset	Visual Domain	Source	Subtitles	Inference	Task	# images/videos	# samples
Movie-QA [54]	video	movie	✓	✗	QA	6.8K	6.5K
MovieFIB [44]	video	movie	✗	✗	QA	118.5K	349K
TVQA [35]	video	TV show	✓	✗	QA	21.8K	152.5K
VCR [72]	image	movie	✗	✓	QA	110K	290K
GQA [25]	image	indoor	✗	✓	QA	113K	22M
SNLI-VE [61]	image	natural	✗	✓	Entailment	31.8K	565.3K
NLVR^2 [52]	image	natural	✗	✓	Entailment	127.5K	107.3K
VIOLIN (ours)	video	TV show/movie	✓	✓	Entailment	15.9K	95.3K

Table 2. Comparison between VIOLIN and other existing vision-and-language datasets.

3.1. Dataset Collection

We collect videos from two sources: (i) 4 popular TV shows, and (ii) movie clips from YouTube channels[4] covering thousands of movies. Both sources contain rich human interactions and activities. Each episode of the TV shows is 20-40 minutes long, which we split into clips of 90 seconds long (while avoiding splitting dialogues in the middle). These 90 second-long clips may contain more than one scene, which are then presented to crowdworkers to select a video segment containing a single, self-contained scene for which they can write the statements. Additionally, we restrict the length of the selected interval to 15-40 seconds long, to maintain a reasonable difficulty level for the task. For movie clips from YouTube channels, the original lengths are around two minutes, which by nature usually contain only one scene of the movie. Thus, there is no need for the workers to manually select a video segment from the provided movie clips. We just select the first 40 seconds from every movie clip for annotation, to keep it consistent with TV show clips. Figure 2 shows the interface for AMT workers. By dragging the slider below the video player, users can adjust the start and end timestamps of the segment they want to select (for movie clips the slider is disabled).

After video segments are selected, they are presented to another group of annotators to write positive/negative statements. Each worker is assigned with one video clip, and is required to write three pairs of positive/negative statements describing the video (in the text boxes in Figure 2). We do not require AMT workers to follow any templates, as our goal is to collect diversified and natural expressions. We do have several rules/guidelines for writing positive statements: (i) We do not allow annotators to refer to characters in the video by name. Instead, they should use grounded referring expressions (*e.g.*, "the man with blonde hair wearing grey shirt", "the girl sitting in the sofa holding a cup of coffee"). The purpose of this is to keep the dataset consistent across different video sources (not all video clips have character names), and to reduce potential bias (in TV shows, the number of character names is very small). (ii) We ask workers to keep to a minimum level of copying from subtitles (*e.g.*, "somebody says ...") or describing explicit visual information (*e.g.*, object, color), and encourage them to write statements combining information from both the video clip and subtitles. (iii) We encourage workers to write about different aspects of the given video clip in different statement pairs, which may require different types of reasoning, such as inferring character emotions/relations/intentions and inferring causal relations in complex events.

[4] https://www.youtube.com/user/movieclips

Figure 2. User interface for annotators. Each annotator is provided with a video clip and required to first drag the slider below the video player to select a single-scene clip from the video, then write three pairs of positive/negative statements in the text boxes

In practice, we observe that when letting human annotators write negative statements without any constraint, the resulting statements show serious bias (*i.e.*, models can learn to classify positive/negative statements without even absorbing information from the video or subtitles). When intentionally writing fake content without any reference, humans tend to use subtle patterns that statistical models can easily pick up. Therefore, when collecting negative statements, we propose two strategies to alleviate the bias issue. First, we ask annotators to use a positive statement as reference, and only modify a small portion of it to make it negative. In this case, most part of the statement remains true to the video content, and human-introduced bias is kept to minimum. This rigorous setting makes the statements more challenging to distinguish by the model, and in-depth reasoning is required to identify the fake content. For quality control, only workers located in English-speaking countries with a lifetime task approval rate greater than 98% can participate in our study. Also, during data collection, we manually check every worker's submissions to ensure the quality of the video segments and statements.

VCR [72] proposes adversarial matching to construct wrong answers for multiple-choice QA, by selecting a correct answer (from another question) that is most similar to the current question. In our task, we use a similar strategy. For a human-generated positive statement H_i for video V_i, we select a positive statement H_j collected for another video V_j, which is most similar to H_i, and use (H_i, H_j) as a pair of positive/negative statements for video V_i. Using this strategy, a portion of the collected statements serve as both positive and negative samples, which helps removing artificial bias. Unlike the first strategy aforementioned, statement pairs constructed this way focus more on the global understanding of the video. For example, in Figure 1, the first two negative statements are written by modifying positive statements (the modified part is marked in red), and the third negative statement is obtained by adversarial matching. In the final dataset, $2/3$ of the negative statements are constructed following the first strategy, and the remaining $1/3$ with the second strategy.

3.2. Dataset Analysis

The VIOLIN dataset contains 15,887 video clips, and each video clip is annotated with 3 pairs of positive/negative statements, resulting in 95,322 (V, S, H) triplets in total. Statistics on the full dataset is provided in Table 1. Each statement has 18 words on average, and the lengths of positive and negative statements are almost the same, showing no significant bias in length.

As discussed in Sec. 3.1, we use two strategies to collect negative statements: one is adversarial matching that tests a model's ability of global video understanding; the other is modifying a small part of a positive statement for the video clip, which requires in-depth reasoning skills for a model to distinguish between positive and negative statements. To investigate in more detail, for each pair of positive and negative statements, we categorize it into 6 types of reasoning skills required, as shown in Figure 3. The types of "visual recognition", "identifying character", and "action recognition" are more focused on explicit information and require relatively low-level reasoning. "Human dynamics" includes inferring human emotions/relations/intentions, etc. "Conversation reasoning" requires performing inference over characters' dialogues and other forms of interactions (body language, hand gestures, etc.). And "inferring reasons" is about inferring causal relations in complex events. These 3 types of statement require in-depth understanding and commonsense reasoning. Overall, "explicit information recognition" makes up 54% of the dataset, and "commonsense reasoning" makes up the remaining 46%, making our dataset a balanced one, imposing new challenges on multi-facet video-and-language understanding. Compared to other datasets, our VIOLIN dataset is more focused on reasoning rather than surface-level grounding (*e.g.*, in TVQA [35], only 8.5% of the questions require reasoning).

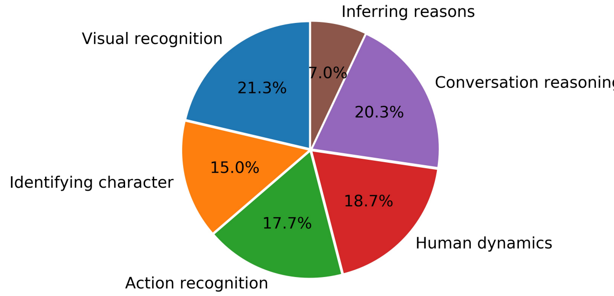


Figure 3. Distribution of reasoning types. "Visual recognition", "identifying character" and "action recognition" focus on explicit visual information; the other three require high-level inference.

4. Model

In this section, we introduce our baseline model used for benchmarking the VIOLIN dataset and evaluating the effectiveness of different feature choices. An overview of the model is illustrated in Figure 4.

4.1. Video and Text Encoders

We first extract a sequence of visual features from video frames as $\mathbf{V} \in \mathbb{R}^{T \times d_v}$, where T is the number of time steps, and d_v is the dimension of each feature. Choices of visual features will later be discussed in Sec. 5.1. The video encoder is implemented by a bi-directional LSTM, to capture the temporal correlation among consecutive frames. By passing video features into the video encoder and stacking hidden states from both directions, we obtain the video representations as $\mathbf{H}_V \in \mathbb{R}^{T \times 2d}$, where d is the hidden-state dimension of the LSTM encoder.

Statements and subtitles share the same text encoder. Statements are tokenized into a word sequence $\{w_i\}_{i=1}^{n_{stmt}}$. Each line in the subtitle is tokenized, and all the lines are concatenated together into one single word sequence $\{u_i\}_{i=1}^{n_{subtt}}$. Here, n_{stmt} and n_{subtt} are the lengths of statement and subtitle, respectively. We experiment with two types of text encoder: LSTM encoder and BERT [14] encoder. For LSTM encoder, every word token is converted to its word embedding and then fed to the LSTM encoder, producing text representations $\mathbf{H}_{stmt} \in \mathbb{R}^{n_{stmt} \times 2d}$ and $\mathbf{H}_{subtt} \in \mathbb{R}^{n_{subtt} \times 2d}$. For BERT encoder, we use pretrained BERT-base model, finetuned on VIOLIN training statements and subtitles. The output of BERT encoder at each position is 768-dimensional, which is then projected to $2d$ dimensions, also denoted as $\mathbf{H}_{stmt}$ and $\mathbf{H}_{subtt}$.

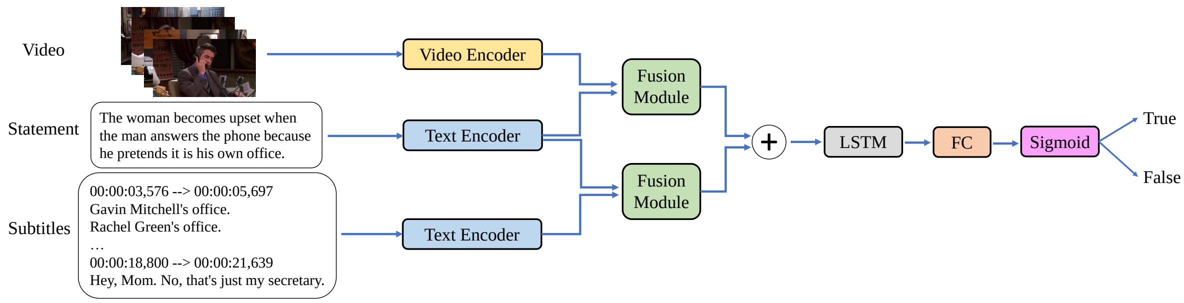


Figure 4. Overview of the proposed model for the Video-and-Language Inference task. The model takes a video (a sequence of frames), its aligned subtitles and a statement hypothesis as input, and produces a scalar measuring the probability of the input statement being positive.

4.2. Combining Multimodality Streams

The model takes three streams of information as input: video, subtitles and statement. The goal is to determine whether the statement entails or contradicts with the video and subtitles. In our model, statement representations are jointly modeled with video and subtitles via a shared fusion module. The fusion module is implemented with bidirectional attention, adopted from [50, 67, 35], where it is used for query-context matching. For simplicity, we only describe the process of combining the video and the statement streams. Subtitles and statement are fused in a similar way. Statement representations $\mathbf{H}_{stmt} \in \mathbb{R}^{n_{stmt} \times 2d}$ are used as context, and video representations $\mathbf{H}_V \in \mathbb{R}^{T \times 2d}$ as query. Each word in the statement thus attends to every time step in the video representations. Let $\mathbf{a}_i \in \mathbb{R}^T$ be attention weights for the i-th word in the statement, $\sum_{j=1}^{T} \mathbf{a}_{i,j} = 1$ for all $i = 1, \ldots, n_{stmt}$, $\mathbf{a} \in \mathbb{R}^{n_{stmt} \times T}$. The output is a video-aware statement representation: $\mathbf{M}^{V}_{stmt} = \mathbf{a}\mathbf{H}_V \in \mathbb{R}^{n_{stmt} \times 2d}$. Similarly, we combine subtitles and statement streams to obtain a subtitle-aware statement representation $\mathbf{M}^{subtt}_{stmt} \in \mathbb{R}^{n_{stmt} \times 2d}$. These two sets of representations are further fused via:

$$\mathbf{M}^{all}_{stmt} = [\mathbf{H}_{stmt}; \mathbf{M}^{V}_{stmt}; \mathbf{M}^{subtt}_{stmt}; \mathbf{H}_{stmt} \odot \mathbf{M}^{V}_{stmt}; \mathbf{H}_{stmt} \odot \mathbf{M}^{subtt}_{stmt}],$$

where $\odot$ stands for element-wise product. The resulting matrix $\mathbf{M}^{all}_{stmt} \in \mathbb{R}^{n_{stmt} \times 10d}$ combines information from all three modality streams, which is then fed into another bidirectional LSTM. The last hidden states from both directions are concatenated and passed through a fully-connected layer with 1-dimensional output followed by a sigmoid activation function, predicting the probability of the input statement being positive.

The proposed baseline model is similar to the one in [35]. The main difference is that our model uses statement representations as context and video/subtitle representations as query in the fusion module. The intuition is that, in our video-and-language inference task, the full statement needs to be supported by evidence from either the video or subtitles, in order to judge the statement to be positive/negative, instead of just locating the position in the video/subtitles that is most relevant to the query (as in TVQA [35]). Thus, in our model, every word in the statement is attended to the video and subtitles in the fusion module, then combined and fed to the final bi-LSTM to make the prediction.

5. Experiments

For evaluation, we compare our model with several baselines on the dataset and provide detailed analysis on the results. In all the experiments, we split the VIOLIN dataset into 80% for training (76,122 (V, S, H) triplets), 10% for validation (9,600 triplets) and 10% for testing (9,600 triplets). Model performance is evaluated via binary classification accuracy.

5.1. Compared Models

First, we define the following combinations of input sources, to evaluate the importance of different modality streams:

Statements Only: Using statements only, without absorbing information from video or subtitles. This option is to test the innate bias of positive/negative statements.

Video: Using video features only.

Subtitles: Using subtitles only.

Video+Subtitles: Using both video and subtitle features, which is the full setting for the task.

Single Frame+Subtitles: Using subtitle features plus only one middle frame from the video. This option is to test the usefulness of temporal information in the video.

Different visual features are also evaluated on the VIOLIN task: (i) Image feature: we use ResNet101 [22] trained on ImageNet [13] to extract the global image feature for each frame; (ii) C3D feature: we use 3-dimensional convolutional neural network (C3D) [55] to extract video features; (iii) Detection feature: we run Faster R-CNN [48] trained on Visual Genome [34] to detect objects in each frame and use their regional features as the input. For image features, we first down-sample each video to 3 frames per second,

#	Method	Visual	Text	Accuracy
0	Random	-	-	50.00
1	Stmt	-	GloVe	53.94
2	Stmt	-	BERT	**54.20**
3	Stmt+Subtt	-	GloVe	60.10
4	Stmt+Subtt	-	BERT	**66.05**
5	Stmt+Vis	Img	GloVe	55.30
6	Stmt+Vis	Img	BERT	59.26
7	Stmt+Vis	C3D	GloVe	55.91
8	Stmt+Vis	C3D	BERT	58.34
9	Stmt+Vis	Det	GloVe	56.15
10	Stmt+Vis	Det	BERT	**59.45**
11	Stmt+Subtt+SglFrm	Img	BERT	66.60
12	Stmt+Subtt+Vis	Img	GloVe	60.33
13	Stmt+Subtt+Vis	Img	BERT	67.60
14	Stmt+Subtt+Vis	C3D	GloVe	60.68
15	Stmt+Subtt+Vis	C3D	BERT	67.23
16	Stmt+Subtt+Vis	Det	GloVe	61.31
17	Stmt+Subtt+Vis	Det	BERT	**67.84**
18	Stmt+Subtt+Vis	LXMERT		66.25

Table 3. Accuracy of different methods on VIOLIN test set. Subtt = Subtitle, Vis = Video, Stmt = Statement, SglFrm = single frame, Img = Image features, Det = Detection features, C3D = C3D features, BERT = BERT features, LXMERT = LXMERT features.

and then extract the 2048-dim feature for each frame. Similarly, for detection features, we use the same sampling rate and extract features followed by a pooling layer outputting the 2048-dim feature for each frame. For C3D features, we extract 4096-dim features for every 16 frames on the original video (without down-sampling). To encode text input as features, we use (i) pre-trained BERT-base model [14] fine-tuned on VIOLIN statements and subtitles in the training set, and (ii) GloVe [46] embeddings. For thorough evaluation, we also test a large-scale pre-trained model LXMERT [53] that jointly learns multimodal features.

5.2. Experimental Results

Table 3 summarizes results from baseline methods and our proposed model (using full-length video clips, subtitles and statements). We also run a set of experiments with different visual/text features and compare the results in Table 3.

Baseline Comparison Row 0 is the random guess baseline with an accuracy of 50%. When using only the statement to decide whether itself is positive or negative, the best model with BERT features only achieves 54.20, presenting little bias in the dataset. By adding subtitles or video, all the models obtain significant gains over the "statement only" versions. Notably, Stmt+Subtt with BERT and Stmt+Vis with Det+BERT achieve 66.05 (row 4) and 59.45 (row 10), respectively. From row 3-4 and 12-17, we can observe that adding subtitles improves the performance significantly. However, the gain of adding video (row 5-10 and 12-17) is not as significant as adding subtitles. This might be due to visual features not capturing video information well. Using only one frame as video features (row 11) is worse than using full video (row 13), showing the importance of exploiting temporal information in the video. Overall, the best performance is achieved by using all the sources, with BERT and Detection features (row 17).

Source	Test Accuracy (%)
Statement	51.38
Subtitle + Statement	73.85
Video + Statement	77.19
Video+Subtitle+Statement	85.20

Table 4. Accuracy in human evaluation on test set over different input sources.

Method	Annotated	Adversarial matching
Stmt+Subtt	61.05	66.05
Stmt+Vis	57.08	59.26
Stmt+Subtt+Vis	61.99	67.60

Table 5. Accuracy (%) on test set containing negative statements collected via different strategies. Image and BERT features are used in this experiment.

Model Variants We first evaluate the effectiveness of different visual features. In most settings, Detection features work better than Image and C3D features, indicating that the extracted regional information and external knowledge from Visual Genome are useful for this task. Among all the textual features, BERT [14] is the strongest as expected. In all the settings, BERT-based versions generally improve the accuracy by 3% to 6%, compared with non-contextualized embedding such as GloVe [46]. Joint multimodal embedding (LXMERT, row 18) achieves 66.25, which is slightly worse than the best baseline model (row 17), showing that VIOLIN imposes more challenges on existing single-image-based joint pre-trained models.

Human Evaluation Human performance via AMT is presented in Table 4. As expected, humans achieve the best performance when provided with both video and subtitles (85.20)[5]. Without context (video and subtitles), humans only achieve 51.38% accuracy. Interestingly, we find that adding video brings in more gain than adding subtitles, showing the importance of visual information in VIOLIN task.

5.3. Further Analysis

Accuracy on Different Question Types To have a better understanding of the dataset, we examine the accuracy of models on different statement types on test set in Table 6. Compared to Stmt+Subtt, Stmt+Subtt+Vis models improve mostly on "visual recognition" and "action recognition".

[5] We repeated the human evaluation ourselves, and the accuracy is 93%.

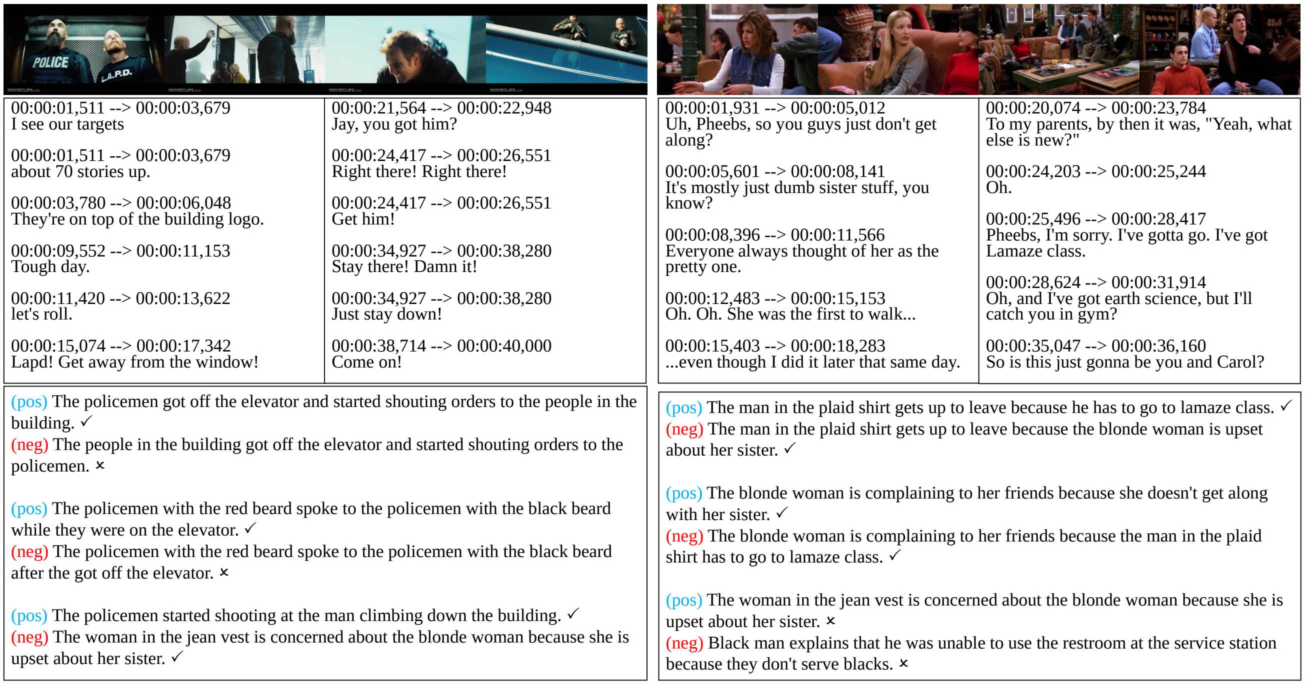


Figure 5. Qualitative analysis results. Pos/neg at the beginning of each statement indicates ground truth. ✓ or ✗ at the end of each statement represents model prediction. ✓ means the system judges the statement as positive, and ✗ means negative.

Statement Reasoning Type	Stmt+ Subtt	Stmt+Vis		Stmt+Subtt+Vis	
		Img	Det	Img	Det
Visual recognition	67.19	67.41	67.41	67.97	67.97
Identify character	57.78	64.44	65.18	62.22	62.22
Action recognition	70.75	66.04	66.04	73.58	73.58
Human dynamics	63.39	58.04	58.04	60.71	61.48
Conversation reasoning	76.23	58.20	58.20	76.23	76.23
Inferring reasons	59.52	50.00	50.31	59.52	60.18

Table 6. Accuracy (%) on each statement type in VIOLIN test set. All the methods use BERT feature.

For categories such as "inferring reasons" and "identify character", including video gains some improvement. On "conversation reasoning" and "human dynamics", adding video features does not help.

Human-Written vs. Adversarially-Sampled Negatives For comparison, we create a new statement set by replacing the adversarially-sampled negative statements with original human-written negative statements. Results are presented in Table 5. Performance on the sampled negatives is higher than that on human-written ones. Our interpretation is that human-written content has higher propensity for intent understanding and in-depth reasoning, which makes the statements more challenging to the model.

Qualitative Analysis Figure 5 presents some prediction examples from our model using statement, video and subtitles. The correct cases in Figure 5 (a) demonstrate the model's ability to recognize action, infer emotion, identify referred person, and understand temporal dynamics in the video. In (b), the error cases show that our model does not work well on inferring reasons and human relations.

6. Conclusion

We introduce a new task, video-and-language inference (VIOLIN), which requires intelligent systems to capture rich temporal signals about activities/events in video and text, in order to acquire reasoning skills for multimodal inference. We provide thorough baseline experiments for benchmarking different models on the large-scale dataset, as well as a comprehensive analysis of the dataset. The gap between the baseline models and human performance is significant. We encourage the community to participate in this task and invent stronger methods to push the state of the art on multimodal inference. Possible future directions include developing models to localize key frames, as well as better utilizing the alignment between video and subtitles to improve reasoning ability.

Acknowledgement We would like to thank Yandong Li, Liqun Chen, Shuyang Dai, Linjie Li, Chen Zhu, Jiacheng Xu and Boyi Li for providing useful feedback on the project and their help in collecting and annotating data. We thank all the reviewers for their helpful comments. The first author is supported in part by NSF under grant IIS-1546329.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 3

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016. 3

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 3

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1, 3

[5] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015. 1, 2, 3

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 3

[7] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019. 3

[8] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018. 3

[9] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. *arXiv preprint arXiv:1910.03230*, 2019. 3

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 1

[11] Yu Cheng, Quanfu Fan, Sharath Pankanti, and Alok Choudhary. Temporal sequence modeling for video event detection. In *CVPR*, 2014. 3

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017. 1

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 6

[14] 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. 2, 5, 7

[15] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 3

[16] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *CVPR*, 2017. 3

[17] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*, 2019. 3

[18] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017. 3

[19] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 3

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1, 3

[21] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6

[23] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017. 3

[24] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. 3

[25] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. In *CVPR*, 2019. 1, 3, 4

[26] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 1, 3

[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1, 3

[28] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, 2017. 3

[29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[30] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 3

[31] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 3

[32] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. In *IJCAI*, 2017. 1

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 2, 3

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1, 6

[35] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 1, 3, 4, 5, 6

[36] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 1, 3

[37] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. *arXiv preprint arXiv:2001.09099*, 2020. 3

[38] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019. 3

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2

[40] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019. 2

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1

[43] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016. 3

[44] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, 2017. 1, 4

[45] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*, 2018. 3

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 7

[47] Yunchen Pu, Martin Renqiang Min, Zhe Gan, and Lawrence Carin. Adaptive feature abstraction for translating video to text. In *AAAI*, 2018. 3

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 6

[49] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017. 3

[50] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016. 6

[51] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1

[52] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 1, 3, 4

[53] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 7

[54] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 1, 3, 4

[55] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 6

[56] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015. 3

[57] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. 2

[58] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, 2016. 3

[59] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 2, 3

[60] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, 2018. 1, 2

[61] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1, 3, 4

[62] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3

[63] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 2

[64] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 3

[65] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020. 3

[66] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 3

[67] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018. 6

[68] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 1

[69] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1

[70] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 3

[71] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, 2017. 3

[72] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 1, 2, 3, 4, 5

[73] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*, 2019. 2

A. Additional Data Analysis

A.1. Statement Length Distribution

The length distribution for positive and negative statements are presented in Figure 6 and Figure 7, respectively. There is no significant bias in statement lengths for positive and negative statements.

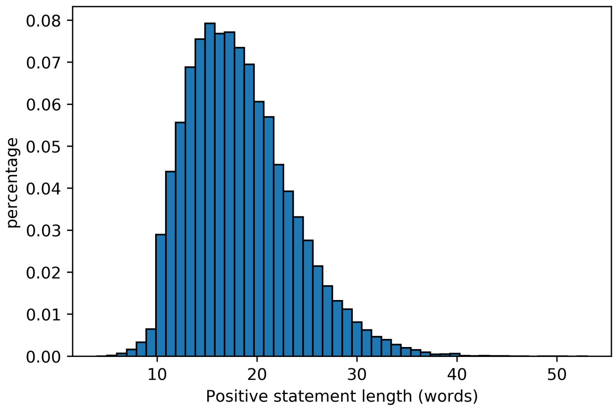


Figure 6. Distribution of positive statement lengths.

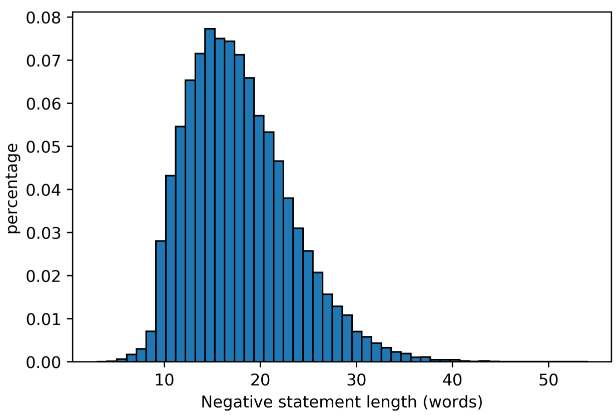


Figure 7. Distribution of negative statement lengths.

A.2. Statement Content

Table 7 shows the most common nouns, verbs and adjectives in positive statements, respectively.

A.3. Video Length Distribution

The video clips collected from MovieClips are all 40 seconds long. For video clips collected from TV shows, their lengths vary from 15 to 40 seconds, shown in Figure 8.

Type	Most Common Words
Noun	man, woman, shirt, suit, hair, jacket, girl, lady, boy, dress, sweater, friend, brunette, room, guy, people, tie, glass, table, car, coat, door, hat, phone, hand, top, bed, house, couch, group
Verb	tell, wear, ask, want, sit, try, say, talk, go, explain, walk, get, make, look, see, think, take, give, will, hold, can, stand, know, come, leave, feel, have, find, put, like
Adj	black, blue, blonde, red, white, brown, green, haired, young, dark, grey, old, other, pink, purple, upset, plaid, gray, yellow, long, little blond, happy, good, excited, surprised, striped, light, angry, short

Table 7. Most common words in positive statements.

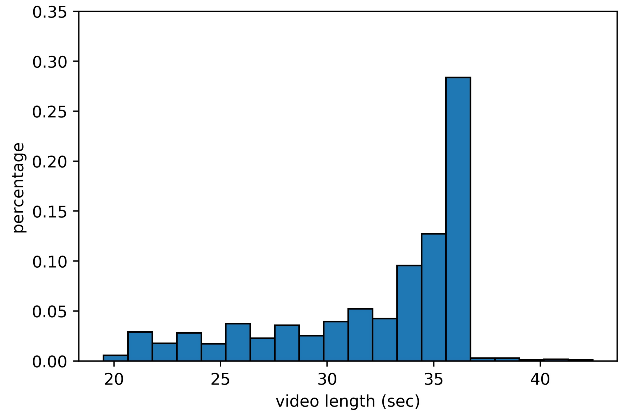


Figure 8. Distribution of video lengths from TV shows.

B. Instructions for Human Annotators

Figure 9 through 12 show the detailed instructions and user interface for human annotators.

C. More Examples

Figure 13 and Figure 14 show some more examples of predictions from our model on movie and TV show clips. The model used in these examples is Stmt+Subtt+Vis with BERT and Img features.

Instructions

Given a movie clip, you are supposed to (1) **drag the start/end slider below the video** to cut out an interval of interest that **contains a single meaningful scene**; (2) **based solely on the interval you cut**, write down a **fact/statement** inferred from the interval, and then slightly **rewrite it to contradict** the interval you cut.
The goal of the HIT task is to collect pairs of real/fake statements that are inferred from the video clip.

Task 1: Cut out an interesting and self-contained **interval** from the given video clip:

- The interval you cut should contain a single meaningful scene, use the two handlers in the slider to mark the beginning and end of the scene.
- The **hard requirement** for the length of the interval you cut is **between 15 and 30 seconds**. Do not cut your interval out of this range.
- The interval you cut should be a **single meaningful scene**
  - It **does not have to contain everything** in the scene. You can cut a scene in half to meet the length requriement, as long as the interval you cut is interesting and meaningful so that you can write real/fake statements about it.
  - Your interval should **contain only one scene**. That means, your interval should not contain flashback, change of scenes (locations), opening songs, etc.
  - In this [example video], there are three separate scenes: 0:04->0:40, 0:40->1:10, and 1:10->1:30. Each of these three intervals is a meaningful scene, since each of them contains a coherent interaction among characters. However, the first interval exceeds the time limit, so we need to cut it short. For example, we can cut it into 00:10->00:40, or cut it even shorter to 00:25->00:40. (the interval does not have to contain everything, as long as it is self-contained)
  - **In this job you only need to cut one scene and write three pairs of statements for it.**

Task 2: Infer "real" and "fake" statements from the interval you just cut:

Real statements:

- Should be **inferred** from the interval you cut: e.g. human interaction, dialog, human relation, intention, common sense, reasoning.
- **The statement you write should not directly appear in subtitles or video scenes.** Write something that is implicitly expressed in the video clip.
- Your statemetns should be only based on the scene (what you see) and the subtitles. **Do not write about the audio information** (e.g. sound in the background, character raising voice, character speaking another language that is not reflected in the subtitles)

Fake statements:

- Should be contradictory to the interval you cut, by **twisting details** in your real statement to make it fake: e.g. switching characters, changing the cause, describing a wrong activity, changing the mood, intention, facts, etc.
- However, your fake statement should be **confusing with the real one given the interval you cut**. That means, your fake statement should **not be completely out-of-context** with the real one. It should be very relevant to the video and the real statement. (more examples later)
- Your fake statement should look **very similar** to the real statement, in terms of length, structure, writing style, wording, etc.

Figure 9. Overall instructions for human annotators.

How to write real statements:

The real statement should be inferred from both the video and subtitles. The followings are some examples for accepted real statements.

- **Describing an event with its cause**
  - The people in the living room are trying to make the brunette in the kitchen calm down and get dressed because she is upset about an ex-boyfriend calling. [example video]
  - The woman a black pleads with her friend because she doesn't want to work a humiliating job. [example video]
  - The brunette girl is sitting at the counter when her friends walk in acting all sad when they're actually excited because they won the game. [example video]
  - The woman in the blue top did not want to come downstairs because it was raining and she wasn't dressed.[example video]
- **Describing a complex event**
  - The brunette in the kitchen starts freaking out and yelling at her friends about a possibly old message she heard on the answering machine.[example video]
  - The lady in the kitchen is talking to her friends about when a message was left on their answering machine.[example video]
  - A woman in a black shirt begs her friend to give her money as soon as she opens the apartment door. [example video]
  - The man in the kitchen is breaking up with his roommate's sister via letter and he's running his plan by his other friend first. [example video]
- **Describing human emotions, intentions, relations, etc.**
  - The woman in cream feels sorry for her friend but she can't help her. [example video]
  - The woman seated on the sofa is upset and crying because she did not get the apartment she wanted.[example video]
  - The man wearing the striped shirt is actually relieved he is interrupted from jumping. [example video]
  - The man with the white shirt is willing to give away the entertainment center for free. [example video]

The followings are examples of rejected real statements. Make sure to avoid them when writing your own statements.

- **Ambiguous referring**
  - **[Wrong]** She feels sorry for her friend but she can't help her. [example video]
  - **[Wrong]** The man is breaking up with his roommate's sister via letter and he's running his plan by his other friend first. [example video]
- **Use character names to refer to characters**
  - **[Wrong]** Rachel feels sorry for her friend but she can't help her. [example video]
  - **[Wrong]** Chandler is breaking up with his roommate's sister via letter and he's running his plan by his other friend first. [example video]

  You need to make your referring unambiguous given the interval you cut (by describing the characters' attribute, like the examples in the upper half of this section). Do not use names to refer to characters.
- **Directly appears in the subtitles, simple copy or rephrase**
  - **[Wrong]** The woman in black tells the woman in cream that she lost all her money and ask her for 100 dollars to get back in the game. [example video]
  - **[Wrong]** The man in black says the letter the man in grey is reading is the hardest letter he has ever written. [example video]

  Your submissions will be immediately rejected if your statements are only based on subtitles (like the ones above). Avoid writing things like "somebody says", "somebody tells somebody", "somebody thinks that", etc.
- **Unjustifiable statments (cannot be inferred from the interval)**
  - **[Wrong]** The woman in black asks for money to buy a better car. [example video]

  Even if you are familiar with the TV shows, you should only write what is presented in the clip, not to use your own knowledge about the plot.
- **Too trivial or simple statements**
  - **[Wrong]** A group of people are hanging out in the living room. [example video]
  - **[Wrong]** The man in suite walk up to the woman in black and talks to her. [example video]
  - **[Wrong]** The group of people in the living room looks at the woman in black when she yells at them. [example video]

  Please write statement that is inferred from the clip, combining information from both the video and the dialogue (subtitles).

Figure 10. Instructions for writing real statements.

How to write fake statements:

When you finish your real statements, please twist a detail in your real statements to make them into fake ones. **Most important requirments** are (1) make fake statements look similar to real ones (length, structure, style); (2) make fake statements confusing with real ones given the clip you cut (they should be in the same context of the video and subtitles).

- **Switch (change) characters**
  [Real] The woman in cream feels sorry for the woman in black but she can't help her. [example video]
  [Fake] The woman in black feels sorry for the woman in cream but she can't help her.

  [Real] The brunette in the kitchen starts freaking out and yelling at her friends about a possibly old message she heard on the answering machine.[example video]
  [Fake] The man in suite starts freaking out and yelling at her friends about a possibly old message she heard on the answering machine.
- **Change the cause of the event**
  [Real] The woman in the blue top did not want to come downstairs because it was raining and she wasn't dressed. [example video]
  [Fake] The woman in the blue top did not want to come downstairs because her camping trip was cancelled.

  [Real] The woman a black pleads with her friend because she doesn't want to work a humiliating job. [example video]
  [Fake] The woman a black pleads with her friend because she has to pay for the diner and the costumes.
- **Change a fact**
  [Real] A woman in a black shirt begs her friend to give her money as soon as she opens he apartment door. [example video]
  [Fake] A woman in a black shirt begs her friend to give her money as soon as the woman in cream steps into the apartment.

The followings are examples of rejected real statements. Make sure to avoid them when writing your own statements.

- **Fake statements too different from real ones**
  [Real] The brunette in the kitchen starts freaking out and yelling at her friends about a possibly old message she heard on the answering machine.[example video]
  [Fake] **[Wrong]** The brunette woman yells at her friends for not leaving her the message. (Lengths differ too much. Real and fake statements should roughly have the same length)

  [Real] The woman a black pleads with her friend because she doesn't want to work a humiliating job. [example video]
  [Fake] **[Wrong]** The woman has to pay for her diner and costume, and so she pleads with her friend. (Sentence structure and style differ too much. See good examples in the first half of this section)
- **Fake statements not confusing with real ones IMPORTANT: this is the most common mistake that turkers made**
  [Real] The people in the living room are trying to make the brunette in the kitchen calm down and get dressed because she is upset about an ex-boyfriend calling. [example video]
  [Fake] **[Wrong]** The people in the living room are trying to make the brunette in the kitchen calm down and get dressed because she was robbed.
  [Fake] **[Wrong]** The people in the living room are trying to make the brunette in the kitchen calm down and get dressed because she was fired by her boss.
  [Fake] **[Wrong]** The people in the living room are trying to make the brunette in the kitchen calm down and get dressed because her friends did not ask her out for dinner.
  [Fake] **[Wrong]** The people in the living room are talking about the baseball game last night because they are huge baseball fans.
  All of the above three fake statements are not acceptable. Even though they are plausible, they are completely **out-of-context** given the video clip. A good fake statement is:
  [Fake] **[Good]** The people in the living room are trying to make the brunette in the kitchen calm down and get dressed because she is upset about breaking up with the man in suite.

  Suggestion: If your real statement has some overlap with the subtitles (e.g. 'message', 'calling'), make sure your fake statement also has the same level of overlap with the subtitles (e.g. 'break up'). For example, if your real statement and subtitles has 8 words in common while your fake statement and subtitles only have 2 words in common, then your real and fake statements are not confusing at all since you can tell which one is real by simply comparing the subtitles with the statements without really understanding the scene.

Figure 11. Instructions for writing fake statements.

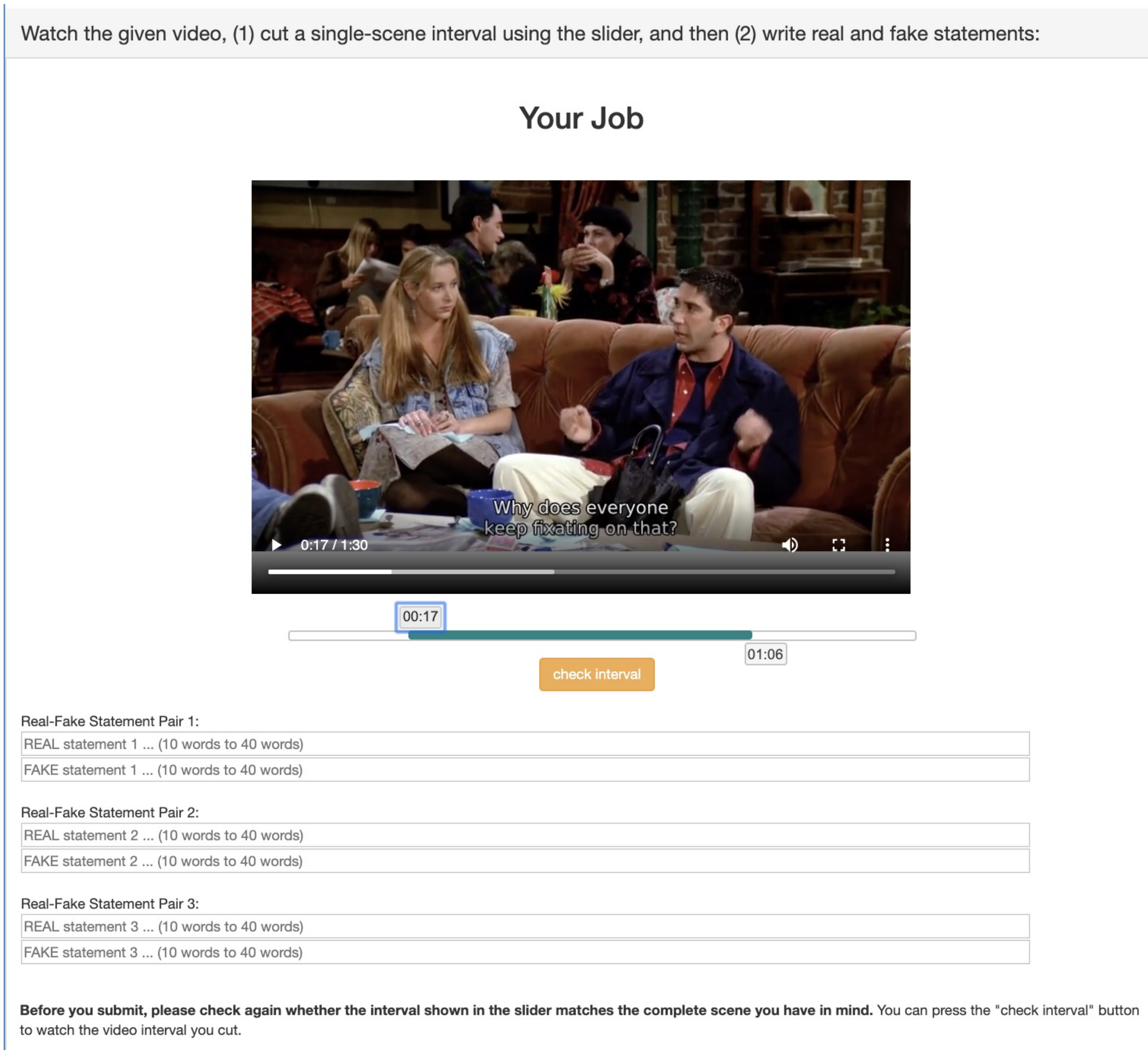


Figure 12. User interface for workers to cut an interval from the video and write three pairs of real and fake statements.

00:00:13,640 --> 00:00:20,740 Monsieur Dilek Anthony	00:00:29,789 --> 00:00:33,729 such as I require from your good self I
00:00:20,750 --> 00:00:24,800 Madame Sabine de Barra I'm very happy to	00:00:33,739 --> 00:00:35,720 brought with me a selection of
00:00:24,810 --> 00:00:27,800 meet you I am directed by the office of	00:00:35,730 --> 00:00:40,000 perennials I thought we might exchange
00:00:27,810 --> 00:00:29,779 buildings of Versailles to order shrubs	

(pos) A woman in a leather jacket approaches a man as he is resting on a bench. ✓
(neg) A woman in a leather jacket approaches a man on a bench alone crying. ×

(pos) A woman in a leather does a curtsy to the man resting on the bench in front of her. ✓
(neg) A woman in a leather jacket gives a hand salute to the man resting on the bench. ×

(pos) The man relaxing on the bench looks annoyed at the woman who has disturbed him. ×
(neg) The man in grey shirt talks with the man in green t-shirt that he hope he didn't throw him off his game. ×

(a)

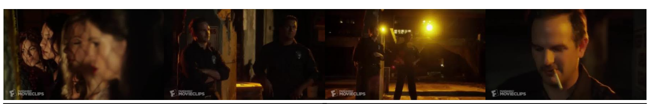
00:00:03,085 --> 00:00:08,328 Uh, I'll just... fire away then,	00:00:21,897 --> 00:00:24,020 you ever thought of having, um,
00:00:03,085 --> 00:00:08,328 shall I ?	00:00:24,108 --> 00:00:27,856 more, uh, horses in it.
00:00:09,925 --> 00:00:11,965 Right.	00:00:30,991 --> 00:00:34,526 Uh, well, we would have liked to,
00:00:15,431 --> 00:00:18,718 The film's great, and, um,	00:00:34,620 --> 00:00:39,959 but it was, um, difficult,
00:00:18,811 --> 00:00:21,812 I just was wondering whether...	00:00:34,620 --> 00:00:39,959 obviously, being set in space.

(pos) Brunette man sits across a table from a woman in a suit and informs her he is going to ask several questions. ✓
(neg) Brunette man sits across a table from a woman in a suit and informs her he is going to sit in silence. ×

(pos) Brunette man asks the woman in a suit if she would have preferred to add more animals to something and she confirms. ✓
(neg) Woman in suit asks the brunette man if she would have preferred to add more animals to something and he confirms. ✓

(pos) Woman in suit explains to brunette man that having many animals present is difficult. ✓
(neg) The man in a black and white top is curious about if the boy in bed has finished what he needed to do for the day. ×

(b)



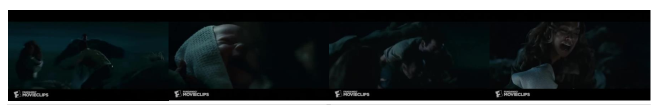
00:00:00,000 --> 00:00:03,230 cops never bother iron John something	00:00:18,020 --> 00:00:31,510 you heard the woman
00:00:03,240 --> 00:00:06,680 must be going on it's red	00:00:31,520 --> 00:00:36,460 place creeps me out man hot we think
00:00:06,690 --> 00:00:10,510 we don't know that you have to be sure	00:00:36,470 --> 00:00:38,100 they're gonna do it iron Jonathas boys
00:00:10,520 --> 00:00:18,010 they look weak let's take them all right	00:00:38,110 --> 00:00:40,000 cares hold over sir

(pos) The woman are suspicious of the reasons why a man is being bothered by the cops. ✓
(neg) The woman are assured by the presence of the cop in the company of the man. ✓

(pos) The woman evaluate the vulnerability of the men and assume they can prevail. ✓
(neg) The woman are fearful of the men they watch, and they're scared to act. ✓

(pos) One man is apprehensive by the location he is shares in the company of the other man. ✓
(neg) The woman in the purple and blue pajamas jumps on-top of the man laying in the bed. ×

(c)



00:00:16,210 --> 00:00:18,210 Give me the child	00:00:27,520 --> 00:00:29,520 Then Death will come to you both
00:00:22,570 --> 00:00:24,570 No	00:00:32,549 --> 00:00:34,550 No

(pos) A blonde woman feels really scared and screams when a man comes flying and hits a man in a white shirt. ✓
(neg) A blonde woman feels relief and laughs when a man comes flying to rescue her and hits a man in a white shirt. ×

(pos) A blonde woman is running away from a man who came flying trying to protect a baby from him, the baby is upset and cries. ✓
(neg) A blonde woman is running away from a man who came flying trying to protect a baby from him, the baby just smiles unaware of everything. ×

(pos) A man wearing a white t-shirt sacrifices himself for a woman and a baby, a blonde woman gets really upset when he falls of the cliff and cries. ✓
(neg) The woman and the man come close to fighting over the man's desires. ✓

(d)

Figure 13. Examples on movie clips. The pos/neg at the beginning of each statement indicates its ground truth. The ✓ or ✗ at the end of each statement indicates the system's prediction. ✓ means the system judges the statement as positive, and ✗ means negative.

00:00:01,169 --> 00:00:02,429
Block me,come on!

00:00:02,459 --> 00:00:04,949
Damn,they must have snuck out again.

00:00:05,049 --> 00:00:07,819
Again? How often
does this happen?

00:00:08,189 --> 00:00:10,469
Boys,get in here now!

00:00:10,489 --> 00:00:11,579
Honey,you know how slippery they are.

00:00:11,629 --> 00:00:13,249
It's like trying to herd cats.

00:00:13,309 --> 00:00:15,869
Tom,it's 9:00 at night.

00:00:15,899 --> 00:00:18,519
Guys,upstairs now.
Let's go.

00:00:18,549 --> 00:00:22,349
Into pj's,into bed.
Honey,they're fine.

00:00:22,389 --> 00:00:24,549
You just... you worry too much.

00:00:24,579 --> 00:00:26,199
And for good reason.

00:00:26,229 --> 00:00:27,959
Someone could have driven off with
them,

00:00:27,989 --> 00:00:28,949
and you wouldn't have even noticed.

(pos) The man in the blue shirt is frustrated because the kids keep running outside when he isn't looking. ✓
(neg) The man in the blue shirt is frustrated because the kids keep getting into cookies when he isn't looking. ×

(pos) The woman in the grey suit is upset that the kids weren't in house because it was dark out. ✓
(neg) The woman in the grey suit is upset that the kids weren't in the house because they are sick. ×

(pos) The man in the blue shirt thinks the woman in the grey suit worries too much about the kids. ✓
(neg) The man in a black shirt and the lady in a brown hair and black dress sit outdoor during their conversation. ×

(a)

00:00:00,000 --> 00:00:01,009
Me, too.

00:00:01,500 --> 00:00:02,839
I miss my wife.

00:00:04,599 --> 00:00:08,129
Hey, let's go around the table and say what
our favorite part was.

00:00:08,229 --> 00:00:10,539
Mine was that thing with the typewriter.

00:00:10,919 --> 00:00:13,869
I mean, she made some spelling mistakes,
but still.

00:00:14,799 --> 00:00:17,439
Ooh, and you guys were all, "Barney, put
out the cigar!

00:00:17,449 --> 00:00:20,959
It's a non-smoking room!" And I was all,
"Hell, no, this is a Cuban!"

00:00:21,699 --> 00:00:24,329
Of course, eventually, I did put it out.

00:00:25,369 --> 00:00:26,389
Did I put it out?

00:00:27,129 --> 00:00:28,149
I put it out.

00:00:29,209 --> 00:00:30,509
Did I put it out?

00:00:31,750 --> 00:00:33,000
I put it out.

00:00:34,699 --> 00:00:35,259
Did I put it out?

(pos) The man with the black top and the man with the black hair react with disgust when the blonde man recounts his favorite act. ✓
(neg) The man with the black top and the man with the black hair react with amusement when the blonde man recounts his favorite act. ✓

(pos) The blonde man is telling the men at the table about his favorite parts from the event they attended. ✓
(neg) The man in the green shirt is telling the men at the table about his favorite parts from the event they attended. ×

(pos) The man with black hair reacts with annoyance when the man with blonde hair is trying to remember if he put out his cigar. ✓
(neg) The man in the suit and the woman in the blue shirt reminisce about how times have been since they saw each other last. ×

(b)

00:00:00,000 --> 00:00:00,240
Uh, 9.

00:00:02,030 --> 00:00:03,410
But it's dark out.

00:00:03,573 --> 00:00:06,783
Um, well, that's because you always
sleep till noon, silly.

00:00:08,161 --> 00:00:10,750
This is what 9 looks like.

00:00:11,831 --> 00:00:14,250
I guess I'll get washed up then.

00:00:14,542 --> 00:00:16,881
Watch that sunrise.

00:00:19,380 --> 00:00:22,170
I'm really getting tired of sneaking around
all the time.

00:00:22,341 --> 00:00:26,641
I know, me too. Hey, what if we went
away for the whole weekend? No
interruptions.

00:00:26,804 --> 00:00:28,933
And we could be naked the entire time.

00:00:29,515 --> 00:00:30,555
All weekend?

00:00:30,725 --> 00:00:32,424
- That's a whole lot of naked.
- Mm-hm.

00:00:33,102 --> 00:00:36,610
Yeah, I could say I have a conference and
you can say you have a chef thing.

(pos) A man in a grey shirt has a confused conversation about what time of the day it is. ✓
(neg) A man in a grey shirt has a confused conversation about where he currently living. ✓

(pos) A man and a women in a robe have a conversation about going away for a weekend. ✓
(neg) A man and a women in a robe have a conversation about a week long stay-cation. ×

(pos) A confused man in a grey shirt goes into to the bathroom to wash-up. ✓
(neg) The woman in the red dress is thrilled that she got divorced and took all of her ex-husband's money in the divorce. ×

(c)

00:00:00,000 --> 00:00:00,780
by looking through their personal things?

00:00:00,790 --> 00:00:03,480
Not just now. Since the invention of things.

00:00:03,480 --> 00:00:05,310
Is that one of Claire's brownies?

00:00:05,310 --> 00:00:07,130
No, they're delicious. Must be Cam's.

00:00:07,130 --> 00:00:09,000
- Oh, no. - Don't worry. There's more.

00:00:09,000 --> 00:00:11,540
No! That's where my backpack is.

00:00:11,550 --> 00:00:13,520
Cam drove me and Luke home.

00:00:13,740 --> 00:00:15,060
I must've left it in his car.

00:00:15,060 --> 00:00:16,390
Why are you so scattered lately?

00:00:16,390 --> 00:00:20,040
Aw, he's just nervous because of his
poetry reading tonight.

00:00:20,410 --> 00:00:21,960
But don't worry, papi.

00:00:21,960 --> 00:00:24,320
I am going to be there to support you.

00:00:24,990 --> 00:00:26,570
I don't want my mom there.

00:00:26,930 --> 00:00:28,490
I'm exploring some darker themes

00:00:28,490 --> 00:00:28,510
I'm not sure she's ready for.

(pos) The kid in white is worried because he can't find his backback. ×
(neg) The kid in white is worried because he can't find his homework. ×

(pos) The brunette woman reassures the kid in white that she'll go to his poetry reading. ✓
(neg) The brunette woman reassures the man in blue that she'll go to his poetry reading. ✓

(pos) The brunette woman is curious who made the brownie that the man in blue is eating. ✓
(neg) The woman in the blue shirt reacts with confusion when the man in the red shirt informs her that he didn't order any lemonade. ×

(d)

Figure 14. Examples on TV show clips. The pos/neg at the beginning of each statement indicates its ground truth. The ✓ or ✗ at the end of each statement indicates the system's prediction. ✓ means the system judges the statement as positive, and ✗ means negative.