Explainable Video Entailment with Grounded Visual Evidence

Junwen Chen and Yu Kong
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY, USA
{jc1088, yu.kong}@rit.edu

Abstract

Video entailment aims at determining if a hypothesis textual statement is entailed or contradicted by a premise video. The main challenge of video entailment is that it requires fine-grained reasoning to understand the complex and long story-based videos. To this end, we propose to incorporate visual grounding to the entailment by explicitly linking the entities described in the statement to the evidence in the video. If the entities are grounded in the video, we enhance the entailment judgment by focusing on the frames where the entities occur. Besides, in the entailment dataset, the entailed/contradictory (also named as real/fake) statements are formed in pairs with subtle discrepancy, which allows an add-on explanation module to predict which words or phrases make the statement contradictory to the video and regularize the training of the entailment judgment. Experimental results demonstrate that our approach outperforms the state-of-the-art methods.

1. Introduction

Bridging the gap between computer vision and natural language processing is a rapid growing research area in various tasks including visual captioning [40, 34], VQA [20, 1, 33], and visual-textual retrieval [22, 23]. Liu et al. [25] introduced a new video entailment problem to infer the semantic entailment between a premise video and a textual hypothesis. As shown in Fig. 1, video entailment [25] task aims at determining whether a textual statement is *entailed* or *contradicted* by a video. In Fig. 1, the label for the first statement with the premise is *entailment* because the statement can be concluded from the dialog of the first clip in which "the woman wearing jeans" appears. On the contrary, the second statement is labeled as *contradiction*, because the premise does not have evidence to conclude the statement. In this paper, we aim to address the video entailment with a faithful explanation.

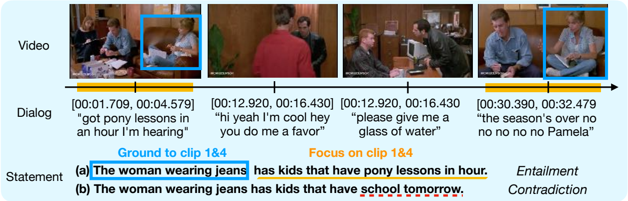

Video
Dialog
[00:01.709, 00:04.579] "got pony lessons in an hour I'm hearing"
[00:12.920, 00:16.430] "hi yeah I'm cool hey you do me a favor"
[00:12.920, 00:16.430 "please give me a glass of water"
[00:30.390, 00:32.479 "the season's over no no no no Pamela"
Ground to clip 1&4
Focus on clip 1&4
Statement
(a) The woman wearing jeans has kids that have pony lessons in hour.
Entailment
(b) The woman wearing jeans has kids that have school tomorrow.
Contradiction

Figure 1. Video entailment aims at judging if a statement is entailed or contradicted by a video and its aligned textual dialog. A pair of real and fake statements have the similar structure and subtle difference (marked by the red dot line). We incorporate visual grounding into the entailment judgment. The entity grounding, *e.g.*, "A woman wearing jeans" guides the entailment judgment module to focus on the entity-relevant frames and the corresponding sentences in the dialog (marked by blue in the temporal axis) to make a correct judgment. Best viewed in color.

The main challenge of video entailment is that it requires fine-grained reasoning to understand the complex story-based videos and then make a correct judgment. The story-based videos are also accompanied by the textual dialog (subtitles) (see Fig. 1). In the existing method for video entailment [25], video frames are less exploited than dialog, because it lacks of a fine-grained understanding of the video and the model does not know which frames in the long video are related to the statement. However, the entities in the textual statement are usually people with their attributes, *e.g.*, "A woman wearing jeans" (see Fig. 1), which should be implied in the video frames instead of the dialog.

To this end, we propose to enhance the entailment judgment by introducing a visual grounding model that links the entity described in a statement to the evidence in the video. This is motivated by the fact that the statement is usually only related to a small subset of the long and untrimmed video. Based on this, a visual grounding module for the entities described in the statement is developed to localize the clips where the entity appears and guide the judgment to focus on the entity's occurring clips as well as the aligned sentences in the dialog. For example, the statements in Fig. 1 are linked to the first and fourth clips and sentences, considering the entity ”The woman wearing jeans”. By highlighting the relevant clips and sentences, the details can be better understood compared to [25] that does not have grounding guidance and equally considers all of the frames.

Visual grounding has been attempted in many video+language tasks, such as image captioning [38] and VQA [21]. However, it cannot be directly generalized to the entailment task, because the bounding box annotations of grounding are not provided in the entailment dataset. Therefore, we resort to the existing weakly-supervised object grounding methods [15, 5] to address the training of the grounding module. But these methods are limited to explicit natural objects (*e.g*., “apple”, “river”). Our grounding is more demanding, as we target at the described entities with fine-grained attributes, such as hair, clothes and gender, to be grounded to the challenging story-telling videos.

Furthermore, we aim at improving the faithfulness of the entailment model by evaluating if the entailment is judged based on correct evidence. A faithful entailment model should tell not only *whether the statement is contradictory to the video* but also *which words or phrases in the statement make it contradictory to the video*. A pair of real/fake statements usually have a similar structure and only have very subtle differences, with only a small number of words' replacement, *e.g*. “pony lessons in hour” and “school tomorrow” marked by the red dot line in Fig. 1. Thus, we propose to regularize the training of the entailment judgment module by encouraging the local explanation on the contribution of the words in the statement to conform to the subtle difference.

Our main contribution is threefold. First, we propose a novel approach to address video entailment with visually grounded evidence. Second, we exploit the pairwise real/fake statements to add the explainability to the entailment model, which can tell the specific words or phrases that make the statement contradictory to the video. Third, extensive results demonstrate that our method outperforms the state-of-the-art video entailment method.

2. Related Work

2.1. Visual Entailment

Natural language inference [9, 8, 26, 3] is the task of understanding if a hypothesis sentence is entailed or contradicted by a premise sentence, which is a fundamental task in natural language understanding. Inspired by the textual entailment, recently visual entailment is proposed to extend NLI to the visual domain. In visual entailment, the premise is an image or a video. And the goal is to predict if the textual hypothesis can be confirmed in the visual premise.

Recently, researchers began to solve visual entailment mainly on image premise. SNLI-VE [35] is a visual entailment dataset combining the textual entailment [2] and Flickr30k image caption [36]. It also provides a solution model that utilizes ROI generation and models the fine-grained cross-modal information. However, the hypothesis (*e.g*., “The two women are holding packages”) is much more straightforward compared to the hypothesis in our video entailment. e-SNLI-VE-2.0 [12] appends and corrects SNLI-VE [35] by the human-written language hypothesis. It also provides the explanation ground-truth of why the hypothesis is entailed/contradicted by the premise. NLVR2 [32] is another image entailment dataset that requires quantitative and comparing reasoning. But similar to SNLI-VE [35], it also mainly focuses on objects in the natural images.

Recently, Liu et al. [25] proposed VIOLIN dataset that focuses on video entailment. Video entailment is a challenging task as the complex temporal dynamics occur in the video. A fine-grained reasoning of the social relations, human motions and intentions is necessary to understand the story-based content and make a correct judgment.

2.2. Grounding for Video+Language Reasoning

Recently, many video+language tasks have been trying to explicitly link the language sentence to the evidence in the video. Zhou et al. [38] proposed a video description dataset with the annotation of the bounding boxes of the referred objects. With this dataset, a good captioning model is desirable by attending to appropriate video regions. For video question answering, Lei et al. [21] built a dataset with the spatio-temporal grounding annotation, which requires the model to localize the temporal moments, detected the referred object, and answer the questions.

Different from captioning and VQA, video entailment needs a fine-grained understanding of the entities with detailed attributes. Meanwhile, the existing video entailment does not provide the grounding annotation. Thus, we propose to achieve the entity grounding in a weakly-supervised manner.

2.3. Weakly-supervised Entity Grounding

Visual grounding is to localize the described entity to its occurring regions visually. Since the annotation of bounding boxes is very expensive, sundry efforts have been made to achieve object grounding in a weakly-supervised manner [15, 5, 29], mainly based on multiple instance learning. It also has been extended to video domain [39, 30, 14, 7, 6, 4], to achieve spatio-temporal grounding of entities in an untrimmed video.

In the video entailment task, the visually related entities are mainly characters, while the existing grounding methods aim at grounding natural objects. Our grounding requires a fine-grained understanding of human gender, dress, hair and other attributes. Therefore, we cannot directly gen-

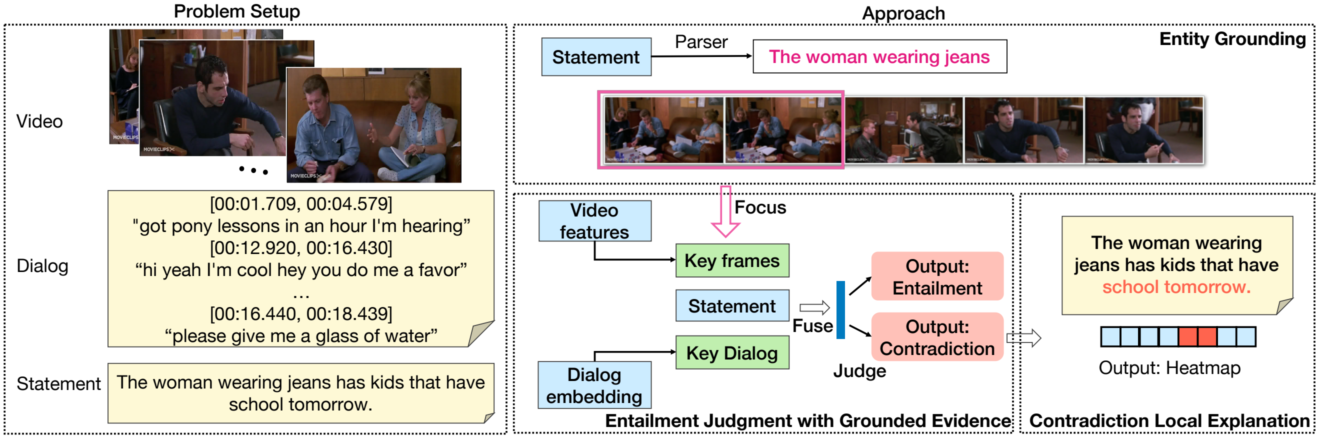

Problem Setup
Video
Dialog
[00:01.709, 00:04.579]
"got pony lessons in an hour I'm hearing"
[00:12.920, 00:16.430]
"hi yeah I'm cool hey you do me a favor"
[00:16.440, 00:18.439]
"please give me a glass of water"
Statement
The woman wearing jeans has kids that have school tomorrow.
Approach
Statement
Parser
The woman wearing jeans
Entity Grounding
Focus
Video features
Key frames
Statement
Key Dialog
Dialog embedding
Fuse
Judge
Output: Entailment
Output: Contradiction
The woman wearing jeans has kids that have school tomorrow.
Output: Heatmap
Entailment Judgment with Grounded Evidence
Contradiction Local Explanation

Figure 2. Given a video, its aligned dialog in text, and a textual statement for the video as input, our goal is to predict if the statement is *entailed* or *contradicted* by the video and dialog. Our model consists of three sub-networks: Entity Grounding, Entailment Judgment with Grounded Evidence, and Contradiction Local Explanation. The entity grounding module helps to find if the described entity occurs in the video clips. Moreover, entity grounding guides the judgment module to focus on the entity-relevant clips and the corresponding sentences in the dialog (marked as "Key"), to make a correct judgment. If judged as "contradiction", our model can also explain which words or phrases in the statement make it contradictory to the video by generating an explanation heatmap.

eralize the existing grounding methods to video entailment.

2.4. Multi-modal VQA

Different from image entailment, video entailment is supposed to understand story-based video content, such as movies. This is more challenging than the plain videos as multiple factors such as human interactions, emotions, motivation, and scenes appear. Similar to existing videoQA datasets [21, 22], the input to our entailment task is multi-modal, including both videos and textual subtitles. For multi-modal VQA, early fusion was commonly used in merging different modalities [27]. Recent methods mainly leverage late fusion approaches [18, 16]. Another aspect [17] is to utilize the content of QA pairs to shift to the relevant modality and constrain the contribution of the irrelevant ones.

Video entailment requires a fine-grained understanding. The statement may only relate to the details in a long and untrimmed video. Thus, we propose to ground the described entities to their occurring clips and highlight the dialog sentence aligned to those clips for entailment judgment.

3. Our Approach

Given a story-like video aligned with a textual dialog (subtitles) and a hypothesis statement, the entailment task is to predict if the hypothesis statement is *entailed* or *contradicted* by the premise video (see the left of Fig. 2). The right part of Fig. 2 shows the overall pipeline of the proposed method. We decompose our model into three sub-networks: entity grounding, entailment judgment with grounded evidence, and contradiction local explanation, to address entailment in a modularized manner.

The motivation of grounding entities described in the statement (*e.g.*, "a woman wearing a red cape") to frames comes from the observation that video modality is not well exploited compared to dialog modality in the existing method [25]. However, many contradictory statements such as the incorrect attributes should be determined from the frames instead of the dialog, (*e.g.*, "a woman wearing a blue cape") in Fig. 3. Moreover, the statements are written about different aspects of a video [25], and a statement is usually related to a small subset of video frames. The entity grounding helps to find the entity-relevant frames and then guides the entailment judgment module to highlight these frames. To learn a credible entailment judgment model, we propose to not only judge the semantic entailment but also explain which words or phrases make the statement contradictory to the video by a heatmap that indicates the contribution of each word in the statement to the model prediction.

3.1. Preliminaries

Text Representation. Following VIOLIN [25], we use BERT encoder [10] provided by VIOLIN to represent the statement and dialog, resulting in a 768-dimension vector for each word. Then using a bi-directional LSTM for both statement and dialog, each word is also embedded to d-dimension. A statement is tokenized into a word sequence, in the length of N_l. A textual dialog is also tokenized and represented as a word sequence. Then, by encoding, the statement is represented as $R = \{r_i\}_{i=1}^{N_l}$ in which r_i indicates the i-th word's representation. The dialog is represented as $H = \{h_j\}_{j=1}^{N_s}$, in which h_j indicates the j-th

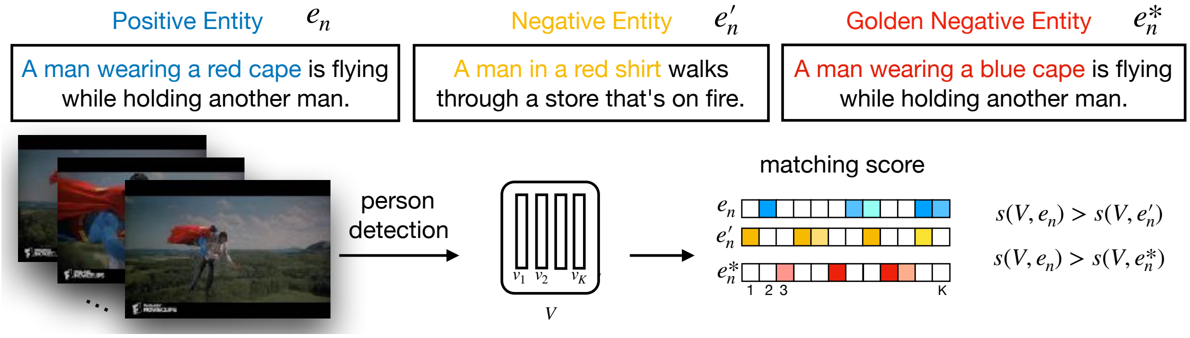

Positive Entity
e_n
Negative Entity
e'_n
Golden Negative Entity
e^*_n
A man wearing a red cape is flying while holding another man.
A man in a red shirt walks through a store that's on fire.
A man wearing a blue cape is flying while holding another man.
person detection
v_1 v_2 v_K
V
matching score
e_n
e'_n
e^*_n
1 2 3
K
$s(V, e_n) > s(V, e'_n)$
$s(V, e_n) > s(V, e^*_n)$

Figure 3. Training of the entity grounding module. We extract the positive entity e_n from the statement aligned with the video and the negative entity e'_n from a statement unaligned with the video. Besides, real and fake statements are formed in pairs. Thus, the entity in the fake statement can be utilized as a golden negative entity e^*_n, which is slightly different from e_n and is a hard sample to enhance the grounding model's training. The training process encourages the matching score of the positive entity to be larger than any negative entity. Best viewed in color.

word's representation. N_s denotes the number of words in the long dialog. The starting time t^j_s and the ending time t^j_e of the j-th sentence are also provided, which can be aligned with the video frames.

Video Representation. Following VIOLIN [25], we extract a sequence of visual features from video frames and then encode the visual features by a bi-directional LSTM layer. The video is then represented as $\mathbf{C} \in \mathbb{R}^{T \times d}$, where T is the number of frames, and d is the feature dimension of each frame.

To realize grounding, we first detect the people in the input video. Specifically, we extract the frames of the middle timestamps corresponding to each sentence $(t^j_s + t^j_e)/2$ and apply Faster R-CNN [28] pretrained on COCO [24] to detect all of the people from each frame and extract their features. Each person is represented by a 4096-dimension vector, denoted as v_k. Then each video is formed as a set of persons $V = \{v_k\}^K_{k=1}$, where v_k encodes the k-th person.

3.2. Entity Grounding Module

In the existing video entailment method [25], the performance gain of video modality is limited compared to dialog modality. Visual information needs fine-grained understanding, but the existing work equally considers all of the frames even if the frames are not relevant to the statement. Video modality should be responsible for a lot of information described in the statement such as entity attributes (*e.g.*, gender and clothes). We propose to leverage entity grounding in the video modality to improve the entailment judgment in a modularized manner (see Fig. 2). First, our grounding module is developed to achieve spatio-temporal grounding of the subject entity described in the statement. The predicted temporal occurrences of the entity are used to guide the following cross-modal entailment judgment.

However, two technical challenges need to be handled to leverage visual grounding for the entailment task. First, spatial-temporal annotations of entities are typically not available for the entailment task so that existing fully-supervised grounding-based VideoQA methods [21] cannot be directly leveraged. We resort to multiple instance learning [39] to achieve entity grounding in a weakly-supervised fashion. Second, detailed visual attributes (*e.g.*, clothes and hair) of entities are essential for the entailment task but they are typically ignored by the existing object grounding methods [30, 39, 4].

To extract the entity and its attributes from a textual statement, we employ a constitute parsing method [19]. For example, in Fig. 2, "The woman wearing jeans" is an entity extracted from the corresponding statement "The woman wearing jeans has kids that have pony lessons in hour". The extracted entities in a statement are denoted as $E = \{e_n\}^{N_e}_{n=1}$, where N_e is the total number of entities and e_n indicates the n-th entity.

To ground the entity to its occurring frames, we compute the matching score $s(V, e_n)$ between video V and an entity e_n as:

$$s(V, e_n) = \frac{1}{K} \sum_{k=1}^{K} \sigma(FC_1(v_k || e_n)) \qquad (1)$$

where FC_1 is a fully-connected layer and σ is the sigmoid activation. We take average of the scores of the K people as the entity-video matching score $s(V, e_n)$.

Following the existing visual-textual matching work [23, 4, 39], we formulate the weakly-supervised learning of grounding as:

$$\mathcal{L}_{ga} = -\log(1 - s(V, e'_n)) - \log(s(V, e_n)), \qquad (2)$$

where e'_n is a "negative entity" extracted from a randomly sampled statement from another video, which is different from e_n. Eq. 2 encourages that the aligned video-entity pair (V, e_n) to better matched and the unaligned pair (V, e'_n) to be less matched.

Different from weakly-supervised video grounding [4, 39], the entailment task consists of the real/fake statements in pairs. Thus, we have the opportunity to obtain hard negative samples, which is the entity described in the fake statement but NOT described in the real statement. As shown in Fig. 3, the negative version is “a man wearing a blue cape”, which is very similar to the positive one “a man wearing a red cape” but is contradicted by the video. We name it as “golden negative entities” e_n^* and use it in training the grounding module:

$$\mathcal{L}_{gb} = -\log(1 - s(V, e_n^*)) - \log(s(V, e_n)), \tag{3}$$

$\mathcal{L}_{gb}$ encourages the video V to match more to its aligned entity e_n and less to the golden negative entity e_n^*. To sum up, we train the grounding model by the grounding loss $\mathcal{L}_g$ which balances the negative entities and the golden negative entities by β.

$$\mathcal{L}_g = \mathcal{L}_{ga} + \beta \mathcal{L}_{gb}, \tag{4}$$

During the inference, if the matching score $s(v_k, e_n) = \sigma(FC_1(v_k || e_n))$ between a person v_k and an entity e_n exceeds a threshold, we consider that the k-th people is e_n. The temporal grounding result will be used to guide the entailment judgment in Sec 3.3.

3.3. Entailment Judgment with Grounded Evidence

Statements are usually related to a small subset of the video, instead of the entire video. For example, in Fig. 2, the clause in the statement “kids that have school tomorrow” should be judged from the first sentence in the dialog. Thus, we utilize the entity grounding result to highlight the frames and the corresponding textual dialog in the temporal range that the entity occurs, since the frames and dialog are aligned by temporal boundaries. The highlighted frame and dialog embeddings are concatenated and marked as key embedding C_O, H_O.

The model takes three streams in different modalities as input: video frames, dialog, and statements. We leveraged the visually grounded evidence to make our model fixate its attention on the frames where the entity appears. Then, we fuse the multi-modal data and predict whether the statement is entailed or contradicted by the video.

To bridge the modal discrepancy between the video frames and textual content, we use heterogeneous reasoning [37] to fuse the statement representation R with different context embedding, including video embeddings C, dialog embeddings H and key embeddings C_O, H_O (see Fig. 4) respectively. The heterogeneous reasoning is based on a graph convolution layer [7]:

$$P_* = A_{*\to s} X_* W_{*s}, \tag{5}$$

where $*$ denotes one of the context among video C, dialog D and key C_O, H_O and Adjacency matrix $A_{*\to s}$ contains the similarity between the statement R and the context embedding X_*. Eq. 5 projects the context X_* to an R-shaped embedding P_* by a learnable linear layer W_{*s}. Then, to avoid forgetting, we learn a gating function z_* by a linear operation W_*, b_* and constrained activation $sigmoid$,

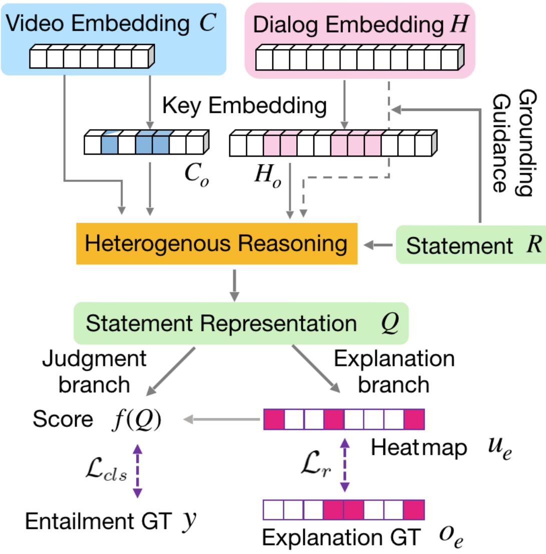


Figure 4. Our multi-task learning framework for entailment judgment and its explanation. Given the video and dialog embedding, we use heterogeneous reasoning to fuse them and update the statement representation. Then, the statement representation is incorporated into two branches: the judgment branch to predict if it is entailed or contradicted and the explanation branch to generate a heatmap that shows the contribution of words in the statement in making it fake. GT abbreviates ground-truth.

$$z_* = sigmoid(W_* [R, P_*] + b_*), \tag{6}$$

and incorporate the projected embedding P_* of different context into the statement representation by:

$$Q_{*s} = z_* \odot R + (1 - z_*) \odot P_*. \tag{7}$$

Eq. 7 respectively results in three statement representations Q_{cs}, Q_{hs}, Q_{cos}, Q_{hos} specific to the video, dialog and key context. $\odot$ indicates element-wise product. We concatenate them and update the statement representation as:

$$Q = [R; Q_{hs}; Q_{cs}; Q_{hos}; Q_{cos}], \tag{8}$$

The updated statement representation Q is passed through a function f that contains a linear layer with 1-dimensional output and a sigmoid activation to predict the score of the statement to be real.

3.4. Explainable Entailment

The local explanation for judging a textual statement is defined as the contribution of each word, which is in form

Table 1. Entailment Accuracy Comparison. We report the Accuracy (%) of all statements, real statements, fake statements, human-written statements, and adversarially sampled statements. $2/3$ of fake statements are human-written and the remaining $1/3$ are adversarially sampled. Not that "Visual" column denotes the visual features used in the entailment judgment stage.

Method	Visual	Accuracy	Real	Fake	Human-written	Adv-sampled
VIOLIN [25]	C3D	67.23	74.66	57.73	61.99	67.60
Ours	C3D	68.15	79.21	57.08	61.33	79.43
VIOLIN [25]	Resnet	67.60	79.10	56.10	59.15	84.49
Ours	Resnet	68.39	79.52	57.25	60.11	84.94

of a heatmap for a sentence. Our method aims to regularize the training of entailment judgment with its local explanation to promote the model's faithfulness and generalization ability (see the explanation branch in Fig. 4) [13]. We encourage the entailment model to focus more on the words that actually make the statement contradictory to the video, instead of memorizing the dataset-specific artifacts.

In VIOLIN dataset [24], more than half of the fake statements were collected by modifying a small subset of the real statement to be contradicted by the video [25], which makes the difference between the real and fake statements subtle and alleviates the bias. We propose to exploit the subtle difference as a kind of supervision signal for the local explanation. During training, we have access to the real/fake statements that are formed in pairs. For example, a pair of real and fake statements are: "A man in a black jacket gets off his white motorcycle" and "A man in a black jacket gets off the bell towel." respectively. By a simple "*diff*" operation between them, the contradictory items are "the bell towel". The indexes of the different words between the real and fake statements obtained by the "*diff*" operation are defined as the ground-truth of local explanation. We mark it as a binary vector $o_e \in \mathbb{R}^{N_l \times 1}$ that is in length of the statement.

Specifically, we form the entailment judgment (see 3.3) and its explanation as multi-task learning. The explanation branch in Fig. 4 takes the updated statement representation Q as input and generates a heatmap $u_e \in \mathbb{R}^{N_l}$ that indicates the contribution of each word to the model prediction $f(Q)$. The explanation loss $\mathcal{L}_r$ is defined as:

$$\mathcal{L}_r = \sum_{i=1}^{N_l} o_e^i(-\log(u_e)) + (1 - o_e^i)(-\log(1 - u_e)), \tag{9}$$

which aligns the generated heatmap u_e with the local explanation ground-truth o_e. The overall objective function $\mathcal{L}_e$ is defined as:

$$\mathcal{L}_e = \mathcal{L}_{cls} + \lambda \mathcal{L}_r, \tag{10}$$

in which L_{cls} is the binary cross entropy loss for entailment judgment. It balances entailment judgment and its explanation by constraint λ. If a statement is justified as real, each word should be entailed by the premise. Thus, during training, we only regularize the fake statements. During inference, if a statement is predicted as "contradiction", the explanation module will be triggered to generate the heatmap for the statement.

4. Experiments

4.1. Dataset

To our best knowledge, VIOLIN [25] is the only dataset for video entailment task. VIOLIN contains $15,887$ video clips and each video clip is annotated with 3 pairs of real/fake statements, resulting in $95,322$ statements in total. Statements are in random lengths and have 18 words on average. The first two fake statements of each video are human-written by modifying a small portion of the corresponding real statements. Thus, the human-written real/fake statements have very subtle differences, such as one or two words replacement. The third negative statement is adversarially sampled and has a relatively larger difference compared to the real statement. Following the original paper, we split the VIOLIN dataset into 80% for training, 10% for validation, and 10% for testing.

4.2. Implementation Details

We use the pre-trained Bert [11] features of both dialog subtitles and statements provided by [25]. For grounding, a Faster R-CNN framework [28] with VGG-Net [31] as backbone pre-trained on COCO [24] is applied to extract persons and their features across frames. The entity grounding threshold is set to 0.5. Both the visual and textual input are embedded into d-dimension for fusion, and d is set as 256. We sample the frames corresponding to the middle timestamp of each sentence for grounding. Adam with a learning rate of $1e{-}3$ is used for optimization. The constraint weight of grounding module β is set to 1. We set batch size as 8 in training. The entities in the statements of other videos in the batch are sampled as the negative samples for training the entity grounding module.

For the contradiction explanation module, we only use the human-written samples for training. Adam with a learning rate of $1e{-}4$ is used for optimization. Constraint weight of multi-task learning λ is set to 1.

Table 2. Ablation Study of Entity Grounding for Entailment (%).

Method	Accuracy	Real Accuracy	Fake Accuracy
v1	66.72	73.60	59.83
v2	67.60	75.50	59.71
v3	66.53	77.78	48.01
Ours	68.39	79.52	57.25

4.3. Comparison Methods

We compare our method with the only existing method proposed for the video entailment task, to our best knowledge. VIOLIN [25] dataset provides a visual/language fusion model to address entailment judgment. The statement representations are jointly modeled with its video and subtitle by an attention-based fusion module.

Experimental results on VIOLIN dataset are shown in Table 1. Our proposed explainable entailment model along with grounded evidence given by our method outperforms the previous video entailment method. Because we precisely model the alignment between the video frames and dialog based on grounded evidence. We also evaluate the influence of different visual features following VIOLIN [25]. The results demonstrate that our method works for both image-based features "Resnet" and motion-based features "C3D".

4.4. Ablation Study

4.4.1 How does grounding help in entailment?

To exhibit the effectiveness of entity grounding in entailment judgment, we compare our proposed method with the following variants. (1) **v1**: Removing the first contradiction judgment from the entity grounding module. Then, entity grounding is only used to provide temporal guidance. (2) **v2**: Removing the temporal grounding guidance on entailment judgment. We substitute the Eq. 8 by $Q = [R; Q_{hs}; Q_{cs}]$. Each frame contributes to the statement without being highlighted. (3) **v3**: Removing $\mathcal{L}_g$. The grounding module is trained without golden negative statements.

Table 2 summarizes the results of the aforementioned variants. Comparing "Ours" and **v3**, adding golden negative entities brings more than 1% performance improvement, as it improves the grounding quality. Comparing "Ours" and **v2**, adding temporal grounding's guidance is necessary for making an accurate judgment. The contradiction judgment from the entity grounding module also brings performance gain by comparing "Ours" to **v1**.

4.4.2 How does explanation help in entailment?

To explore the contribution of the add-on entailment explanation module, we conduct the ablation study with the following variants: (1) **v4**: Using both the adversarial statements and human-written statements in training the explanation model. (2) **v5**: Removing the explanation regularizer $\mathcal{L}_r$ and only use $\mathcal{L}_{cls}$.

Table 3. Ablation Study of the Add-on Explanation Module for Entailment (%).

Method	Accuracy	Real Accuracy	Fake Accuracy
v4	67.65	78.75	56.54
v5	67.32	80.63	54.02
Ours	68.39	79.52	57.25

Table 4. Quantitative Result for Contradiction Explanation (%).

Method	Explanation Accuracy
v6	72.42
Ours	75.20

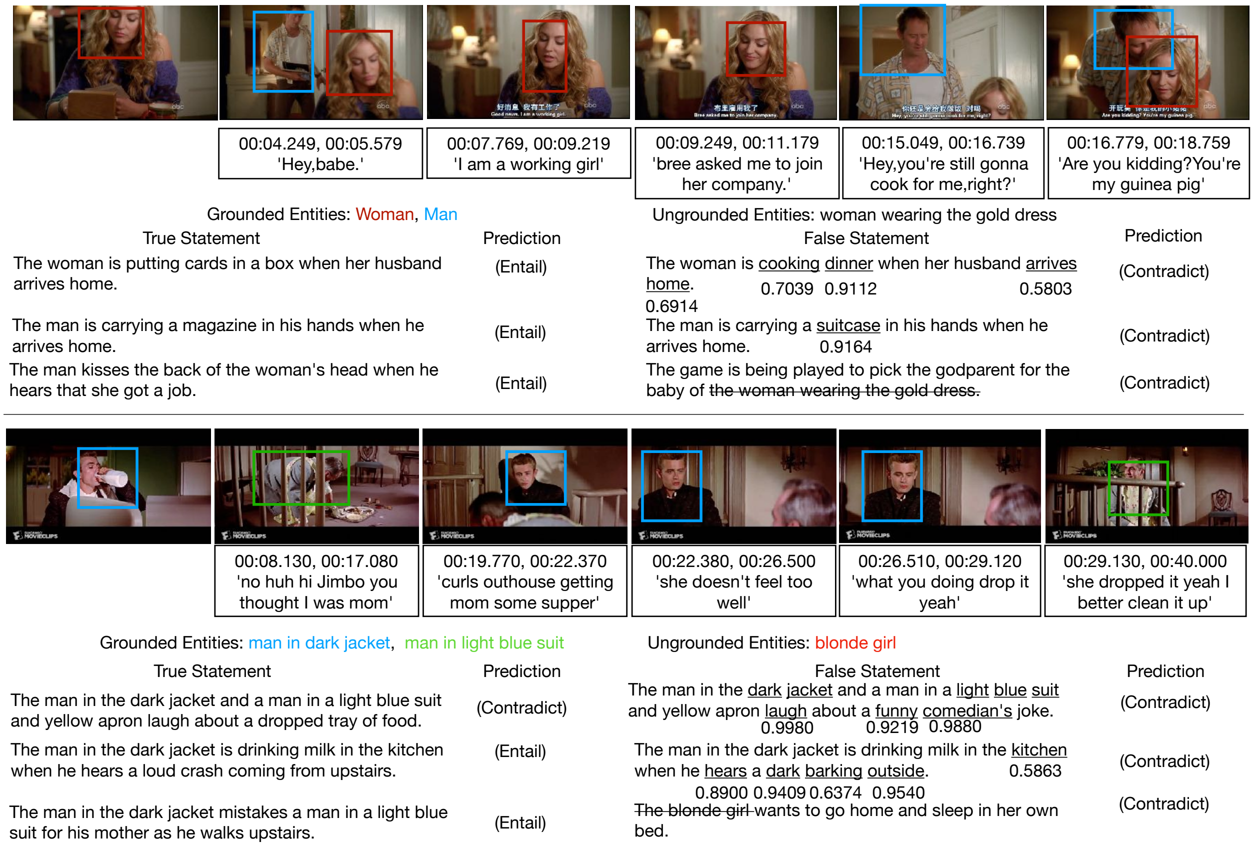
Table 3 illustrates the results of the ablation study on the explanation module. The proposed method outperforms the variant **v5** without explanation module by 0.83%, which shows that the multi-task learning boosts the performance of entailment judgment. By the outperformance to the variant **v4**, it is wise to train the explanation model with only human-written samples instead of the adversarial samples, since the adversarial samples are very different from its paired real statement in sentence structure.

4.5. Contradiction Explanation Result

Since the real and fake statements are formed in pairs, we can get access to the ground-truth of the items (in words or phrases) that make the statement contradictory to the video. For human-written fake statements, the annotators manually change a small portion of words or phrases in the real statement, which makes the paired real and fake statements have similar grammar and very tiny differences. Thus, the ground-truth of the contradictory items can be obtained by a simple "*diff*" operation between a real/fake pair. But in the adversarial sampled pairs, the real and fake statements are mostly different in structure. Thus, we only use the human-written pairs for training the explanation module. But we test all of the statements either human-written or adversarially sampled.

We quantitatively evaluate the local explanation on the fake statements that are human-written. The evaluation metric is defined as the percentage of the number of words that are correctly explained over the overall number of words in the statement. The explanation results are exhibited in Table 4. We achieve 75.2% accuracy in contradiction explanation, which indicates that more than three-quarters of fake words can be found by our explanation model.

We also compare the proposed explanation method with a variant **v6**. **v6** is the variant that explains the entailment of the statement by finding the contradictory constitutes instead of the contradictory words. Constitute parsing method [19] that was used in obtaining entities in Sec.3.3


00:04.249, 00:05.579 'Hey,babe.'
00:07.769, 00:09.219 'I am a working girl'
00:09.249, 00:11.179 'bree asked me to join her company.'
00:15.049, 00:16.739 'Hey,you're still gonna cook for me,right?'
00:16.779, 00:18.759 'Are you kidding?You're my guinea pig'
Grounded Entities: Woman, Man
Ungrounded Entities: woman wearing the gold dress
True Statement
Prediction
False Statement
Prediction

The woman is putting cards in a box when her husband arrives home. (Entail)

The man is carrying a magazine in his hands when he arrives home. (Entail)

The man kisses the back of the woman's head when he hears that she got a job. (Entail)

The woman is cooking dinner when her husband arrives home. 0.7039 0.9112 0.5803 0.6914 (Contradict)

The man is carrying a suitcase in his hands when he arrives home. 0.9164 (Contradict)

The game is being played to pick the godparent for the baby of ~~the woman wearing the gold dress.~~ (Contradict)

00:08.130, 00:17.080 'no huh hi Jimbo you thought I was mom'
00:19.770, 00:22.370 'curls outhouse getting mom some supper'
00:22.380, 00:26.500 'she doesn't feel too well'
00:26.510, 00:29.120 'what you doing drop it yeah'
00:29.130, 00:40.000 'she dropped it yeah I better clean it up'
Grounded Entities: man in dark jacket, man in light blue suit
Ungrounded Entities: blonde girl
True Statement
Prediction
False Statement
Prediction

The man in the dark jacket and a man in a light blue suit and yellow apron laugh about a dropped tray of food. (Contradict)

The man in the dark jacket is drinking milk in the kitchen when he hears a loud crash coming from upstairs. (Entail)

The man in the dark jacket mistakes a man in a light blue suit for his mother as he walks upstairs. (Entail)

The man in the dark jacket and a man in a light blue suit and yellow apron laugh about a funny comedian's joke. 0.9980 0.9219 0.9880 (Contradict)

The man in the dark jacket is drinking milk in the kitchen when he hears a dark barking outside. 0.5863 0.8900 0.9409 0.6374 0.9540 (Contradict)

~~The blonde girl~~ wants to go home and sleep in her own bed. (Contradict)

Figure 5. Visualization of the entailment judgment and its explanation with grounded evidence. Strikethrough indicates that the video does not contain the described entity and thus is judged as "contradiction". The contradictory items are marked by the underline with the predicted scores.

is applied to extract constitutes from statement. The result demonstrates that a plain word-level explanation is better than using the constitute.

4.6. Explainable Entailment Result

Fig. 5 presents several entailment judgment examples using our method. Our model can successfully ground the described entities to the specific regions and the relevant frames, even if the grounding annotation is not provided in training. Our model also has the resilience to the entities in the fake statements that are absent in the video. The two fake statements contain the entities that are missing (*e.g.*, "the blonde girl", "the woman wearing the golden dress"), marked by strikethrough, and are judged as fake in the grounding stage. The predicted fake items are marked by the underline with explanation scores. We find if the statement is correctly judged as fake, the explanation result is more reliable.

5. Conclusion

In this paper, we present a novel approach for video entailment and its local explanation. Entity grounding is highly incorporated into our task from two aspects. First, we train a weakly-supervised entity video grounding module to judge a statement as "contradiction" if the statement consists of an entity absent in the video. Then if the entity is present in the video, we infer the temporal occurrence of that entity to guide the entailment judgment module focusing on the entity-relevant clips. In addition to entailment judgment, our method is also developed to explain which words or phrases make the statement contradictory to the video. We formulate the local explanation as a regularizer to the decision-making of entailment to improve the model's faithfulness. Extensive results on VIOLIN dataset demonstrate the resulting model consistently outperforms the existing methods.

Acknowledgement: This project is supported in part by the NSF SaTC award 1949694, and the Army Research Office under grant number W911NF-21-1-0236. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government.

References

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

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

[3] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In *NIPS*, 2018.

[4] Junwen Chen, Wentao Bao, and Yu Kong. Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In *ACM Multimedia*, pages 3789–3797, 2020.

[5] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*, 2018.

[6] Lei Chen, Mengyao Zhai, Jiawei He, and Greg Mori. Object grounding via iterative context reasoning. In *ICCV Workshops*, 2019.

[7] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *ACL*, 2019.

[8] Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel Bobrow. Entailment, intensionality and text understanding. In *NAACL*, 2003.

[9] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Springer, 2005.

[10] 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.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.

[12] Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve-2.0: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020.

[13] Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. Learning credible deep neural networks with rationale regularization. In *ICDM*, 2019.

[14] De-An Huang*, Shyamal Buch*, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised, reference-aware visual grounding in instructional videos. In *CVPR*, 2018.

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[16] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *CVPR*, 2019.

[17] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *CVPR*, 2020.

[18] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. In *ECCV*, 2018.

[19] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *ACL*, pages 2676–2686, 2018.

[20] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[21] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *ACL*, 2019.

[22] 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.

[23] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019.

[24] 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.

[25] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *CVPR*, 2020.

[26] Bill MacCartney and Christopher D Manning. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, 2009.

[27] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *ICCV*, 2017.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[29] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*. Springer, 2016.

[30] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *CVPR*, 2019.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] 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.

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

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[35] 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.

[36] 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.

[37] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[38] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019.

[39] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *BMVC*, 2018.

[40] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *CVPR*, 2020.