

# Visual Abductive Reasoning

Chen Liang<sup>1,4\*</sup>, Wenguan Wang<sup>2</sup>, Tianfei Zhou<sup>3</sup>, Yi Yang<sup>1</sup>

<sup>1</sup>CCAI, Zhejiang University <sup>2</sup>ReLER, AAIL, University of Technology Sydney <sup>3</sup>ETH Zurich <sup>4</sup>Baidu Research

<https://github.com/leonnop/VAR>

## Abstract

Abductive reasoning seeks the likeliest possible explanation for partial observations. Although abduction is frequently employed in human daily reasoning, it is rarely explored in computer vision literature. In this paper, we propose a new task and dataset, Visual Abductive Reasoning (VAR), for examining abductive reasoning ability of machine intelligence in everyday visual situations. Given an incomplete set of visual events, AI systems are required to not only describe what is observed, but also infer the hypothesis that can best explain the visual premise. Based on our large-scale VAR dataset, we devise a strong baseline model, REASONER (causal-and-cascaded reasoning Transformer). First, to capture the causal structure of the observations, a contextualized directional position embedding strategy is adopted in the encoder, that yields discriminative representations for the premise and hypothesis. Then, multiple decoders are cascaded to generate and progressively refine the premise and hypothesis sentences. The prediction scores of the sentences are used to guide cross-sentence information flow in the cascaded reasoning procedure. Our VAR benchmarking results show that REASONER surpasses many famous video-language models, while still being far behind human performance. This work is expected to foster future efforts in the reasoning-beyond-observation paradigm.

## 1. Introduction

*Abduction . . . consists in studying facts and devising a theory to explain them.*

– Charles Sanders Peirce (1839 – 1914)

Abductive reasoning [51] was coined by Charles Sanders Peirce, the founder of American pragmatism, around 1865. It is inference to the most likely explanation or conclusion for an incomplete set of observations. Abductive reasoning is invariably employed in our everyday life; the generated hypothesis ( $H$ ) is expected to best explain what happens before, after, or during the observation ( $O$ ). Fig. 1 gives some examples. If you see  $O$ : “the road is wet”, abduction

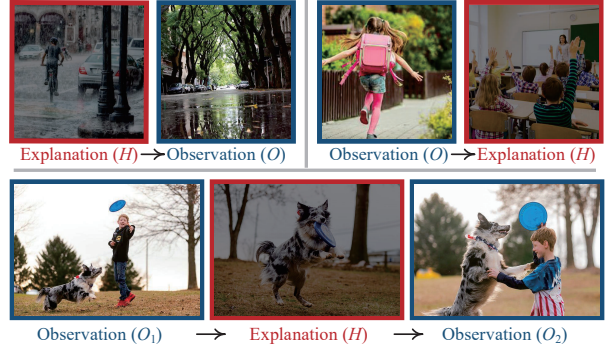


Figure 1. Abductive reasoning is inference to the most likely explanation for an incomplete set of observations.

will lead you to the best explanation  $H$ : “it rained earlier” (i.e.,  $H \rightarrow O$ ). One morning you find  $O$ : “sister leaves home hurriedly”, then you conclude  $H$ : “she will be late for class” (i.e.,  $O \rightarrow H$ ). You see  $O_1$ : “a boy throws a frisbee out and his dog is running after it”. One minute later you find  $O_2$ : “frisbee is in the boy’s hand”. You can imagine  $H$ : “the dog just caught the frisbee back” (i.e.,  $O_1 \rightarrow H \rightarrow O_2$ ). Although abductive reasoning has long been considered as a core ability of everyday human cognition [40, 57, 59], it still remains an untouched domain in computer vision literature.

In this article, we propose *Visual Abductive Reasoning* (VAR), a novel task and dataset for investigating the abductive reasoning ability of AI systems in daily visual situations. In particular, inspired by the recent advance of causal reasoning in NLP community (i.e., abductive text generation [5] and counterfactual story revision [52]), we explore the use of natural language as the expression form to fully capture the complexity of real situations. This also better reflects the nature of human mind, which relies on linguistic thinking [38, 39]. VAR requires the AI systems to describe the incomplete observation (i.e., visual premise) and write down the hypothesis that can best explain the premise. This allows to thoroughly evaluate the entire abduction procedure, as accurate understanding of the premise is the basis of abductive reasoning. Moreover, this can hasten the development of this new field, by comparing and embracing ideas for a relevant, well-established, yet different task – dense video captioning (DVC) [30]. In contrast to DVC that focuses only on *describing the observation*, VAR yields a new

\*Part of this work was done when Chen Liang was an intern at Baidu.

visual-linguistic reasoning paradigm – *inference beyond observation*. Three characteristics make VAR uniquely challenging: **i)** VAR needs imagination to find hypotheses *outside* the observation. **ii)** VAR seeks to discover the plausible causal structure among the observed events. **iii)** VAR is more related to the kind of human reasoning in daily situations where the information at hand is often incomplete [26] and absolutely certain conclusions cannot be reached [5,27].

Our dataset is collected to address the characteristics of the VAR task (*cf.* §3). It contains 9K examples from 3,718 videos. Each example consists of several chronologically-ordered events, most of which are logically related. For each event, abduction oriented description is written by people, and its role of *premise* or *explanation* is also annotated. When presenting each example to the AI system, the explanation event is masked and premise events are visible. The AI system is required to understand the partial, noisy observations (*i.e.*, premise events) and construct the most plausible explanatory hypothesis – accurately describing both the observable premise events and the masked explanation event. The examples are on average 37.8s long, with 4.17 events, and harvested from diversely event-rich sources, *i.e.*, YouTube Lifestyle videos, movies and TV shows.

To lay a solid foundation for future efforts, a new model, named REASONER (causal-and-cascaded reasoning Transformer), is proposed (*cf.* §4). Specifically, REASONER is building upon a Transformer encoder-decoder architecture. In the encoder of REASONER, a *contextualized directional position embedding* strategy is adopted to capture causal dependencies among the premise events. Hence the context of the premise events can be gathered in a causality-aware manner, enabling REASONER to learn discriminative representations for the premise and explanatory hypothesis. Then REASONER cascades a set of decoders for premise/hypothesis sentence production and refinement. For each generated sentence, the associated prediction score is viewed as the confidence and embedded into the next decoder as a signal for inspiring more information to flow from high-confident sentences to the low-confident ones. This leads to a *confidence-guided multi-step reasoning* strategy, boosting the reasoning power of REASONER eventually.

Extensive experimental results are provided in §5. First, to comprehensively probe deep neural models on this challenging task, we establish a group of baselines based on current top-leading DVC models. The benchmarking results on VAR dataset show that REASONER outperforms the best competitor by a large margin, *e.g.*, 33.44 *vs* 28.71 in terms of BERT-S, but is still far behind human performance (42.96). This shows that VAR is especially challenging for current video-language models. Then a set of user studies and ablative experiments are conducted for a thorough evaluation. For completeness, we further test REASONER on the DVC task and confirm again its superiority.

Concurrent to us, [17] studies *image*-based abductive reasoning: AI systems are required to identify, ground, or compare *given* inferences. Overall, we feel vision-based abductive reasoning is an intriguing problem worthy of exploring.

## 2. Related Work

**Dense Video Captioning (DVC).** Different from the classic video description task [29, 46, 67, 68, 76, 80] that aims to describe a short video clip using a single sentence, DVC is to comprehensively describe all the events in an untrimmed video through a multi-sentence paragraph [30]. Typical DVC models [30, 43–45, 60, 74, 81, 82] follow a *two-stage, bottom-up* paradigm: first parse a video into several temporal events and then decode a description from each detected event. As the problem of event detection is ill-defined [10], some alternative solutions either adopt a *single-stage* strategy to simultaneously predict events and descriptions [35, 71], or turn to a *top-down* regime: first generate paragraphs, and then ground each description to a video segment [10, 37]. A few other methods [23, 32, 50] focus purely on generating better paragraph captions from a provided list of events.

Both VAR and DVC are concerned with video-based text generation; a part of our dataset is sourced from ActivityNet Captions [30], a famous DVC dataset. However, DVC is aware of general fact based plain narrative, while VAR addresses cause-effect chain based abductive reasoning. Rather than accurately understanding what is observed, VAR further requires invoking what might have happened or will happen. In our experiments, we involve several recent DVC models as baselines for our VAR task and also report the performance of our REASONER on the DVC task.

**Context-Aware Text Generation.** Our work is also related to some context-aware text generation tasks in the NLP literature. For instance, *text infilling* [83], also known as the *cloze task* [64], is to generate a span of missing tokens in a text chunk, while *sentence/story infilling* [19, 21] aims to generate missing sentences in long-form text. The generated tokens/sentences are expected to smoothly blend into and fit the context syntactically [83], semantically [19, 21], and logically [24]. *Counterfactual story revision* [52] requires generating a new ending, given a story context altered by a counterfactual condition. Our work draws inspiration from *abductive text generation* [5], which investigates abductive reasoning via a natural language inference task: write an appropriate reason that could explain observations described by narrative text. Unlike these language tasks addressing inter-sentential relationship understanding only, our VAR task requires abduction and narrative for a sequence of partially observable visual events. Moreover, our VAR task setting is more general; it is not limited to the strict form of abductive reasoning in [5], *i.e.*, generate a hypothesis ( $H$ ) of what happened between the observed past ( $O_1$ ) and future ( $O_2$ ) contexts:  $O_1 \rightarrow H \rightarrow O_2$ , but involves

$O \rightarrow H$  and  $H \rightarrow O$  abductive reasoning cases.

**Visual Future/State Prediction.** Our work is, to some degree, relevant to future prediction – a popular research area in computer vision. In this area, a huge spectrum of topics/tasks are put forward, including forecasting future frames [42, 70], future features [62, 69], future actions [1, 28, 31, 55, 61], future human motions [14, 22, 41], future human trajectories [2], future goals [13], *etc.* Rather than studying the future generation at the semantic-category or color-pixel level, event-level prediction was recently addressed in [34] and [49]. However, [34] only requires choosing from two candidates for future event prediction, making the task less challenging. [49] targets to describe past, present, and future events for a single image, while our VAR task requires making full use of the information from a set of premise events. There are also some efforts made towards understanding the dynamics/transformations between states [18, 48, 75] or goal-conditioned procedure planning [7], while either relying on a pre-defined, limited action prediction space [7, 18], or using simulated environments [7, 18]. Our VAR task is essentially to discover and describe the causal relations in real visual environments. It is not constrained to a narrow view of predicting either future events or between-state changes, but tries to infer the missing parts in the cause-effect chains, even with some unrelated premise events. All of these together make VAR a unique and challenging visual reasoning task.

**Position Encoding in Transformers.** Due to the permutation invariant nature of the attention operation, [58] learns and encodes position embeddings into Transformer tokens. Subsequent language-Transformers hence explore further variations, like incorporating sinusoid prior with more parameters [9], simplifying position embeddings as learnable scalars [53], disentangling special tokens ([CLS]) [25], *etc.* Some recent vision-Transformers [8, 73] consider directional relative distance between 2D positions, and/or the interactions between visual tokens and position embeddings. REASONER encodes the relations of premise events in a directional and contextualized manner for causal relation modeling, and leverages the prediction scores of descriptions for confidence-guided multi-step reasoning.

### 3. Our VAR Task and Dataset

#### 3.1. VAR Task

Our visual abductive reasoning (VAR) task is designed to test the abductive reasoning ability of machine intelligence in everyday visual environments  $\mathcal{E}$ . Formally, given a video example  $\mathcal{V} \subset \mathcal{E}$  that contains a set of  $N$  events, *i.e.*,  $\mathcal{V} = \{O_1, \dots, O_{n-1}, H, O_n, \dots, O_{N-1}\}$ , which are logically related and chronologically organized, we denote  $\{O_n\}_{n=1}^{N-1}$  as *premise* events – partial observation of  $\mathcal{E}$ , and  $H$  as *explanation* event that can best explain the premise events. Conditioning on the past *and/or* future visual con-

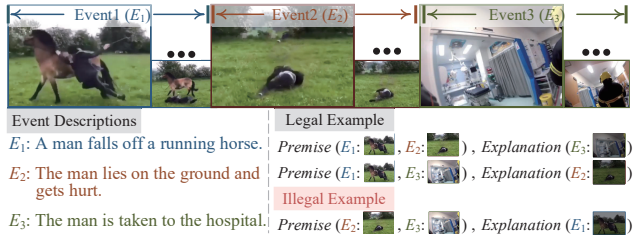


Figure 2. An illustrative example of our VAR dataset (§A).

text  $\{O_n\}_{n=1}^{N-1}$  *only*, the AI system is asked to describe these premise events, and, more importantly, infer and write down the most plausible explanatory hypothesis for the premise. Naturally, such a hypothesis is expected to be consistent with the content of the explanation event  $H$ . The abduction ability can thus be thoroughly examined by assessing the quality of both the premise-based descriptions and explanatory hypothesis sentences – adequate understanding of the premise is a necessary prerequisite for abductive reasoning.

#### 3.2. VAR Dataset

Guided by the above task setup, we build a large-scale dataset for VAR. Fig. 2 depicts an illustrative example.

##### 3.2.1 Dataset Collection

**Data Source.** VAR dataset is collected from three sources:

- 23,457 *Youtube* lifestyle Vlog videos from ActivityNet Captions [30] and VLEP [34] datasets. These videos cover rich social scenarios and human activities.
- 13,799 *TV show and movie* videos from TVC dataset [33] and a famous Youtube channel, Fandango MovieClips<sup>1</sup>. These videos are key scenes in popular TV shows and films containing wide-ranging genres.

YouTube videos include diverse daily events, but last relatively short durations with short intervals (about minutes). While TV shows and movie videos usually have limited scenarios, they contain rich artificial cause-effect chains in their story-lines and last relatively long durations with long intervals (spanning even years). Thus gathering these videos together makes our dataset a good testbed for VAR.

**Data Cleaning.** The collected videos are accompanied by event labels, and videos containing only one single event are first dropped. Then, for each of the rest videos, three human experts are invited to examine if there exists cause-effect relations between the video events. We only preserve qualified ones with more than two votes in the affirmative, finally resulting in 3,718 videos in total for further annotation.

##### 3.2.2 Dataset Annotation

For each video in VAR, the annotation contains three steps:

**Step 1: Event Type Annotation.** For an event  $E$  of video  $\mathcal{V}$ , if  $E$  can well explain some other events in  $\mathcal{V}$ ; or in other words, if we can imagine that  $E$  could happen by only considering the other events  $\mathcal{V}/E$ , event  $E$  will be labeled as

<sup>1</sup><https://youtube.com/user/movieclips>

Split	#Example	#Event	#Video
Train	7,053	12,582	3,000
Val	460	860	205
Test	1,093	2,044	513
Total	8,606	15,486	3,718



Table 1 & Figure 3. Summative statistics of VAR dataset (§3.2.3).

*explanation* and the other events  $\mathcal{V}/E$  will be labeled as *premise*. Fig. 2 gives an example. For the video containing three events:  $E_1$  “a man falls off a running horse”,  $E_2$  “the man lies on the ground and gets hurt”,  $E_3$  “the man is taken to the hospital”, we can derive two legal examples for our VAR dataset:  $\{\textit{premise}(E_1, E_2), \textit{explanation}(E_3)\}$ , and  $\{\textit{premise}(E_1, E_3), \textit{explanation}(E_2)\}$ .

**Step 2: Abductive Reasoning Aware Description Annotation.** Although some videos are collected with event-level descriptions/plot summaries, we re-annotate all the events with abductive reasoning oriented descriptions. Specifically, instead of capturing all the visual details in video captioning, like {“a boy in a black jacket plays frisbee with his white dog in a park”, “the dog catches the blue frisbee and runs back”, “the boy smiles, takes the frisbee and pats the dog”}, our descriptions are only aware of describing the visual content related to abductive reasoning, like {“a boy throws a frisbee out and his dog is running after it”, “the dog catches the frisbee back”, “the boy gets the frisbee”}.

**Step 3: Validation.** Finally, each annotated example is examined by three verifiers: the verifiers are shown with only the *premise* events and language-based explanation (*i.e.*, description on the *explanation* event), and vote for: “Is the explanation sound?”. If an example wins majority approval, it will be accepted; otherwise, it will be relabeled or dropped.

### 3.2.3 Dataset Features and Statistics

To offer deeper insights into our VAR dataset, we next discuss its distinctive properties and detailed statistics.

**Abductive Reasoning Orientated.** VAR is the first dataset that underpins machine intelligence study of abductive reasoning in visual daily scenarios. It is designed to reason beyond visual *premise* for a plausible *explanation*, distinguishing it from existing video-language datasets/tasks.

**Diversity.** To capture diverse cause-effect relations and abduction cases, our VAR dataset covers **i)** various daily events/activities, *e.g.*, work, leisure, household; **ii)** rich scenarios, *e.g.*, lifestyle recording, scripted drama; **iii)** different durations and intervals, ranging from minutes to years.

**Large-Scale.** As shown in Table 1, VAR consists of 8,606 data examples, collected from 3,718 unique videos that span over 153 hours in total. On average, each video in VAR contains 4.17 events that last 37.8 seconds, resulting in a total of 15K corresponding descriptive sentences of 13.5 words.

**Dataset Split.** We separate the VAR dataset into train/val/test sets and arrive at a unique split of 7,053/460/1,093 examples with no overlapping video between val/test and train sets. We provide more detailed statistics

in both Table 1 & Fig. 3 and the supplement.

## 4. Methodology

**Problem Statement.** Given a video  $\mathcal{V}$  with  $N$  temporally ordered events, *i.e.*,  $\mathcal{V} = \{O_1, \dots, O_{n-1}, H, O_n, \dots, O_{N-1}\}$ , the *premise* events, *i.e.*,  $\{O_n\}_{n=1}^{N-1}$ , and *explanation* event, *i.e.*,  $H$ , are logically related. The AI system is only presented with a partially observable version of  $\mathcal{V}$ , *i.e.*,  $\tilde{\mathcal{V}} = \{O_1, \dots, O_{n-1}, \tilde{H}, O_n, \dots, O_{N-1}\}$ , where  $\tilde{H}$  is obtained by setting all the pixel values of  $H$  as zero. The AI system is required to not only describe the premise, but also reason about the most likely explanation for the premise, *i.e.*, generate  $N$  sentences  $\mathcal{S} = \{S_n^O\}_{n=1}^{N-1} \cup S^H$  that describe the content of the  $N$  events in  $\mathcal{V}$ , while conditioning on  $\tilde{\mathcal{V}}$  only:

$$\begin{aligned} P(\mathcal{S}|\tilde{\mathcal{V}}) &= P(S^H|\tilde{\mathcal{V}}) \prod_n P(S_n^O|\tilde{\mathcal{V}}) \\ &= \prod_l P(w_l^H|w_{<l}^H, \tilde{\mathcal{V}}) \prod_n \prod_l P(w_l^{O_n}|w_{<l}^{O_n}, \tilde{\mathcal{V}}); \end{aligned} \quad (1)$$

where  $w_l$  is the  $l$ -th word in a generated sentence  $S \in \mathcal{S}$ . It is worth mentioning that, when  $H = \emptyset$ , our VAR task is degraded into a classic DVC task [30] which focuses only on describing the content of observed events  $\{O_n\}_{n=1}^{N-1}$ .

**Core Idea.** Building upon a Transformer encoder-decoder architecture (Fig. 4), our REASONER is aware of two core challenges posed by the VAR task: **i)** inferring cause-effect relations, and **ii)** reasoning beyond the partial observation. To address **i)**, a *contextualized directional position embedding* strategy is adopted to capture causal relations residing in the input video  $\tilde{\mathcal{V}}$ , leading to a *Causality-aware encoder* (§4.1). To accommodate **ii)**, a *confidence-guided multi-step reasoning* strategy is developed, *i.e.*, utilize the prediction scores of sentences to guide cross-sentence information flow, yielding a *cascaded-reasoning decoder* (§4.2).

### 4.1. Causality-Aware Encoder

For notational simplicity, we redefine the partially observable video  $\tilde{\mathcal{V}} = \{O_1, \dots, O_{n-1}, \tilde{H}, O_n, \dots, O_{N-1}\}$  as  $\tilde{\mathcal{V}} = \{E_n\}_{n=1}^N$ , where  $E_h$  refers to the masked explanation event  $\tilde{H}$ , and  $\{E_n\}_{n \neq h}$  indicates the visible, premise events  $\{O_n\}_{n=1}^{N-1}$ . Let us denote the initial features of the  $N$  events as  $\{\mathbf{E}_n \in \mathbb{R}^d\}_{n=1}^N$ . For each premise event  $E_{n \neq h}$ , corresponding feature  $\mathbf{E}_{n \neq h}$  is obtained by aggregating the visual features of its frames. For the masked explanation event  $E_h$ , we set  $\mathbf{E}_h = \mathbf{0}^d$ . The Causality-aware encoder is to leverage the context from the past and/or future observable events  $\{E_n\}_{n \neq h}$  to reinforce their own representations as well as posit a meaningful representation for the most likely explanatory hypothesis, *i.e.*, the masked explanation event  $E_h$ .

The attention operation is the core of Transformer:

$$\mathbf{A} \sim \mathbf{X} \mathbf{W}^q (\mathbf{X} \mathbf{W}^k)^\top, \mathbf{Y} = \text{softmax}(\mathbf{A}) \mathbf{X} \mathbf{W}^v. \quad (2)$$

where the output  $\mathbf{Y} \in \mathbb{R}^{N \times d}$  is with the same length  $N$  and embedding dimension  $d$  as the input  $\mathbf{X} \in \mathbb{R}^{N \times d}$ , and

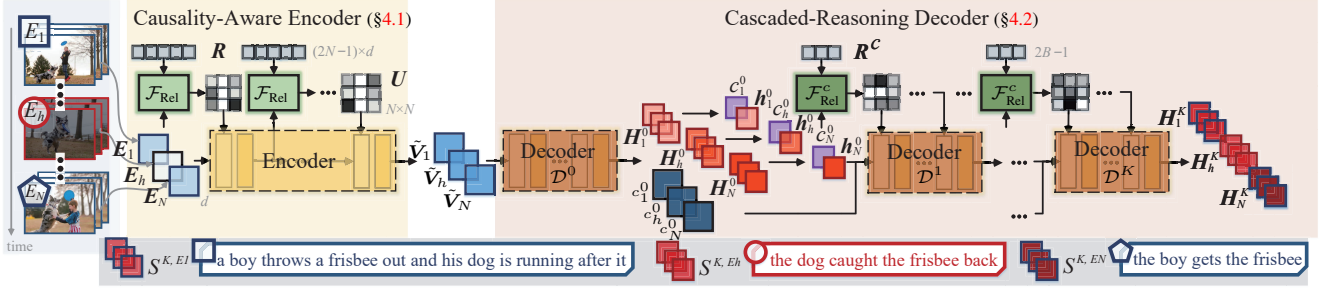


Figure 4. Network architecture of REASONER. See §4 for more details.



Figure 5. Illustration of our contextualized directional position embedding  $U$  (§4.1). Darker color indicates larger attention.

$W^{q,k,v} \in \mathbb{R}^{d \times d}$  project the input into *query*, *key*, and *value* matrices, respectively. As the attention computation is invariant with respect to reordering of the inputs, explicit position encoding is widely adopted, in two typical ways: **i) Absolute position encoding** [65]: each position  $n$  is assigned an embedding, *i.e.*,  $U_n = \mathcal{F}_{\text{Abs}}(n) \in \mathbb{R}^{1 \times d}$ , and the position embeddings are directly added to the input, *i.e.*,  $X \leftarrow X + U$ .  $\mathcal{F}_{\text{Abs}}(\cdot)$  can be a linear projection [12], a sinusoidal function [65], *etc.* **ii) Relative position encoding** [58]: the position relationships are constructed considering the pairwise relationships between positions, *i.e.*,  $U_{nm} = \mathcal{F}_{\text{Rel}}(n, m) \in \mathbb{R}$ .

**Contextualized Directional Position Embedding.** Since the VAR task is essentially aware of the plausible chains of cause-effect, the relative ordering of the input events matters. We continue in the vein of relative position encoding [58, 73] and adopt a *contextualized directional position embedding* strategy, *i.e.*,  $U_{nm} = \mathcal{F}_{\text{Rel}}(n, m, X_n) \in \mathbb{R}$ :

$$\begin{aligned} \mathcal{F}_{\text{Rel}}(n, m, X_n) &= X_n R \ell(n, m), \\ \ell(n, m) &= n - m + N, \end{aligned} \quad (3)$$

where  $R \in \mathbb{R}^{(2N-1) \times d}$  is a learnable matrix, and  $\ell(\cdot, \cdot)$  is a directional indexing function, *i.e.*,  $\ell(n, m) \neq \ell(m, n)$ . The directional projection  $\mathcal{F}_{\text{Rel}}$  is conditioned on the visual context, *i.e.*,  $X_n$ , since the causal dependency between events is typically related to specific content, *e.g.*, when we see people are laughing, we tend to look back only a short time into the past to figure out the reason; when we see a man falls off his horse, we worry about whether he gets hurt and the impact on his future life. Some more visual examples regarding our contextualized directional position embedding strategy can be found in Fig. 5. Then,  $U \in \mathbb{R}^{N \times N}$  is injected by manipulating on the attention matrix  $A \in \mathbb{R}^{N \times N}$ :

$$A_{nm} \sim X_n W^q (X_m W^k)^\top + U_{nm}. \quad (4)$$

We further set  $A_{nh} = 0$  to encourage leveraging the context from the observable events  $\{E_n\}_{n \neq h}$  to infer the masked explanation event  $E_h$ , rather than vice versa. The Causality-aware encoder in REASONER is therefore achieved by

stacking several Transformer encoder blocks [65] with our contextualized directional position embedding strategy. We denote the output event representations as  $\{\tilde{V}_n \in \mathbb{R}^d\}_{n=1}^N$ .

## 4.2. Cascaded-Reasoning Decoder

With the discriminative representations  $\{\tilde{V}_n\}_{n=1}^N$  of the observable premise events  $\{O_n\}_{n=1}^{N-1}$  as well as the explanatory hypothesis  $\tilde{H}$ , the cascaded-reasoning decoder first generates a descriptive sentence for each event/hypothesis individually, and then refines all the sentences in a comprehensive, confidence-guided, and step-by-step manner.

**Initial Description Generation.** For each event representation  $\tilde{V}_n \in \mathbb{R}^d$ , a multi-modal, *masked* Transformer decoder is first adopted for initial description generation:

$$[\tilde{V}_n^0, H_n^0] = \mathcal{D}^0([\tilde{V}_n, H_n]), \quad (5)$$

where  $H_n \in \mathbb{R}^{L_n \times d}$  is a set of  $L_n$  words embeddings. During training, it is computed over the groundtruth description, *i.e.*,  $\hat{S}^{E_n}$ , and masked attention [65] is adopted to prevent the leakage of future words. During inference, it is recurrently generated. Learnable modal-type embeddings [11, 32] are also added into the input yet omitted for brevity. By fusing visual and linguistic representations as the input,  $\mathcal{D}^0$  conducts cross-modal reasoning, and hence generates improved event representation, *i.e.*,  $\tilde{V}_n^0 \in \mathbb{R}^d$ , and updated visual-linguistic state, *i.e.*,  $H_n^0 \in \mathbb{R}^{L_n \times d}$ , for each event  $E_n$ . Then a captioning head is adopted to map  $H_n^0$  into word distribution. The probability of  $l$ -th word is given as:

$$\begin{aligned} P(w_l^{E_n} | w_{<l}^{E_n}, \tilde{V}_n) &= P(w_l^{E_n} | w_{<l}^{E_n}, H_n^0) \\ &= \text{softmax}(H_n^0(l) \Omega^\top), \end{aligned} \quad (6)$$

where  $\Omega \in \mathbb{R}^{|\Omega| \times d}$  is the embedding matrix of the word vocabulary  $\Omega$ , and  $H_n^0(l) \in \mathbb{R}^d$  denotes  $l$ -th vector of  $H_n^0$ . As standard, the description  $S^{0, E_n} = \{w_l^{E_n}\}_{l=1}^{L_n}$  for event  $E_n$  is generated by greedy prediction, and we set the averaged prediction score as the confidence:  $c_n^0 = \frac{1}{L_n} \sum_i P(w_i^{E_n})$ .

**Iterative Description Refinement.** To better respond to the fundamental challenge of VAR task in reasoning beyond observation, we further cascade several Transformer decoder blocks over  $\mathcal{D}^0$  for iterative description refinement. This allows REASONER to make full use of both visual and linguistic context from the past and/or future observable events, and improves the explanatory hypothesis in a step-by-step manner, boosting the reasoning ability eventually.

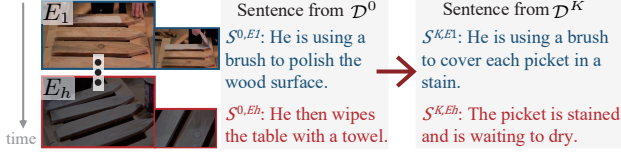


Figure 6. Sentences from the **cascaded-reasoning decoder** (§4.2).

Specifically, our whole refinement procedure can be defined in a recursive, confidence-guided form:

$$\begin{aligned}
 [\tilde{\mathbf{V}}_n^k, \mathbf{H}_n^k] &= \mathcal{D}^k([\tilde{\mathbf{V}}_n^{k-1}, \mathbf{H}_n, \{\mathbf{h}_n^{k-1}\}_{n=1}^N]), k = \{1, 2, \dots, K\} \\
 P(w_l^{En} | w_{<l}^{En}, \tilde{\mathbf{V}}) &= P(w_l^{En} | w_{<l}^{En}, \mathbf{H}_n^k) \\
 &= \text{softmax}(\mathbf{H}_n^k(l) \Omega^\top),
 \end{aligned} \tag{7}$$

where  $\mathcal{D}^k$  refers to  $k$ -th refinement module and all the refinement modules are weight-sharing Transformer decoders;  $\mathbf{h}_n^{k-1} \in \mathbb{R}^d$  indicates a condensed representation of  $\mathbf{H}_n^{k-1} \in \mathbb{R}^{L_n \times d}$ :  $\mathbf{h}_n^{k-1} = \text{maxpool}(\mathbf{H}_n^{k-1})$ . In this way, each  $\mathcal{D}^k$  can leverage inter-sentential relationship in previously generated descriptions  $\{\mathbf{h}_n^{k-1}\}_{n=1}^N$  for refinement and better reason about the explanatory hypothesis. Moreover, we introduce the event confidence, *i.e.*,  $\{c_n^k\}_{n=1}^N$ , as a kind of bias into the refinement procedure: leverage the information from those more confident descriptions to help improve the predictions with relatively lower confidence. Without causing ambiguity, we denote  $\mathbf{X}$  as the input of the decoder  $\mathcal{D}^k$ , *i.e.*,  $\mathbf{X} = [\tilde{\mathbf{V}}_n^{k-1}, \mathbf{H}_n, \{\mathbf{h}_n^{k-1}\}_{n=1}^N]$  and omit the superscript  $k$ . For each input “token”  $\mathbf{X}_i$ , its confidence score  $c_{n_i}$  is the one of its sourced event  $E_{n_i}$ , and we normalize  $\{c_n^k\}_{n=1}^N$  over all the  $N$  events. Analogous to Eq. 4, the attention computation in  $\mathcal{D}^k$  is modified as:

$$\begin{aligned}
 \mathbf{A}_{ij} &\sim \mathbf{X}_i \mathbf{W}^q (\mathbf{X}_j \mathbf{W}^k)^\top + \mathcal{F}_{\text{Rel}}^c(c_{n_i}, c_{n_j}), \\
 \mathcal{F}_{\text{Rel}}^c(c_{n_i}, c_{n_j}) &= \mathbf{r}_{\iota(c_{n_i}, c_{n_j})}^c,
 \end{aligned} \tag{8}$$

where the learnable vector  $\mathbf{r}^c \in \mathbb{R}^{2B-1}$  can be viewed as a bucket to store the relative confidence weight; and the directional indexing function  $\iota(\cdot, \cdot)$  is given as  $\iota(c_{n_i}, c_{n_j}) = \lceil c_{n_i} \cdot B \rceil - \lfloor c_{n_j} \cdot B \rfloor + B$ . With such confidence-guided decoding scheme, descriptions are refined by intelligently gathering context from more reliable sentences, while ignoring noisy cues from less confident ones. By stacking several such decoders  $\{\mathcal{D}^k\}_k$ , outputs will be progressively improved (Fig. 6). Related experiments can be found in §5.2.

### 4.3. Training Objective

Given the groundtruth sentences  $\{\hat{S}^{En}\}_{n=1}^N$  corresponding to the  $N$  events  $\{E_n\}_{n=1}^N$  of video  $\mathcal{V}$ , REASONER is trained by minimizing the negative log-likelihood over the outputs of the cascaded-reasoning decoder  $\{\mathcal{D}^k\}_{k=0}^K$ :

$$\mathcal{L}_{\text{Main}} = - \sum_{k=0}^K \sum_{n=1}^N \sum_{l=1}^{L_n} P(\hat{w}_l^{En} | \hat{w}_{<l}^{En}, \mathbf{H}_n^k), \tag{9}$$

where  $\hat{S}^{En} = \{\hat{w}_l^{En}\}_{l=1}^{L_n}$ . As the teacher forcing scheme [72] is used for training,  $\mathbf{H}_n$  in Eq. 5 and 7 is embedded over one-hot encoded groundtruth words  $\{\hat{w}_l^{En}\}_l$ . We further

adopt a *hypothesis reconstruction* based optimization criterion, to provide the encoder with more explicit supervision signals for explanatory hypothesis reasoning:

$$\mathcal{L}_{\text{Aux}} = \|\mathcal{F}_{\text{Proj}}(\tilde{\mathbf{V}}_h) - \mathcal{F}_{\text{Proj}}(\hat{\mathbf{V}}_h)\|_2, \tag{10}$$

where  $\tilde{\mathbf{V}}_h$  and  $\hat{\mathbf{V}}_h$  are embeddings for the explanatory hypothesis obtained from the masked and original videos, *i.e.*,  $\tilde{\mathcal{V}}$  and  $\mathcal{V}$ , respectively, and  $\mathcal{F}_{\text{Proj}}$  is a projection head, based on a small multi-layer perceptron. This auxiliary training objective forces REASONER to “imagine” an effective representation  $\tilde{\mathbf{V}}_h$  that better aligns with the original content of  $E_h$ .  $\hat{\mathbf{V}}_h$  is from the momentum version of the encoder.

## 4.4. Implementation Details

Details on implementing the algorithm are as follows:

- *Detailed network architecture*: The encoder (§4.1) of REASONER is implemented as two Transformer encoder blocks, and each decoder module (§4.2), *i.e.*,  $\mathcal{D}^k$ , is implemented as two Transformer masked decoder blocks. They have  $d = 768$  hidden size and 12 attention heads. We use a bucket size  $B = 10$  to quantize confidence scores (Eq. 8). We stack a total of  $K = 3$  decoders for cascaded reasoning.
- *Data preprocessing*: For each video event, action/appearance features are pre-extracted using ActivityNet [6] pre-trained ResNet200 [15]/BN-Inception [20], as in [32, 71, 82]. We uniformly sample 50 frames per event and concatenate their features as the corresponding event representation which is denoted in a vector form in §4.1-4.3 for ease of notation. Sentences are padded or truncated into 20 words.
- *Training/Inference*: For the first decoder  $\mathcal{D}^0$ , we adopt scheduled sampling [4] to make the later decoders fully trained. The coefficient between the main and auxiliary training objectives is set as 0.2. During inference, the final descriptive sentences are generated from the last decoder  $\mathcal{D}^K$ , using deterministic decoding, *i.e.*, greedy search. All the experiments are conducted on 2 NVIDIA GeForce RTX 2080 Ti GPUs with a 11GB memory per-card.

## 5. Experiments

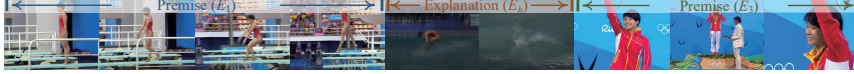
We first provide benchmarking results on our VAR dataset (§5.1). Then, to verify the efficacy of our core model designs, we conduct a set of diagnostic studies (§5.2). Finally, for comprehensive evaluation, we test our REASONER on the classic, dense video captioning (DVC) task [30] (§5.3).

### 5.1. Performance on VAR Task

**Competitor.** We benchmark five top-leading DVC models on VAR to reveal the abductive reasoning ability in existing techniques. They include three Transformer-based [9, 32, 82] and two RNN-based [71, 74] models, which are trained on train set of our VAR dataset with pre-provided event segments using their original training protocols.

Method	Encoder Decoder		Premise Event					Explanation Event				
			BLEU@4	METEOR	ROUGE	CIDEr	BERT-S	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
Human	-	-	13.26	21.27	39.47	155.72	45.33	11.35	19.36	36.92	147.79	40.59
VTrans [82] CVPR18	Trans.	Trans.	4.20	9.94	21.13	31.09	29.05	0.71	6.92	19.12	7.11	22.13
MFT [74] ECCV18	RNN	RNN	3.93	9.69	20.81	30.96	27.41	1.81	7.16	19.16	17.67	25.90
Trans-XL [9] ACL19	Trans.	Trans.	3.98	9.53	21.02	30.87	29.12	2.96	7.51	20.94	24.54	27.23
MART [32] ACL20	Trans.	Trans.	3.74	9.48	21.17	29.22	29.03	2.86	7.47	20.87	24.05	27.77
PDVC [71] ICCV21	Trans.	RNN	4.28	9.95	21.19	33.59	29.37	3.00	8.54	20.71	25.14	27.80
<b>REASONER</b>	Trans.	Trans.	<b>5.03</b> $\uparrow 0.72$	<b>10.75</b> $\uparrow 0.80$	<b>24.81</b> $\uparrow 3.62$	<b>38.27</b> $\uparrow 4.68$	<b>34.88</b> $\uparrow 5.51$	<b>3.44</b> $\uparrow 0.44$	<b>9.05</b> $\uparrow 0.49$	<b>22.89</b> $\uparrow 1.95$	<b>30.75</b> $\uparrow 5.61$	<b>30.64</b> $\uparrow 2.84$

Table 2. **Quantitative results** on the test set of our VAR dataset. ‘Trans.’ indicates Transformer-based architecture. See §5.1 for details.



**Vtransformer** [82]: [A young girl walks around to the pool and jumps.] [A man is seen speaking to the camera and begins playing the instrument.] [A man is talking to the camera.]

**MART** [32]: [A female gymnast walks up to a beam, ready to perform.] [She does several flips and tricks.] [She dismounts, throwing her arms into the air.]

**PDVC** [71]: [Woman is standing at the end of a diving board and then she falls off of the diving board.] [Jumps off of the diving board.] [Girl is now standing on the trampoline and is shown smiling and talking.]

**REASONER(Ours)**: [A woman stands on a diving board and stretches out on a pool.] [The diver jumps off the diving board and does a flip into the water making a small splash.] [She smiles and waves hands on the podium.]

**Groundtruth**: [A woman walks on a springboard and jumps in the air.] [She then flips and dives in the water with a small splash.] [She stands on the podium waving her hands to the audience.]

Figure 7. **Qualitative comparison** (§5.1) of REASONER and [32, 71, 82] on VAR test.

**Evaluation Metric.** Five well-known automated metrics, *i.e.*, BLEU@4 [47], CIDEr [66], METEOR [3], ROUGE-L [36], and BERTScore [79], are used for evaluation.

**Quantitative Result.** Table 2 summarizes the benchmarking results on the test set of our VAR dataset. For detailed analysis, we report the performance over the observable premise events and invisible explanation events separately. Moreover, to probe the upper bound of model performance, we evaluate human performance by asking ten volunteers to perform VAR. Specifically, we randomly sample 500 examples from unique videos in VAR test. The volunteers are only provided with partially observable videos and requested to write down the corresponding descriptions and hypotheses. The human-written descriptions and hypotheses are evaluated by the automatic metrics, and evaluation scores are shown in the first row of Table 2. Several essential conclusions can be drawn from Table 2: **i)** Humans are good at VAR; although human-written hypotheses for explanation scored lower than the descriptions for the visual premise, they are still very plausible in absolute terms. **ii)** All traditional DVC models [9, 32, 71, 74, 82] struggle with VAR that humans excel at. Their generated hypotheses are usually untrusted, and far worse than their created premise narratives. This suggests that existing video-based language generation models are not good at reasoning beyond observation. **iii)** Our REASONER outperforms other AI models [9, 32, 71, 74, 82], in both explanatory hypothesis reasoning and premise description, demonstrating the effectiveness of our whole model design. Compared to other AI models, REASONER also yields a relatively smaller performance drop, from premise description to hypothesis reasoning. This suggests that REASONER can make better use of the context of observed events to infer the explanatory hy-

pothesis. **iv)** Although our REASONER shows more promising results, there still remains a significant gap from human performance, that is waiting for more sophisticated abductive reasoning models to conquer.

**User Study.** For comprehensive performance assessment, we further carry out a subjective evaluation, based on pairwise model comparison. Specifically, we randomly sample 500 examples from unique videos in VAR test. Three volunteers are presented the outputs of a pair of systems (*i.e.*, REASONER vs PDVC [71] or human) on the sampled examples, and requested to do a comparison about which one is better, or ‘‘equally good’’ or ‘‘equally bad’’. The human preference results are collected in Table 3, and again the statistics for premise events and explanation events are presented separately. The human subjective judgments are generally accordant with the trends reflected by Table 2. Specifically, the human pairwise comparison results confirm the superiority of REASONER over PDVC, the second-best model in Table 2: REASONER receives 34.2 and 29.9 percent preference votes on the premise description and explanatory hypothesis, respectively. However, human-written hypotheses and descriptions are much more favorable than our results, showing again VAR is a very challenging task.

**Qualitative Analysis.** A test video example in VAR dataset is shown in Fig. 13. It contains the explanatory hypotheses and premise descriptions from our REASONER and other competitors [32, 71, 82] as well as groundtruth sentences. We can find that our REASONER is able to discover and correctly describe the cause-effect chain, and hence generate a plausible hypothesis, *i.e.*, *making a small splash*, that well explains the observed events, *i.e.*, *standing on the podium*. In contrast, other competitors typically produce unsatisfactory results, especially for the explanatory hypothesis.

Premise Event		
Prefer A	Neutral	Prefer B
<b>REASONER 34.2</b>	41.4	15.9 PDVC [71]
<b>REASONER 16.0</b>	35.3	<b>39.5</b> Human
Explanation Event		
Prefer A	Neutral	Prefer B
<b>REASONER 29.9</b>	13.7	10.4 PDVC [71]
<b>REASONER 8.9</b>	22.1	<b>64.8</b> Human

Table 3. **User study** of pairwise model preference (%). ‘‘Neutral’’ means A and B models are ‘‘equally good’’. Percentage of ‘‘equally bad’’ are omitted. See §5.1 for details.

#	Causality-Aware Encoder (§4.1)	Cascaded-Reasoning Decoder (§4.2)	BLEU@4	CIDEr	BERT-S
1			3.39	30.04	26.35
2	✓		3.91	32.32	29.85
3		✓	4.05	33.71	29.94
4	✓	✓	4.66	36.13	33.44

(a) Key components

$U_n/U_{nm}$ (§4.1)	Formulation	BLEU@4	CIDEr	BERT-S
Absolute	$U_n = \mathcal{F}_{\text{Abs}}(n)$	4.20	33.27	29.95
Directional	$U_{nm} = \mathcal{F}_{\text{Rel}}(n, m)$	4.35	34.25	31.79
Contextualized Directional	$U_{nm} = \mathcal{F}_{\text{Rel}}(n, m, \mathbf{X}_n)$	4.66	36.13	33.44

(b) Position embedding strategy

## 5.2. Diagnostic Experiment

A set of ablative studies is conducted on VAR test for indepth analyzing each component in our REASONER, using BLEU@4, CIDEr and BERT-S metrics, averaged over all the events.

**Key Component Analysis.** We first study the efficacy of core model components. The first row in Table 8a gives the performance of a basic Transformer model, which simply uses absolute position embedding in the encoder and only adopts one single decoder, *i.e.*,  $\mathcal{D}^0$ . The results in the first two rows reveal that contextualized directional position embedding (§4.1) consistently improves the performance over the three metrics. Moreover, from the first and third rows we can observe that confidence-guided multi-step reasoning (§4.2) indeed boosts the performance. By further considering the scores in the last row, we can safely conclude that combining the two model designs together leads to the best results.

**Contextualized Directional Position Embedding.** Next, to thoroughly study the impact of our contextualized directional position embedding strategy (§4.1), we report the performance of two alternatives in Table 8b. Specifically, “absolute” refers to the widely used, learnable absolute position embedding, while “directional” indicates learning relative position embedding without considering any input context. As seen, our contextualized directional position embedding is significantly better than the two alternatives.

**Cascaded Reasoning.** Table 8c reports the performance with different steps of our cascaded reasoning (§4.2), *i.e.*,  $K = \{0, 1, \dots, 5\}$ . When  $K = 0$ , only one decoder  $\mathcal{D}^0$  is adopted and the CIDEr score is just 32.72. However, after adding an extra refinement decoder, the score is greatly improved to 36.13. The increasing trend is gradually saturated until  $K > 3$ . We therefore use  $K = 3$  as our default setting for balancing performance and inference efficiency.

**Confidence Embedding.** We inject sentence scores into the cascaded reasoning for guiding information flow (Eq. 8). As shown in Table 8d, removing confidence embedding hinders the performance, *e.g.*, 36.13→35.22 in terms of CIDEr.

**Training Objective.** Finally we examine our training objective design (§4.3). Table 8e demonstrates a beneficial impact of the hypothesis reconstruction loss  $\mathcal{L}_{\text{Aux}}$  (Eq. 10).

## 5.3. Performance on DVC Task

For completeness, we report performance on DVC task.

**Dataset.** As a gold-standard dataset for DVC, ActivityNet Captions [30] contains a total of 20k untrimmed videos

$\mathcal{D}^k$ (§4.2)	B@4	CIDEr	BERT-S
$K = 0$	3.91	32.72	29.50
$K = 1$	4.34	34.89	31.60
$K = 2$	4.61	35.53	32.57
$K = 3$	4.66	36.13	33.44
$K = 4$	4.66	36.05	33.51
$K = 5$	4.60	35.90	33.32

(c) Cascaded reasoning

$\mathcal{F}_{\text{Rel}}^c$ (Eq. 8)	BLEU@4	CIDEr	BERT-S
✓	4.45	35.22	33.17
	4.66	36.13	33.44

(d) Confidence embedding

Loss (§4.3)	BLEU@4	CIDEr	BERT-S
$\mathcal{L}_{\text{Main}}$	4.40	35.51	32.83
$\mathcal{L}_{\text{Main}} + \mathcal{L}_{\text{Aux}}$	4.66	36.13	33.44

(e) Training objective

Table 4. A set of **ablation studies** (§5.2) on the test set of our VAR dataset.

Method	BLEU@4	METEOR	CIDEr
HSE [78] <small>ECCV18</small>	9.84	13.78	18.78
Trans-XL [9] <small>ACL19</small>	10.39	15.09	21.67
VTrans [82] <small>CVPR18</small>	9.75	15.64	22.16
MART [32] <small>ACL20</small>	10.33	15.68	23.42
PDVC [71] <small>ICCV21</small>	11.80	15.93	27.27
<b>REASONER</b>	<b>12.45</b> $\uparrow 0.65$	<b>16.43</b> $\uparrow 0.50$	<b>30.08</b> $\uparrow 2.81$

Table 5. **Quantitative results** (§5.3) on the ae-val set of ActivityNet Captions [30]. The scores are mainly borrowed from [71].

(10,009/4,917/5,044 for train/val/test). Each video lasts 120s and is annotated with 3.65 temporally-localized sentences on average. Following [32, 60, 82], val set is further split into two subsets: ae-val with 2,460 videos and ae-test with 2,457 videos without overlapping.

**Evaluation Metric.** As in [32, 60, 82], BLEU@4 [47], METEOR [3], and CIDEr [66] metrics are used for evaluation.

**Quantitative Result.** REASONER is trained on the train set and evaluated on ae-val set in paragraph-level. Since we focus only on descriptive quality, the sentences are generated from a provided list of events, like in [23, 32, 50]. As shown in Table 5, REASONER outperforms state-of-the-art DVC models over all the metrics, *e.g.*, +2.81 performance gain in CIDEr. This proves the strong reasoning ability of REASONER and emphasizes the value of our VAR task in promoting innovations of powerful video-language models.

## 6. Conclusion

We introduce VAR (Visual Abductive Reasoning) – a novel task that investigates the abductive reasoning ability of machine intelligence in the visual world. We establish REASONER, a new Transformer based visual-language model, which captures the context from visual premise in a causality-aware manner, and generates premise descriptions and hypothesis sentences in a confidence-guided, step-by-step fashion. REASONER shows promising results on both VAR and dense video captioning tasks. We also observe a remaining large headroom for AI systems in VAR, which is expected to encourage exciting avenues in the future.





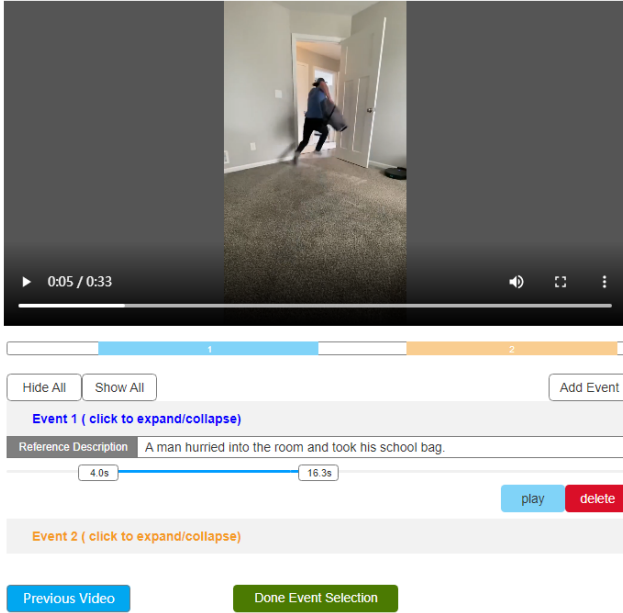


Figure 9. Interface for event selection. See §A.2 for more details.

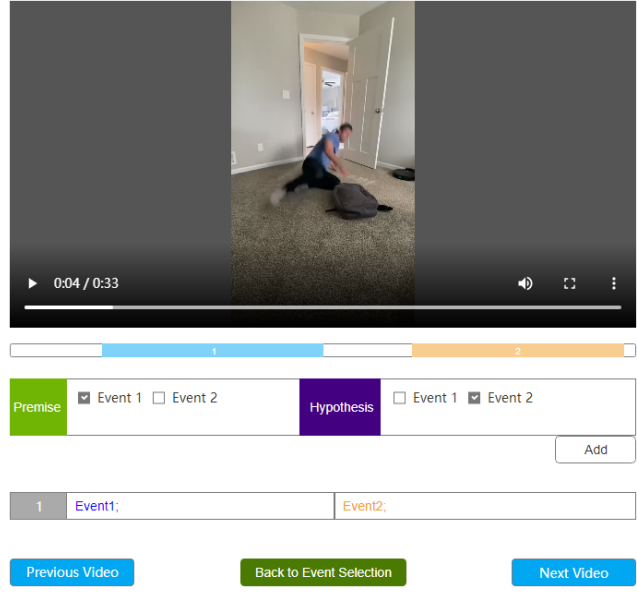


Figure 11. Interface for event type annotation. Details are in §A.2.

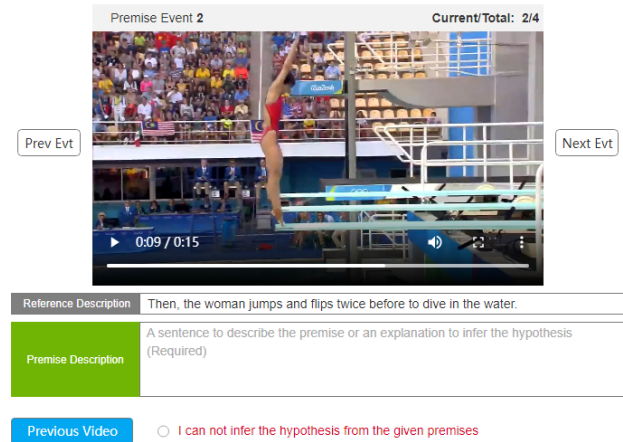


Figure 10. Interface for describing premises. Details are in §A.2.

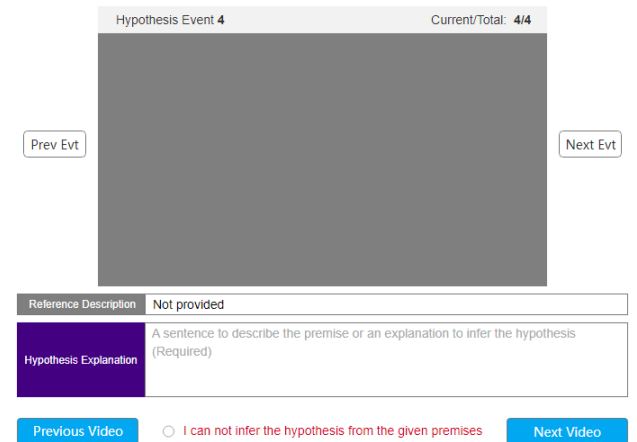


Figure 12. Interface for explaining hypotheses. Details are in §A.2.

**MFT** MFT [74] is an LSTM-based method that consists of a selection LSTM for relevant event filtering and a captioning LSTM for coherent sentence generation. We adapt it to VAR task by recurrently passing the given events into MFT and the selection LSTM is transformed into a visually coherent maintaining module. The unidirectionally causal structure can be captured that enables the inference on potential effect along the temporal order.

**PDVC** Similarly, PDVC [71] employs an LSTM-based captioning decoder, while it is conditioned on a deformable soft attention aggregated visual event. And the visual events are embedded with a Transformer-based encoder that could also capture bidirectional causal dependencies within it.

**VTrans** VTrans [82] is a fully attentional model that originates from the vanilla Transformer proposed in [65]. We follow the implementation in [32], which serves as a baseline that only considers a single event and independently generates a single sentence describing the given event. Thus causal structure can not be formulated in this method.

**Trans-XL** Transformer-XL (Trans-XL) [9] is originally proposed for modeling unlimited longer-term dependencies with a segment-level recurrent strategy. It can capture the intrinsic unidirectional causal structure within recurrent steps. Following the implementation in [32], gradients can flow through recurrent steps instead of being stopped. This enables stronger long-term modeling.

Method	Setting	Premise Event					Explanation Event				
		BLEU@4	METEOR	ROUGE	CIDEr	BERT-S	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
REASONER	no hidden event	5.26	11.32	24.94	39.52	35.09	-	-	-	-	-
	with premise text only	-	-	-	-	-	1.72	8.37	18.10	15.80	25.28
	w/o external knowledge (reported in the main paper)	<b>5.03</b>	<b>10.75</b>	<b>24.81</b>	<b>38.27</b>	<b>34.88</b>	<b>3.44</b>	<b>9.05</b>	<b>22.89</b>	<b>30.75</b>	<b>30.64</b>

Table 6. **Additional quantitative results** on the `test` set of our VAR dataset. See §C for details.

**MART** MART [32] is also built on a fully Transformer-based encoder-decoder architecture, that maintains a summarized memory module to model dependencies among events. Similar to Trans-XL, the unidirectional causal structure is preserved in the memory. Whereas, experimental results show that it suffers more on our VAR potentially due to the content drift brought by masked visual hypotheses.

MFT and PDVC are benchmarked following the original training protocols. We adapt VTrans, Trans-XL and MART to the VAR task with the implementation provided by [32]. All of these baselines are trained with given events from our VAR `train` and evaluated on VAR `test` under the same setting of REASONER as reported in our main paper.

## C. Additional Experimental Results

To shed light on the essence of both the Visual Abductive Reasoning task and our VAR dataset, we study two edge cases of the main setting: **i)** First, all events are made available to the model, so that no abductive reasoning is needed. And the VAR task is degraded to a basic Dense Video Captioning task. **ii)** Second, only ground-truth linguistic descriptions of premise events are supplied to the models. Therefore, models are expected to conduct abductive reasoning with and only with linguistic cues.

In Table 6, we summarize the quantitative results of these two settings. As seen, when there is no hidden event, *i.e.*, the incomplete causal structure is directly provided, REASONER achieves even better performance. It reveals that fulfilling explanation events through abductive reasoning is indeed challenging and the causal structure understanding is also helpful to basic visual recognition. And for the next setting, when only premise texts are given, comparing to fully utilize both visual and linguistic cues, REASONER can not well-infer the hypothesis within linguistic modality only, which proves that the visual-based abductive reasoning is indispensable in the VAR task.

In our main paper, for the benchmarking results, we report the average scores of ten trained models with different random seeds. To prove the statistical significance of our results, here we further provide the corresponding standard deviations of REASONER and two representative methods [32, 71] in Table 7.

Method	Premise Event			Explanation Event		
	BLEU@4	CIDEr	BERT-S	BLEU@4	CIDEr	BERT-S
[32]	3.74±0.07	29.22±0.39	29.53±0.12	2.86±0.07	24.05±0.27	27.77±0.16
[71]	4.28±0.04	33.59±0.30	29.37±0.18	3.00±0.05	25.14±0.21	27.80±0.17
REASONER	5.03±0.02	38.27±0.15	34.88±0.10	3.44±0.01	30.75±0.24	30.64±0.08

Table 7. Average scores and their standard deviations of REASONER and two representative methods [32, 71] (§C).

## D. Additional Qualitative Visualization

In Fig. 13-14, we show more qualitative examples from VAR `test` following Fig. 7 in the main paper. Generated sentences from competitors [32, 71, 82] along with our REASONER are presented in Fig. 13. In contrast to the competitors, both adequate descriptions on premises and plausible inferences for hypotheses are observed for our proposed method, REASONER, which demonstrates a superior abductive reasoning ability for capturing causal structures among visual events. Some failure cases and gold human-written explanations are shown in Fig. 14. Even though REASONER shows impressive performance on inferring with abduction, VAR task is still a mostly unsolved technical problem. And there remains a large headroom for future works to conquer.

## E. Limitation and Reproducibility

### E.1. Dataset Limitation

During annotation process, we observe a bias against women and minorities due to the highly biased nature of movie and web sourced video data [54, 56, 63]. VAR, derived from these data, inevitably runs into the same problem [16, 77]. We thus suggest that models trained on VAR dataset should be cautiously examined before being deployed onto real-world applications. And we will devote further efforts to mitigating the issue in our later works.

### E.2. Details of BERTScore Evaluation

BERTScore [79] leverages the pre-trained contextual embeddings from BERT-based models for similarity measurement. Thus the evaluated scores vary a lot with different model settings. In this paper, all reported BERTScores are evaluated under a hash code version: `roberta-large_L17_no-idx_version=0.3.0(hug_trans=2.3.0)-rescaled`. **We encourage later works to follow the same setting for a fair comparison.** A static version of BERTScore is released at: <https://github.com/leonnop/VAR>.

## F. Legal and Ethical Considerations

### F.1. Asset License

Videos in VAR dataset are collected from four main assets: (1) ActivityNet Captions [30]<sup>2</sup>, 2017 version, under CC-BY 4.0 license<sup>3</sup>; (2) VLEP [34]<sup>4</sup>, 2020 version, under CC-BY 4.0 license<sup>3</sup>; (3) TVC [33]<sup>5</sup>, 2020 version, under CC-BY 4.0 license<sup>3</sup>; (4) MovieClips<sup>6</sup>, copyright © 2021 Fandango. The site and services are available for non-commercial use. Detailed terms of use are available online<sup>7</sup>. VAR dataset will be released under CC-BY 4.0 license<sup>3</sup>, respecting the licences of all its videos.

### F.2. Concerns on Personal Data Collection

VAR is annotated by human experts and we conduct user studies to evaluate the human-subjective generation quality. All human experts are noticed that the annotation and evaluation will be used for academic research and individual consents are reached with signed agreements. The annotation will not leak any personal information about the experts.

### F.3. Potential Societal Impact

Endowing an AI system with human intelligence has long been dreamed by AI researchers, which could fundamentally change the experience of human-machine interaction. VAR takes an important step towards more human-like AI systems that are endowed with abductive reasoning ability, while it might provoke concerns about disinformation, *e.g.*, fabricating deceptive facts. We encourage more technical researching efforts devoted to fake content detection, and at the same time, we will organize a gated release of our dataset and model to prevent potentially malicious abuses.

---

<sup>2</sup><https://cs.stanford.edu/people/ranjaykrishna/densevid/>

<sup>3</sup><https://creativecommons.org/licenses/by/4.0/>

<sup>4</sup><https://github.com/jayleicn/VideoLanguageFuturePred>

<sup>5</sup><https://github.com/jayleicn/TVCaption>

<sup>6</sup><https://www.movieclips.com/>

<sup>7</sup><https://www.fandango.com/policies/terms-of-use>



**Vtransformer** [82]: [The person uses a flat to clean the part of the wood.] [The man then rubs down the wood with a cloth.] [A person is putting objects on a table and leads into a person painting the wood.] [A man is holding a razor and begins playing the instrument.]

**MART** [32]: [A person puts a piece of wood over the wood floor.] [The person uses a brush to rub the surface of the wood.] [Then, the person paints a wooden fence with white paint.] [After, the person cleans the borders of the borders with a cloth.]

**PDVC** [71]: [Person is using a sander to paint a wooden table.] [He is using a spray bottle to clean the board.] [He then takes a rag and runs the ski over the surface.] [Man then takes a paper towel and wipes off the table.]

**REASONER(Ours)**: [A person puts a wood on a board.] [The person uses a brush to clean the wood floor.] [He is using a brush to cover each picket in a stain.] [He cleans the wood with a brush.]

**Groundtruth**: [A man uses a power sander to sand fence pickets.] [He rubs a bare hand over the picket to make sure it is smooth.] [He checks each picket and then covers them in a stain.] [The man waits the stain to dry and shows what each picket looks like stained.]



**Vtransformer** [82]: [A group of people are running around a bull and leads into several clips of people running around a.] [A man is seen speaking to the camera while holding a razor and begins playing the instrument.] [The bull is running in the ring and the bull is running and the bull is running in the ring.]

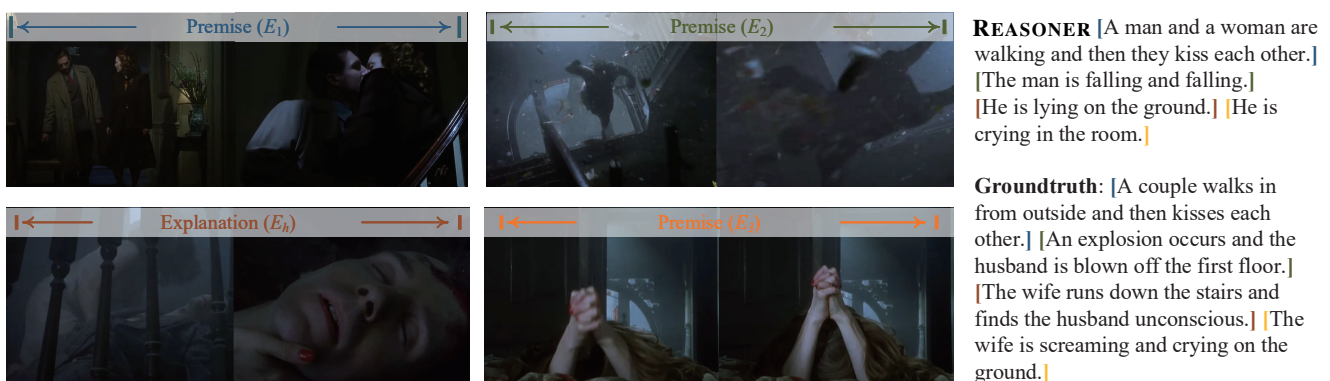
**MART** [32]: [A bull is running around a ring, trying to get the bull from the ground.] [The bull doesn't get hurt, but he isn't able to get it from the ground.] [The bull charges at the center of the ring, and gets off the bull fighting.]

**PDVC** [71]: [Large group of people are running around a bull and leads into a bull running into a pit.] [Bull continues running around the bull and the bull is running into the pit.] [People continue to fight with one another while the crowd cheers on the sides.]

**REASONER (Ours)**: [A large group of people are running around a bull with a bull running around the field.] [Several people are seen running around the bull and lead into them chasing a person.] [More people are seen running around the bull and end with a man taking away.]

**Groundtruth**: [A large group of people are running down a street with bulls chasing them one behind.] [Several people taunt the bull with sticks while someone is hurt by the bull.] [People hold up blankets and run away from one another while the ambulance comes to take injured people away.]

Figure 13. Additional qualitative comparisons of REASONER and [32, 71, 82] on VAR test. See §D for more details.



**REASONER** [A man and a woman are walking and then they kiss each other.] [The man is falling and falling.] [He is lying on the ground.] [He is crying in the room.]

**Groundtruth**: [A couple walks in from outside and then kisses each other.] [An explosion occurs and the husband is blown off the first floor.] [The wife runs down the stairs and finds the husband unconscious.] [The wife is screaming and crying on the ground.]

Figure 14. Additional failure cases of REASONER on VAR test. See §D for more details.

## References

- [1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *CVPR*, 2018. 3
- [2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 3
- [3] Satyanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, 2005. 7, 8
- [4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*, 2015. 6
- [5] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *ICLR*, 2020. 1, 2
- [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. 6
- [7] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *ECCV*, 2020. 3
- [8] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. In *NeurIPS*, 2021. 3
- [9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019. 3, 6, 7, 8, 9, 10
- [10] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *CVPR*, 2021. 2
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018. 5
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5
- [13] Dave Epstein and Carl Vondrick. Learning goals from failure. In *CVPR*, 2021. 3
- [14] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, 2015. 3
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [16] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 11
- [17] Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. *arXiv preprint arXiv:2202.04800*, 2022. 2
- [18] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *CVPR*, 2021. 3
- [19] Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. Inset: Sentence infilling with inter-sentential transformer. In *ACL*, 2020. 2
- [20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6
- [21] Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, 2019. 2
- [22] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, 2016. 3
- [23] Lei Ji, Xianglin Guo, Haoyang Huang, and Xilin Chen. Hierarchical context-aware network for dense video event captioning. In *ACL*, 2021. 2, 8
- [24] Dongyeop Kang and Eduard Hovy. Linguistic versus latent relations for modeling coherent flow in paragraphs. In *EMNLP-IJCNLP*, 2019. 2
- [25] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *ICLR*, 2020. 3
- [26] Frank C Keil. Folkscience: Coarse interpretations of a complex reality. *Trends in cognitive sciences*, 2003. 2
- [27] Frank C Keil. Explanation and understanding. *Annual review of psychology*, 2006. 2
- [28] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *ECCV*, 2012. 3
- [29] Atsuhiko Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 2002. 2
- [30] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 2, 3, 4, 6, 8, 9, 12
- [31] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *ECCV*, 2014. 3
- [32] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, 2020. 2, 5, 6, 7, 8, 9, 10, 11, 13
- [33] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 3, 9, 12
- [34] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *EMNLP*, 2020. 3, 9, 12
- [35] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018. 2

- [36] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *ACL*, 2002. 7
- [37] Hui Liu and Xiaojun Wan. Video paragraph captioning as a text summarization task. In *ACL*, 2021. 2
- [38] RK Logan and PO Izabella. A topology of mind – spiral thought patterns, the hyperlinking of text, ideas and more, 2018. 1
- [39] Robert K Logan and Marlie Tandoc. Thinking in patterns and the pattern of human thought as contrasted with ai data processing. *Information*, 9(4):83, 2018. 1
- [40] Tania Lombrozo. Explanation and abductive inference. *The Oxford Handbook of Thinking and Reasoning*. 1
- [41] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 3
- [42] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016. 3
- [43] Luke Melas-Kyriazi, Alexander M Rush, and George Han. Training for diversity in image paragraph captioning. In *EMNLP*, 2018. 2
- [44] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, 2019. 2
- [45] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020. 2
- [46] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, 2016. 2
- [47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 7, 8
- [48] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, 2019. 3
- [49] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*, 2020. 3
- [50] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial inference for multi-sentence video description. In *CVPR*, 2019. 2, 8
- [51] Charles Sanders Peirce. *Collected papers of charles sanders peirce*. Harvard University Press, 1931. 1
- [52] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *EMNLP-IJCNLP*, 2019. 1, 2
- [53] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 3
- [54] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *ACL workshops*, 2017. 11
- [55] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011. 3
- [56] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *EMNLP*, 2017. 11
- [57] Murray Shanahan. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive science*, 2005. 1
- [58] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL-HLT*, 2018. 3, 5
- [59] Cameron Shelley. Visual abduction in anthropology and archaeology. In *Systematic Methods of Scientific Discovery: Papers from the 1995 Spring Symposium*, 1995. 1
- [60] Yuqing Song, Shizhe Chen, and Qin Jin. Towards diverse paragraph captioning for untrimmed videos. In *CVPR*, 2021. 2, 8
- [61] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *CVPR*, 2019. 3
- [62] Didac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *CVPR*, 2021. 3
- [63] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, 2021. 11
- [64] Wilson L Taylor. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953. 2
- [65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5, 10
- [66] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 7, 8
- [67] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015. 2
- [68] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL-HLT*, 2015. 2
- [69] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016. 3
- [70] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *NIPS*, 2016. 3
- [71] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021. 2, 6, 7, 8, 9, 10, 11, 13
- [72] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1989. 6
- [73] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *ICCV*, 2021. 3, 5

- [74] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, 2018. [2](#), [6](#), [7](#), [9](#), [10](#)
- [75] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 2021. [3](#)
- [76] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015. [2](#)
- [77] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. [11](#)
- [78] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, 2018. [8](#)
- [79] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2019. [7](#), [11](#)
- [80] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020. [2](#)
- [81] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019. [2](#)
- [82] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. [2](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [13](#)
- [83] Wanrong Zhu, Zhiting Hu, and Eric Xing. Text infilling. *arXiv preprint arXiv:1901.00158*, 2019. [2](#)