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

Abductive reasoning is considered a core ability of human cognition [39, 54, 56] and 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 [51]), 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 [37, 38]. 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) [29]. 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 [25] and absolutely certain conclusions cannot be reached [5, 26].

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, [16] 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 [28, 45, 63, 64, 72, 75] 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 [29]. Typical DVC models [29, 42–44, 57, 70, 76, 77] 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 [34, 67], or turn to a top-down regime: first generate paragraphs, and then ground each description to a video segment [10, 36]. A few other methods [22, 31, 49] 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 [29], 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 [78], also known as the cloze task [60], is to generate a span of missing tokens in a text chunk, while sentence/story infilling [18, 20] aims to generate missing sentences in long-form text. The generated tokens/sentences are expected to smoothly blend into and fit the context syntactically [78], semantically [18, 20], and logically [23]. Counterfactual story revision [51] 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 (O₁) and future (O₂) contexts: O₁ → H → O₂, 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 [41, 66], future features [59, 65], future actions [1, 27, 30, 53, 58], future human motions [14, 21, 40], 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 [33] and [48]. However, [33] only requires choosing from two candidates for future event prediction, making the task less challenging. [48] 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 [17, 47, 71] or goal-conditioned procedure planning [7], while either relying on a pre-defined, limited action prediction space [7, 17], or using simulated environments [7, 17]. 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, [55] 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 [52], disentangling special tokens ([CLS]) [24], etc. Some recent vision-Transformers [8, 69] 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., $\mathcal{V} = \{O_1, \cdots, O_{n-1}, H, O_n, \cdots, O_{N-1}\}$, 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 content $\text{text} \ \{O_n\}_{n=1}^{N-1}$ 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 of our VAR dataset (§3.2).

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

- 23,457 *Youtube* lifestyle Vlog videos from ActivityNet Captions [29] and VLEP [33] datasets. These videos cover rich social scenarios and human activities.
- 13,799 *TV show and movie* videos from TVC dataset [32] and a famous Youtube channel, Fandango MovieClips\(^1\). 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.

### 3.2.1 Dataset Collection

**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

---

\(^1\)https://youtube.com/user/movieclips
Table 1 & Figure 3. Summative statistics of VAR dataset (§3.2.3).

**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 V with N temporally ordered events, i.e., \( V = \{O_1, \ldots, O_{n-1}, H, O_n, \ldots, O_{N-1}\} \), the premise events, i.e., \( \{O_{n+1}^{N-1}\} \), and explanation event, i.e., \( H \), are logically related. The AI system is only presented with a partially observable version of \( V \), i.e., \( \tilde{V} = \{O_1, \ldots, O_{n-1}, \tilde{H}, O_n, \ldots, O_{N-1}\} \), 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 \( \tilde{S} = \{S_n^{O_{n+1}^{N-1}}\} \) that describe the content of the \( N \) events in \( V \), while conditioning on \( \tilde{V} \) only:

\[
P(S|\tilde{V}) = P(S|H) \prod_{n} P(S_n^{O_n}|H) \\
= \prod_{n} P(w_{H}^{T}|w_{C_{H}}^{T}, \tilde{V}) \prod_{n} P(w_{O_n}^{T}|w_{C_{O_n}}^{T}, \tilde{V});
\]

where \( w_{l} \) is the \( l \)-th word in a generated sentence \( S \in S \). It is worth mentioning that, when \( H = \emptyset \), our VAR task is degraded into a classic DVC task [29] 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 \( V \), 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 \( \tilde{V} = \{O_1, \ldots, O_{n-1}, \tilde{H}, O_n, \ldots, O_{N-1}\} \) as \( \tilde{V} = \{E_n\}_{n=1}^{N-1} \), 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 \( \{E_{n}|_{n=1}^{N-1}\} \). For each premise event \( E_{n \neq h} \), corresponding feature \( E_{n \neq h} \) is obtained by aggregating the visual features of its frames. For the masked explanation event \( E_h \), we set \( E_h = 0^{d} \). 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 \( \tilde{E}_h \).

The attention operation is the core of Transformer:

\[
A \sim XW^{T}(XW^{h})^{T}, \ Y = \text{softmax}(A)XW^{w}.
\]
EhEN, is typically related to specific content, directional projection relationships between positions, i.e., \( \ell(n, m) = n - m + N \), and the position encodings 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 \([61]\), etc.

Absolute position encoding is widely adopted, in two typical ways: i) Absolute position encoding \([61]\): 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., \( X \leftarrow X + U \). ii) Relative position encoding \([55]\[]): the position embeddings 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 \([55, 69]\) and adopt a contextualized directional position embedding strategy, i.e., \( U_{nm} = \mathcal{F}_{\text{Rel}}(n, m, X_n) \in \mathbb{R} \):

\[
\mathcal{F}_{\text{Rel}}(n, m, X_n) = X_n \mathcal{R}_\ell(n, m),
\]

where \( \mathcal{R} \in \mathbb{R}^{(2N-1) \times d} \) 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 \( \mathcal{A} \in \mathbb{R}^{N \times N} \):

\[
\mathcal{A}_{nm} \sim X_n W^\gamma (X_m W^k)^\top + U_{nm}.
\]

We further set \( \mathcal{A}_{nh} = 0 \) to encourage leveraging the context from the observable events \( \{E_n\}_n \neq \emptyset \) to infer the masked explanation event \( E_n \), rather than vice versa. The Causality-aware encoder in REASONER is therefore achieved by stacking several Transformer encoder blocks \([61]\) 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 \( \{V_n\}_{n=1}^N \) of the observable premise events \( \{O_n\}_{n=1}^{N-1} \) as well as the explanatory hypothesis \( H \), 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 \( V_n \in \mathbb{R}^d \), a multi-modal, masked Transformer decoder is first adopted for initial description generation:

\[
[\tilde{V}_n, H_n] = D^\theta([V_n, H_n]),
\]

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., \( S^{E_n} \), and masked attention \([61]\) is adopted to prevent the leakage of future words. During inference, it is recurrently generated. Learnable modal-type embeddings \([11, 31]\) are also added into the input yet omitted for brevity. By fusing visual and linguistic representations as the input, \( D^\theta \) conducts cross-modal reasoning, and hence generates improved event representation, i.e., \( \tilde{V}_n^\theta \in \mathbb{R}^d \), and updated visual-linguistic state, i.e., \( H_n^\theta \in \mathbb{R}^{L_n \times d} \) for each event \( E_n \). Then a captioning head is adopted to map \( H_n^\theta \) into word distribution. The probability of \( l \)-th word is given as:

\[
P(w_{l|E_n}^{|E_n|} | w_{<l|^E_n|}, \tilde{Y}) = P(w_{l|E_n}^{|E_n|} | w_{<l|^E_n|}, H_n^\theta) = \text{softmax}(H_n^\theta(l)\Omega),
\]

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

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 \( D^\theta \) 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.
Specifically, our whole refinement procedure can be defined in a recursive, confidence-guided form:

\[
[V_n^k, H_n^k] = D^k\left((V_{n-1}^{k-1}, H_n^{k-1}, \{h_n^{k-1}\}_{n=1}^N), k = 1, 2, \ldots, K\right)
\]

where \(D^k\) refers to \(k\)-th refinement module and all the refinement modules are weight-sharing Transformer decoders; \(h_n^{k-1} \in \mathbb{R}^d\) indicates a condensed representation of \(H_n^{k-1} \in \mathbb{R}^{L \times d}\), \(h_n^{k-1} = \text{maxpool}_1(H_n^{k-1})\). In this way, each \(D^k\) can leverage inter-sentential relationship in previously generated descriptions \(\{h_n^{k-1}\}_{n=1}^N\) for refinement and better reason about the explanatory hypothesis. Moreover, we introduce the event confidence, i.e., \(\{c_n\}_{n=1}^N\), 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 \(X\) as the input of the decoder \(D^k\), i.e., \(X = [V_n^{k-1}, H_n, \{h_n^{k-1}\}_{n=1}^N]\) and omit the superscript \(k\). For each input “token” \(X_i\), its confidence score \(c_n\) is the one of its sourced event \(E_n\), and we normalize \(\{c_n\}_{n=1}^N\) over all the \(N\) events. Analogous to Eq. 4, the attention computation in \(D^k\) is modified as:

\[
\begin{align*}
A_{ij} & \sim X_i, W^V(X_i, W^K)^T + F^c_{\text{Rel}}(c_n_i, c_n_j), \\
F^c_{\text{Rel}}(c_n_i, c_n_j) & = r_{\text{c}}(c_n_i, c_n_j),
\end{align*}
\]

where the learnable vector \(r_{\text{c}} \in \mathbb{R}^{2B-1}\) can be viewed as a bucket to store the relative confidence weight; and the directional indexing function \(l(\cdot, \cdot)\) is given as \(l(c_n_i, c_n_j) = [c_n_i]B - [c_n_j]B + B\). 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 such decoders \(\{D^k\}_k\), outputs will be progressively improved (Fig. 6). Related experiments can be found in §5.2.

4.3. Training Objective

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

\[
\mathcal{L}_{\text{Main}} = - \sum_{k=0}^K \sum_{n=1}^N \sum_{l=1}^{L_n} P(\hat{w}_l^{E_n} | w_{\hat{c}_l}^{E_n}, H_n^k),
\]

where \(\hat{S}E_n = \{\hat{w}_l^{E_n}\}_{l=1}^{L_n}\). As the teacher forcing scheme [68] is used for training, \(H_n\) in Eq. 5 and 7 is embedded over one-hot encoded groundtruth words \(\{w_l^{E_n}\}_1\). We further adopt a hypothesis reconstruction based optimization criterion, to provide the encoder with more explicit supervision signals for explanatory hypothesis reasoning:

\[
\mathcal{L}_{\text{Aux}} = \|F^c_{\text{Proj}}(\hat{V}_h) - F^c_{\text{Proj}}(\hat{V}_h)\|_2,
\]

where \(\hat{V}_h\) and \(\hat{V}_h\) are embeddings for the explanatory hypothesis obtained from the masked and original videos, i.e., \(\hat{V}\) and \(\hat{V}\), respectively, and \(F^c_{\text{Proj}}\) is a projection head, based on a small multi-layer perceptron. This auxiliary training objective forces \textsc{Reasoner} to “imagine” an effective representation \(\hat{V}_h\) that better aligns with the original content of \(E_n\). \(\hat{V}_h\) 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 \textsc{Reasoner} is implemented as two Transformer encoder blocks, and each decoder module (§4.2), i.e., \(D^k\), 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 [19], as in [31,67,77]. 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 \(D^0\), 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 \(D^K\), 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 \textsc{Reasoner} on the classic, dense video captioning (DVC) task [29] (§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, 31, 77] and two RNN-based [67, 70] models, which are trained on train set of our VAR dataset with pre-provided event segments using their original training protocols.
Evaluation Metric. Five well-known automated metrics, i.e., BLEU@4 [46], CIDEr [62], METEOR [3], ROUGE-L [35], and BERTScore [74], 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, 31, 67, 70, 77] 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, 31, 67, 70, 77], 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 hypothesis. 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 [67] 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. 7. It contains the explanatory hypotheses and premise descriptions from our REASONER and other competitors [31, 67, 77] 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.
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., $D^0$. 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., $K \in \{0, 1, \cdots, 5\}$. When $K = 0$, only one decoder $D^0$ 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 [29] contains a total of 20k untrimmed videos (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 [31, 57, 77], 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 [31, 57, 77], BLEU@4 [46], METEOR [3], and CIDEr [62] 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 [22, 31, 49]. 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.
References


[29] 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


[34] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In CVPR, 2018. 2


