Supplemental Material: Visual Abductive Reasoning

Chen Liang[1,4] , Wenguan Wang[2] , Tianfei Zhou[3] , Yi Yang[1]

[1]CCAI, Zhejiang University [2]ReLER, AAII, University of Technology Sydney [3]ETH Zurich [4]Baidu Research

https://github.com/leonnnop/VAR

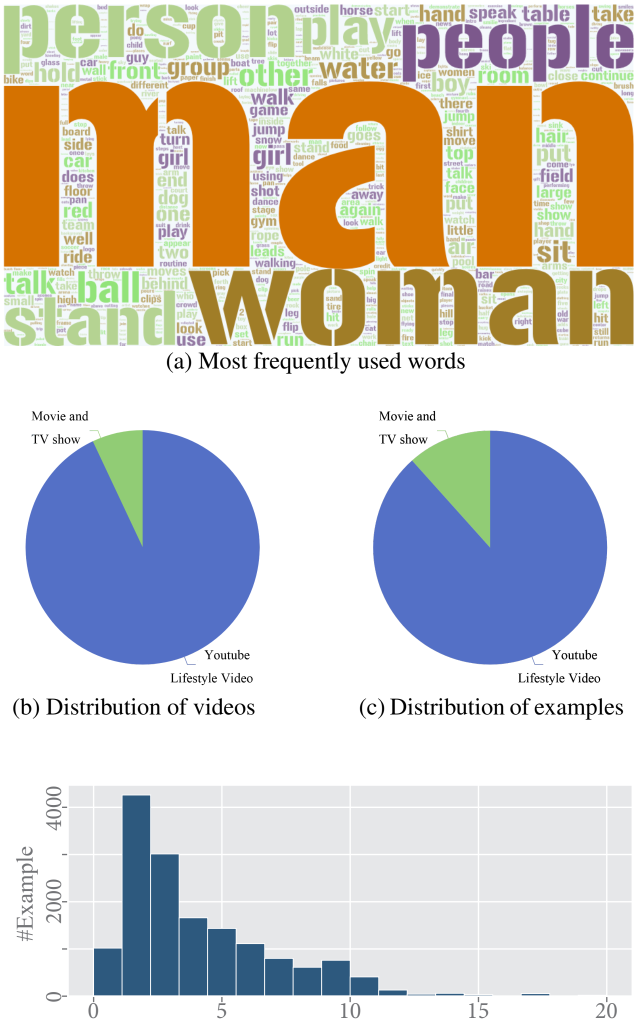
In this work, we build a headway of Visual Abductive Reasoning (VAR) as a new task and introduce a large-scale dataset for VAR that scaffolds the investigation of Abductive Reasoning ability in AI systems. Along with them, REASONER *is further presented as a modest solution. In the supplemental material, we provide the following items that shed deeper insight on the aforementioned contributions:*

- *Additional dataset analysis (§A)*
- *Additional details of benchmarked baselines (§B)*
- *Additional experimental results (§C)*
- *Additional qualitative visualization (§D)*
- *Discussion of limitation and reproductibility (§E)*
- *Discussion of legal and ethical considerations (§F)*

A. Additional Dataset Analysis

A.1. Detailed Dataset Statistics

Our VAR dataset is curated from three main sources, *i.e.*, YouTube Lifestyle video, movie and TV show, in Fig. 1b and Fig. 1c, we illustrate the detailed distribution of videos and examples by collected sources. As seen, most videos in VAR are from YouTube Lifestyle video, while movie videos tend to have more events, and thus leads to more complicated causal structures and more examples. We then study the distribution of frequently used words in VAR descriptions, which is illustrated as a word cloud in Fig. 1a. More frequent words are shown in larger font size. Finally, the distribution of premise events is shown in Fig. 1d.



(a) Most frequently used words

(b) Distribution of videos

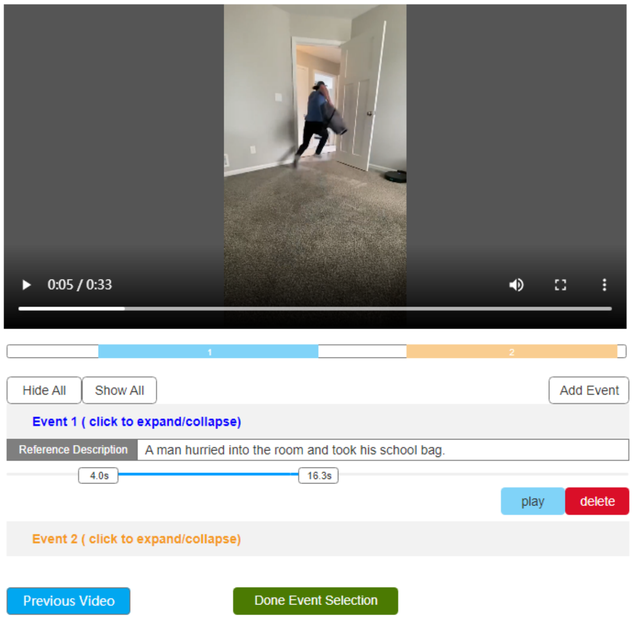
(c) Distribution of examples

(d) Distribution of premise events

Figure 1. Additional summative statistics of VAR dataset (§A.1).

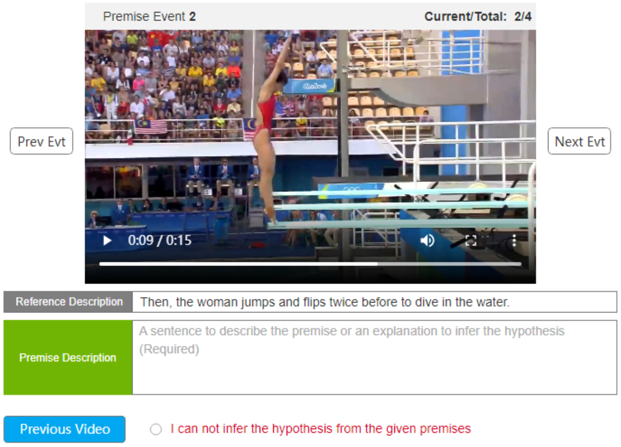
A.2. Detailed Annotation Process

In the annotation process, each video is passed at least four times: (1) A first quick pass for filtering out videos without cause-effect relations; (2) The second pass for annotating the event type, *i.e.*, the premise and explanation. In this phase, the entire video with initialized events and descriptions are all shown to the experts. We use the original event annotation provided in [3, 5, 6] to initialize event boundaries, while they might contain noise annotations. We thus request human experts to i) edit event boundaries when the initial separation can not well fit the description; ii) delete duplicate events when they are overlapped; iii) add additional events when they find a missing part in the cause-effect chain. And the annotation interface is shown in Fig. 2. After that, experts are requested to further annotate the event type based on the events he selected in the previous step, as shown in Fig. 4. (3) The third pass for abductive rea-



Reference Description: A man hurried into the room and took his school bag.

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



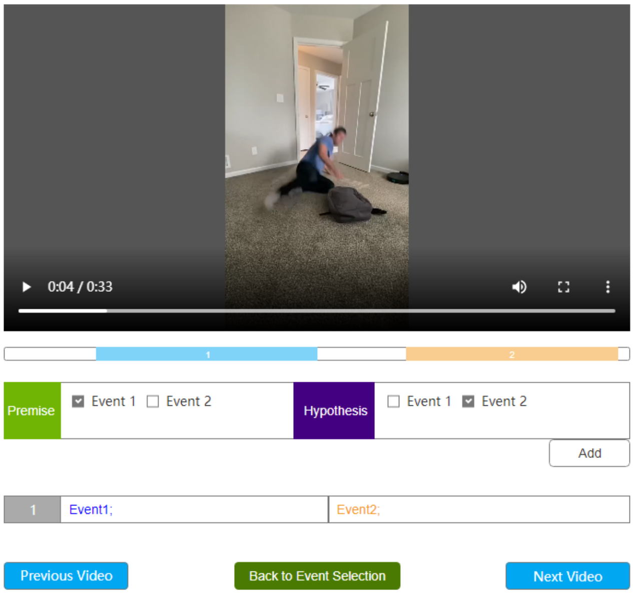
Reference Description: Then, the woman jumps and flips twice before to dive in the water.

A sentence to describe the premise or an explanation to infer the hypothesis (Required)

○ I can not infer the hypothesis from the given premises

Figure 3. Interface for describing premises. Details are in §A.2.

soning oriented description annotation. Human experts are only shown with premise visual events while the hypothesis event is hidden. Annotation interfaces for annotating the premise and hypothesis are shown in Fig. 3 and Fig. 5 respectively. Experts might vote to delete an example if they find that a plausible explanation can not be inferred from the premise. And the remaining examples are re-annotated with abductive reasoning oriented descriptions. Notably, a video might contain multiple examples (candidate cause-effect chains), and we manually control the example distribution to make sure the same video will not be shown to the same expert twice. (4) Forth, the final pass for validation. Both the annotated descriptions for premises and explanations for hypotheses are shown to another group of human experts. And they will vote for the validity.

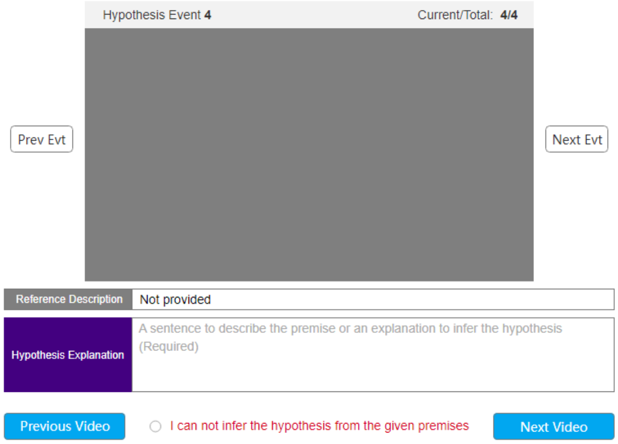


Premise ☑ Event 1 □ Event 2

Hypothesis □ Event 1 ☑ Event 2

Event1; Event2;

Figure 4. Interface for event type annotation. Details are in §A.2.



Reference Description: Not provided

A sentence to describe the premise or an explanation to infer the hypothesis (Required)

○ I can not infer the hypothesis from the given premises

Figure 5. Interface for explaining hypotheses. Details are in §A.2.

B. Additional Details of Baselines

We benchmark five top-leading DVC models [1,4,11,12, 15] on VAR. In this section, we detail the implementation and training protocol of these baseline methods.

MFT MFT [12] 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 casual structure can be captured that enables the inference on potential effect along the temporal order.

PDVC Similarly, PDVC [11] employs an LSTM-based captioning decoder, while it is conditioned on a deformable soft attention aggregated visual event. And the visual events

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**)	**5.03**	**10.75**	**24.81**	**38.27**	**34.88**	**3.44**	**9.05**	**22.89**	**30.75**	**30.64**

Table 1. **Additional quantitative results** on the `test` set of our VAR dataset. See §C for details.

are embedded with a Transformer-based encoder that could also capture bidirectional causal dependencies within it.

VTrans VTrans [15] is a fully attentional model that originates from the vanilla Transformer proposed in [10]. We follow the implementation in [4], 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) [1] 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 [4], gradients can flow through recurrent steps instead of being stopped. This enables stronger long-term modeling.

MART MART [4] 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 [4]. 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 1, 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.

Method	Premise Event			Explanation Event		
	BLEU@4	CIDEr	BERT-S	BLEU@4	CIDEr	BERT-S
[4]	3.74±0.07	29.22±0.39	29.53±0.12	2.86±0.07	24.05±0.27	27.77±0.16
[11]	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 2. Average scores and their standard deviations of REASONER and two representative methods [4, 11] (§C).

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 [4, 11] in Table 2.

D. Additional Qualitative Visualization

In Fig. 6-7, we show more qualitative examples from VAR `test` following Fig. 7 in the main paper. Generated sentences from competitors [4, 11, 15] along with our REASONER are presented in Fig. 6. 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. 7. 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 [7–9]. VAR, derived from these data, inevitably runs into the same problem [2, 13]. 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 [14] 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-idf_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/leonnnop/VAR.

F. Legal and Ethical Considerations

F.1. Asset License

Videos in VAR dataset are collected from four main assets: **(1)** ActivityNet Captions [3][1], 2017 version, under CC-BY 4.0 license[2]; **(2)** VLEP [6][3], 2020 version, under CC-BY 4.0 license[2]; **(3)** TVC [5][4], 2020 version, under CC-BY 4.0 license[2]; **(4)** MovieClips[5], copyright © 2021 Fandango. The site and services are available for non-commercial use. Detailed terms of use are available online[6]. VAR dataset will be released under CC-BY 4.0 license[2], 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.

References

[1] 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. 2, 3

[2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 3

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 4

[4] 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, 3, 5

[5] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 1, 4

[6] 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. 1, 4

[7] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *ACL workshops*, 2017. 3

[8] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *EMNLP*, 2017. 3

[9] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, 2021. 3

[10] 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. 3

[11] 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, 3, 5

[12] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, 2018. 2

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

[14] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2019. 4

[15] 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, 3, 5

[1] https://cs.stanford.edu/people/ranjaykrishna/densevid/

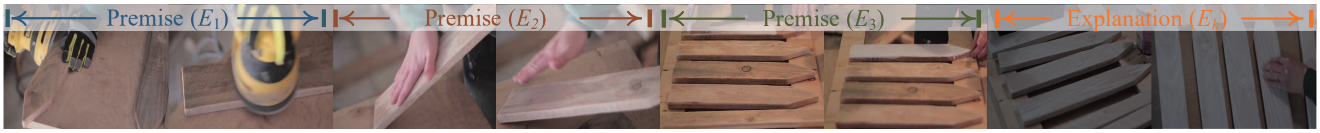
[2] https://creativecommons.org/licenses/by/4.0/

[3] https://github.com/jayleicn/VideoLanguageFuturePred

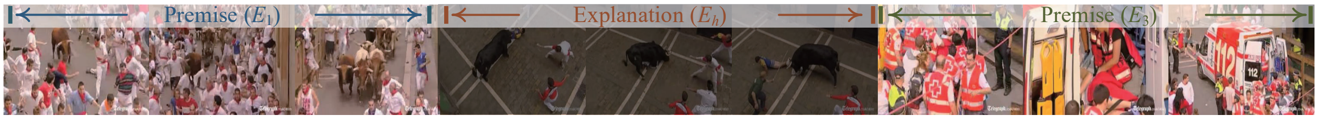
[4] https://github.com/jayleicn/TVCaption

[5] https://www.movieclips.com/

[6] https://www.fandango.com/policies/terms-of-use

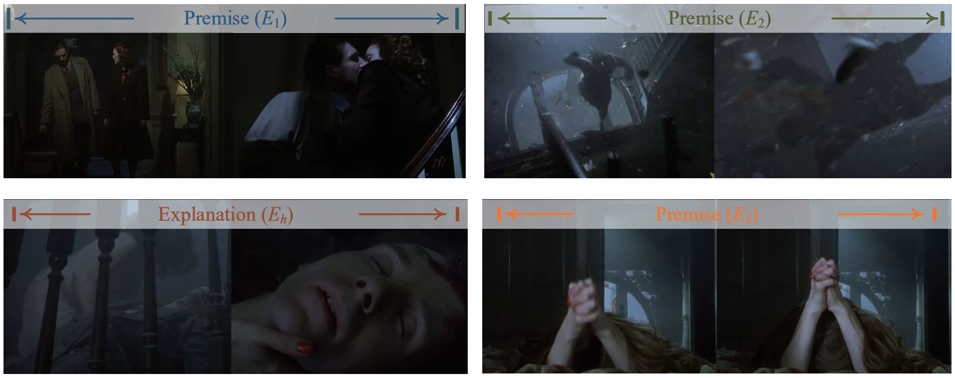


Vtransformer [15]: [The person uses a flat to clean the part of the wood.] [The man then rubs down the wood with a cloth.] [A person is putting objects on a table and leads into a person painting the wood.] [A man is holding a razor and begins playing the instrument.]
MART [4] : [A person puts a piece 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 [11]: [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 [15]: [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 [4] : [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 [11]: [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 6. Additional qualitative comparisons of REASONER and [4, 11, 15] 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 7. Additional failure cases of REASONER on VAR `test`. See §D for more details.