Unleashing Language Models by Summarizing the Causal Factors for Counterfactual Video Question Answering

Anonymous EACL submission

Abstract

While pretrained vision and language models have shown improvement in many high-level tasks, causal and counterfactual video question-answering (QA) still remains challenging. In contrast, pretrained (Large) Language Models have shown superior counterfactual reasoning abilities. In this work, therefore, we propose a natural language representation of videos that can capture objects, spatial relationships, and temporal events involving the objects; which can be more efficiently processed by pretrained Language Models for Video QA. We use this representation to capture important causal factors as additional context for Language-only models. We perform extensive experiments over two synthetic video QA datasets, namely CRIPP-VQA and CLEVERER – which covers counterfactual questions. We observe state-of-the-art results on CRIPP-VQA (upto 20% improvement in some categories) and improvement over state-of-the-art on counterfactual questions for CLEVRER. Our analysis shows the effectiveness of explicit entity-event representations for counterfactual video QA.

1 Introduction

Counterfactual Reasoning revolves around hypothetical scenarios and explores possible outcomes or consequences if certain conditions have been different. Counterfactual QA dataset is principally comprised of queries designed to delve into counterfactual or "What if" scenarios related to the intrinsic properties of objects. Properties of an object can be divided into extrinsic and intrinsic properties. Extrinsic properties of objects, encompassing shape, size, color and textures, can be easily identified with computer vision and found utility in applications such as captioning, retrieval etc. However, for Causal or Counterfactual reasoning tasks the intrinsic properties of objects such as mass, friction, elasticity, etc. must be taken into consideration. For example the Figure ...

With Video QA the modality of the data increases compared to a Language QA. Thus, the reasoning models need to preserve and understand the spatio-temporal relations between the objects. The spatio-temporal events depicted in videos such as collisions between objects hides visual cues about the intrinsic properties of these objects. Learning these video representations in multimodal setting is a struggle for the current Vision-Language models. On the contrary, extensive experiments have shown that Language models such as BERT, RoBERTa, T5 with their vast Natural Language understanding have out performed Vision-Language counterparts in every reasoning tasks. Models such as QUARTET (Rajagopal et al., 2020) and RGN (Zheng and Kordjamshidi, 2021) trained on Language only Counterfactual dataset, WIQA (Tandon et al., 2019), also provided significant evidence on this fact.

To explicitly capture such causal factors(objects, relations and events) we proposed a structured and causal model to extract video representations and convert them into a Language only structure. Extracting these video representations involves extracting each of the object's spatial relations and information as well as extracting the spatio-temporal collision events. We employed the YOLOv8 object detector to frame-wise extract each object's region proposals. These region proposals helped in building the spatial relationship graph and discover the collision events. To convert these representations into text format we proposed a visual dependency grammar. For the spatial information of an object, this grammar involves the direction of the object with respect to the origin and the nearest object. For the collision temporal event we use the collision dyanmics involving the direction of approach - collision - direction of separation. We concatenate all this together to build the entity-event graph. This graph, then serves as a context for the QA based training of Language models.

In summary, we make the following contributions:

- We proposed a method that could extract video representations and convert them into Language only format, preserving the spatial and spatio-temporal relations.
- Our pipeline provided a Language only solution for counterfactual reasoning in Video QA dataset

2 Related Work

Image and Video Knowledge Representation. Graph-based Knowledge Representations such as scene graphs (Krishna et al., 2017) and Visual relationships (Lu et al., 2016) are efficient representations to capture the complete set of objects and pairwise relationships (including spatial and semantic) between objects in an image. Researchers have also extended such graph-based representations in the context of videos, such as video scene graph (Yang et al., 2023b), video relationship detection (Ji et al., 2023, 2021). However, the method for capturing such graphs are not easily generalizable to capture spatio-temporal events in synthetic scenarios. For example, most methods target natural images, and videos; capturing commonsense events or actions, such as *walking towards*, *lifts*; whereas in synthetic videos, events are primarily collisions among objects. In our work, we take inspiration from visual dependency relations (Elliott and Keller, 2013) to capture the spatial relationships among objects. Primarily, this work proposes to calculate the spatial relationships based on certain geometric properties, overlap among objects, angle and distance between regions. Temporally, we model the collision events by detecting the collision dynamics consisting of *approach-collide-separate* sub-events.

Image Difference captioning and Video Captioning. Image Difference Captioning is the task of identifying the difference between a pair of images. Models such Robust Change Captioning (Park et al., 2019) and CLIP4IDC (Guo et al., 2022) can be trained to identify changes between a pair of images. However, the task and the datasets are not targeted to capture temporal or spatial relationship changes. Similarly state-of-the-art image captioning methods fall short for two reasons: i) they overtly concentrate on capturing a few salient aspects of the image, and ii) they are not trained for arbitrary synthetic images. We have qualitatively explored dense captioning (Johnson et al., 2016) as it attempts to describe all objects present in the scene. The dense captioning methods suffer from similar issues such as image captioning methods for lack of generalization to arbitrary videos.

A video dense captioning model Vid2seq (Yang et al., 2023a) is an interesting baseline, that is targeted to capture important saptio-temporal events in natural videos. This has similar limitations as compared to image captioning, focusing on capturing salient events and trained for general domain commonsense natural videos. Most of these techniques do not capture necessary spatial relationships and all important temporal events in a video.

Representation and Reasoning with Pretrained Vision and Language Models. The Vision and Language models (Kim et al., 2021; Li et al., 2023a) and Language Models (BERT, RoBERTa, T5 and LLMs such as GPT-3.5, GPT-4) have seen to successfully learn task-agnostic vectorized representations that demonstrated superior performance for both text-only, and vision and language tasks. However, for counterfactual reasoning (Pearl, 2009), we ideally need a structural causal model describing the relationship among causal factors. The task also concerns reasoning about aspects not present in the observed data. Thus, several work has demonstrated shortcomings in counterfactual reasoning over vision and language (both images and videos) (Sampat et al., 2021; Patel et al., 2022) Especially, for videos, Patel et al. (2022) shows the best performing models achieve around 70% accuracy for counterfactual questions. In contrast, extensive experiments have shown that smaller language models such as BART, T5 (Frohberg and Binder, 2022; Li et al., 2023b) and Large Language Models such as GPT-4 (Kiciman et al., 2023) can perform causal and counterfactual reasoning with impressive performance. We, therefore, look towards utilizing language models for counterfactual video QA by capturing the important causal factors in videos through text, primarily capturing objects, spatial relationships and temporal events.

3 Extracting Entity-Event Representation from Videos

The key challenge for causal and counterfactual Reasoning from video datasets is preserving as much of the spatial and spatio-temporal relations as possible. The method should encompass extracting informations from all the frames of the video

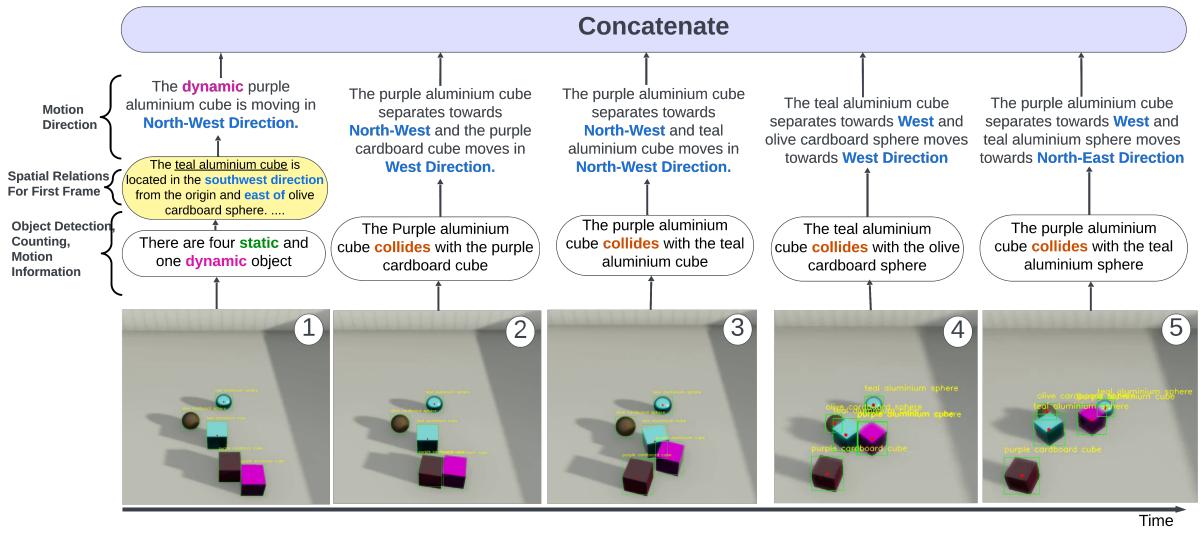

Figure 1: An example of the SUMMARY GENERATOR generating video summaries. For the initial frame, we summarize the number and type of each object, spatial relationships among pairs of objects and any motion information (with respect to next frame). For subsequent frames, we summarize the motion information and the collision information.

and not just the salient feature of the video. In this Section, we propose an efficient video representation that can be processed by powerful Language Models, and discuss how to efficiently extract from the videos. Consider Figure 1 as a running example. We are breaking down the example into key frames and their representations.

- **Frame 1:** Description of the static and dynamic objects and spatial relationships among the objects.
- **Frame 2:** Collision between purple aluminium cube and purple cardboard cube along with their directions of approach and separations.
- **Frame 3:** Collision between purple aluminium cube and teal aluminium cube along with their directions of approach and separations.
- **Frame 4:** Collision between teal aluminium cube and olive cardboard sphere along with their directions of approach and separations.
- **Frame 5:** Collision between purple aluminium cube and teal aluminium sphere along with their directions of approach and separations.

3.1 Extracting Video Representations

We aim to create an efficient and comprehensive representation of visual information in video to reason about spatio-temporal events in a causal and counterfactual manner. To this end, our representation involves per-frame representation of objects, spatial relations and temporal events across frames.

Representing & Extracting Object, Spatial Relationships. To create a graph for video representation, the first thing we need is to get both the information and the location of the objects present in the video. Extracting this information is done by object detection models such as YOLO (Redmon et al., 2016). After the extraction process, the next job is to efficiently represent the objects as well as their spatial relationships. Our proposed approach hinges on structured dependency representations of images.

The localization of an object is done by establishing its direction from the nearest two objects or it's direction from the nearest object and the direction with respect to the origin. An illustration of this concept can be encapsulated in this statement, *The pen is positioned upon the table and adjacent to the box*. This method will convert the spatial information about each of the objects into structured textual format. For example, consider the Figure

This figure shows the spatial information of one such object. To build the graph we need the spatial information of every static and dynamic object.

Representing & Extracting Collision Events. The next part of building the video representation graph is preserving the spatio-temporal relations of all the objects. The temporal event for the synthetic datasets such as CRIPP-VQA (Patel et al., 2022) and CLEVRER (Yi et al., 2020) is "Collision" between the various objects. To extract out

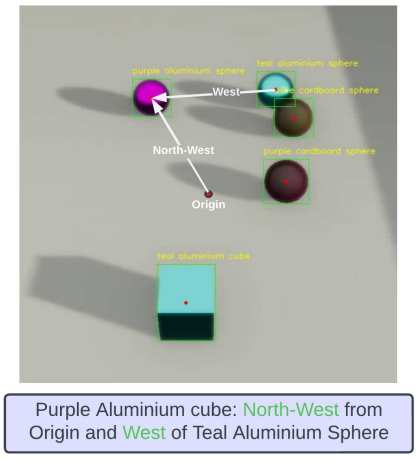

Figure 2: For each pair of objects, based on the bounding boxes, we capture spatial relationships. We use ordinal directions to describe the relationships, as LMs may acquire the commonsense understanding of such directions during pretraining (Tarunesh et al., 2021).

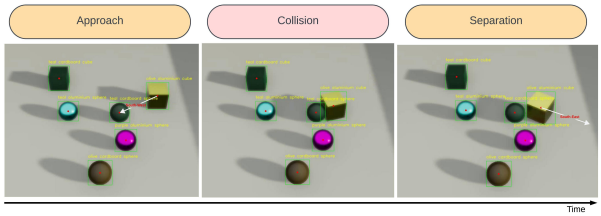

Figure 3: Any collision event can be decomposed into *approach*, *collide* and then *separation* subevents.

the collision events we need to observe the locations of each objects. A collision event will result in an overlapping of the location coordinates. The Figure below demonstrates the difference between the collisions and no-collisions. It also marks the direction of the dynamic object which is required for the next stage of the process.

3.2 Efficient Video Representation

From the previous sections we got the building blocks for the video representation graph. We have done this representation by concatenation of two parts: spatial information of the objects and spatio-temporal relations of the objects.

Building the graph. After the extraction of the collision events, the spatio-temporal relations are obtained by delineating those events for both the "collider" and "collidee". We need to preserve the dynamics of approach and seperation. This entails the consideration of the respective motion and directionality for both "collider" and "collidee" objects. Figure 3 shows the dynamics of Collisions. We have implemented the same approach of directional cues into textual format encoding. Finally, to build the graph we need to concatenate the spatial informations with these spatio-temporal relations.

4 Experiments

4.1 Entity-Event Extraction

The extraction of salient entities and events helps to generate an efficient object-centric representation of the visual context in videos to help the model focus on the information relevant to question answering. To this end, we employ a YOLOv8 (Redmon et al., 2016) (Jocher et al., 2023) detector to extract frame-wise object proposals. In order to add temporal context, we analyze the change in position of objects across successive frames and note the variation of motion directions over time. An illustration of our context generation approach is shown in Figure 1. We describe the initial state of objects by their relative position in terms of direction with respect to the center of the frame. Moreover, for dynamic objects, we incorporate the initial movement directions while entering the scene or at the beginning of motion. In addition to an initial description of the static and dynamic objects, we also detect key collision events under the assumption that **objects collide when their bounding boxes overlap significantly**. Our description of a collision event involves the colliding objects involved (*collider* and *collidee*) and the motion directions after collision.

Specifying the relative initial state of objects and motion directions along with the extraction of collisions allows for a succinct representation of the key events and the causal factors which govern them, allowing text-only language models to reason effectively even in case of more challenging counterfactual scenarios. Further analysis on the context generation quality is provided in Section 5.2.2.

4.2 Datasets

We evaluate the performance of our approach on the CRIPP-VQA (Patel et al., 2022) and CLEVRER (Yi et al., 2020) datasets.

CRIPP-VQA aims to test the model's understanding of implicit physical properties based on learning from videos. It consists of 5000 videos and 93,962 questions depicting standard-shaped

moving objects with implicit physical properties like mass and friction interacting with each other through collisions. Questions are divided into three types - DESCRIPTIVE (41,761 questions), COUNTERFACTUAL (41,761 questions) and PLANNING (10,440 questions), in which counterfactual questions involve three possible object manipulation scenarios - *Add* (27,016 questions), *Remove* (9603 questions) and *Replace* (5142 questions) objects in videos.

CLEVRER evaluates causal and temporal reasoning for videos based on physics simulations. It consists of 20,000 videos and 305,910 questions depicting object shapes interacting through collisions in a simulation environment. The dataset aims to test model reasoning through four types of questions - DESCRIPTIVE (219,918 questions), EXPLANATORY (33,811 questions), COUNTERFACTUAL (37,253 questions) and PREDICTIVE (14,298 questions).

4.3 Baseline methods

Pre-trained transformer-based models (Devlin et al., 2019), in conjunction with visual encoders (He et al., 2018), have been extensively used in creating state-of-the-art models for joint visual and textual reasoning owing to their ability to handle inputs pertaining to diverse modalities with large context lengths (Liu et al., 2021b) (Carion et al., 2020). However, since our approach involves the representation of video features using textual summaries, we employ the text-only T5 (Raffel et al., 2023) architecture, a state-of-the-art large language model which shows strong performance in various natural language tasks. We conduct our evaluation on two T5 variants: T5-*base* (223M parameters) and T5-*large* (783M parameters) and explore both *rule*-based and *triplet*-based contextual configurations.

We compare our approach to a combination of text-only and visual-language models on the CRIPP-VQA dataset :

Blind-T5 refers to the T5-base model, which processes only questions as input. We employ this baseline to analyze textual biases in the dataset.
YOLOv8-T5$_{summary}$ consists of the YOLOv8 and T5 models with rule-based summaries as video context.
YOLOv8-T5$_{triplets}$ consists of the YOLOv8 and T5 models with triplet-based summaries as video context.
Aloe*-BERT is a state-of-the-art video-language model which consists of a modified MaskRCNN-based Aloe visual encoder (Ding et al., 2021) (He et al., 2018) to detect object proposals, a pretrained BERT (Devlin et al., 2019) encoder to generate question embeddings and a learnable BERT component for question answering. We use the Aloe*-BERT model as specified by (Patel et al., 2022) and benchmark our approach's results against it for the CRIPP-VQA dataset tasks.

Similarly, for the CLEVRER dataset, we contrast the performance of our approach against the following baseline models specified in (Yi et al., 2020) which rely on both visual and language inputs.

CNN+LSTM extracts visual video features via a convolutional neural network (CNN) and uses pretrained Word2vec (Mikolov et al., 2013) embeddings for encoding questions, which are then sent to an LSTM (Hochreiter and Schmidhuber, 1997) for answer prediction.
MAC (Hudson and Manning, 2018) incorporates a joint attention mechanism on both the image feature map and the question. (Yi et al., 2020) employ a modified version of the model by integrating a temporal attention unit across frames to generate a latent video embedding **(MAC (V))**.
NS-DR (Yi et al., 2020) is a state-of-the-art video-language model which uses a dynamics predictor, a learnable physics engine which inputs object proposals from a visual encoder and learns the variation of object dynamics across frames for predicting motion directions and collision events.

4.4 Implementation details

We employ the following procedures for testing our approach's performance on the CRIPP-VQA and CLEVRER datasets :

T5: For the CRIPP-VQA dataset, we leverage a randomly-sampled 67.5:22.5:10 train:val:test split of the set of question-answer pairs for each task in the IID set. Moreover, for CLEVRER we employ a 3:1 ratio of the train set for train and validation splits and the original validation set as the held-out test set. The video frames are sampled at 25FPS and passed through a YOLOv8

model **finetuned on a small set of 150 manually annotated training examples?** to extract the visual features. The extracted visual features were then passed through the SUMMARY GENERATOR to generate tuples $\{c_i, q_i, a_i\}$, where c_i is the generated video context, q_i is the question and a_i is the answer. In case of multi-choice questions, we apply the same procedure to create tuples for each option, effectively reformulating the multi-choice question into a set of option-wise binary single choice questions. For both the *base* and *large* versions, we use the corresponding pretrained T5 tokenizer, setting the maximum question length $q_{len} = 1024$ and answer length $t_{len} = 32$. with lr = $1e^{-5}$ is used. For each task, the model was trained using Adam optimizer (Kingma and Ba, 2017) for 10 epochs with a batch size of 2 on a single 32GB Nvidia V100 GPU for 16 hours.

Aloe*-BERT: The Aloe*-BERT module consists of three components - (1) a visual encoder, (2) a pretrained textual encoder and (3) a large language model (LLM). The model utilizes extracted visual features from videos and textual embeddings from questions in the CRIPP-VQA dataset to predict answers. We use a modified version of the Aloe (Ding et al., 2021) module called Aloe* for visual feature extraction as specified by (Patel et al., 2022) in which the MONet (Burgess et al., 2019) module is replaced by Mask-RCNN (He et al., 2018). The video frames are sampled at 25FPS and forwarded through the Aloe* module with a linear layer f_v to generate frame embeddings u_v of dimension $d = 768$. The questions and answers are passed through a pretrained BERT encoder and a linear layer f_t to generate textual embeddings u_t of dimension $d = 768$. The visual and textual features are then concatenated and passed as input to a BERT model for question answering.

We finetune the BERT model with RAdam (Liu et al., 2021a) optimizer using a linear schedule with lr = $5e^{-6}$ and warmup steps = 4000. The model is initially trained for 6 hours on the descriptive task and then finetuned for 14 hours on a combined task of descriptive and counterfactual questions, each for 25 epochs. Subsequently, the trained descriptive checkpoint is further finetuned on the counterfactual and planning tasks each for 50 epochs and 6 hours. The training is done with a batch size = 64 on two 32 GB Nvidia V100 GPUs.

4.5 Evaluation metrics

For the CLEVRER and CRIPP-VQA datasets, we adopt the per-question (PQ) accuracy for all types of tasks. Moreover, for explanatory, counterfactual and predictive tasks which involve multiple-choice answers, we reformat the questions in terms of binary True-False sub-questions for each option, employing the per-option (PO) accuracy for evaluating option-wise correctness. For CRIPP-VQA, we report the results on the aforementioned test split, while for CLEVRER, we report the results on the original validation split.

5 Results

In this section, we demonstrate the relative performance of our approach over various state-of-the-art baselines on the CRIPP-VQA and CLEVRER datasets. Subsequently, we conduct an ablation study detailing the rationale behind the design choices employed for the SUMMARY GENERATOR.

5.1 Finetuning performance on Video-QA

We employ two types of video summaries - (1) rule-based, and (2) triplets and evaluate their effectiveness with the T5-*base* and *large* versions for video-based question answering. We contrast our approach to Aloe*-BERT on CRIPP-VQA tasks and the baselines mentioned in Section 4.3 on CLEVRER tasks.

Table 1 shows the experimental results on the CRIPP-VQA dataset for the aforementioned scenarios. Our approach outperforms Aloe*-BERT by 20.7% in terms of PO accuracy for the descriptive task, and by 10.2% and 17.2% on the counterfactual and planning tasks respectively. This indicates that the extraction of key entities and events for context generation significantly improves the large language model's ability to develop an efficient object-centric comprehension of the video events without explicitly using visual features. Consequently, it not only helps the model attend to relevant visual information for answering descriptive queries but also allows it to engage in counterfactual reasoning and deduce potential actions aimed at specific objectives in the context of planning tasks.

A similar trend is observed for the *Add*, *Remove* and *Replace* sub-tasks where our method outperforms Aloe*-BERT by $7 - 10\%$ and $10 - 13\%$ on PO and PQ accuracy respectively. The results also

	Descriptive	Remove		Replace		Add		Counterfactual	Planning
Model		PO	PQ	PO	PQ	PO	PQ	Avg. PO	
Blind-T5	56.03	52.44	21.13	47.95	14.86	51.26	15.24	49.8	7.44
Aloe*+BERT	71.04	65.46	33.64	56.76	22.07	67.43	39.71	63.21	32.61
T5+Metadata (base)	81.42	70.94	40.36	60.19	31.23	68.92	41.56	68.35	41.64
YOLOv8-T5$_{triplets}$ (base)	90.73	70.81	42.63	67.03	35.95	73.68	46.01	72.18	46.39
YOLOv8-T5$_{summary}$ (base)	88.98	72.56	44.88	65.72	34.6	74.24	47.12	72.78	47.94
YOLOv8-T5$_{summary}$ (large)	**91.73**	**72.61**	**45.95**	**65.91**	**35.73**	**75.20**	**48.59**	**73.43**	**49.85**

Table 1: Video-QA accuracy of visual reasoning models on CRIPP-VQA

Model	Descriptive	Explanatory PO	Explanatory PQ	Predictive PO	Predictive PQ	Counterfactual PO	Counterfactual PQ
Blind-T5	34.41	59.56	16.57	50.35	23.28	51.31	8.26
CNN+LSTM	51.8	62.0	17.5	57.9	31.6	61.2	14.7
MAC (V)	85.6	59.5	12.5	51.0	16.5	54.6	13.7
NS-DR	**88.1**	87.6	79.6	**82.9**	**68.7**	74.1	42.2
YOLOv8-T5$_{base}$	81.96	93.22	83.13	76.51	40.94	74.45	35.74
YOLOv8-T5$_{large}$	83.31	**93.24**	**83.18**	76.54	42.31	**78.67**	**45.40**

Table 2: Video-QA accuracy of visual reasoning models on CLEVRER

Model	Descriptive	Remove PO	Replace PO	Add PO	Counterfactual PO	Planning
Blind-T5	56.03	52.44	47.95	51.26	49.8	7.44
collisions-only	89.64	69.13	58.96	69.39	66.88	42.2
add-positions	88.96	70.27	63.74	71.4	68.89	44.35
collisions-directions	**89.66**	69.39	62.8	70.72	68.01	45.76
YOLOv8-T5$_{base}$	88.98	**72.56**	**65.72**	**74.24**	**72.78**	**47.94**

Table 3: Ablation results for the T5-base model tested on rule-based summaries

indicate that models generally perform better on the *Add* and *Remove* tasks compared to the *Replace* task, since it involves the *in-situ* change in physical properties of an object in the video frames. This implies that the baseline models, in general, are somewhat able to reason spatially, but are found lacking at reasoning about changes in physical properties.

A comparison between both summary generation approaches as highlighted in Table 1 shows that the rule-based summary approach performs on par with the triplet-based approach on the descriptive and counterfactual tasks but achieves a significant 3.5% gain for the planning task, suggesting that LLMs like T5 are better at comprehending fully textual contexts compared to triplet-based scene descriptions.

The experimental results of visual reasoning models on the CLEVRER dataset is shown in Table 2. Our approach outperforms the state-of-the-art NS-DR model on the explanatory and counterfactual tasks but surprisingly underperforms by a significant margin on the descriptive and predictive tasks. This can be attributed to the fact that the NS-DR approach incorporates a dynamics planner which is a learnable physics engine for explicitly modelling the motion of objects over short time frames and predicting future motion traces. This not only allows for a more intricate scene representation which positively impacts the model's performance on the descriptive tasks but also enables it to predict future events with better accuracy. Nevertheless, our approach still outperforms NS-DR by 6% and 4% in PO accuracy on the explanatory and counterfactual tasks respectively while employing a relatively simpler and efficient approach for visual context representation, suggesting that large language models trained for video-QA tasks using fully textual descriptions of key entities and events can exhibit competitive (and sometimes superior) causal and counterfactual reasoning capabilities over more sophisticated event modelling approaches.

5.2 Ablation Studies

To investigate the overall effectiveness of our proposed approach, we analyze the OBJECT DETECTOR and the SUMMARY GENERATOR and their effect on the overall context quality.

5.2.1 OBJECT DETECTOR

To extract moving objects from video frames, we employ YOLOv8 as an object detector by

finetuning the model on a small set of manually annotated video frames from the CRIPP-VQA and CLEVRER datasets. Since the context quality depends on the accuracy of localization of objects per frame, we analyze the performance of our object detector by evaluating the mAP scores for both datasets as shown in Table 4. Our visual extractor achieves a high mAP score of **88.6%??** on 12 classes of the CRIPP-VQA dataset and 93.4% on 48 classes of the CLEVRER dataset indicating the accurate detection of salient objects in video frames.

Dataset	# classes	mAP
CRIPP-VQA	12	0.886 **FIX!**
CLEVRER	48	0.934

Table 4: mAP scores for YOLOv8

5.2.2 SUMMARY GENERATOR

We analyse the impact of including different types of spatio-temporal information in video contexts on the overall context quality. We consider the following scenarios.

i) ***Blind-T5*** : T5-base model with no context provided, only questions as input.

ii) ***collisions-only*** : context with no relative initial positions or motion directions provided, only the initial objects and collider-collidee pairs mentioned.

iii) ***add-positions***: The context consists of the initial objects, their relative initial positions with respect to the frame center and the collider-collidee pairs mentioned. We remove motion directions.

iv) ***collisions-directions***: The context consists of only the initial objects, motion directions and collider-collidee pairs mentioned. We remove relative initial positions.

The PO accuracy results for each of these scenarios are compared with the T5-base model for rule-based summaries, as shown in Table 3. Blind-T5 exhibits close to random PO accuracy performance (49.8%) on counterfactual tasks and subpar performance (7.44%) on the planning task due to the absence of contextual information. A slightly better performance is observed for the descriptive task (56.03%), indicating that the descriptive task suffers from slight textual biases allowing the model to predict correct answers based on the question itself.

The addition of video summary as context leads to a steep increase in accuracy across all tasks (17% for counterfactual, 33% for descriptive and 34.8% for planning), highlighting the importance of contextual information in question answering. Furthermore, for dynamic scenarios involving moving constituents and collisions, the presence of spatio-temporal information like relative initial positions and motion directions greatly improves the model's ability to understand the complex physical interactions and reason about possible steps towards a planning objective. This is substantiated by both *add-positions* and *collisions-directions* outperforming *collisions-only* by $1-2\%$ on counterfactual tasks and by $2-3\%$ on the planning tasks. However, the performance on descriptive questions remains relatively similar across all context variations, suggesting that the descriptive task relies less on spatio-temporal information and more on textual information for question answering.

Furthermore, a more significant performance gain is observed for the *Add* and *Replace* tasks when the complete context is provided (6% and 5% on the *Replace* and *Add* tasks respectively) compared to the *Remove* task (3%). This corroborates our experimental results on the CRIPP-VQA dataset in Section 5.1 that models typically find it challenging to analyze scenarios involving in-place changes in physical properties and suggests that the *Add* and *Replace* tasks benefit from a combination of

6 Conclusion

References

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. 2019. Monet: Unsupervised scene decomposition and representation.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers.

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

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. 2021. Attention over

learned object embeddings enables complex visual reasoning.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, USA. Association for Computational Linguistics.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association.

Zixin Guo, Tzu-Jui Wang, and Jorma Laaksonen. 2022. CLIP4IDC: CLIP for image difference captioning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 33–42, Online only. Association for Computational Linguistics.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2018. Mask r-cnn.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning.

Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. 2021. Detecting human-object relationships in videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8086–8096. IEEE.

Xiaofeng Ji, Jin Chen, and Xinxiao Wu. 2023. Counterfactual inference for visual relationship detection in videos. In *IEEE International Conference on Multimedia and Expo, ICME 2023, Brisbane, Australia, July 10-14, 2023*, pages 162–167. IEEE.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. YOLO by Ultralytics.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4565–4574. IEEE Computer Society.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *CoRR*, abs/2305.00050.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023b. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2665–2679. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2021a. On the variance of the adaptive learning rate and beyond.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows.

Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 852–869. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2022. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Dheeraj Rajagopal, Niket Tandon, Peter Clark, Bhavana Dalvi, and Eduard Hovy. 2020. What-if I ask you to explain: Explaining the effects of perturbations in procedural text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3345–3355, Online. Association for Computational Linguistics.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection.

Shailaja Keyur Sampat, Akshay Kumar, Yezhou Yang, and Chitta Baral. 2021. CLEVR_HYP: A challenge dataset and baselines for visual question answering with hypothetical actions over images. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3692–3709, Online. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Lonli: An extensible framework for testing diverse logical reasoning capabilities for NLI. *CoRR*, abs/2112.02333.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023a. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10714–10726. IEEE.

Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. 2023b. Panoptic video scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18675–18685. IEEE.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning.

Chen Zheng and Parisa Kordjamshidi. 2021. Relational gating for "what if" reasoning. In *International Joint Conference on Artificial Intelligence*.

A Example Appendix

This is a section in the appendix.