Implicit Hypothetical Reasoning about Intrinsic Physical Properties

by

Maitreya Jitendra Patel

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2022 by the
Graduate Supervisory Committee:

Yezhou Yang, Chair
Chitta Baral
Kookjin Lee

ARIZONA STATE UNIVERSITY

December 2022

ABSTRACT

Multimodal reasoning is one of the most interesting research fields because of the ability to interact with systems and the explainability of the models' behavior. Traditional multimodal research problems do not focus on complex commonsense reasoning (such as physical interactions). Although real-world objects have physical properties associated with them, many of these properties (such as mass and coefficient of friction) are not captured directly by the imaging pipeline. Videos often capture objects, their motion, and the interactions between different objects. However, these properties can be estimated by utilizing cues from relative object motion and the dynamics introduced by collisions. This thesis introduces a new video question-answering task for reasoning about the implicit physical properties of objects in a scene, from videos. For this task, I introduce a dataset – CRIPP-VQA (Counterfactual Reasoning about Implicit Physical Properties - Video Question Answering), which contains videos of objects in motion, annotated with hypothetical/counterfactual questions about the effect of actions (such as removing, adding, or replacing objects), questions about planning (choosing actions to perform to reach a particular goal), as well as descriptive questions about the visible properties of objects. Further, I benchmark the performance of existing video question-answering models on two test settings of CRIPP-VQA: i.i.d. and an out-of-distribution setting which contains objects with values of mass, coefficient of friction, and initial velocities that are not seen in the training distribution. Experiments reveal a surprising and significant performance gap in terms of answering questions about implicit properties (the focus of this thesis) and explicit properties (the focus of prior work) of objects.

DEDICATION

Dedicated to, my loving parents, family and friends for the love, patience, and faith in this short journey and in much more to come...

ACKNOWLEDGEMENTS

Writing this thesis was a quite a journey, formulating a research problem, performing hypothesis testing, and developing engineering infrastructure. I am really grateful to have Dr. Yezhou Yang as my advisor. I want to express my sincere gratitude towards Dr. Yang for guiding me throughout my masters' journey as this would not have been possible without his constant support. Special thanks to Dr. Chitta Baral for doing regular discussions to formulate the problem statement. Also, thanks to Dr. Kookjin Lee for agreeing to become the presiding member of the thesis.

I would also like to thank Tejas Gokhale for mentoring me on this journey. And helping me with ups and downs. Moreover, thanks to Shailaja Sampat and Pratyay Banerjee for their involvement in brainstorming sessions.

I would like to express my serious gratitude towards my parents, family, and friends to being there for me and keeping me motivated. This would not have been possible without them!

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

With the advancement in deep learning, many fields (including, Computer Vision, Natural Language Processing, Biomedical, Trading, etc.) are benefiting from it day by day. Recently an intersection of Computer Vision (CV) and Natural Language Processing (NLP), also known as Vision and Language (V&L), is becoming more popular because of its innate ability to express and interact with the system using natural language. V&L has been applied to various fields such as Robotics, Guidance to the visually impaired, etc. V&L consists of many tasks such as Image Captioning, Visual Question Answering, Image-Text Retrieval, etc. Apart from Images, V&L also consists of tasks having video understanding requirements such as Video Question Answering, Video Retrieval, Video Summarization, etc.

Vision and Language research problems mainly involve real-world datasets such as COCO (Lin *et al.*, 2014), Visual Genome (Krishna *et al.*, 2017), etc. However, these datasets cannot test the models' ability to reason. Datasets created using the internet do not evaluate systems for commonsense or complex reasoning abilities. To take a step towards Human-Level Artificial Intelligence (HLAI), existing benchmarks and datasets are very limited. Because of that, there has been decent progress in creating challenges for AI systems using synthetic environments. The synthetic environment allows us to create a close-world scenario, where everything is under control. The benchmark has a synthetic dataset that focuses on various complex reasoning tasks involving physics and compositional behavior. Some of these problems are nearly solved and for some, we are nowhere near to human level performance.

The main aim of this thesis work is to propose a new dataset (which requires visually

hidden properties-based reasoning) and benchmark it using SotA methods.

1.1 Vision and Language Research

Vision and Language (V&L) is a term coined for multi-modal filed having applications at the intersection of CV and NLP. Each V&L application can be further categorized based on the type of visual input (i.e., image or video). V&L is growing significantly over the past few years as it brings forth a new set of challenges which is required to overcome to achieve the HLAI. Mogadala et. al. summarizes the recent trends in visual grounding including the various important tasks and corresponding datasets (Mogadala *et al.*, 2021). This research direction started with image-text embedding models (Barnard *et al.*, 2003), (Frome *et al.*, 2013), and (Kiros *et al.*, 2014). With the advancement of deep learning, Visual description generation (also known as captioning) tasks started getting attention. Some of the earliest work involves image and video-specific captioning methodologies, which uses Convolutional Neural Networks (CNNs) as visual feature extractor and adopts seq2seq-based strategies to get better performance. Visual Question Answering (VQA) is a task that takes natural language and visuals as input and predicts the answer corresponding to the question, where the answer is present in the image. Earlier approaches for VQA were very similar to captioning, where instead of text generation the problem of answer classification. Other sets of important problems include referring expression (where the task is to identify the focused region within the image), and visual entailment (where the task is to identify whether the image and text are entailed or not). There are several datasets proposed for each of these tasks. MSCOCO (Lin *et al.*, 2014), Flickr8k, Flickr30k (Young *et al.*, 2014), Conceptual Captions (Sharma *et al.*, 2018), etc. contains the image-caption pairs, while YouCook (Zhou *et al.*, 2018), MSR-VTT (Xu *et al.*, 2016), etc. are especially created for video-captioning. In the case of the referring expression, RefCOCO, and RefCOCO+ are standard datasets (Yu *et al.*, 2016). For VQA tasks, there are several

datasets with different focuses such as VQA1.0/2.0 (Antol *et al.*, 2015; Goyal *et al.*, 2017), OK-VQA (Marino *et al.*, 2019), KVQA (Shah *et al.*, 2019), GQA (Hudson and Manning, 2019), etc.

1.2 Physical Reasoning

The exciting possibility of V&L is to gain the ability to reason about commonsense knowledge. Visual Commonsense Reasoning is an important benchmark proposed for this task. Commonsense reasoning is a very abstract term itself and this makes it even more difficult to create the tasks. Therefore, to evaluate the models' capability to do complex reasoning, the current trend is moving toward the synthetic environment. These synthetic environments are closed worlds and allow full control with scalability. Another problem in regular real-world datasets is that they are highly biased and there are many spurious correlations within the dataset. Because of that model trained on these datasets learns this correlation and gets better performance. Hence, it becomes hard to learn the limitations of any proposed approaches. This is another reason which gave an invitation to synthetic datasets for better analysis of the various proposed approaches.

In visual grounding, CLEVR (Johnson *et al.*, 2016) first introduced a block-world environment for VQA to evaluate the models' compositional reasoning ability. Similarly, CLEVR-Ref+ was proposed for referring expression (Liu *et al.*, 2019). On the other hand, CLEVRER (Yi *et al.*, 2020), CATER (Girdhar and Ramanan, 2019), etc. focused on video-based complex reasoning. The video-based benchmark studies mainly involve physics and their goal is to evaluate whether models can learn the physical dynamics or not. These studies involve various levels of physics, for example, they consider the influence of gravity, complex object motion, and a combination of both. Studies showed that SotA systems perform very low compared to human upper-bound on these benchmarks. Dataset-specific methods are developed to achieve better performance. Here, neuro-symbolic approaches

seem to out-perform the pure deep learning methods. However, they cannot be applied outside the close-world cases; suggesting that more research on making these methods successful is required.

1.3 Causal Reasoning

Causality is the relationship between cause and its effects/consequences. Humans do causal reasoning easily in day-to-day activities because of various factors. For example, we know the differences between the consequences of kicking a football vs. a brick. This knowledge allows us to do various tasks without putting in more effort. There has been a lot of work on causality in medicine, economics, etc. (Pearl, 2009) gives a nice summary of how causal reasoning can be applied via Structural Causal Models and *do*-calculus. Another important aspect of Human Intelligence is counterfactual reasoning. Given a scenario, we can think of another imaginary case where certain conditions change (like, what if ...?). This affects our decision-making ability and makes it more reliable.

In recent years, causal/counterfactual reasoning is being applied to various machine-learning problems. Consider an example, where during evaluations there is an image with a toothbrush on the dining table, while during the training there isn't one. Here, deep learning systems learn this spurious correlation and fail at test time. (Wang *et al.*, 2020) attempts to resolve this spurious correlation issue for robust object detection using causal interventions. Similarly, (Zhang *et al.*, 2020) proposes DeVLBert for learning deconfounded vision-language representations. (Niu *et al.*, 2021) attempts to remove the language bias in V&L models using causal interventions as well. These and many other similar methods are highly customized based on the problem statement and cannot be generalized. To get a step closer to HLAI, it is necessary to make the system self-sufficient to these biases and learn meaningful world knowledge to act accordingly.

It has been shown that counterfactual thinking is one of the main sources of our intelli-

gence. It governs many of our daily decisions. For example, if someone fails in math tests then s/he thinks that what if s/he had studied one hour more each day? (Epstude and Roese, 2008) shows that there are two types of counterfactual thinking: 1) Subtractive, and 2) Additive. Removing the type of counterfactual thinking (i.e., after getting the exam score) increases performance, while additive thinking (i.e., before the exam) improves creativity. Several studies are focusing on counterfactuals in deep learning. CLEVRER (Yi *et al.*, 2020) focuses on remove action-based counterfactual questions-answering. CoPhy (Baradel *et al.*, 2020) focuses on displacement-based counterfactual consequences estimation. Both CLEVRER and CoPhy are developed in a simulated environment; allowing scalable experimentation. While CoSci focuses on real-world question-answering on hypothetical conditions.

1.4 Motivation and Contribution

Many of our day-to-day life requires commonsense reasoning having a different level of physical properties. I learn these properties by interesting in the world or observing someone interact. In other terms, I perform various actions in the environment to gain/use different knowledge. (Sampat *et al.*, 2022) collected the various problems in multi-modality with the focus on actions to do complex reasoning. Motivated by this, this thesis attempt to propose a new benchmark with a specific focus on physical properties, which can only be learned by performing actions on imaginary scenarios.

Videos often contain objects, each having their own properties; for instance objects belong to certain categories, have shapes, sizes, and colors. These visible properties can be estimated by using computer vision algorithms for object recognition, detection, color recognition, shape estimation, etc. However, objects also have physical properties which in many cases are not captured by cameras. For example, cameras can capture the shape and color of an object, but not its mass. Consider the frames in Figure 1.1 that contain objects

with different shapes, textures, and colors, existing video question-answering datasets ask questions about these visible properties, but it is hard to reason about the masses of these objects or their coefficients of friction.

Collisions between objects, however, do offer visual cues about mass and friction. When objects collide, their resulting velocities and direction of motion depend upon their mass and friction coefficient, according to fundamental Newtonian dynamics. By observing the change in velocities and directions, it is possible to reason about the relative physical properties of colliding objects.

In many cases, When humans watch objects in motion and under collision, we do not accurately know the masses, friction, or other properties of objects. Yet, when we interact with these objects, for example in a game of billiards, we can reason about the effect of actions such as hitting one ball with another, removing an object, replacing an object with a different one, or adding an object to the scene. In this thesis, I consider the task of reasoning about such implicit properties of objects, via the use of language, without having annotations for the true values of mass and friction of objects. Based on (Patel *et al.*, 2022), I propose a video question answering dataset called CRIPP-VQA, short for **C**ounterfactual **R**easoning about **I**mplicit **P**hysical **P**roperties. CRIPP-VQA contains videos annotated with question-answer pairs. Each video contains several objects with at least one object in motion. The object in motion causes collisions which changes the spatial configuration of the scene. The consequences of the collision are directly impacted by the physical properties of objects. CRIPP-VQA asks questions about these consequences.

As shown in Figure 1.1, questions in the dataset require understanding the current configuration as well as counterfactual situations, i.e. the effect of actions such as removing, adding, and replacing objects. The dataset also contains questions that require the ability to plan in order to achieve certain configurations, for example producing or avoiding particular collisions. It is important to note that both tasks can not be performed without an

understanding of the relative mass. For example, replace action can lead to a change in mass inside the reference video, which can drastically change the consequences (i.e., set of collisions).

I benchmark existing state-of-the-art video question-answering models on the new CRIPP-VQA dataset. Key finding, in this thesis, is that compared to performance on questions about visible properties ("descriptive" questions), the performance on counterfactual and planning questions is significantly low. This reveals a large gap in understanding the physical properties of objects from video and language supervision. Detailed analysis shows that the models can predict the first collision on counterfactual questions with high accuracy compared to the subsequent collisions. Models struggle at answering questions about the effect of *"replace"* action. In the case of the *"add"* action, models can predict which collisions won't happen, but fail to predict the collisions that indeed happen.

I found Aloe (**?**) one-of-the stat-of-the-art baselines to be unstable on the proposed dataset, CRIPP-VQA, primarily because one of its pre-processing modules of identifying objects fails on CRIPP-VQA videos. Although this module works on previous datasets, it leads to close-to-random performance on CRIPP-VQA due to the presence of complex textures, reflections, and shadows. To mitigate this problem, I modify Aloe by adapting the Mask-RCNN (He *et al.*, 2017) for the object segmentation module. Moreover, I found that adding pre-trained BERT-based word embedding significantly improves the performance over the simple Aloe.

CRIPP-VQA also allows us to evaluate trained models on out-of-distribution test sets, where the videos vary from the training data in terms of a single physical property at test time (such as a change in mass, friction, and velocity). This OOD evaluation reveals a further degradation in performance and a close-to-random accuracy for most state-of-the-art models.

I summarize main contributions of the thesis below:

1. I introduce a new benchmark, CRIPP-VQA, for video question answering which requires reasoning about the implicit physical properties of objects in videos.
2. CRIPP-VQA contains questions about the effect of actions such as removing, replacing, and adding objects, which have not been considered in prior work on video QA.
3. Performance evaluation on both in-domain and out-of-domain test sets shows the significant challenge that CRIPP-VQA brings to video understanding systems.

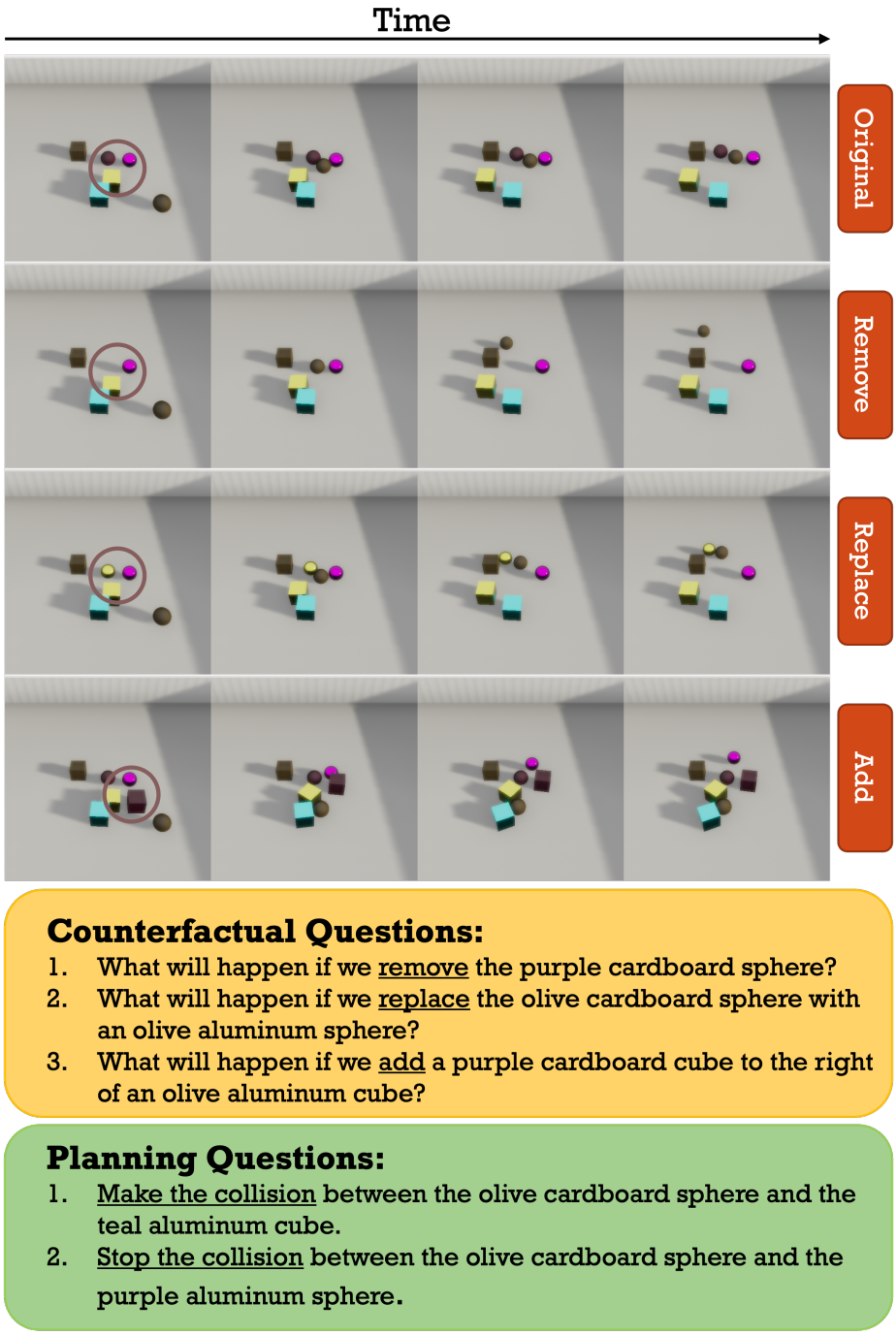

Figure 1.1: The CRIPP-VQA Dataset Contains Questions about the Future Effect of Actions (Such as Removing, Adding, or Replacing Objects) as well as Planning-based Questions. A Few Frames of a Video Are Shown above, with the Red Highlighted Area Depicting the Objects on Which Actions Are Performed.

Chapter 2

RELATED WORK

Video Question Answering is growing a lot in the past few years. There are several datasets/benchmarks proposed for video-based QA tasks. VideoQA is different than ImageQA as video is a temporal task and there are other important aspects such as conversations, actions, consequences, etc. within a video. While ImageQA only contains the questions whose answer is inside a single image. MovideQA (Tapaswi *et al.*, 2016) dataset contains 14,944 question-answer pairs from 408 movies. TGIF (Li *et al.*, 2016) dataset contains 100k GIFs with the corresponding description in the text. TVQA/TVQA+ (Lei *et al.*, 2020) is a large-scale VideoQA dataset having 152.5k question-answering pairs. AGQA (Grunde-McLaughlin *et al.*, 2021) is another dataset for spatial-temporal reasoning within the video as question answering task. This work on video question answering has mainly focused on real-world scenes taken from movies and television shows. Ego4D (Grauman *et al.*, 2022) is another huge dataset that contains ego-centric videos taken from volunteers with a diverse set of applications.

Textual Commonsense Reasoning type of tasks involves the understanding of the hidden aspects within the input textual data which are not explicitly present. For example, we cannot fit a person inside a toy car means that the toy car is relatively smaller than the person. There have been many studies in the field of NLP to incorporate commonsense at different levels of difficulty. PIQA (Bisk *et al.*, 2020) is proposed for physical commonsense reasoning for natural language understanding (NLU) systems as multiple choice question-answering. CommonsenseQA (Talmor *et al.*, 2019) is another QA dataset that focuses on the relationship between entities, which are not given as a part of the input

Dataset	Video QA	Physical Reasoning	Visually Hidden Properties	Counterfactual Actions			Planning	Physics OOD	Implicit reasoning
				Add	Replace	Remove			
MovieQA (Tapaswi *et al.*, 2016)	✓	-	-	-	-	-	-	-	-
TGIF-QA (Li *et al.*, 2016)	✓	-	-	-	-	-	-	-	-
TVQA/TVQA+ (Lei *et al.*, 2020)	✓	-	-	-	-	-	-	-	-
AGQA (Grunde-McLaughlin *et al.*, 2021)	✓	-	-	-	-	-	-	-	-
CoPhy (Baradel *et al.*, 2020)	-	✓	✓	-	-	-	-	-	✓
CLEVR_HYP (Sampat *et al.*, 2021)	-	-	-	✓	✓	✓	-	-	-
IntPhys (Riochet *et al.*, 2018)	✓	✓	-	-	-	-	✓	-	-
ESPRIT (Rajani *et al.*, 2020)	✓	✓	-	-	-	-	✓	-	-
CATER (Girdhar and Ramanan, 2019)	✓	-	-	-	-	-	-	-	-
CRAFT (Ates *et al.*, 2020)	✓	✓	-	-	-	✓	-	-	-
CLEVRER (Yi *et al.*, 2020)	✓	✓	-	-	-	✓	-	-	-
ComPhy (Chen *et al.*, 2022)	✓	✓	✓	-	-	-	-	-	-
CRIPP-VQA (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2.1: A Comparison of CRIPP-VQA with Prior Work on Video Question Answering, in Terms of Different Aspects of Visual Reasoning That Are Tested.

(i.e., that is hidden). Verb Physics (Forbes and Choi, 2017) is a dataset with an emphasis on learning relative physical knowledge (size, weight, strength, etc.). PROST attempts to learn the consequences of several actions on different objects, which depend upon their physical properties. QuaRel dataset focuses on the relationship between different physical properties and how they impact each other.

Visual Commonsense Reasoning adds another layer of challenges for commonsense reasoning on multi-modality. In the case of a multimodal setup, the commonsense meaning is hidden inside the image, and text is used to infer this commonsense knowledge. (Sampat *et al.*, 2022) gives a survey of previous works on action reasoning. VisualCOMET (Park *et al.*, 2020) is a dataset for inferring commonsense concepts with the main emphasis on learning the relationship between future events and their cause/effects from the images and textual descriptions. Video2Commonsense (Fang *et al.*, 2020) is a video captioning task that seeks to include intentions behind the event to learn the effects of human actions to

derive the attribute of the subject. VCR (Zellers *et al.*, 2019) is another standard dataset that introduces a VQA task that requires commonsense and understanding of the scene context to answer questions and to justify the answer. CLEVR_HYPE (Sampat *et al.*, 2021) proposes a benchmark having hypothetical conditions/actions on CLEVR (Johnson *et al.*, 2016) environment.

Physical Reasoning is the most relevant and focus of interest in this thesis. With the recent advancement in physics-based simulators, the application of these systems is becoming quite important and an alternative to a real-world environment. MuJuCO (Todorov *et al.*, 2012) physics engine is designed for model-based control. PyBullet[1] is another physics engine similar to MuJuCo. iGibson (Shen *et al.*, 2021) is a rigid body-specific engine for home simulations. ThreeDWorld (Gan *et al.*, 2020) is a high-level API that uses unity as a physics engine to render and control the environment, where the user can control different nobs of physical properties. Because of the flexibility of using these and many other simulators, there have been many use cases and several of the benchmarks use one or another engine for various tasks where it is much hard to replicate the same set of experiments in real-world scenarios. CATER (Girdhar and Ramanan, 2019) seeks to solve the temporal reasoning on different actions such as pick-place, slide, rotate, etc. CLEVRER benchmark (Yi *et al.*, 2020) proposed the challenge of counterfactual reasoning of object dynamics over remove action. However, all objects in CLEVRER have identical physical properties leading to the same set of consequences. CoPhy (Baradel *et al.*, 2020) attempted to predict the consequences with respect to causal intervention (i.e., displacement of a single object). It does not involve the change in physical as well for counterfactual reasoning instead focuses on predicting these properties using a single example. Filtered-CoPhy (Janny *et al.*, 2022) extends the CoPhy to perform the same task in pixel space. Recent

[1] `https://pybullet.org/wordpress/`

work, ComPhy (Chen *et al.*, 2022), is the most relevant/closest to this study, intending to do physics-based counterfactual reasoning after inferring physical properties in a few-shot setting. ComPhy focuses on explicit question-answering regarding the physical properties (*"What if object A was heavier?"*). In contrast, in this thesis, physical properties need to be learned from the video and are not mentioned in any type of questions, with three types of questions (descriptive, counterfactual, and planning). And they should be learned in an implicit setting.

Planning based reasoning is important aspect in robotics. Planning and decision-making require performing the action to achieve the end goal. Planning-based tasks such as object navigation with and without the various obstacles are previously explored. Physical reasoning-based decision-making adds another layer to complex reasoning. A handful of attempts have shown that indeed this is challenging. Visual planning (i.e. inferring the required actions to reach the desired goal state) has been explored in Chang *et al.* (2020) and Gokhale *et al.* (2019). IntPhy (Riochet *et al.*, 2018) and ESPRIT (Rajani *et al.*, 2020) require planning-based reasoning under the influence of gravity. In the case of IntPhy and ESPRIT, the system needs to predict the initial conditions to achieve the given goal. In contrast, CRIPP-VQA focuses on a planning-based task that requires the understanding of intrinsic physical properties.

Chapter 3

THE CRIPP-VQA DATASET

CRIPP-VQA, short for Counterfactual Reasoning about Implicit Physical Properties via Video Question Answering, focuses on understanding the consequences of different hypothetical actions (i.e., remove, replace, and add) in the presence of mass and friction as visually hidden properties. CRIPP-VQA further involves the planning task which requires the system to do reasoning involving the learned physical properties. This chapter, explains the proposed dataset creation process. First, we explain the simulation setup and how each videos are generated. Second, we describe the flow-chart for creating the useful videos/annotations for training and evaluations. Third, we focus on question-answer pair generation for each category of tasks. At last, we show the statistics of dataset to gain the better understanding.

3.1 Simulation Setup

Physics Simulator. *ThreeDWorld* (*TDW*) (Gan *et al.*, 2020) is used as default CRIPP-VQA physics simulator. *TDW* is a multimodal physics simulator, having physics-based interaction between environment objects and the agents. *TDW* uses Unity as a physics engine, which allows better physical interaction and photo-realistic rendering. Furthermore, *TDW* allows the ease of scalability for data creation. Many of the previous studies use CLEVR style rendering, leading to global overfitting (i.e., across the models/problem statements) on visuals (as shown in section 4.3). *TDW* allows us to control the environment and play with physical parameters such as mass, friction, velocity, bounciness, size, etc. in a realistic setting.

Property	IID	Mass	Friction	Number of Objects	Velocity
Shape	(sphere, cube)	-	-	-	-
Color	(purple, teal, olive)	-	-	-	-
Texture	(cardboard, aluminum)	-	-	-	-
Mass	(2,14)	(2,8,14)	-	-	-
Friction	(0.25)	-	(0.0)	-	-
# of moving objects	1	-	-	2	-
Initial velocity	(14)	-	-	-	(18)

Table 3.1: The Key Difference Between the *i.i.d.* And Various OOD Evaluation Settings in CRIPP-VQA. Here, "-" Indicates the No Change in Particular Property from the *i.i.d.* Setting.

Video creation. In each video instance, I first initialize it with N (where, $N \in \{5, 6\}$) number of randomly chosen objects $O = \{o_1, ..., o_N\}$. Here, $o_i \in \{o^1, ..., o^M\}$, where M is the predefined number of objects with fixed physical properties consistent in all videos. Out of these N objects, one object (o_i) will be initialized with a directional velocity and with another random object (o_k, where $k \neq i$) as target. The magnitude of the velocity is defined in such a way that initial acceleration is same for either lighter or heavier objects. This can be achieved my applying applying variable force to the object based on its mass from $F = ma$ (where, F is force, m is mass of the object, and a is the acceleration). This allows us to create the setting where all objects are perceived as having the similar velocity which will be impacted differently according to the friction and mass through collisions. Each video in CRIPP-VQA is 5 seconds long, with a frame rate of $25fps$. We provide annotation and metadata for each video which contains object locations, velocities, orientation, and collision info at each frame. These annotations are further used to generate

the different types of question-answer pairs.

Objects and States. Table 3.1 summarizes the different properties in CRIPP-VQA. Each object o_i in the CRIPP dataset has four visible properties: a shape ($S \in \{cube, sphere\}$), color ($C \in \{olive, purple, teal\}$), texture ($T \in \{aluminum, cardboard\}$), and state ($T \in \{stationary, inmotion, undercollision\}$). Each object also has two invisible properties: mass ($m \in \{2, 14\}$) and coefficient of friction ($\mu \in \{0.25\}$). Three actions can be performed on each object – "remove", "replace", and "add".

This work focuses on mass and friction as intrinsic physical properties of objects. Other intrinsic properties such as bounciness or charge can be further introduced; however, this creates unnecessarily complicated scenarios which become hard to quantify during the evaluations. Each object having a unique combination of visible properties (i.e., {SHAPE, COLOR, TEXTURE}) has a pre-assigned value of mass; for instance, all teal aluminum cubes have mass 2. Note that these values are not provided as input to the VQA model and need to be inferred in order to perform counterfactual and planning tasks. In the training set and *i.i.d.* test set, the coefficient of friction for all objects with the surface is identical and non-zero.

Out-Of-Distribution properties. Previous studies mainly focus on visible properties based on OOD settings. This thesis proposes another dimension and seeks to observe the models' behaviors in OOD settings involving a change in physical properties. I consider four types of OOD scenarios: 1) *Mass*: where the mass of a few objects is changed to 8, 2) *Friction*: where the surface friction is changed to zero, 3) *Number of objects*: where two objects are moving instead of one when the scene is initialized, and 4) *Velocity*: initial object velocity is increased to 18 from 14. There can be an infinite amount of possible values for OOD physical scenarios. Doing so leads to a large number of variations within

the dataset, which becomes hard to quantify. Therefore, I chose this fix set of OOD values such that it leads to maximum deviations and at the same time the results are explainable.

Instance filtering. While creating each instance I randomly place N object within the defined bounds based on the camera position. Because of this randomness, there can be many reasons for the given set of collisions. This leads to hard-to-learn training/evaluation examples. For instance, if we "replace" an object with another object then the resulting collisions may or may not be because of the change in mass but it depends upon where each object is located. Although such instances are not useful, they are valid examples and should not be removed from the training/evaluation set. Therefore, I carefully design the instance filtering strategy, which leads to a high number of deviations and can be explained as mass as the major affecting factor. This pipeline can be summarized in the following steps:

- *Step 1:* Randomly initialize the video with objects and their properties to render the video. Record the video and generate annotations.
- *Step 2:* Filter the objects that were in collisions except for the first moving object.
- *Step 3:* Randomly select one object from the filtered set of objects and replace it with another object from a predefined list of objects. Render the video and record the annotations.
- *Step 4:* Change the mass of the newly introduced object and record the collisions.
- *Step 5:* If there are differences between collisions in "replace" based counterfactual settings (i.e., with different mass) that means that mass is the important factor for change in collisions.

- *Step 6:* If there are differences in collisions then store this data instance in the database else remove it and continue the loop.

Using this flow, the CRIPP-VQA dataset is created with a good balance of different types of scenarios. Only "replace" action-based counterfactual scenarios are considered during this filtering process because replace action directly leads to a change in intrinsic properties. While "remove" and "add" actions depend upon the properties of existing objects and their spatial location.

3.2 QA Generation

CRIPP dataset focuses on three categories of tasks: 1) Descriptive, 2) Counterfactual, and 3) Planning. Table 3.5 and 3.6 shows the examples of the questions asked in the CRIPP-VQA dataset. This study uses various annotations files generated using *TDW* to create the question-answer pairs.

Descriptive: These questions involve understanding the visual properties of the scene, including:

1. Counting the number of objects having a certain combination of visually seen properties,
2. Yes/No questions requiring object recognition
3. Finding the relationship between two objects under collision
4. Counting the number of collisions
5. Finding the maximum/minimum occurring object properties.

CRIPP-VQA does not include questions that require reasoning over mass, to avoid the introduction of spurious correlation which may influence counterfactual and planning-based questions.

Counterfactual. These questions focus on action-based reasoning (i.e., remove, replace, and add). I generate a hypothetical situation based on one of these actions, and the task is to predict which collisions may or may not happen if we perform the action on an object. To do this, I first note the differences in collisions between original and corresponding counterfactual scenarios. This gives us the list of new collisions and the list of collisions that didn't happen. They are used to create multiple-choice questions such that there are no visual biases, which can be learned. Therefore, bias experiments do not get more than 50% per-option accuracy (as shown in section 5.4).

Remove action focuses on a counterfactual scenario where a certain object is removed from the original video. **Replace** action focuses on a counterfactual scenario where one object is replaced with a different object. Replace action does not only change the object but it may also lead to a change in the hidden property. **Add** action-based questions focus on evaluating the system's understanding of spatial relationship along with the hidden property, where I create a new hypothetical condition by placing a new object to the $left/right/front/back$ at a fixed distance from the reference object.

Planning. CRIPP also contains planning-based questions, where the task is to perform an action on objects within the given video to either $make/stop$ collisions. Here, the system needs to predict which action needs to be performed and on which object, to achieve the goal. There can be so many solutions for a given scenario (if not infinite), which is a time-consuming process to get all of them. To avoid this, the proposed dataset uses existing counterfactual annotations to get the list of collisions that were new or didn't happen in a counterfactual setting. This allowed us to create balanced planning task questions. It is worth noting that during the evaluations, we need to perform the actions over the environment and render the video again to check whether the predicted action achieves the goal or not.

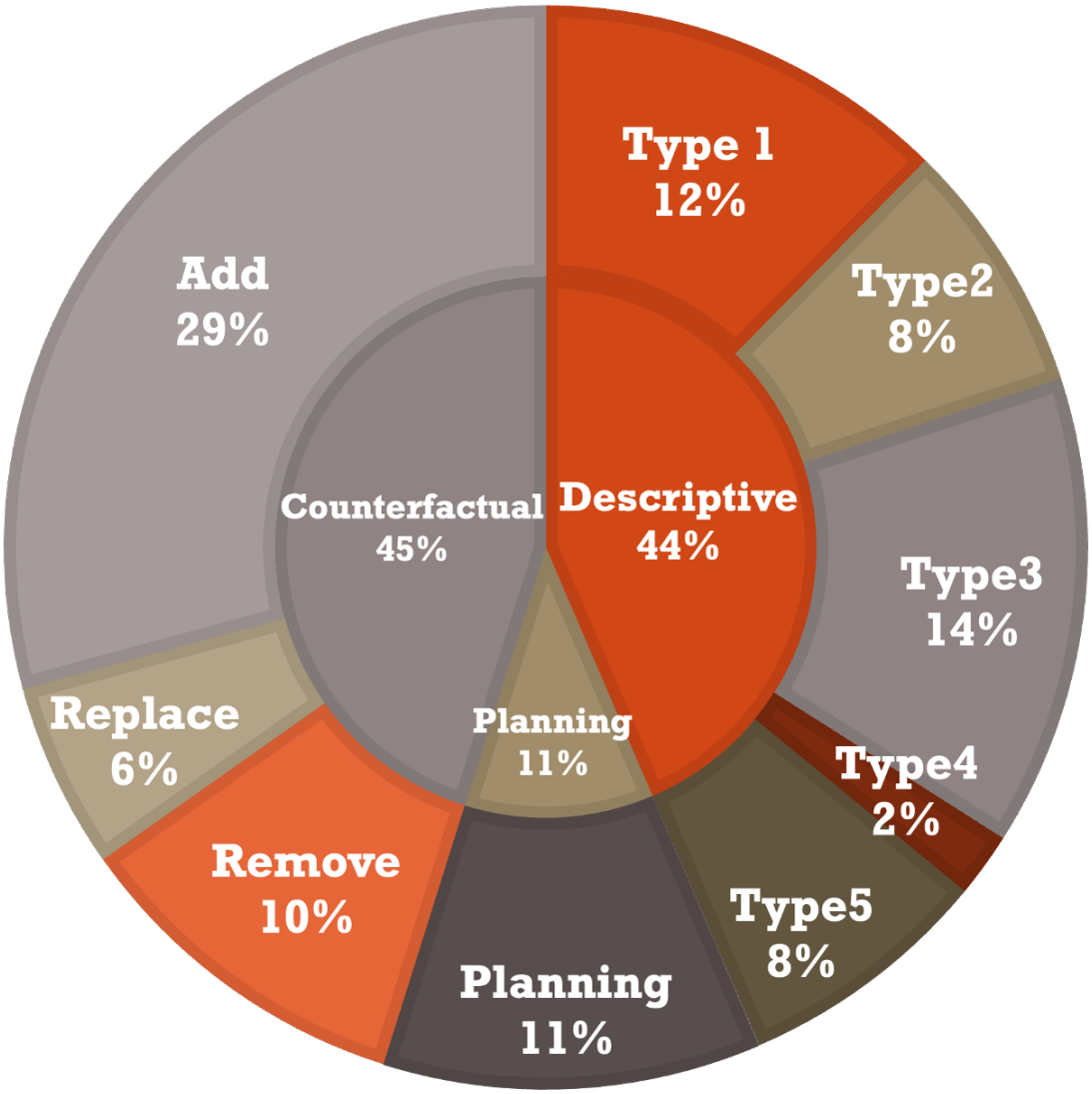

Figure 3.1: A Pie-chart Showing the Distribution of Various Question Types in the CRIPP-VQA Dataset. The Inner Pie Chart Shows the Three Broad Categories of Questions (Counterfactual, Descriptive, and Planning), While the Outer Pie-chat Shows a Fine-grained Categorization.

3.3 Dataset Statistics

CRIPP contains 4000, 500, and 500 videos for training, validation, and testing, respectively. Furthermore, it has about 2000 videos focused on evaluation for physical out-of-distribution scenarios. CRIPP training dataset has about 41761 descriptive questions, 41761 counterfactual questions (9603, 5142, and 27016 questions for remove, replace, and add actions, respectively), and 10440 planning-based questions. Figure 3.1 shows the percentages of each subcategory within the dataset. As described before, I used specific pipeline to filter the data. To generate the most useful data, it is important to cover different scenarios without ignoring anything. For example, counterfactual action should not always lead to different consequences then the original reference video. Because of the distance between two nearby objects even if we perform hypothetical actions, the set of collisions can be the same. These type of observations are not helpful during the training but they are valid. To get more insights, on how generated data look-like and how each physical properties are important and can be learned with given generated dataset, I calculate various statistics:

- Mass of the objects is randomly assigned using a random operator. This allows us to assign weights that are not highly correlated by color/shape/material. Table (3.2) shows the list of objects with the corresponding physical properties.

- From data annotations, the observations can be made that there is a high number of the first collision between two different objects, which itself should be sufficient to learn the mass distribution.

- Table (3.3) shows the average number of collisions when one type (in terms of mass) of objects collides with another. Here, it can be inferred that when the same mass objects collide, the avg. a number of collisions are close. While, when a lighter and heavier object collides this number either increases or decreases. This suggests that

Object number	Shape	Color	Material	Mass
1	Cube	Olive	Aluminium	2
2	Cube	Teal	Aluminium	2
3	Cube	Purple	Aluminium	14
4	Cube	Olive	Cardboard	14
5	Cube	Teal	Cardboard	14
6	Cube	Purple	Cardboard	2
7	Sphere	Olive	Aluminium	2
8	Sphere	Teal	Aluminium	14
9	Sphere	Purple	Aluminium	2
10	Sphere	Olive	Cardboard	2
11	Sphere	Teal	Cardboard	14
12	Sphere	Purple	Cardboard	14

Table 3.2: Average Number of Collisions When One Type of Object Collides with Another in Terms of Mass.

mass impacts the number of collisions in the video.

- Furthermore, Table (3.4) shows that the average number of collisions in different counterfactual settings is different. This creates another challenging task to apply physical reasoning where the effect of hypothetical action and mass is important.

To summarize, the above statistics suggest that the CRIPP-VQA dataset does not contain any visual cues. At the same time, it covers the different types of collisions, and learning physical properties is essential to do counterfactual/planning-based reasoning.

First collision type	**Light → Light**	**Heavy → Heavy**	**Light → Heavy**	**Heavy → Light**
Avg. Number of collisions	3.12	3.23	1.78	4.03

Table 3.3: Average Number of Collisions When One Type of Object Collides with Another in Terms of Mass.

	# of moving objects	**Vanilla # of collisions**	**Remove # of collisions**	**Replace # of collisions**	**Add # of collisions**
CLEVRER	2.34	2.46	N/A	-	-
CRIPP-VQA	1	3	2.06	3.31	4.15

Table 3.4: General Statistics Comparison Between CLEVRER and CRIPP-VQA. Here, N/A Shows That Annotations Are Missing to Derive the Number, and − Represents That Action Is Not Present in the Dataset.

Question Type	Examples
Descriptive - Type 1	How many teal cardboard cube objects are there ? How many cardboard sphere objects are static when video ends ?
Descriptive - Type 2	Do teal cardboard cube objects exist in the video ? Do purple aluminium cube objects exist in the video ?
Descriptive - Type 3	What is the color of the collidee of purple aluminium cube in collision number 1? What is the material of the collider of purple cardboard cube in collision number 2?
Descriptive - Type 4	How many collisions are there between teal sphere objects and teal aluminium objects ? How many collisions are there between purple cardboard cube objects and teal aluminium cube objects ?
Descriptive - Type 5	What is the maximum occurring shape of objects in the video ? What is the minimum occurring material of objects in the video ?

Table 3.5: Descriptive Question Examples of the CRIPP-VQA Dataset, Asked from Different Types of Question Categories as Shown in Figure 3.1.

Question Type	Examples
Counterfactual - Remove	What will happen, if the teal cardboard sphere is removed ? Choice: purple cardboard sphere would collide with purple cardboard cube Choice: teal cardboard cube would collide with purple cardboard cube
Counterfactual - Replace	What will happen, if the purple cardboard sphere is replaced by the purple aluminium sphere? Choice: purple aluminium sphere would collide with olive aluminium sphere Choice: teal cardboard sphere would collide with purple aluminium sphere
Counterfactual - Add	What will happen, if the purple cardboard sphere is added to the right of teal aluminium sphere? Choice: teal aluminium sphere would collide with purple cardboard cube Choice: olive aluminium cube would collide with teal aluminium sphere
Planning	Make the collision between olive cardboard cube and olive aluminium sphere. Make the collision between teal cardboard sphere and olive cardboard sphere .

Table 3.6: Counterfactual and Planning Task Examples from the CRIPP-VQA Dataset.

Chapter 4

MODELING STRATEGIES

There are several methodologies proposed for Visual Question Answering. This chapter describes the three state-of-the-art deep learning-based VideoQA systems and how I modified them for the CRIPP-VQA dataset. Apart from these deep learning models, Neuro-Symbolic systems are also being studied for similar tasks. Neuro-Symbolic approaches assume that some prior knowledge about the environment is already known. In the case of the CRIPP-VQA tasks, it proposes the challenge of learning visually hidden properties without external knowledge. Therefore, I select the following three approaches: 1) Memory, Attention, and Composition (MAC) (Hudson and Manning, 2018), 2) Hierarchical Conditional Relation Network (HCRN) (Le *et al.*, 2020), and 3) Attention over learned embeddings (Aloe) (**?**).

4.1 Memory, Attention, and Composition (MAC)

(Hudson and Manning, 2018) proposed MAC for compositional visual question answering task. MAC network consists of several MAC cells focusing on basic reasoning steps. MAC cell allows the system to decompose the input into sequences of attention, which can be used to perform the reasoning without the need for strong supervision. Unlike LSTMs, each MAC cells are independent of the other (in terms of weights sharing) and they rely on previous cells' output to do the reasoning. Figure (4.1) shows the example of the MAC network and what each MAC cell looks like. MAC cell contains three submodules: 1) control unit, 2) read unit, and 3) write unit. The Control unit takes the textual input and previous step control unit output to update the state. The read unit takes the image as input and memory state to extract the important features, which is controlled by the control unit.

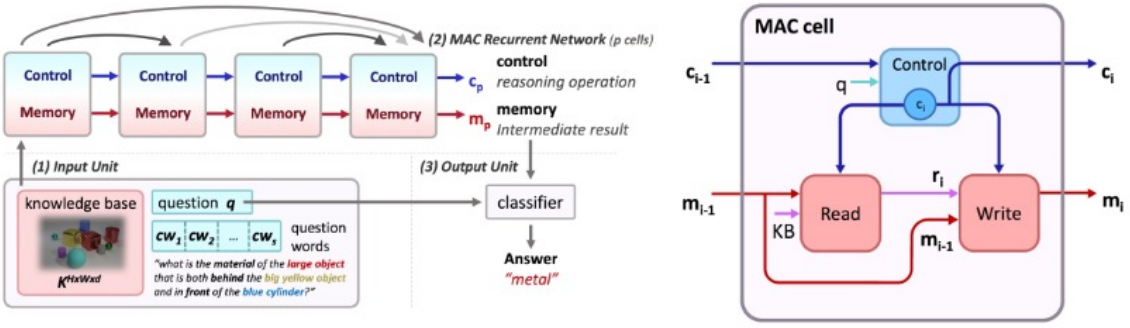

Figure 4.1: Mac Network Model Architecture with MAC Cell Network From (Hudson and Manning, 2018).

Write unite processes the read unit output and previous memory state to update the existing memory state, which is again maintained by the control unit.

Experiments on CLEVR show that MAC can achieve 98.9% accuracy. Furthermore, it has been observed that based on the input question, MAC attentions can successfully identify the relevant objects within the image (Hudson and Manning, 2018). As MAC has shown its potential for compositional reasoning, I adapt it for benchmarking the CRIPP-VQA dataset. The original MAC network only takes the single image feature vector extracted using the pre-trained classifier. I modify the image feature extractor to adapt the image-frame sequences from a video. To do this, we can first get features for each frame in $(batch\_size, frames, channels, h, w)$ dimensions and convert it to $(batch\_size, frames, h, w)$ by taking a mean of channel-wise features for each frame. The rest of the network is kept unchanged during the experiments

4.2 Hierarchical Conditional Relation Network (HCRN)

I use HCRN as another baseline to benchmark the CRIPP-VQA dataset. HCRN was designed specifically for Spatio-temporal video question answering by (Le *et al.*, 2020). The hierarchy in the HCRN model is divided into two steps: 1) Clip-level, and 2) Video-level.

Each video is divided into many clips with some overlap frames. First, HCRN performs this clip-level modeling as a part of the first layer. Second, output feature representation from the clip level is used as input to the second layer (i.e., video level). This hierarchy style modeling allows the HCRN to attend at a different granular level and achieve better performance.

HCRN extracts two different feature representations from the set of frames from the video/clip: 1) Frame-specific features using pre-trained ResNet, and 2) Video/Clip motion features extracted from pre-trained 3D CNN. Textual inputs are encoded using pre-trained context-free GloVe embeddings. As HCRN is designed for spatial-temporal reasoning, I adapt the model as it is without changes [1].

4.3 Attention Over Learned Embeddings (Aloe)

Aloe (**?**) is one of the best-performing models on the CLEVRER (Yi *et al.*, 2020) benchmark. It is a transformer-based model, designed for object trajectory-based complex reasoning over synthetic datasets. Figure (4.2) summarizes the Aloe model from the (**?**). First, Aloe uses MONet (Burgess *et al.*, 2019) for obtaining object features. MONet extracts these features by performing an unsupervised decomposition of each frame into observed objects. Like, other vision-language-based models, Aloe takes the object-specific features from the images (in this case, MONet features) as a visual representation of the video. Furthermore, Aloe trains the text embedding lookup from the scratch. Aloe takes these frame-wise object features to predict the answers to the input question, using the *[CLS]* token and self-supervised training strategy. (**?**) explores several different self-supervised learning strategies, where per-frame object mask prediction seemed to be more useful during experimentation.

[1] `https://github.com/thaolmk54/hcrn-videoqa`

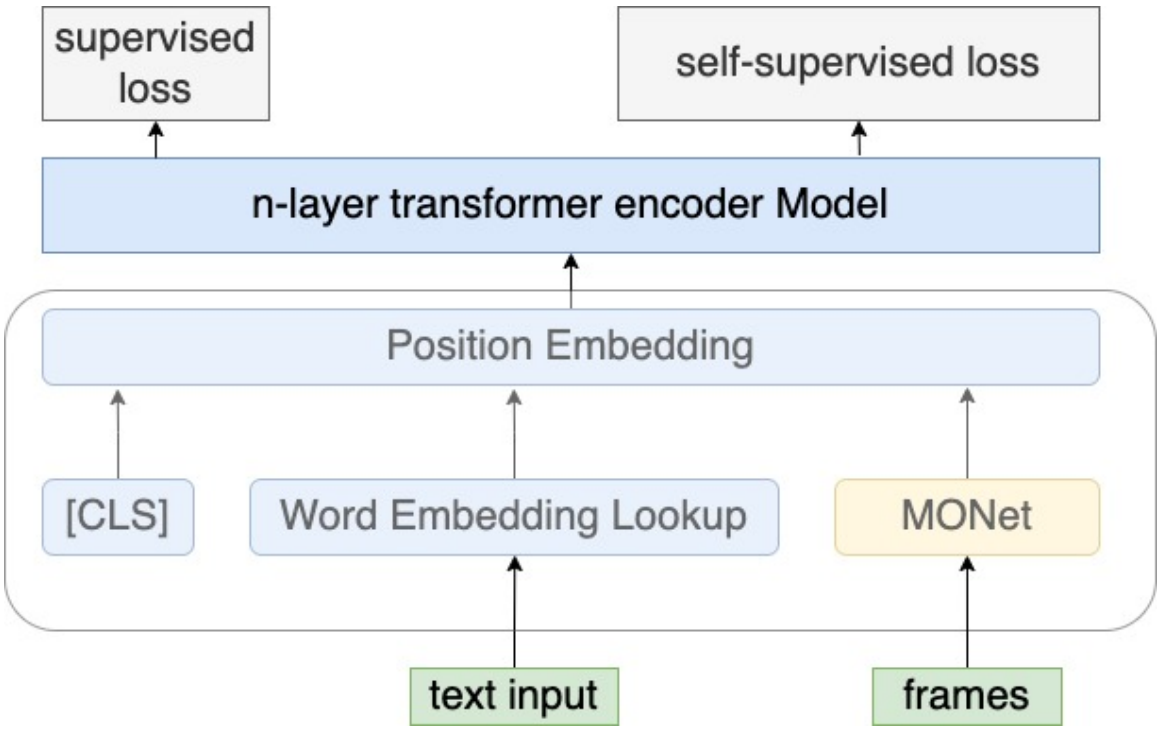

Figure 4.2: Baseline Aloe Model Architecture from (Ding *et al.*, 2021). Here, MONet Is Pre-trained Separately on the given Training Dataset. The Rest of the Modules Are Trained from Scratch.

4.3.1 Drawbacks of Aloe

From experiments, I find that the MONet module used in Aloe is very unstable and fails to produce reliable frame-wise features on complex visuals from CRIPP. MONet is not able to recognize object properties such as color and is not able to decompose the image into several images containing individual objects. It was observed that MONet-based unsupervised object decomposition results in failure on complex realistic visuals and it is hard to guarantee that it will decompose each object on independent images/features. In Figure (**??**), I show three failure cases of MONet on the CRIPP-VQA dataset. Here, we can observe that MONet is not only able to decompose the objects independently, but it is also not able to learn the color of the objects. That said, at least MONet can learn the texture

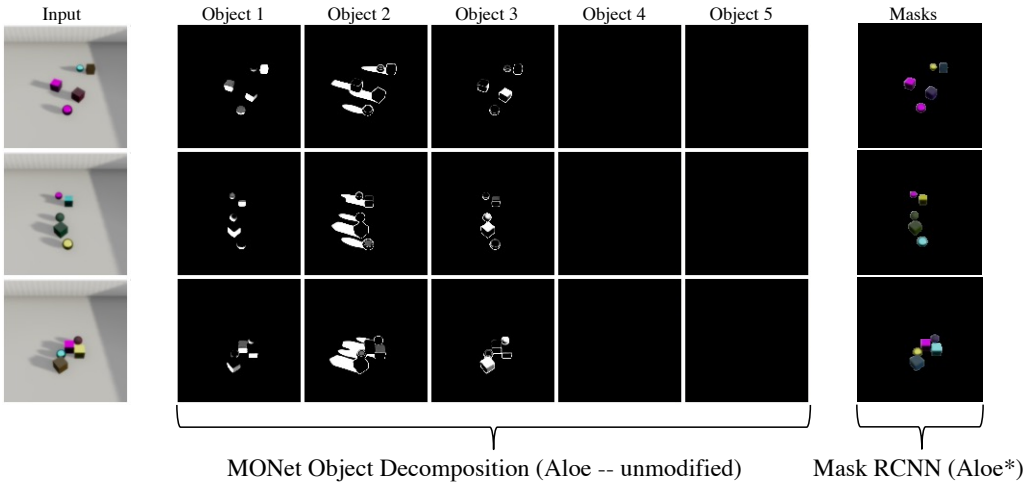

Figure 4.3: Illustration of the Failure of MONet (the Object Decomposition Module in Aloe (Ding *et al.*, 2021)) on CRIPP-VQA Videos. The Intended Functionality of MONet Is to Decompose Individual Objects into Separate Masks. However, as Shown above, the Predicted Masks Contain Areas Corresponding to More than One Object. We Modified Aloe by Replacing Monet with Mask-RCNN, and This Approach (Aloe*) Leads to More Reliable Object Detection Which Can Be Used by the Downstream Question-answering Module.

(i.e., metal or cardboard). As a result, we can see that the re-generated images lack greatly in terms the important features. However, the same MONet model achieves remarkably good results. Since the CLEVR was proposed, many algorithms and SotA challenges were derived from similar visuals setup. This leads to temporal overfitting on visual biases on proposed methodologies.

4.3.2 Aloe* (Modified Aloe)

As MONet fails measurably on CRIPP-VQA visuals, the Aloe baseline also exhibits close-to-random performance because of the lack of information. I propose additional modifications to Aloe to make it more widely applicable beyond prior datasets that are

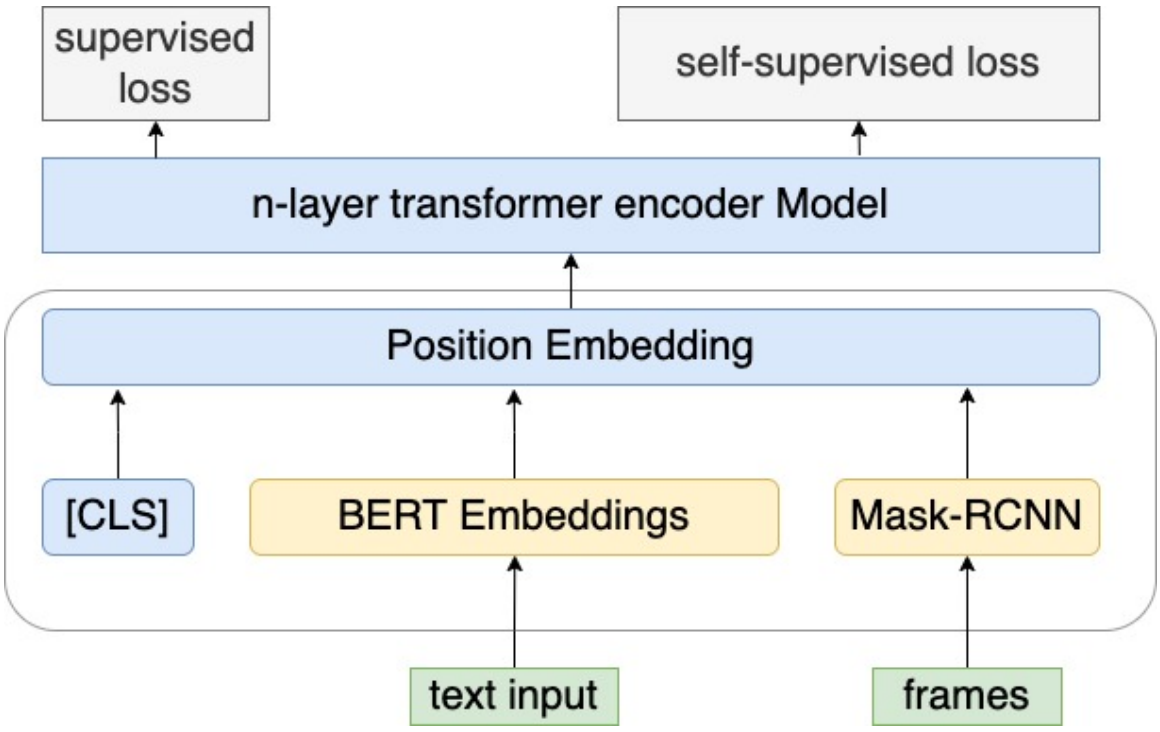

Figure 4.4: Aloe*+BERT Model Architecture. Here, Mask-RCNN and BERT Models Are Pre-trained for Instance Segmentation and Masked-Language-Modeling, Respectively. These Two Modules Are Kept Frozen During the Training and the Rest Are Trained from the Scratch.

built using the CLEVR (Johnson *et al.*, 2016) rendering pipeline. I replace unsupervised MONet with supervised Mask-RCNN object detector (He *et al.*, 2017) to perform instance segmentation. Then I train an auto-encoder to compress the mask-based object-specific features to make it compatible with the Aloe feature requirement. I call this version of Aloe with Mask-RCNN object detector Aloe*.

With this change, it was observed that Aloe* improves the performance over Aloe. However, in my experiments on CRIPP-VQA, I learn that instead of learning the word embedding from the scratch, given that the text embeddings are extracted from a pre-trained BERT model. Therefore, I further modify the Aloe* to use BERT-based word embeddings

as an alternative to learn embedding lookup from scratch. Aloe*+BERT leads to faster and stable convergence on CRIPP-VQA. Figure (4.4) shows the model diagram of the Aloe*+BERT model.

Chapter 5

EXPERIMENTS AND RESULTS

5.1 Problem Statement

Given an input video (v), and a question (q) the task is to predict the answer (a). Each video v contains the m number of objects randomly selected from the set $O = \{o_1, o_2, ..., o_n\}$. Here, object o_i has several associated properties (i.e., $o_i = (m_i, c_i, s_i, t_i, l_i, v_i)$), where color ($c_i$), shape ($s_i$), texture ($t_i$), location ($l_i$), and velocity ($v_i$) are visually observable properties alongside with mass (m_i) as hidden property. More formally, we need to learn the probability density function p such that we maximize the $p(a|v, q)$.

Evaluation Metrics. Accuracy is adapted as an evaluation metric for different categories of QAs. In the case of the descriptive question, we simply need to check whether the predicted answer is in the set of correct answers. To evaluate the models on counterfactual questions, this study uses two accuracy metrics – per-option (PO) and per-question (PQ) accuracy. Here, each counterfactual questions have multiple options describing the collisions. Therefore, per-option accuracy refers to the option-wise performance and per-question accuracy considers whether all options of correctly predicted or not. Each planning task involves performing an action over objects within a video. Because of that, to achieve the given goal there can be multiple possible solutions, which are hard to predict. Therefore, *TDW* is used to re-simulate the models' predictions on the original video to check whether the given planning goal is achieved or not. This creates a new evaluation pipeline for planning-based questions for interactive evaluations.

Model	Descriptive	Remove		Replace		Add		Planning
		PQ	PO	PQ	PO	PQ	PO	
Frequency	8.21	0.00	50.18	0.00	50.00	0.00	50.00	3.49
Random	8.51	7.21	49.58	3.34	49.40	9.39	50.04	7.39
Blind-BERT	53.82	20.18	54.67	17.57	50.45	15.86	51.55	8.11
MAC	48.72	16.41	50.68	17.31	50.21	16.29	49.83	6.26
HCRN	64.98	27.20	59.04	19.87	55.97	20.49	56.06	21.38
Aloe*	68.94	31.10	62.90	9.91	52.10	18.13	56.55	31.76
Aloe*+BERT	71.04	33.64	65.46	22.07	56.76	39.71	67.43	32.61

Table 5.1: Results on the *i.i.d.* Test Set Showing Performance of Models Evaluated in Terms of Per-Question (PQ) Accuracy and Per-Option (PO) Accuracy. For Descriptive and Planning Questions, Only One of the Answer Options Are True, Therefore Per-Question and Per-Option Accuracies Are Identical. Here, Both Aloe Variants Are Modified Version over Aloe Baseline.

5.2 Benchmark Model Details

As described in the previous chapter, for this study, I consider three different deep learning-based state-of-the-art models for the video question answering task: 1) MAC (Hudson and Manning, 2018), 2) Hierarchical Conditional Relation Network (HCRN) (Le *et al.*, 2020), and 3) Attention over learned embeddings (Aloe) (**?**).

In addition to these strong baselines, it is also important to consider a *"random"* baseline which randomly selects one answer from a possible set of answers, and a *"frequent"* baseline which always predicts the most frequent label. To analyze textual biases, a text-only QA model *"Blind-BERT"* is used. Blind-BERT is a pre-trained language model (BERT (Devlin *et al.*, 2019)) which takes only questions as input to predict the answer

and ignores the visual input.

5.3 Experimental Setup

I follow the training guidelines provided by the authors of each baseline study. All systems are trained on Quadro RTX 8000 GPUs. Each model is trained with a maximum of 200 epochs. And I select the best model based on average performance accuracy on the validation set. I follow the below instructions to support each model which are MAC, HCRN, Aloe*, and Aloe*+BERT. Moreover, for planning based task, we need to add extra four classifier heads on top of all models which predicts: 1) the type of the action, 2) an object on which action needs to be performed, 3) an object which needs to be added through replace or add action, and 4) relative direction of the object if we are adding a new object.

MAC: I modify the public implementation of MAC from `https://github.com/rosinality/mac-network-pytorch` to adapt the video frames as input. First, it is important to resize the each 125 frames leading $(125, 3, 224, 224)$ video dimension. Later, ResNet101 is used to extract the features $(125, 512, 14, 14)$. After taking the channel-wise mean of features, this leads to the final video re-presentation of $(125, 14, 14)$ dimension matrix supportable for the rest of the pipeline. I also do the necessary changes described for the planning task as well.

HCRN: As HCRN is the VideoQA model and official implementation is available at: `https://github.com/thaolmk54/hcrn-videoqa`, I use the source code as it is. Except again additional classifier heads are introduced to do the planning based tasks.

Aloe*/Aloe*+BERT: I first reproduce the Aloe on PyTorch based on the architecture details from the research paper by Ding et. al. (**?**) and their public available demo at `https://github.com/deepmind/deepmind-research/tree/master/object_attention_`

Hyper-parameter	Value
# of layers	28
# of attention heads	128
embedding size	768
visual feature size	512
text embedding size	768
Batch Size for descriptive	96
Batch Size for Counterfactual	32
Batch Size for Planning	16
Learning rate	0.00005
Optimizer	RAdam

Table 5.2: Aloe*+BERT Architecture and Hyper-Parameter Details.

`for_reasoning`. Moreover, the code base from transformers[1] library (as it is well tested and used across the industry and academia) is used and modified to support the VideoQA in the same way as Aloe does. My initial experiments on CLEVRER showed that Aloe cannot reproduce the results on CLEVRER with the specified set of architecture details and hyper-parameters from the original paper. Therefore, I do extensive experiments on Aloe architecture and hyper-parameter search to reproduce similar results. After achieving a similar performance from the paper, this new reproducible Aloe architecture is used in experiments. Table (5.2) shows the hyper-parameter details to reproduce the results. Moreover, the Aloe* source code from experiments is available at `https://github.com/Maitreyapatel/CRIPP-VQA/`.

[1] `https://github.com/huggingface/transformers`

5.4 Results

Table 5.1 summarizes the performance comparisons of baselines on the CRIPP-VQA *i.i.d.* test set. On **Descriptive** questions, the *"random"* and *"frequent"* baselines achieve around only 8% accuracy, while Blind-BERT gets 53.82% which suggests the existence of language bias associated with correlations between question types and most likely answers for each. Surprisingly, MAC achieves only 48.72% which is lower than Blind-BERT. This implies that the video feature representations learned by MAC hurt performance compared to text-only features. HCRN, and both Aloe variants improve performance indicating that visual features are crucial for descriptive questions. Aloe*+BERT is the best performing model which implies that proposed modifications helps to improve the performance.

Counterfactual questions involve a total of three types of actions. Table (5.1) shows the action-wise performances. The performance of MAC is again close to Blind-BERT. HCRN performs slightly better than Blind-BERT. This shows that even though visual features in HCRN are better than the MAC but it is not sufficient enough to do such complex reasoning. Aloe*+BERT achieves much better results only in terms of remove and add actions. However, Aloe*+BERT is close to random for questions with the *"replace"* action as it directly involves the change in physical properties (i.e., mass and shape) of an existing object within the given scenario. This implies that Aloe*+BERT is able to do spatial reasoning to some extent, but is not good at reasoning about changes in physical properties. While it can also be seen that Aloe*+BERT outperforms the Aloe across the actions, this implies that BERT-based embedding helps the model to learn the relation between the objects and action.

Planning task can have more than one possible answer and it is created from the counterfactual reasoning tasks. Therefore, we can observe a similar trend in results, and Aloe*+BERT

Model	**Descriptive**	**Remove – PO**	**Replace – PO**	**Add – PO**	**Planning**
Aloe*+BERT	71.04%	65.46%	56.76%	67.43%	32.61%
Human Evaluations	90.00%	86.67%	73.33%	76.67%	58.87%

Table 5.3: Comparison of Aloe*+BERT with Human Evaluations. Results Show That There Is a Huge Gap Compared to the Human Evaluations.

performs better than the other baselines. Further analysis on Aloe*+BERT predictions shows that model predicts "remove", "replace", and "add" actions for planning tasks with 70.52%, 10.6%, and 18.87%, respectively. This also suggests that the model finds it easy to reason when "remove" hypothetical action is present.

Human evaluations: To learn more about the upper bound of the CRIPP-VQA dataset and what to expect from the different proposed systems, I perform human evaluations. There was a total of 6 people participated as volunteers. All were given 5 videos and corresponding QA pairs to get habituated with the environment. Then they are asked to answer total of 30 questions. As shown in Table 5.3, Human evaluations achieved 90.00%, 78.89%, and 58.87% on descriptive, counterfactual, and planning tasks, respectively.

5.5 Physical out-of-distribution Experiments

Most of the previous studies focus on feature-based OOD cases (like the rotation of the entities within the image).

Figures (5.1, 5.2, 5.3, and 5.4) shows the comparison of VideoQA models on *i.i.d.* and different OOD scenarios for remove, replace, and add action, and planning questions, respectively. From Figure 5.1, it can be seen that in the case of the OOD settings the model performance becomes close to random which is around 50%, for all models. This

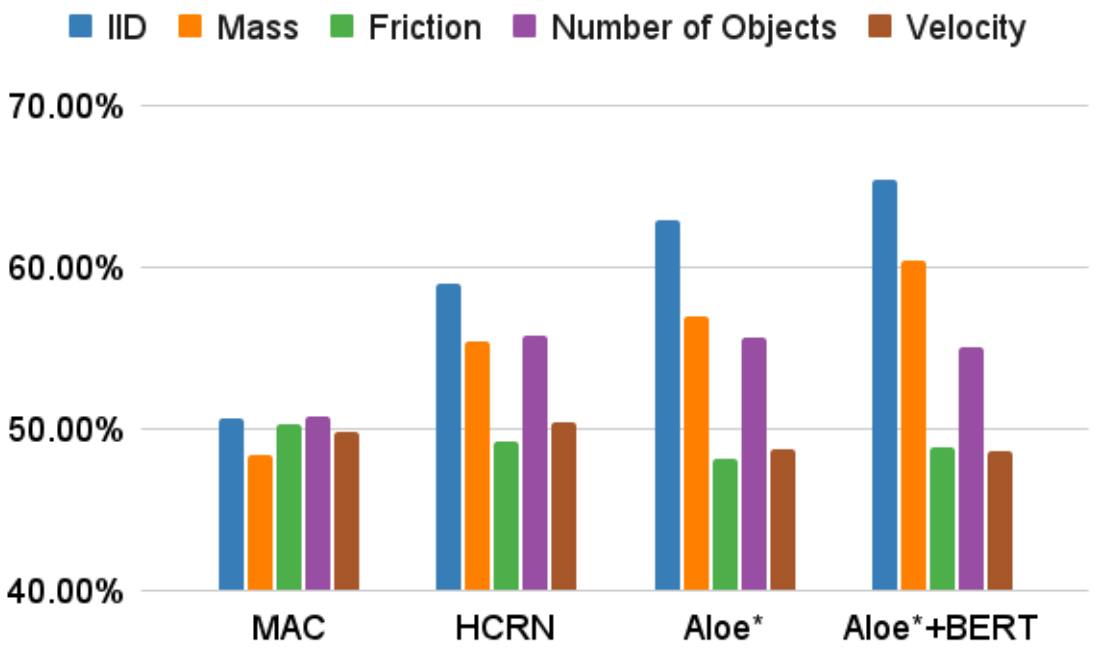

Figure 5.1: Comparison of Performance of Models (Per-option Accuracy) for "remove" Questions When Tested Using the *i.i.d.* Test Set and Each OOD Test Set.

suggests that models are very sensitive to such small physical perturbations, especially for the *"remove"* action. Furthermore, Figure 5.2 shows that all models perform close to random for the replace action. From Figure 5.3, we can observe that the performance drop is negligible across the OOD sets for the add action, especially for Aloe*+BERT. The reason behind this is that Aloe*+BERT is not able to predict the actual set of collisions but it can learn that based on the direction of the new object which collisions won't happen (more details are in the next section). Moreover, Figure 5.4 shows that the performance increases on several OOD scenarios for planning task. In case of the remove action, Friction and Velocity OOD settings is the hardest for models to perform. While, for replace action, Number of Objects OOD setting is the hardest but Aloe*+BERT improves the performance on Velocity. Number of Objects – OOD setting is also tough for models to understand for

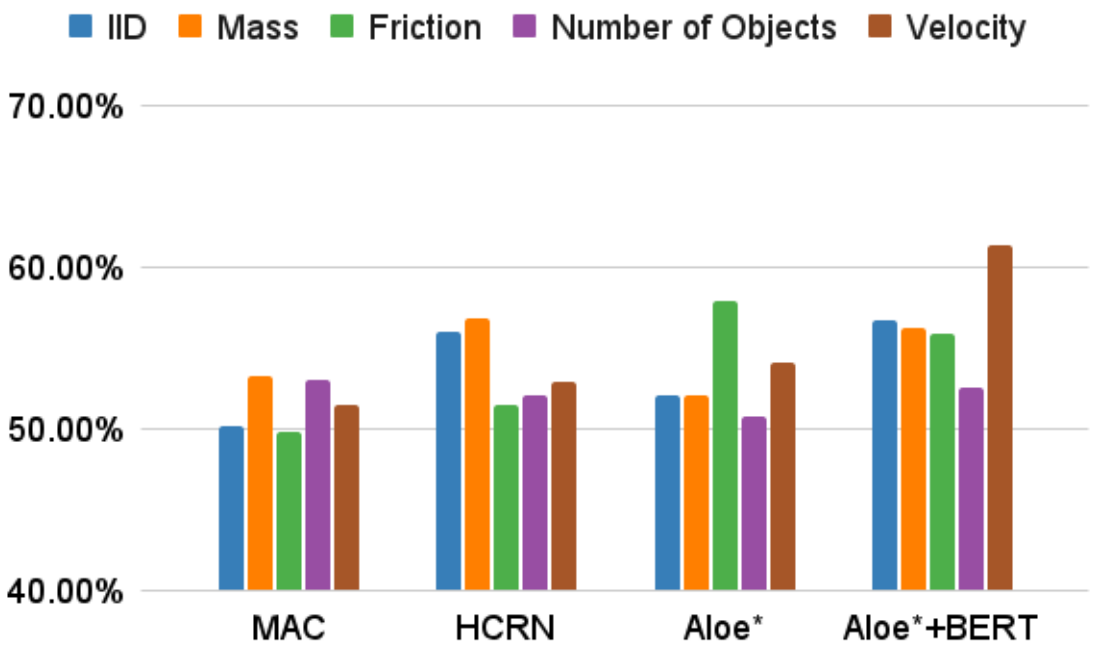

Figure 5.2: Comparison of Performance of Models (Per-option Accuracy) for "replace" Questions When Tested Using the *i.i.d.* Test Set and Each OOD Test Set.

add action based questions.

Descriptive questions are based on the observable reference video. In ideal scenario, irrespective of the OOD tasks the performance on OOD descriptive questions should be similar to the *i.i.d.* setting. Table 5.4 shows the performance of Aloe*+BERT model on descriptive evaluation set from different settings. Here, it can be observed that performance of Aloe*+BERT is consistent across the evaluation sets. Except for the "Number of Objects" OOD setting where there are two moving objects instead of only one moving object, leading to slight drop in performance.

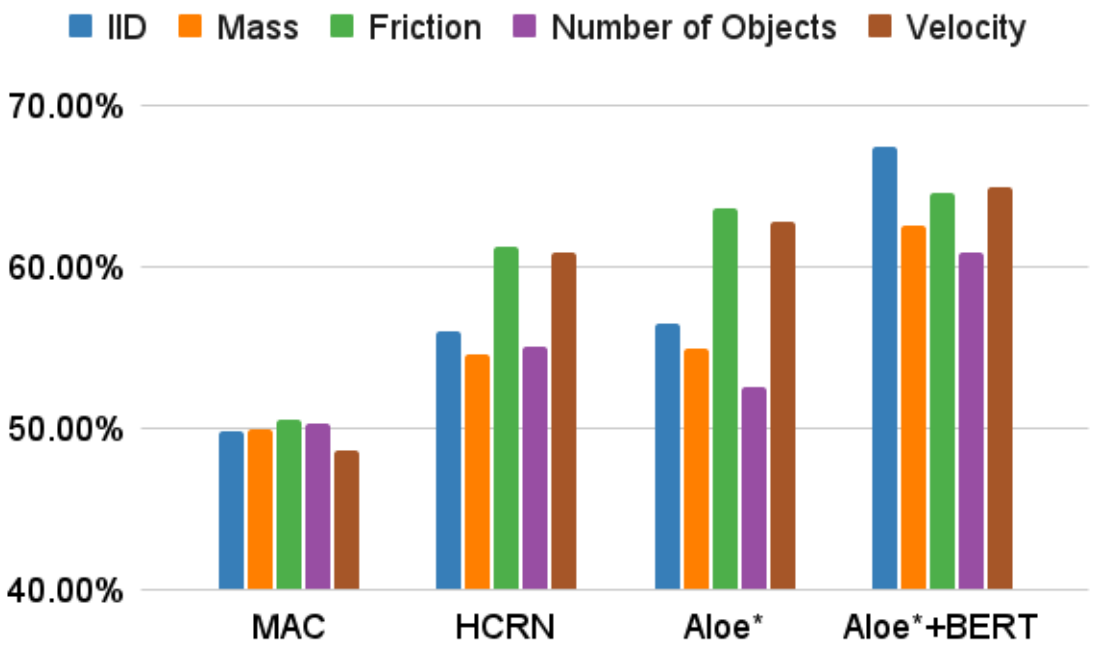

Figure 5.3: Comparison of Performance of Models (Per-option Accuracy) for "add" Questions When Tested Using the *i.i.d.* Test Set and Each OOD Test Set.

	i.i.d.	**Mass**	**Friction**	**Number of Objects**	**Velocity**
Aloe*+BERT	71.04	70.85	70.62	66.57	71.16

Table 5.4: Accuracy of Aloe*+BERT on Descriptive Questions from Different (*i.i.d.* and OOD) Evaluations Sets.

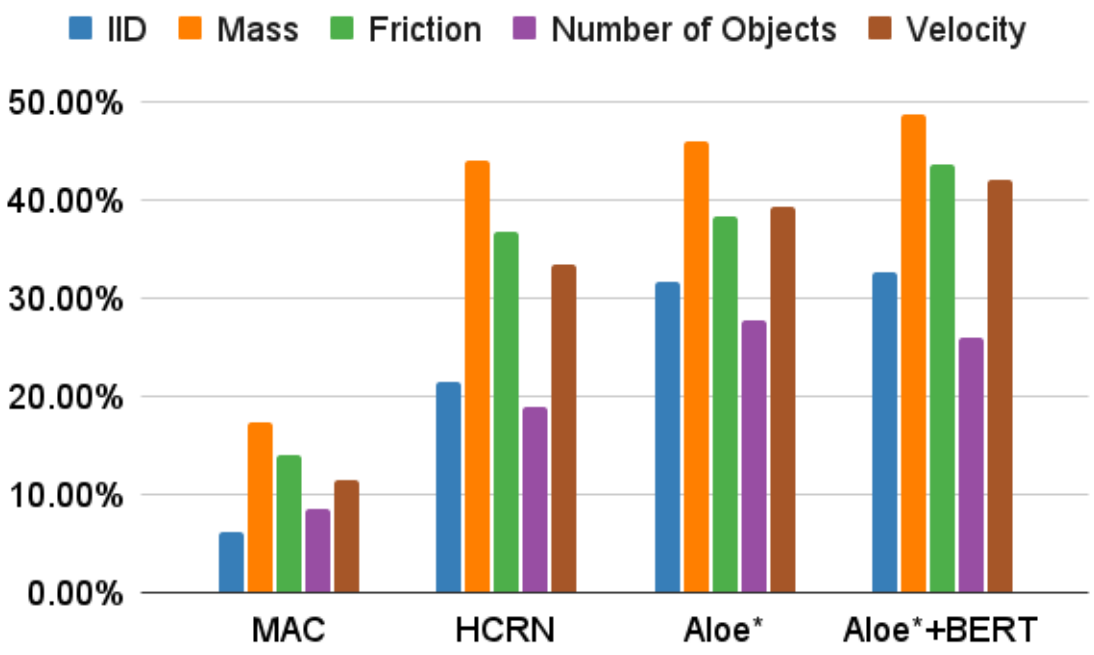

Figure 5.4: Comparison of Performance of Models "planning" Questions When Tested Using the *i.i.d.* Test Set and Each OOD Test Set.

Chapter 6

ANALYSIS AND DISCUSSION

This section raises several important questions and derive the insights accordingly.

Performance for true vs. false collision detection Consider the scenario with three objects (A,B,C), where only A collides with B. In this case, the collisions are categorized between A & B as the actual collision (i.e., prediction label *true*), and collisions are categorized between B & C and A & C as an absent collision (i.e., prediction label *false*). I further check the performance of detecting all occurring collisions and the collisions that never happened. For this analysis, all occurring collisions are selected if they are captured in annotation files. Table 6.1 shows the action-based performance of Aloe*+BERT on these two categories. It can be inferred that detecting the actual set of collisions is easy in the case of the *"remove"* action but we can observe the opposite results for *"add"* action. This reveals that Aloe*+BERT is learning the object trajectories as it can detect collisions correctly for removing action but is not able to perform spatial reasoning. However, in the case of the replace action, the model is failing in both categories. This implies that it is hard for existing models to learn the concept of mass in an implicit setting.

Performance for First Collision vs Subsequent Collisions. In the CRIPP-VQA dataset, a collision between a pair of objects may lead to subsequent collisions between other objects. We analyze the performance of the best model (Aloe*+BERT) on counterfactual questions, by comparing the accuracy on questions about the first collision, with the accuracy on questions about subsequent collisions. To correctly predict subsequent collisions, models need to understand the mass of the objects involved in the first collision to learn the consequences (i.e., sequence of future events). From Table (6.2), we can observe that

Action	Present collisions	Absent collisions
Remove	78.27	52.81
Replace	65.74	60.23
Add	46.41	79.47

Table 6.1: Per-option Accuracy of Aloe*+BERT for Detecting Present Vs.Absent Collisions Correctly.

for all three actions, there is a drop in performance on subsequent collisions; the drop is highest (28.48%) for *"remove"* action.

Importance of mass as intrinsic property. There are many hidden factors (i.e., mass, friction, object shape, velocity) that play roles in object trajectories and collisions in any scenario. Therefore, to understand the some dynamics, I analyze the number of collisions in different counterfactual scenarios and collisions between two different types (in terms of mass) of objects. Table 6.3 shows that if first collision is between either two light or two heavy objects then it leads to almost same number of collisions. However, if first collision is between light and heavy objects then the number of collisions either decreases or increases based on the scenarios. Analysis on the number of collisions in different counterfactual settings shows that there are on an average 3.0, 2.06, 3.31, and 4.15 collisions in vanilla, "remove", "replace", and "add" counterfactual settings, respectively.

To summarize, these analyses suggests that each counterfactual scenarios are unique and contains different challenges. Furthermore, these strengths the argument that models fail to learn various reasoning capabilities including but not limited to intrinsic physical properties, and consequences of the actions.

Action	First Collision	Subsequent collisions	Difference
Remove	90.52	62.45	28.07
Replace	75.38	66.03	9.35
Add	55.45	41.01	14.44

Table 6.2: Per-option Accuracy of Aloe*+BERT for Detecting First Collision Vs. Subsequent Collisions from the Set of Occurring Collisions in Counterfactual Scenario.

First collision type	L → L	H → H	L → H	H → L
	3.12	3.23	1.78	4.03

Table 6.3: Average Number of Collisions in Ground Truth Videos (i.e., Vanilla) When Different Types of Objects Participate in First Collision. "$x \rightarrow y''$, Where $x, y \in \{Light, Heavy\}$, Means That x Mass Object Collides with y Mass Object. Moreover, H: Heavy Object and L: Light Object.

Chapter 7

CONCLUSION AND FUTURE WORK

This chapter summarizes the thesis, the limitations of the study and shows the potential future research directions.

7.1 Summary

The current imaging pipeline cannot determine the visually hidden properties (such as mass, and friction). However, these properties can be identified by using visual cues from the video (like, collisions and change in velocity). Humans do not require one-to-one mapping to estimate such properties of the surroundings. We infer these properties in implicit manner just by observations. Therefore, in this thesis, I propose CRIPP-VQA to benchmark the state-of-the-art models' ability to learn such properties in implicit settings.

CRIPP-VQA contains the three types of tasks: 1) Descriptive QA, 2) Counterfactual QA, and 3) Planning. Descriptive question is about the visually seen properties of the given video. Counterfactual QA is about an hypothetical scenario where we perform remove/replace/add action on given reference video. Intrinsic physical properties play a big roles in such counterfactual scenarios. Hence, Counterfactual QAs gives the additional cues about hidden properties. Planning tasks are completely opposite of the counterfactuals. In this case, model needs to perform an action which achieves the given goal to either make or stop the collisions between two objects.

Extensive experiments shows that state-of-the-art models struggle to achieve the human-like performance. Especially, in case of the "replace" action all models hardly improve upon the random baseline. Further analysis suggests that visuals are key component and unsupervised object decomposition method (i.e., MONet) fails to decompose on CRIPP-

VQA visuals, suggesting that previous SotAs are only evaluated on some specific visuals leading to overfitting on some visuals properties. I further extend the studies on physical out-of-distribution, where I vary the different physical properties one at a time and observe the behaviour of the trained models on *i.i.d.* set. Detailed analysis on various kinds of collision predictions showed that Aloe*+BERT can predict the first collision in counterfactual scenarios correctly but it cannot predict the subsequent collision.

7.2 Limitations and Future Work

This study has several limitations and future work is needed to improve the current state of the AI/ML algorithms.

- This thesis focuses on discrete intrinsic physical properties which leads to diverse set of consequences. In future, it would be important to extend this work to continuous variables to make the system more generalize.
- CRIPP-VQA contains fix set of visual properties leading to only handfull of objects. In future, it would be important to expand the types and number of objects.
- Our experiments shows that deep learning based state-of-the-art methods struggles to achieve better performance. Maybe neuro-symbolic methods might be able to achieve better results. It is worth noting it down that CRIPP-VQA proposes the challenge of learning intrinsic properties in unsupervised manner without external resources such as simulators. This also creates another layer of challenge for neuro-symbolic systems.
- CRIPP-VQA is the synthetic dataset on simulated environment. But it is necessary to expand this idea to real-world setting. For example, a person kicking a football vs. brick leads to different consequences and it necessary for systems to learn these physical differences.

REFERENCES

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "VQA: visual question answering", in "2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015", pp. 2425–2433 (IEEE Computer Society, 2015), URL https://doi.org/10.1109/ICCV.2015.279.

Ates, T., M. S. Atesoglu, C. Yigit, I. Kesen, M. Kobas, E. Erdem, A. Erdem, T. Goksun and D. Yuret, "Craft: A benchmark for causal reasoning about forces and interactions", arXiv preprint arXiv:2012.04293 URL https://arxiv.org/abs/2012.04293 (2020).

Baradel, F., N. Neverova, J. Mille, G. Mori and C. Wolf, "Cophy: Counterfactual learning of physical dynamics", in "8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020", (OpenReview.net, 2020), URL https://openreview.net/forum?id=SkeyppEFvS.

Barnard, K., P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei and M. I. Jordan, "Matching words and pictures", The Journal of Machine Learning Research **3**, 1107–1135 (2003).

Bisk, Y., R. Zellers, R. LeBras, J. Gao and Y. Choi, "PIQA: reasoning about physical commonsense in natural language", in "The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020", pp. 7432–7439 (AAAI Press, 2020), URL https://aaai.org/ojs/index.php/AAAI/article/view/6239.

Burgess, C. P., L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick and A. Lerchner, "Monet: Unsupervised scene decomposition and representation", arXiv preprint arXiv:1901.11390 URL https://arxiv.org/abs/1901.11390 (2019).

Chang, C.-Y., D.-A. Huang, D. Xu, E. Adeli, L. Fei-Fei and J. C. Niebles, "Procedure planning in instructional videos", in "European Conference on Computer Vision", pp. 334–350 (Springer, 2020).

Chen, Z., K. Yi, Y. Li, M. Ding, A. Torralba, J. B. Tenenbaum and C. Gan, "Comphy: Compositional physical reasoning of objects and events from videos", arXiv preprint arXiv:2205.01089 URL https://arxiv.org/abs/2205.01089 (2022).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)", pp. 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), URL https://aclanthology.org/N19-1423.

Ding, D., F. Hill, A. Santoro, M. Reynolds and M. Botvinick, "Attention over learned object embeddings enables complex visual reasoning", Advances in neural information processing systems **34**, 9112–9124 (2021).

Epstude, K. and N. J. Roese, "The functional theory of counterfactual thinking", Personality and social psychology review **12**, 2, 168–192 (2008).

Fang, Z., T. Gokhale, P. Banerjee, C. Baral and Y. Yang, "Video2commonsense: Generating commonsense descriptions to enrich video captioning", arXiv preprint arXiv:2003.05162 (2020).

Forbes, M. and Y. Choi, "Verb physics: Relative physical knowledge of actions and objects", in "Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 266–276 (Association for Computational Linguistics, Vancouver, Canada, 2017), URL https://aclanthology.org/P17-1025.

Frome, A., G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato and T. Mikolov, "Devise: A deep visual-semantic embedding model", Advances in neural information processing systems **26** (2013).

Gan, C., J. Schwartz, S. Alter, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, M. Sano *et al.*, "Threedworld: A platform for interactive multimodal physical simulation", arXiv preprint arXiv:2007.04954 URL https://arxiv.org/abs/2007.04954 (2020).

Girdhar, R. and D. Ramanan, "Cater: A diagnostic dataset for compositional actions and temporal reasoning", arXiv preprint arXiv:1910.04744 URL https://arxiv.org/abs/1910.04744 (2019).

Gokhale, T., S. Sampat, Z. Fang, Y. Yang and C. Baral, "Blocksworld revisited: Learning and reasoning to generate event-sequences from image pairs", arXiv preprint arXiv:1905.12042 (2019).

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 6904–6913 (2017).

Grauman, K., A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 18995–19012 (2022).

Grunde-McLaughlin, M., R. Krishna and M. Agrawala, "Agqa: A benchmark for compositional spatio-temporal reasoning", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", pp. 11287–11297 (2021).

He, K., G. Gkioxari, P. Dollár and R. B. Girshick, "Mask R-CNN", in "IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017", pp. 2980–2988 (IEEE Computer Society, 2017), URL https://doi.org/10.1109/ICCV.2017.322.

Hudson, D. A. and C. D. Manning, "Compositional attention networks for machine reasoning", in "6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings", (OpenReview.net, 2018), URL https://openreview.net/forum?id=S1Euwz-Rb.

Hudson, D. A. and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering", in "IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019", pp. 6700–6709 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Janny, S., F. Baradel, N. Neverova, M. Nadri, G. Mori and C. Wolf, "Filtered-cophy: Unsupervised learning of counterfactual physics in pixel space", arXiv preprint arXiv:2202.00368 (2022).

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Zitnick and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. corr, abs/1612.06890", (2016).

Kiros, R., R. Salakhutdinov and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models", arXiv preprint arXiv:1411.2539 (2014).

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations", International journal of computer vision **123**, 1, 32–73 (2017).

Le, T. M., V. Le, S. Venkatesh and T. Tran, "Hierarchical conditional relation networks for video question answering", in "2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020", pp. 9969–9978 (IEEE, 2020), URL https://doi.org/10.1109/CVPR42600.2020.00999.

Lei, J., L. Yu, T. Berg and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering", in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics", pp. 8211–8225 (Association for Computational Linguistics, Online, 2020), URL https://aclanthology.org/2020.acl-main.730.

Li, Y., Y. Song, L. Cao, J. R. Tetreault, L. Goldberg, A. Jaimes and J. Luo, "TGIF: A new dataset and benchmark on animated GIF description", in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", pp. 4641–4650 (IEEE Computer Society, 2016), URL https://doi.org/10.1109/CVPR.2016.502.

Lin, T.-Y., M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft coco: Common objects in context", in "ECCV (5)", (2014).

Liu, R., C. Liu, Y. Bai and A. L. Yuille, "Clevr-ref+: Diagnosing visual reasoning with referring expressions", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 4185–4194 (2019).

Marino, K., M. Rastegari, A. Farhadi and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge", in "Proceedings of the IEEE/cvf conference on computer vision and pattern recognition", pp. 3195–3204 (2019).

Mogadala, A., M. Kalimuthu and D. Klakow, "Trends in integration of vision and language research:: A survey of tasks, datasets, and methods", Journal of Artificial Intelligence Research **71**, 1183–1317 (2021).

Niu, Y., K. Tang, H. Zhang, Z. Lu, X.-S. Hua and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 12700–12710 (2021).

Park, J. S., C. Bhagavatula, R. Mottaghi, A. Farhadi and Y. Choi, "Visualcomet: Reasoning about the dynamic context of a still image", in "European Conference on Computer Vision", pp. 508–524 (Springer, 2020).

Patel, M., T. Gokhale, C. Baral and Y. Yang, "CRIPP-VQA: Counterfactual reasoning about implicit physical properties via video question answering", in "Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)", (2022).

Pearl, J., "Causal inference in statistics: An overview", Statistics surveys **3**, 96–146 (2009).

Rajani, N. F., R. Zhang, Y. C. Tan, S. Zheng, J. Weiss, A. Vyas, A. Gupta, C. Xiong, R. Socher and D. Radev, "ESPRIT: Explaining solutions to physical reasoning tasks", in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics", pp. 7906–7917 (Association for Computational Linguistics, Online, 2020), URL https://aclanthology.org/2020.acl-main.706.

Riochet, R., M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard and E. Dupoux, "Intphys: A framework and benchmark for visual intuitive physics reasoning", arXiv preprint arXiv:1803.07616 URL https://arxiv.org/abs/1803.07616 (2018).

Sampat, S. K., A. Kumar, Y. Yang and C. Baral, "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", pp. 3692–3709 (Association for Computational Linguistics, Online, 2021), URL https://aclanthology.org/2021.naacl-main.289.

Sampat, S. K., M. Patel, S. Das, Y. Yang and C. Baral, "Reasoning about actions over visual and linguistic modalities: A survey", arXiv preprint arXiv:2207.07568 (2022).

Shah, S., A. Mishra, N. Yadati and P. P. Talukdar, "Kvqa: Knowledge-aware visual question answering", in "Proceedings of the AAAI conference on artificial intelligence", vol. 33, pp. 8876–8884 (2019).

Sharma, P., N. Ding, S. Goodman and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning", in "Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 2556–2565 (Association for Computational Linguistics, Melbourne, Australia, 2018), URL https://aclanthology.org/P18-1238.

Shen, B., F. Xia, C. Li, R. Martín-Martín, L. Fan, G. Wang, C. Pérez-D'Arpino, S. Buch, S. Srivastava, L. Tchapmi *et al.*, "igibson 1.0: a simulation environment for interactive tasks in large realistic scenes", in "2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)", pp. 7520–7527 (IEEE, 2021).

Talmor, A., J. Herzig, N. Lourie and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge", in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)", pp. 4149–4158 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), URL https://aclanthology.org/N19-1421.

Tapaswi, M., Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun and S. Fidler, "Movieqa: Understanding stories in movies through question-answering", in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", pp. 4631–4640 (IEEE Computer Society, 2016), URL https://doi.org/10.1109/CVPR.2016.501.

Todorov, E., T. Erez and Y. Tassa, "Mujoco: A physics engine for model-based control", in "2012 IEEE/RSJ international conference on intelligent robots and systems", pp. 5026–5033 (IEEE, 2012).

Wang, T., J. Huang, H. Zhang and Q. Sun, "Visual commonsense r-cnn", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 10760–10770 (2020).

Xu, J., T. Mei, T. Yao and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language", in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", pp. 5288–5296 (IEEE Computer Society, 2016), URL https://doi.org/10.1109/CVPR.2016.571.

Yi, K., C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba and J. B. Tenenbaum, "CLEVRER: collision events for video representation and reasoning", in "8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020", (OpenReview.net, 2020), URL https://openreview.net/forum?id=HkxYzANYDB.

Young, P., A. Lai, M. Hodosh and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", Transactions of the Association for Computational Linguistics **2**, 67–78 (2014).

Yu, L., P. Poirson, S. Yang, A. C. Berg and T. L. Berg, "Modeling context in referring expressions", in "European Conference on Computer Vision", pp. 69–85 (Springer, 2016).

Zellers, R., Y. Bisk, A. Farhadi and Y. Choi, "From recognition to cognition: Visual commonsense reasoning", in "IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019", pp. 6720–6731 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html.

Zhang, S., T. Jiang, T. Wang, K. Kuang, Z. Zhao, J. Zhu, J. Yu, H. Yang and F. Wu, "Devlbert: Learning deconfounded visio-linguistic representations", in "Proceedings of the 28th ACM International Conference on Multimedia", pp. 4373–4382 (2020).

Zhou, L., C. Xu and J. J. Corso, "Towards automatic learning of procedures from web instructional videos", in "Thirty-Second AAAI Conference on Artificial Intelligence", (2018).