CRIPP-VQA: Counterfactual Reasoning about Implicit Physical Properties via Video Question Answering

Maitreya Patel and **Tejas Gokhale** and **Chitta Baral** and **Yezhou Yang**
Arizona State University
{maitreya.patel, tgokhale, chitta, yz.yang}@asu.edu

Abstract

Videos often capture objects, their visible properties, their motion, and the interactions between different objects. Objects also have physical properties such as mass, which the imaging pipeline is unable to directly capture. However, these properties can be estimated by utilizing cues from relative object motion and the dynamics introduced by collisions. In this paper, we introduce CRIPP-VQA[1], a new video question answering dataset for reasoning about the implicit physical properties of objects in a scene. CRIPP-VQA contains videos of objects in motion, annotated with questions that involve counterfactual reasoning about the effect of actions, questions about planning in order to reach a goal, and descriptive questions about visible properties of objects. The CRIPP-VQA test set enables evaluation under several out-of-distribution settings – videos with objects with masses, coefficients of friction, and initial velocities that are not observed in the training distribution. Our experiments reveal a surprising and significant performance gap in terms of answering questions about implicit properties (the focus of this paper) and explicit properties of objects (the focus of prior work).

1 Introduction

Visual grounding seeks to link images or videos with natural language. Towards this goal, many tasks such as referring expressions (Yu et al., 2016), captioning (Vinyals et al., 2015; Xu et al., 2016), text-based retrieval (Vo et al., 2019; Rohrbach et al., 2015), and visual question answering (Antol et al., 2015; Jang et al., 2017) have been studied for both images and videos. Videos often contain objects which can be identified in terms of their visible properties such as their shapes, sizes, colors, textures, and categories. 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 such as mass and coefficient of friction, which are not captured by cameras. For instance, given a video of a stone rolling down a hill, cameras can capture the color of the stone and its trajectory – but how heavy is the stone? It is therefore difficult to reason about such implicit physical properties, by simply watching videos.

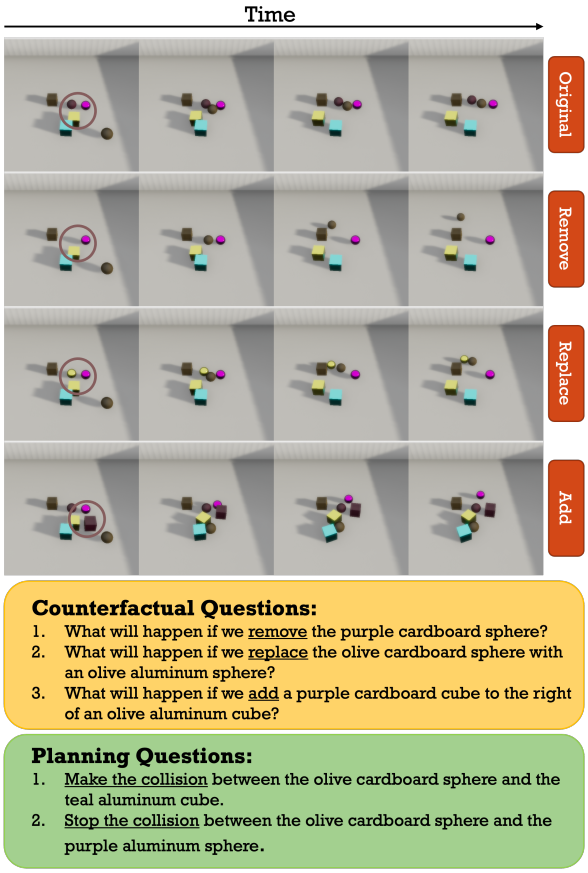

Figure 1: The CRIPP-VQA dataset contains questions about the future effect of actions (such as removing, adding, or replacing objects) as well as planning-based questions. Frames from an example video are shown above with the red highlighted area depicting the objects on which actions (remove, replace, add) are performed.

Collisions between objects, however, do offer

[1] https://maitreyapatel.com/CRIPP-VQA/

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9856 - 9870
December 7-11, 2022

visual cues about mass and friction. When objects collide, their resulting velocities and directions of motion depend upon their physical properties, and are governed by fundamental laws of physics such as conservation of momentum and energy. 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 sports such as billiards, carrom, or curling, 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 paper, we 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. We propose a video question answering dataset called CRIPP-VQA, short for **C**ounterfactual **R**easoning about **I**mplicit **P**hysical **P**roperties. Each video contains several objects with at least one object in motion. The object in motion causes collisions and changes the spatial configuration of the scene. The consequences of these collision are directly impacted by the physical properties of objects. CRIPP-VQA contains videos annotated with question-answer pairs, where the questions are about the consequences of actions and collisions, as illustrated in Figure 1. These questions require an understanding of 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, the *"replace"* action can lead to a change in mass inside the reference video, which can drastically change the consequences (i.e., set of collisions).

We benchmark existing state-of-the-art video question-answering models on the new CRIPP-VQA dataset. Our key finding 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 reveals that models can answer questions about the first collision with higher accuracy compared to questions about subsequent future collisions.

Aloe (Ding et al., 2021) is a strong baseline for video QA tasks and has improved the state of the art on many previous video QA benchmarks such as CLEVRER (Yi et al., 2020) and CATER (Girdhar and Ramanan, 2019). However on CRIPP-VQA, we discovered that the object identification module from Aloe failed to recognize objects in our videos, which we believe is due to the presence of complex textures, reflections, and shadows in our dataset. To mitigate these failures, we modified Aloe by adapting the Mask-RCNN (He et al., 2017) as the object segmentation module. We also found that using pre-trained BERT-based word embeddings significantly improves the performance over our modified Aloe (Aloe*), serving as the strongest model on CRIPP-VQA.

CRIPP-VQA also allows us to evaluate trained models on out-of-distribution (OOD) test sets, where the videos vary in terms of objects having previously unobserved physical properties. There are four OOD test sets in CRIPP-VQA such that one physical property varies at test time – objects with a new mass, zero friction coefficient, increased initial velocity, and two moving objects at initialization. This OOD evaluation reveals a further degradation in performance and a close-to-random accuracy for most state-of-the-art models. Out of all OOD scenarios the results show that the most challenging scenario is the one where there are two objects initially moving.

Contributions and Findings:

- We introduce a new benchmark, CRIPP-VQA, for video question answering which requires reasoning about the implicit physical properties of objects in videos.
- CRIPP-VQA contains questions about the effect of actions such as removing, replacing, and adding objects, as well as a novel planning task, where model needs to perform the three hypothetical actions to either stop or make the collisions between given two objects.
- Performance evaluation on both *i.i.d.* and out-of-distribution test sets shows the significant challenge that CRIPP-VQA brings to video understanding systems.

Dataset	Video QA	Physical Reasoning	Visually Hidden Properties	Counterfactual Actions			Planning	Physical 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., 2022)	✓	✓	-	-	-	✓	-	-	-
CLEVRER (Yi et al., 2020)	✓	✓	-	-	-	✓	-	-	-
ComPhy (Chen et al., 2022)	✓	✓	✓	-	-	-	-	-	-
CRIPP-VQA *(this work)*	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: A comparison of CRIPP-VQA with prior work on video question answering, in terms of different aspects of visual reasoning that are tested.

2 Related Work

Image Question Answering. The VQA dataset (Antol et al., 2015) has been extensively for image-based question answering. GQA (Hudson and Manning, 2019) and CLEVR (Johnson et al., 2017a) focus on the compositional and spatial understanding of visual question answering models. CLEVR-HYP (Sampat et al., 2021) extends the CLEVR setup with questions about hypothetical actions performed on the image. OK-VQA (Marino et al., 2019) deals with answering questions where external world knowledge (such as Wikipedia facts) are required for answering questions, whereas VLQA (Sampat et al., 2020) studies image question answering with additional information provided via an input paragraph.

Video Question Answering. Datasets such as MovieQA (Tapaswi et al., 2016), TGIF (Li et al., 2016), TVQA/TVQA+ (Lei et al., 2020), and AGQA (Grunde-McLaughlin et al., 2021), have been introduced for real-world video question answering. However, work on video question answering has largely focused on scenes such as movies and television shows.

Physical Reasoning. Visual planning 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 reasoning under the influence of gravity. CATER (Girdhar and Ramanan, 2019) is a video classification dataset, which proposes the challenge of temporal reasoning on actions such as slide, rotate, pick-place, etc. Recently, the CLEVRER benchmark (Yi et al., 2020) studied the ability to do counterfactual reasoning only with remove action. However, all objects in CLEVRER have identical physical properties. CoPhy (Baradel et al., 2020) studied the problem of predicting consequences in the presence of mass as a confounding variable. It does not involve the change in the physical properties during counterfactual reasoning and only studies displacement-based counterfactual object trajectory estimation. ComPhy (Chen et al., 2022) is a work closest to ours, with the task of learning visually hidden properties in a few-shot setting and performing counterfactual reasoning with a question that explicitly describes the changes in physical properties (*"What if object A was heavier?"*). In contrast, in our work, physical properties need to be learned from video, and are not mentioned in the question, with three types of questions (descriptive, counterfactual, and planning). We position our work in comparison to previous work in Table 1.

Textual Commonsense Reasoning. PIQA (Bisk et al., 2020) is a dataset for physical commonsense reasoning for natural language understanding (NLU) systems . CommonsenseQA (Talmor et al., 2019) is a QA dataset that focuses on inferring associated relations of each entity. Verb Physics (Forbes and Choi, 2017) proposes the task of learning relative physical knowledge (size, weight, strength, etc.) for NLU systems.

Visual Commonsense Reasoning. Visual-COMET (Park et al., 2020) is a dataset for inferring commonsense concepts such as future events and their effects from the images and textual descriptions. Video2Commonsense (Fang et al., 2020) is a video captioning task that seeks to include intentions and effects of human actions in the generated caption. VCR (Zellers et al., 2019) dataset introduces a VQA task that requires commonsense and understanding the scene context

in order to answer questions and also to justify the answer. While, (Sampat et al., 2022) gives the overview of recent advances in multimodal action based reasoning.

Robustness of Multimodal Models. Robustness to distribution shift and language bias has been extensively studied in the VQA domain (Ray et al., 2019; Gokhale et al., 2020; Selvaraju et al., 2020; Kervadec et al., 2020; Li et al., 2020; Agarwal et al., 2020). Shortcuts and spurious correlations have been observed in visual commonsense reasoning (Ye and Kovashka, 2021). For V+L entailment tasks, Gokhale et al. (2022) found that models are not robust to linguistic transformations, while Thrush et al. (2022) found that models were unable to distinguish between subject and object of actions. However most of the work in robust V+L has focused on biases or distribution shift in the language domain. CRIPP-VQA introduces out-of-distribution evaluation in terms of *physical* properties of objects in a scene.

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 mass) in the presence of mass and friction as visually hidden properties.

3.1 Simulation Setup

Objects and States. Table 2 summarizes the different properties in CRIPP-VQA. Each object in the CRIPP dataset has four visible properties: shape (cube or sphere), color (olive, purple, and teal), texture (aluminum and cardboard), and state (stationary, in motion, and under collision). Each object also has two invisible properties: mass and coefficient of friction. Three actions can be performed on each object – "remove", "replace", and "add".

In this work, we focus on mass and friction as intrinsic physical properties of objects. Each unique {SHAPE, COLOR, TEXTURE} combination is preassigned a mass value that is either 2 or 14; 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. For one of the OOD test sets, we make the surfaces and objects frictionless. Table 2 shows the object properties for training videos, *i.i.d.* test set and OOD test set.

Video creation. We render videos using *TDW* (Gan et al., 2021). Firstly, in each instance, we initialize the video with either 5 or 6 randomly chosen objects, out of which a single object will be initialized with a fixed velocity such that it will collide with other objects. Here, we keep a constant initial velocity so that the only way to infer mass is through the impact of subsequent collisions. Each video 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.

3.2 Question and Answer Generation

CRIPP dataset focuses on three categories of tasks: 1) Descriptive, 2) Counterfactual, and 3) Planning.

Descriptive: These questions involve understanding the visual properties of the scene, including:
[Type-1] Counting the number of objects with a certain combination of visible properties,
[Type-2] Yes/No questions about object types
[Type-3] Finding the relationship between two objects under collision
[Type-4] Counting the number of collisions
[Type-5] Finding the maximum/minimum occurring object properties.
We do 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). We 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. *"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

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 2: They key difference between the IID and various OOD evaluation settings in CRIPP-VQA. Here, "-" indicates the no change in particular property from the IID setting.

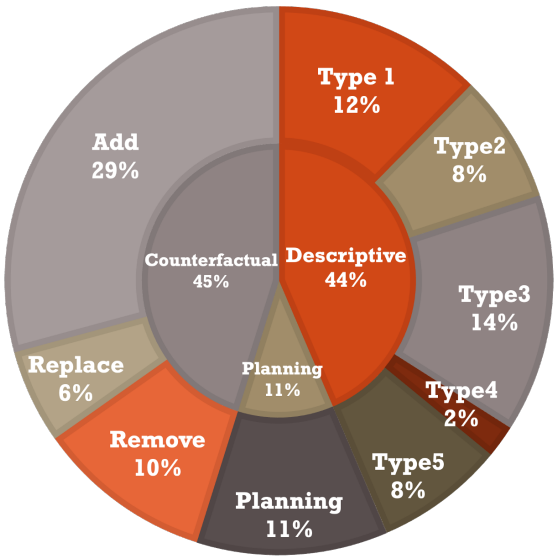

Figure 2: A pie-chart showing the distribution of various question types in the CRIPP-VQA dataset. Inner pie chart shows the three broad categories of questions (counterfactual, descriptive, planning), while the outer pie-chat shows a fine-grained categorization.

the system's understanding of spatial relationship along with the hidden properties, where we create a new hypothetical condition by placing a new object to the LEFT/RIGHT/FRONT/BEHIND 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 has to be performed and on which object, to achieve the goal.

3.3 Dataset Statistics

CRIPP contains 4000, 500, and 500 videos for training, validation, and testing, respectively. Additionally, 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 2 shows the percentages of each subcategory within the dataset.

4 Experiments

4.1 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 F such that we maximize the $F(a|v, q)$.

Evaluation Metrics. To evaluate the models, we use two accuracy metrics – per-option (PO) and per-question (PQ) accuracy. Each counterfactual question has multiple options describing the collisions. Per-option accuracy refers to the option-wise performance and per-question accuracy considers whether all options are 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 solutions. We use *TDW* to re-simulate the models' predictions on the original video to check whether the given planning goal is achieved or not, leading to iterative performance evaluation.

4.2 Benchmark model details

We consider three different state-of-the-art models for the video question answering task: Memory, Attention, and Composition (MAC) (Hudson and Manning, 2018), Hierarchical Conditional Relation Network (HCRN) (Le et al., 2020), and 3)

Model	Descriptive	Remove		Replace		Add		Counterfactual	Planning
		PQ	PO	PQ	PO	PQ	PO	Avg. PO	
Frequency	8.21	0.00	50.18	0.00	50.00	0.00	50.00	50.06	3.49
Random	8.51	7.21	49.58	3.34	49.40	9.39	50.04	49.67	7.39
Blind-BERT	53.82	20.18	54.67	17.57	50.45	15.86	51.55	52.22	8.11
MAC (Hudson and Manning, 2018)	48.72	16.41	50.68	17.31	50.21	16.29	49.83	50.24	6.26
HCRN (Le et al., 2020)	64.98	27.20	59.04	19.87	55.97	20.49	56.06	57.02	21.38
Aloe*	68.94	31.10	62.90	9.91	52.10	18.13	56.55	57.18	31.76
Aloe*+BERT	71.04	33.64	65.46	22.07	56.76	39.71	67.43	63.21	32.61

Table 3: 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. Aloe* refers to our modified Aloe, where we replace the MONet module with a Mask-RCNN object detector.

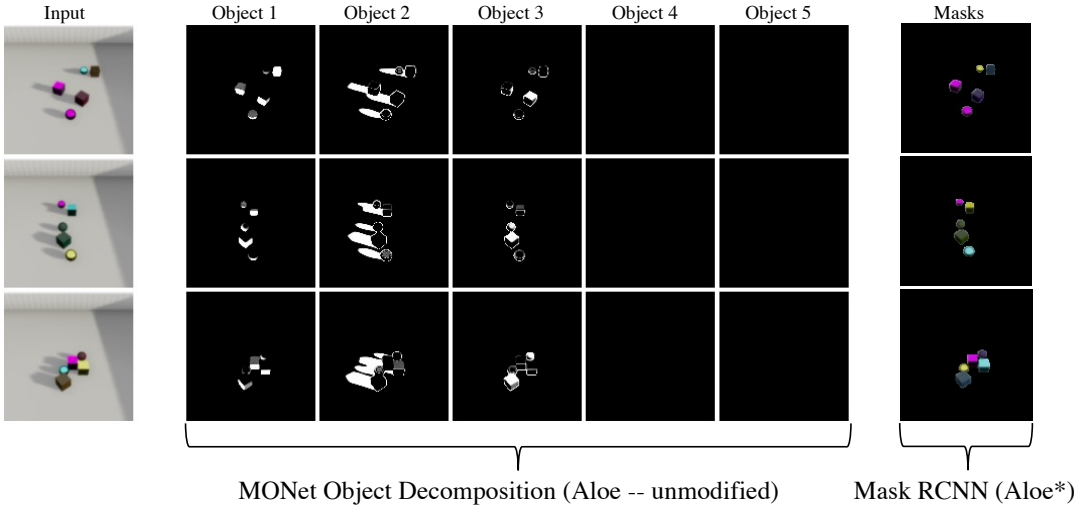

Figure 3: Illustration of the failure of MONet (the object decomposition module in Aloe (Ding et al., 2021)) on CRIPP-VQA videos. The intended functionality of MONet is to decompose individual objects into separate masks. However as shown above, the predicted masks contain areas corresponding to more than one objects. 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.

Attention over learned embeddings (Aloe) (Ding et al., 2021). **MAC** is designed for compositional VQA. We modify it by performing channel-wise feature concatenation of each frame, where the channel will contain temporal information instead of spatial information allowing MAC to adapt to the video inputs. **HCRN** uses a hierarchical strategy to learn the relation between the visual and textual data. **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. Aloe uses MONet (Burgess et al., 2019) for obtaining object features by performing an unsupervised decomposition of each frame into observed objects. Aloe takes these frame-wise object features to predict the answers to the input question, using the *[CLS]* token and self-supervised training.

Drawbacks of Aloe. We found that the MONet module used in Aloe is very unstable and fails to produce reliable frame-wise features on videos from CRIPP. MONet is not able to recognize simple object properties such as color and is not able to decompose the image into masks corresponding to individual objects. This drawback hurts the performance of Aloe on the CRIPP-VQA dataset, even though Aloe is one of the best-performing models on previous video QA benchmarks. An example is shown in Figure 3, and more details can be found in Appendix C. We believe that this failure could be a result of shadows and textures in our dataset that are not found in previous datasets.

Modifying Aloe. Due to the failures of the MONet object decomposition module, the Aloe baseline fails measurably on CRIPP-VQA, exhibit-

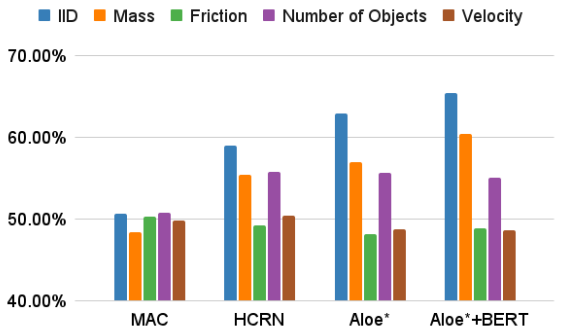

Figure 4: Comparison of performance of models (per-option accuracy) for "remove" questions when tested using the IID test set and each OOD test set.

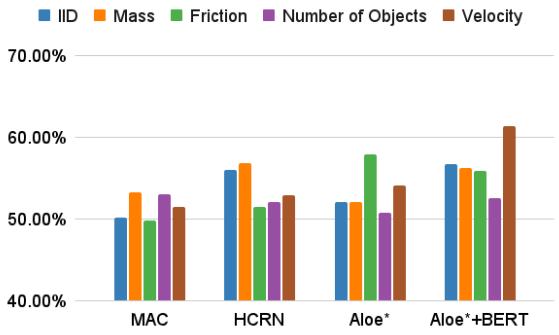

Figure 5: Comparison of performance of models (per-option accuracy) for "replace" questions when tested using the IID test set and each OOD test set.

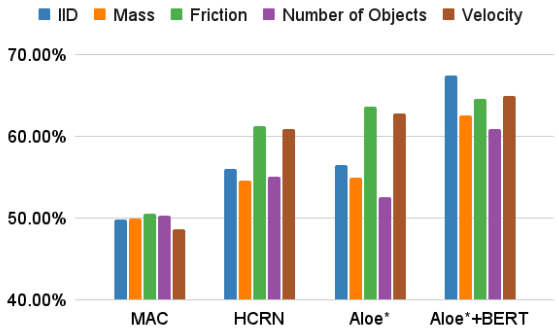

Figure 6: Comparison of performance of models (per-option accuracy) for "add" questions when tested using the IID test set and each OOD test set.

ing close-to-random performance. Therefore, we propose additional modifications to Aloe to make it more widely applicable beyond prior datasets that are built using the CLEVR (Johnson et al., 2017a) rendering pipeline. First, we replace MONet with Mask-RCNN (He et al., 2017) to perform instance segmentation and then train an auto-encoder to compress the mask-based object-specific features to make it compatible with Aloe. Second, instead of learning the word embedding from the scratch, we further propose to use pre-trained BERT-based word embeddings as input to the Aloe, which leads to faster and more stable convergence. Further architecture modifications and hyper-parameter settings are specified in Appendix A.

In addition to these baselines, we also 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, we use a text-only QA model and denote it by *"Blind-BERT"*. 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.

4.3 Results

Table 3 summarizes the performance comparisons of our 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. An unmodified version of the Aloe also achieves only 56% accuracy. HCRN and both Aloe variants (Aloe* and Aloe*+BERT) improve performance indicating that visual features are crucial for descriptive questions. Aloe*+BERT is the best performing model which implies that our modification with BERT features helps performance.

Counterfactual questions involve a total of three types of actions. Table (3) 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. While, unmodified Aloe gets 52% average accuracy on counterfactual questions, which is close-to-random performance. 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

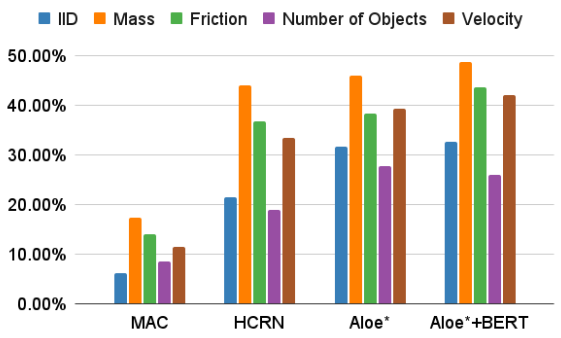

Figure 7: Comparison of performance of models "planning" questions when tested using the IID test set and each OOD test set.

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 multiple possible answers. We observe a similar trend in results and Aloe*+BERT 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 tells us that the model finds it easy to reason when "remove" hypothetical action is present.

Human evaluations: To learn the expected behavior of any models, we conduct a human studies on CRIPP-VQA dataset. There were total 6 people participated as volunteers. All were given 5 videos and corresponding QA pairs to get habituated with the environment. Then we asked them to answer total 30 questions on different set of randomly selected videos. Results shows that Human evaluations achieved 90.00%, 78.89%, and 58.87% on descriptive, counterfactual, and planning tasks, respectively.

4.4 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). We propose a new dimension of OOD evaluation involving physical properties, by considering 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.

Figures 4, 5, 6, 7 shows the comparison of VideoQA models on *i.i.d.* and different OOD scenarios for remove, replace, and add action, and planning questions, respectively. It can be seen that the models' performance becomes close to random which is around 50%. This suggests that models are very sensitive to such small physical perturbations, especially for the *"remove"* action (as shown in Figure 4). From Figure 6, we can observe that the performance drop is negligible across the OOD sets for the add action, especially for Aloe*+BERT. Moreover, Figure 7 shows that the performance increases on several OOD scenarios for planning task. At the same time, the bias-check baselines' performance also improves. This suggests that the expected behavior of the model changes based on the given physical properties. In the case of the remove action, Friction and Velocity OOD settings are the hardest for models to perform. While, for replace action, multiple OOD setting is the hardest for Aloe*+BERT. Number of initial moving Objects based OOD setting is also difficult for models to understand, especially for the add action based questions.

5 Analysis

In this section, we raise several important questions and derive the insights accordingly.

Performance for true vs. false collision detection. Consider the example with three objects (A,B,C), where only object A collides with B. In this case, we categorize the collision between A & B as the actual collision (i.e., prediction label *true*), and we categorize the collision between B & C and A & C as an absent collision (i.e., prediction label *false*). Following this rule, we independently check the performance of detecting all occurring collisions and the collisions that never happened. Table 4 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 except for the *"add"* action, where model mainly predicts that none of the collisions are present in counterfactual scenario. However, in the case of the replace action, the model is failing in both categories.

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

Table 4: Per-option accuracy of Aloe*+BERT for detecting occurring collisions vs. not occurring collisions correctly.

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

Table 5: 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
Remove	3.12	3.23	1.78	4.03

Table 6: 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.

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

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. To understand the dynamics, we analyze the number of collisions in different counterfactual scenarios and collisions between two different types of objects (in terms of mass). Table 6 shows that if first collision is between either two light or two heavy objects then it leads to almost similar number of collisions. If first collision is between light and heavy objects then the number of collisions either decreases or increases based on the intuitive conditions. Analysis on the number of collisions in 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 show that each counterfactual scenarios are unique and contain different challenges. This also strengths our argument that models fail to learn various reasoning capabilities including but not limited to intrinsic physical properties, and consequences of the actions.

6 Conclusion

In this work, we present a new video question answering benchmark: CRIPP-VQA, for reasoning about the implicit physical properties of objects. CRIPP-VQA contains novel tasks that require counterfactual reasoning and planning, over three hypothetical actions (i.e., remove, replace, and add). We evaluate state-of-the-art models on this benchmark and observe a significant performance gap between descriptive questions about visible properties and counterfactual and planning questions about implicit properties. We also show that models can learn the initial dynamics of object trajectories but they fail to detect subsequent collisions, which requires an understanding of relative mass. This result is positioned as a challenge for the V&L community for building robust video understanding systems that can interact with language.

7 Limitations

While CRIPP proposes the implicit reasoning about intrinsic physical properties, it is limited to two physical properties (mass and friction). However, even these fundamental properties are a big challenge for existing systems. While other properties and complex dynamics can be considered, that is beyond the scope of this work. Our benchmark is limited to a synthetic environment in blockworld, and we believe that future work should extend our work with real-world objects and backgrounds.

Acknowledgements

This work was supported by NSF RI grants #1750082, #1816039 and #2132724, and the DARPA GAILA ADAM project.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9687–9695. IEEE.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Tayfun Ates, M. Ateşoğlu, Çağatay Yiğit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. 2022. CRAFT: A benchmark for causal reasoning about forces and inTeractions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2602–2627, Dublin, Ireland. Association for Computational Linguistics.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. 2020. Cophy: Counterfactual learning of physical dynamics. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. 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*, pages 7432–7439. AAAI Press.

Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. 2019. Monet: Unsupervised scene decomposition and representation. *CoRR*, abs/1901.11390.

Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. 2020. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. 2022. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. 2021. Attention over learned object embeddings enables complex visual reasoning. In *Advances in Neural Information Processing Systems*, volume 34, pages 9112–9124. Curran Associates, Inc.

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2Commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860, Online. Association for Computational Linguistics.

Maxwell Forbes and Yejin Choi. 2017. 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)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel M. Bear, Dan Gutfreund, David Cox, Antonio Torralba, James J. DiCarlo, Joshua B. Tenenbaum, Josh H. McDermott, and Daniel L.K. Yamins. 2021. Threedworld: A platform for interactive multi-modal physical simulation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Rohit Girdhar and Deva Ramanan. 2019. Cater: A diagnostic dataset for compositional actions & temporal reasoning. In *International Conference on Learning Representations*.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*. Springer.

Tejas Gokhale, Abhishek Chaudhary, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2022. Semantically distributed robust optimization for vision-and-language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1493–1513.

Tejas Gokhale, Shailaja Sampat, Zhiyuan Fang, Yezhou Yang, and Chitta Baral. 2019. Cooking with blocks: A recipe for visual reasoning on image-pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 5–8.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11287–11297.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2018. 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.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1359–1367. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2020. Roses are red, violets are blue... but should vqa expect them to? *arXiv preprint arXiv:2006.05121*.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. 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*, pages 9969–9978. IEEE.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Linjie Li, Zhe Gan, and Jingjing Liu. 2020. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. 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*, pages 4641–4650. IEEE Computer Society.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. ESPRIT: Explaining solutions to physical reasoning tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7906–7917, Online. Association for Computational Linguistics.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. 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 5860–5865, Hong Kong, China. Association for Computational Linguistics.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. 2018. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212. IEEE Computer Society.

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.

Shailaja Keyur Sampat, Maitreya Patel, Subhasish Das, Yezhou Yang, and Chitta Baral. 2022. Reasoning about actions over visual and linguistic modalities: A survey. *arXiv preprint arXiv:2207.07568*.

Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-linguistic question answering (vlqa) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4606–4616.

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at vqa models: Interrogating vqa models with sub-questions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. 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*, pages 4631–4640. IEEE Computer Society.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6439–6448. Computer Vision Foundation / IEEE.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. 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*, pages 5288–5296. IEEE Computer Society.

K Ye and A Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *AAAI*.

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. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. 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*, pages 6720–6731. Computer Vision Foundation / IEEE.

Appendix

A Training details

We follow the standard training guidelines provided by the authors of each baseline papers. We train all systems on Quadro RTX 8000 GPUs. We train each model with a maximum of 200 epochs. And select the best model based on average performance accuracy. We follow the below instructions to support each model which are MAC, HCRN, Aloe, and Aloe+BERT. For planning based task, we 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: We modify the public implementation of MAC from `https://github.com/rosinality/mac-network-pytorch` to adapt the video frames as input. We first resize the each 125 frames leading $(125, 3, 224, 224)$ video dimension. Later, we use ResNet101 to extract the features $(125, 512, 14, 14)$. After taking the channel-

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 7: Aloe*+BERT architecture and hyper-parameter details.

wise mean of features, we get the final video representation of $(125, 14, 14)$ dimension matrix supportable for the rest of the pipeline. We 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, we use the source code as it is. Except we do important changes to do planning tasks.

Aloe*/Aloe*+BERT: We first reproduce the Aloe on PyTorch based on the architecture details from the research paper by (Ding et al., 2021) and their public available demo at https://github.com/deepmind/deepmind-research/tree/master/object_attention_for_reasoning. However, we use the code base from transformers[2] library (as it is well tested and used across the industry and academia) and modify it to support the VideoQA in the same way as Aloe does. Our 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, we do extensive experiments on Aloe architecture and hyper-parameter to reproduce similar results. After achieving a similar performance from the paper, we use this new reproducible Aloe architecture in our experiments. Table (7) shows the hyper-parameter details to reproduce the results. The Aloe* source code from our experiments is available at https://github.com/Maitreyapatel/CRIPP-VQA/

B Dataset Examples

The demo page contains several examples of the CRIPP-VQA dataset. Apart from that, Table 8 shows the types of questions asked in different subcategories of the QAs.

C MONet failure cases

We learn that MONet-based unsupervised object decomposition is not working on complex realistic visuals and it is hard to guarantee that it will decompose each object on independent images/features. Here, we show three failure cases from the CRIPP-VQA. Basically, from the figures, 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. While MONet can learn the texture (i.e., metal or cardboard). As a result, we can see that the re-generated images lack greatly in terms the important features. Hence, we drop the MONet from the pipeline and adapt mask r-cnn to work on our CRIPP dataset.

D Physical out-of-distribution results

In this section, we provide the accuracy tables for the OOD evaluations. First, Table (9) shows the performance of all models when the mass of few objects are changed (either increased or decreased to 8 from 2 or 14). Second, Table (10) shows the results when the surface friction is removed. Third, Table (11) shows the results where we have two objects initialized with fixed velocity creating more collisions. At last, Table (12) contains the results when we slightly increase the initial velocity of the object. Overall, we observe that for both counterfactual and planning tasks all model performs poorly.

E Neuro-symbolic methods

Recently, a lot of neuro-symbolic approaches are proposed for CLEVRER-like settings. For example, IEP (Johnson et al., 2017b), NS-DR+ (Mao et al., 2019), are CPL (Chen et al., 2022) proposed for physical reasoning. The goal of our study is to evaluate whether systems can learn the implicit relationship from counterfactual tasks. Symbolic approaches either require providing this implicit information or learning through a physics engine,

[2] https://github.com/huggingface/transformers

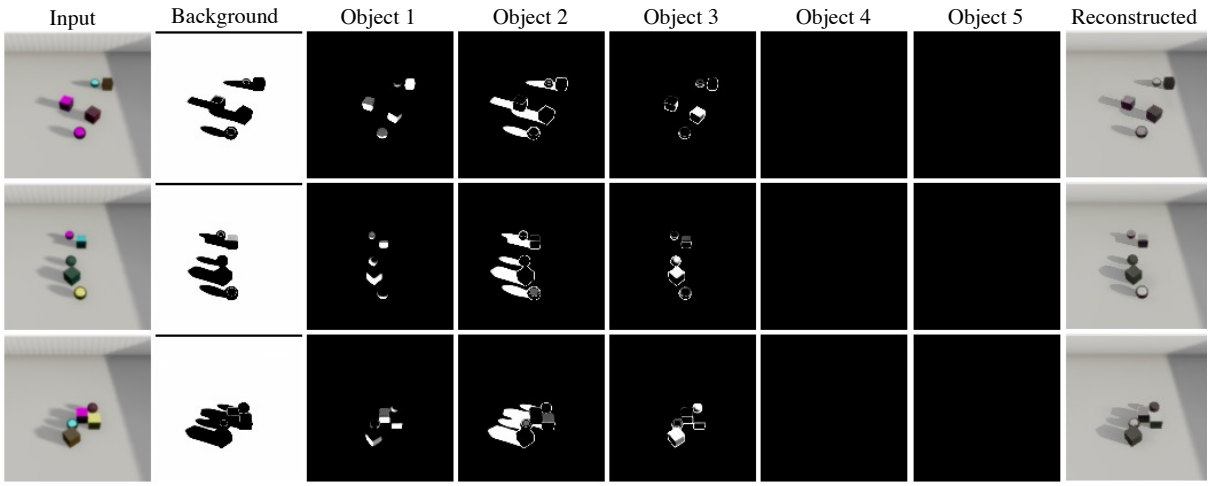

Figure 8: Example outputs of MONet-based scene decomposition failure cases. Left most images represents the input image. The Middle six images represent the predicted masks. And right most images represent the reconstructed input image by MONet.

which is not feasible for real-life situations. Therefore, in this study, we only consider neural models to evaluate their performance where learning implicit information is necessary.

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 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 ?
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 8: Examples of the CRIPP-VQA questions asked from different types of question categories as shown in Figure 2.

Model	Remove		Replace		Add		Planning QA
	PQ	PO	PQ	PO	PQ	PO	
Frequency	0.00	50.27	0.00	50.00	0.00	50.00	21.00
Random	9.61	49.95	10.57	49.71	10.29	49.85	21.16
Blind-BERT	13.52	49.67	13.72	48.71	9.44	50.80	16.43
MAC	12.99	48.36	18.89	53.25	12.21	50.00	17.34
HCRN	18.15	55.47	20.08	56.84	14.03	54.58	43.94
Aloe*	20.46	57.00	12.98	52.05	12.90	54.93	46.07
Aloe*+BERT	26.16	60.67	22.42	56.21	20.34	62.62	48.71

Table 9: Performance evaluations when mass dist. is different than the training.

Model	Remove		Replace		Add		Planning QA
	PQ	PO	PQ	PO	PQ	PO	
Frequency	0.00	50.16	0.00	50.00	0.00	50.00	20.46
Random	10.41	50.51	3.92	50.42	6.58	49.93	20.49
Blind-BERT	10.86	49.90	11.90	49.77	13.96	50.80	12.34
MAC	11.6	50.30	14.23	49.82	7.86	50.53	14.07
HCRN	11.74	49.20	13.74	51.54	13.27	61.32	36.73
Aloe*	11.46	48.11	19.58	57.91	16.83	63.67	38.27
Aloe*+BERT	7.21	48.85	24.87	55.93	18.37	64.66	43.67

Table 10: Performance evaluations with zero surface friction.

Model	Remove		Replace		Add		Planning QA
	PQ	PO	PQ	PO	PQ	PO	
Frequency	0.00	50.57	0.00	50.00	0.00	50.00	5.04
Random	5.09	49.92	4.76	50.75	9.00	49.37	8.15
Blind-BERT	12.64	49.70	12.70	49.64	5.82	51.58	7.57
MAC	14.87	50.84	13.78	52.99	12.17	50.33	8.51
HCRN	19.75	55.80	13.78	52.10	13.10	55.03	18.92
Aloe*	17.39	55.73	12.50	50.82	10.61	52.55	27.76
Aloe*+BERT	12.81	55.10	18.37	52.54	21.10	60.86	25.95

Table 11: Performance evaluations with multiple objects moving.

Model	Remove		Replace		Add		Planning QA
	PQ	PO	PQ	PO	PQ	PO	
Frequency	0.00	50.20	0.00	50.00	0.00	50.00	19.20
Random	10.93	49.47	3.66	50.12	6.99	49.83	19.53
Blind-BERT	10.02	50.27	16.06	52.17	5.73	51.82	13.68
MAC	10.76	49.80	15.09	51.44	6.34	48.66	11.54
HCRN	10.89	50.52	17.62	52.89	12.67	60.93	33.48
Aloe*	11.63	48.79	14.51	54.15	15.36	62.82	39.30
Aloe*+BERT	6.43	48.63	29.53	61.37	15.82	65.02	42.10

Table 12: Performance evaluations with higher initial velocity.