Compositional Physical Reasoning of Objects and Events from Videos

Zhenfang Chen Shilong Dong* Kexin Yi Yunzhu Li Mingyu Ding Antonio Torralba Joshua B. Tenenbaum Chuang Gan*

Abstract— Understanding and reasoning about objects' physical properties in the natural world is a fundamental challenge in artificial intelligence. While some properties like colors and shapes can be directly observed, others, such as mass and electric charge, are hidden from the objects' visual appearance. This paper addresses the unique challenge of inferring these hidden physical properties from objects' motion and interactions and predicting corresponding dynamics based on the inferred physical properties. We first introduce the Compositional Physical Reasoning (ComPhy) dataset. For a given set of objects, ComPhy includes limited videos of them moving and interacting under different initial conditions. The model is evaluated based on its capability to unravel the compositional hidden properties, such as mass and charge, and use this knowledge to answer a set of questions. Besides the synthetic videos from simulators, we also collect a real-world dataset to show further test physical reasoning abilities of different models. We evaluate state-of-the-art video reasoning models on ComPhy and reveal their limited ability to capture these hidden properties, which leads to inferior performance. We also propose a novel neuro-symbolic framework, Physical Concept Reasoner (PCR), that learns and reasons about both visible and hidden physical properties from question answering. Leveraging an object-centric representation, PCR utilizes videos and the associated natural language to infer objects' physical properties without dense object annotations. Furthermore, It incorporates property-aware graph networks to approximate the dynamic interactions among objects. PCR also employs a semantic parser to convert questions into semantic programs, and a program executor to execute the programs based on the learned physical properties and dynamics. After training, PCR demonstrates remarkable capabilities. It can detect and associate objects across frames, ground visible and hidden physical properties, make future and counterfactual predictions, and utilize these extracted representations to answer challenging questions. We hope the proposed ComPhy dataset and the PCR model present a promising step towards more comprehensive physical reasoning in AI systems.

Index Terms—Physical Reasoning, Neuro-Symbolic Models, Hybrid Models.

1 Introduction

WHAT causes apples to float in water while bananas sink? What is the underlying reason for magnets attracting on one side and repelling on the other? Objects in nature frequently manifest complex properties, which delineate their interaction schema within the physical world. For humans, deciphering these *intrinsic* physical properties often represents pivotal milestones in fostering a more profound and precise comprehension of nature. The majority of these properties are intrinsic in nature, as they are not readily apparent through objects' static visual attributes and are only detectable from objects' interactions. Furthermore, these properties influence object motion in a *compositional* manner, where the causal relationships and mathematical laws governing these properties can often be complex.

As depicted in Figure 1, various *intrinsic* physical properties, such as charge and inertia, often result in significantly divergent future trajectories. Objects bearing identical or opposite *charges* will exert either repulsive or attractive forces on one another. The resultant motion is not only related to the magnitude of the charge each object possesses but also to their respective signs, as illustrated in Figure 1-(a). *Inertia* governs the degree of sensitivity of an object's motion to external forces. In scenarios, where a massive object interacts with a lighter one through attraction, repulsion, or collision, the lighter object experiences more substantial alterations in its motion relative to the trajectory of the massive object, as depicted in Figure 1-(b).

Recent research has introduced a suite of benchmarks aimed at assessing and diagnosing machine learning systems across a range of physics-related settings [1]–[3]. These benchmarks present reasoning tasks involving intricate object motion and complex interactions, imposing significant challenges on existing models as they demand an understanding of the underlying physical dynamics to perform well. However, the majority of complexity in the motion trajectories facilitated by these environments arises from alterations or interventions in the initial conditions of the physical experiments. The impacts of objects' intrinsic physical properties, along with the distinct challenges they present, hold significant importance for further research.

However, it is non-trivial to construct a benchmark for compositional physical reasoning. A straightforward approach might involve adhering to the settings established in previous benchmarks [2], [4], wherein a model is required to observe a video and subsequently respond to questions regarding physical properties. Nevertheless, physical prop-

- *Z. Chen and S. Dong contribute equally.*
- *Z. Chen is with MIT-IBM Watson AI lab. E-mail: zfchenzfc@gmail.com*
- *S. Dong is with New York University. E-mail: shilongdong00@gmail.com*
- *K. Yi is with Harvard University. E-mail: kyi@g.harvard.edu*
- *Y. Li is with UIUC. E-mail: yunzhuli@illinois.edu*
- *M. Ding is with UC Berkeley. E-mail: myding@berkeley.edu*
- *A. Torralba and J. B. Tenenbaum are with MIT. E-mail: {torralba, jbt}@mit.edu*
- *C. Gan is with MIT-IBM Watson AI lab and UMass Amherst. E-mail: ganchuang1990@gmail.com*

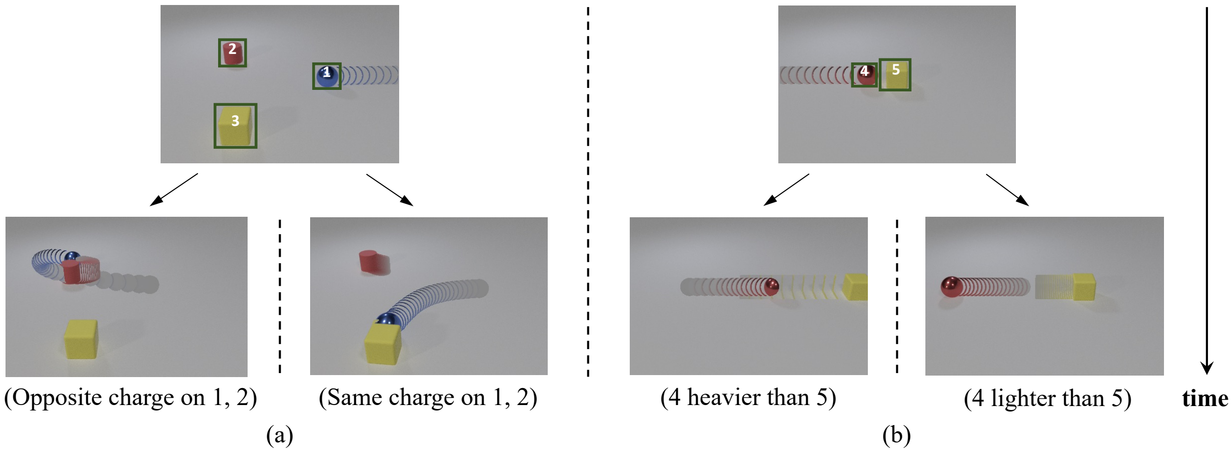

Fig. 1: Non-visual properties like mass and charge govern the interaction between objects and lead to different motion trajectories. a) Objects attract and repel each other according to the (sign of) charge they carry. b) Mass determines how much an object's trajectory is perturbed during an interaction. Heavier objects have more stable motion.

erties are intricate and often cannot be comprehensively elucidated within the confines of a single video. Another approach is to establish correlations between object appearance and physical properties, such as designating all *red spheres* as *heavy*, and subsequently posing questions regarding their dynamics. Nonetheless, this design may lead models to employ shortcuts by merely memorizing appearances rather than comprehending the interconnected physical properties.

In this paper, we present an extended version of the ComPhy benchmark [5], with significant additions, including more diverse simulated scenes, real-world videos, and new experimental settings. It centers on the comprehension of object-centric and relational physics properties not readily discernible from visual appearances. Initially, ComPhy presents a limited number of video examples featuring dynamic interactions among objects. Models are tasked with identifying the physical properties of objects and subsequently answering questions pertaining to these properties and their associated dynamics.

As depicted in Figure 2, the ComPhy is composed of meta-train and meta-test sets, with each data point comprising four reference videos and one target video. In each set, the objects consistently possess the same intrinsic physical properties across all videos. To facilitate the task, we systematically ensure that each object in the query video appears in at least one of the reference videos. Reasoning on the ComPhy is challenging. First, models must infer both the intrinsic and compositional physical properties of the object set using only a limited number of video samples. Moreover, they must predict video dynamics based on the predicted physical properties.

To overcome the challenges in ComPhy, we introduce Physical Concept Reasoner (PCR). Inspired by recent work on neural-symbolic reasoning on images and videos [2], [6], [7], our model is modularized with four disentangled components: perception, physical property learning, physical dynamics prediction, and symbolic reasoning. Our PCR model can learn to infer objects' compositional and intrinsic physical properties, predict their future dynamics, and make counterfactual imaginations by only watching videos and reading question-answer pairs.

To summarize, this paper makes the following contributions. First, we extend the original ComPhy benchmark [5] by introducing new diverse simulated scenes and real-world video data. It is based on a few-shot reasoning setting that integrates physical properties (mass and charge), physical events (attraction and repulsion), and their compositions. Second, we introduce a new neural-symbolic framework PCR, a modularized model that can infer objects' physical properties and predict the objects' movements from watching videos and reading question-answer pairs. Additionally, we collect a real-video dataset to better assess the physical reasoning capabilities of current models in real-world scenarios.

Some preliminary results were presented in our earlier ICLR 2022 paper [5]. In this manuscript, we significantly extend that work in three aspects. First, we introduce a Physical Concept Reasoner, PCR, to learn hidden physical properties like *mass* and *charge* from video and language efficiently without dense property supervision signals during training and perform reasoning in counterfactual and predictive scenes. Second, besides the experiments in the original data [8], we also simulate more diverse physical scenes and collect **real** videos for physical reasoning. We perform experiments in both synthetic and **real** videos and analyze how the new proposed PCR works and fails, while there are only experiments for synthetic data in the original conference version. Third, we also evaluate recent state-of-the-art large vision-language models (LVLMs) [9], [10] on ComPhy, providing a more thorough analysis. Our code, datasets, and models can be found at https://physicalconceptreasoner.github.io.

The rest of the paper is organized as follows. Section 2 reviews the related datasets and models based on physical reasoning, video question answering, and few-shot learning. Section 3 introduces how we construct the dataset and reduce its biases. Section 4 analyze how representative baselines and the recent state-of-the-art models perform on the ComPhy benchmark. Section 5 introduces the new PCR model and its optimization mechanism. Section 6 summarizes the paper's contribution, discusses its limitations, and suggests potential extension directions.

Dataset	Video	Question Answering	Diagnostic Annotation	Composition	Few-shot Reasoning	Physical Property	Counterfactual Property Dynamics	Evaluated on LVLM
CLEVR [11]	-	✓	✓	✓	-	-	-	-
MovieQA [12]	✓	✓	-	✓	-	-	-	-
TGIF-QA [13]	✓	✓	-	-	-	-	-	-
TVQA/ TVQA+ [14]	✓	✓	-	✓	-	-	-	-
AGQA [15]	✓	✓	-	-	-	-	-	-
IntPhys [16]	✓	-	✓	-	-	✓	-	-
PHYRE/ ESPRIT [17]	✓	-	✓	✓	-	✓	-	-
Cater [16]	✓	✓	✓	✓	-	-	-	-
CoPhy [3]	✓	-	✓	-	-	✓	-	-
CRAFT [4]	✓	✓	✓	✓	-	-	-	-
CLEVRER [2]	✓	✓	✓	✓	-	-	-	-
Physion [18]	✓	-	✓	-	-	-	-	-
Physion++ [19]	✓	-	✓	-	-	✓	-	-
ComPhy (ours)	✓	✓	✓	✓	✓	✓	✓	✓

TABLE 1: Comparison between ComPhy and other visual reasoning benchmarks. ComPhy is a physical reasoning dataset with a wide range of reasoning tasks for physical property learning and corresponding dynamic prediction.

2 Related Work

Physical Reasoning. Our research prominently aligns with contemporary advancements in the domain of physical reasoning benchmarks, as delineated by recent studies [4], [16], [18]–[20]. PHYRE [1] and its variant, ESPRIT [17], establish an environment where objects maneuver within a vertical 2D plane, influenced by gravitational forces. Each task within this framework is tethered to a distinct goal state, and the model seeks resolution by delineating initial conditions conducive to achieving said state. Conversely, CLEVRER [2] incorporates videos featuring multiple objects in motion, colliding on a planar surface, and poses natural language questions pertaining to the description, explanation, prediction, and counterfactual reasoning of the resultant collision events. CoPhy [3] encompasses experimental trials involving objects moving in 3D space under the influence of gravity, with a focal point on predicting object trajectories following counterfactual interventions upon initial conditions. CRIPP-VQA [21] introduces a challenge that emphasizes reasoning over physical properties such as mass and friction from a single video with simple primitive shapes, material and colors. Our work builds upon the original ComPhy dataset introduced in our prior work [5], extending it with more diverse physical scenes and real-world videos, which requires models to infer physical properties from a few physical interactions in reference videos. Compared to other previous datasets, ComPhy requires models to infer intrinsic properties from a limited array of video examples and draw dynamic predictions based on the identified properties.

Dynamics Modeling. Modeling the dynamics of physical systems has long been a focal point of research. This issue has been explored by some researchers through physical simulations, drawing inferences regarding crucial system- and object-level properties via statistical methodologies such as MCMC [22]–[24]. In contrast, others have proposed to directly ascertain the forward dynamics employing neural networks [25]. Owing to their object- and relation-centric inductive biases and efficacy, Graph Neural Networks (GNNs) [26] have been broadly applied in predicting forward dynamics across a diverse array of systems [27]–[30]. Our research combines the strengths of both approaches: initially inferring the object-centric intrinsic physical properties and subsequently predicting their dynamics predicated on these intrinsic properties.

Video Question Answering. Our research also pertains to the domain of video question answering, which responds to queries about visual content. Several benchmarks have been posited to address the task of video question answering, such as MarioQA [31], TVQA [32], and AGQA [15]. Nevertheless, these datasets primarily concentrate on comprehending human actions and activities rather than acquiring knowledge regarding physical events and properties, a competency crucial for robotic planning and control.

We summarize the differences between our extended ComPhy benchmark and other prior physical reasoning datasets in Table 1. Compared to our previous version [5], this work introduces more diverse simulated scenes and real-world videos. Notably, ComPhy remains the only dataset requiring models to infer physical properties from a sparse set of video examples, perform dynamics prediction, and answer compositional reasoning questions.

Few-shot Learning. Our research bears relevance to few-shot learning, which learns to classify images utilizing merely a few examples [33]–[36]. ComPhy mandates that models identify object property labels from a limited selection of video examples. Contrasting with the aforementioned works, reference videos in our approach do not furnish labels for objects' physical properties but exhibit more interactions among objects, thereby providing models with information to discern objects' physical properties.

3 Dataset

This section describes the dataset used in our benchmark. We build upon our prior work ComPhy [5], originally introduced in ICLR 2022, and present a significantly extended version. In addition to the synthetic split described in [5], we enrich the synthetic dataset with more diverse physical scenes and include a new real-world video dataset. First, we introduce video details and the task setup in Section 3.1. Subsequently, Section 3.2 delves into the different categories of questions, while Section 3.3 explores the underlying statistics and ensures balance. Finally, in Section 3.4, we introduce how we build the real-world data set.

3.1 Videos

Objects and Events. Following [11], objects in ComPhy are characterized by compositional appearance attributes,

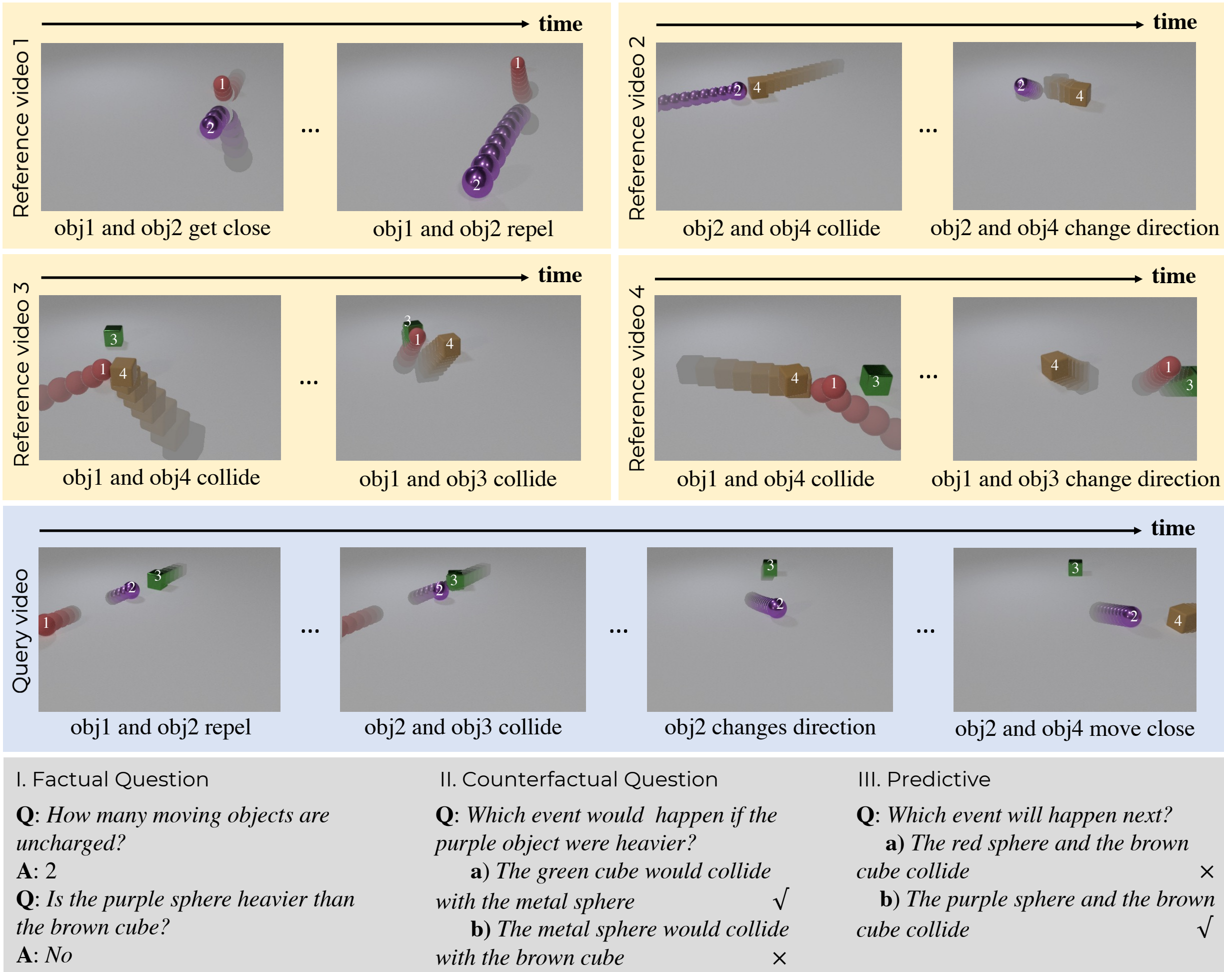

Fig. 2: Sample target video, reference videos and question-answer pairs from ComPhy.

including color, shape, and material. For ease of identification, each object in the videos is uniquely distinguishable based on these three characteristics. The dataset incorporates events such as *in*, *out*, *collision*, *attraction*, and *repulsion*. The basic concepts in ComPhy are derived from these object appearance attributes, events, and their compositionality.

Physical Properties. Previous benchmarks [2], [16] predominantly focused on visually perceptible appearance concepts like color and collision, discernible in a single frame. In contrast, our dataset, ComPhy, additionally explores the *intrinsic* physical properties of *mass* and *charge*, which are not directly discernible from an object's static appearance (Figure 1(a,b)). These properties are independent of visual features and can interact, resulting in more intricate and diverse dynamic scenarios. For simplicity, the dataset categorizes objects into discrete mass groups (*heavy*/ *light*) and charge categories (*positively* / *negatively charged*/ *uncharged*). While introducing additional continuous parameters like bounciness and friction is possible, the complexity could render the dataset overly intricate and hinder intuitive property inference from people.

Video Generation. Each target video designated for question-answering encompasses 3 to 5 objects, integrating a random compositionality of appearance attributes and physical properties. The videos are standardized to a duration of 5 seconds, with an extended simulation of the 6-th and 7-th seconds specifically for annotating questions for future prediction.

As in our prior work [5], the synthetic videos are generated in a two-step process using the Bullet physics engine and rendered via Blender. In the first step, we employ the Bullet physical engine [37] to simulate the movements of objects and their interactions with one another. Since Bullet does not officially support the effect of electronic charges, we add external forces between charged objects, whose values are inversely proportional to the square of the objects' distance, to simulated Coulomb forces. We assign a mass value of 1 to the *light* objects and a mass value of 5 to the *heavy* objects. We manually ensure that every reference video includes at least one interaction, such as collision, attraction, or repulsion, among objects, to provide sufficient information for inferring physical properties. Every object in the target video must appear in the reference videos at least once. The simulated object movements are then transmitted to Blender [38] for high-quality image sequences.

Task Setup. It presents a non-trivial challenge to design an evaluative framework that accurately assesses a model's capacity for physical reasoning because physical properties are not discernible within a static frame. A simplistic approach would involve associating physical attributes directly with object appearances like "*The red object is heavy*", "*The yellow object is light*" and then asking "*What would hap-*

pen if they collide?" However, this setting is flawed, as it fails to ascertain whether the model genuinely comprehends the physical properties or merely relies on memorizing visual cues. An ideal setup would demand a model to demonstrate human-like discernment of objects' properties from their motion and mutual interactions within dynamic scenes, and subsequently formulate relevant dynamic predictions.

To achieve this goal, We introduce a meta-framework for physical reasoning that pairs a target video with a limited set of reference videos, enabling models to infer physical properties. Questions are then formulated regarding these properties and underlying dynamics, as illustrated in Figure 2. Thus, each collection includes a target video, four reference videos, and numerous inquiries related to the target video. Notably, all objects within each collection maintain consistent visual attributes, including color, shape, and material, as well as intrinsic physical properties, specifically mass and charge.

Reference Videos. To enrich the visual content for physical property inferring, we supplement each target video with four reference videos. From the target video, we select 2 to 3 objects, assign them different initial velocities and positions, and orchestrate interactions such as attraction, repulsion, or collision. The reference videos, though lasting 2 seconds each for scalability, follow the same generation criteria as the target videos. These supplementary interactions help models deduce physical properties; for example, observing repulsion in Reference Video 1 of Figure 2 indicates that *object 1* and *object 2* possess the same electrical charges.

3.2 Questions

Inspired by the previous datasets [2], [11], we propose a question engine capable of generating questions that test *factual*, *predictive*, and *counterfactual* reasoning abilities.

Queries. *Factual questions* are open-ended, requiring concise answers in the form of a single word or short phrase, and assess a model's understanding and reasoning about objects' physical properties, visual attributes, events, and relationships. Building upon existing benchmarks [2], [4], our dataset (ComPhy) introduces novel and challenging factual questions focused specifically on the physical properties of charge and mass (See Figure 2 (I)). Predictive and counterfactual questions, conversely, adopt a multiple-choice format that critically evaluates the plausibility of each provided answer option. *Predictive questions* require models to analyze objects' physical properties and dynamics to forecast events in future video frames. *Counterfactual questions* investigate hypothetical scenarios where an object's physical properties (e.g., charge or mass) are altered, focusing on their impact on object dynamics (See Figure 2 (II)). This methodology contrasts with prior research [2], [16] that centered on object removal, emphasizing the divergent implications of changing physical properties for predicting motion instead.

Templates. We present typical question templates in Table 2. Examining the table reveals that these novel question templates incorporate diverse symbolic operators associated with physical properties. For example, phrases such as "*heavy moving spheres*" and "*charged cubes*" demand that models deduce the values of objects' physical properties. For counterfactual questions, we introduce novel conditions, such as "*If the cyan object were uncharged*" and "*If the sphere were lighter*". These conditions are designed to enable reasoning about the dynamics when a particular object possesses an alternative physical property.

Type	Template and Example
CUN1	If the *SA* were *MP*, *Q*? If the sphere were lighter, which event would not happen?
CUN2	If the *SA* were *CP*, *Q*? If the cube were uncharged, which event would happen?
Mass1	Is the *DA1 SA1* heavier than the *DA2 SA2*? Is the blue sphere heavier than the gray cube?
Mass2	Is the *DA1 SA1* lighter than the *DA2 SA2*? Is the blue sphere lighter than the gray cube?
CHR1	Are the *DA1 SA1* and the *DA2 SA2* oppositely charged? Are the blue sphere and the purple sphere oppositely charged?
CHR2	Are the *DA1 SA1* and the *DA2 SA2* with the same type of charge? Are the cube and the cylinder with the same type of charge?
CHR3	What are the *Hs* of the two objects that are charged? What are the colors of the two objects that are charged?
Query	What is the *H* of the *DA SA* that is *PA*? What is the color of the moving cylinder that is heavy?
Exist	Are there any *PA DA SA TI*? Are there any charged moving cube when the video ends?
Count	How many *PA DA SA* are there *TI*? How many heavy stationary spheres are there?

TABLE 2: Typical question templates and examples in ComPhy. *SA* denotes static attributes like "`red`"; *DA* denotes dynamic attributes, "`moving`"; *MP* denotes mass attributes like "`heavier`"; *Q* denotes question phrases like "`which of the following would happen`"; *CP* denotes charge attributes like "`uncharged`"; *H* denotes visible concepts like "`material`"; *PA* denotes physical attributes like `heavy` and `charged`; *TI* denotes time indicators like "`when the video ends`".

3.3 Balancing and Statistics

In total, ComPhy features 8,000 training sets, 2,000 for validation, and 2,000 for testing, with a total of 41,933 factual, 50,405 counterfactual, and 7,506 predictive questions constituting 42%, 50%, and 8% of the dataset, respectively. For simplicity, video sets will include a pair of charged objects only if charged objects are already present, and similarly, a video will contain a heavy object or none at all. We ensure that these few video examples are sufficiently informative to answer questions based on the questions' programs and the properties and interaction annotations in the videos. Specifically, for questions comparing mass or establishing charge relations, we meticulously confirm at least one interaction exhibited between the relevant objects.

3.4 Real-World Datasets

As shown in Fig. 3, we collect a new real-world video dataset to further estimate the capabilities of physical reasoning models. The construction of this dataset involves two key stages: real video collection and question annotation.

Real Video Collection. We capture a dataset consisting of 492 real-world videos using the iPhone's SLO-MO feature, which records high-definition slow-motion footage at 240 frames per second. These videos are organized into 123 sets, with 60 sets designated for training and 20 sets for

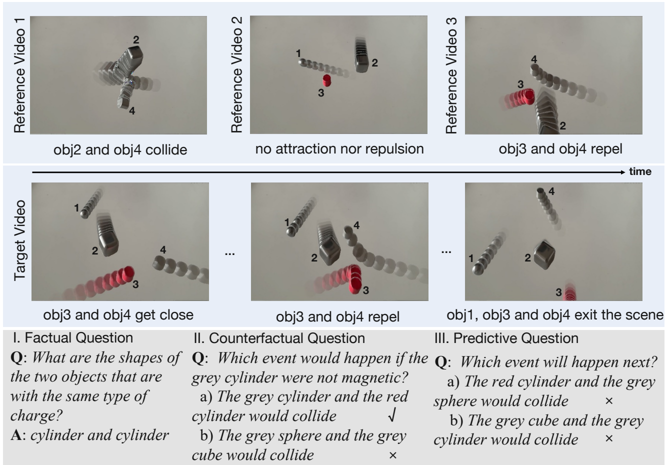

Fig. 3: Samples of real data. We collect real objects of different mass values and magnetism for extensive experiments, which have a significant effect on objects' motion and interaction.

validation. Each set comprises one target video featuring 3-4 objects interacting and 3 associated reference videos containing 2-3 objects each. This design mirrors the simulated video split, focusing on object interactions characterized by attributes like color (red, brown, grey), shape (cylinder, cube, sphere), magnetism (neutral, attractive, repulsive), and mass (heavy and light) in physical world environment.

Question Annotations. The static attributes, physical properties, and events in each video are initially annotated by an annotator and subsequently checked by another to ensure their correctness. We utilize a question engine similar to the one used in the simulated split to generate diverse questions, including counterfactual, predictive, and various property-based inquiries. The engine randomly selects from predefined templates and incorporates video annotations to create questions that explore various aspects of physical interaction, such as magnetism's effect on dynamics, the influence of mass, and the objects' static attributes. We collect 1,068 questions in total, including 776 for physical properties, 134 for counterfactual reasoning, and 158 for predictive future events. We provide more details on real-world videos in the supplementary material.

4 Experiments

In this section, we assess baseline models and conduct an in-depth analysis to comprehensively study ComPhy.

4.1 Baselines

We assess multiple baseline models on ComPhy, as displayed in Table 3. These baselines fall into four categories: bias-analysis models [39], video question answering models [40], [41], compositional reasoning models [42], [43], and large vision-language foundation models [9], [10], [44]. For a comprehensive comparison, we additionally introduce variant models that leverage both the target video and reference videos.

Biased Analysis Models. The first category of models is bias analysis models. These models predict answers without relying on visual input and aim to scrutinize the language bias present in ComPhy. In particular, the **Random** model randomly selects answers based on the question type and requires no training. The **Frequent** model selects the most frequently occurring answer in the training set for each question type, which requires no training phase. **Blind-LSTM** employs an LSTM [39] to encode the question and predict the answers without visual input; it is trained solely on the question-answer pairs from the dataset's training split to isolate language bias.

Visual Question-Answering Models. The second category of models encompasses visual question-answering models. These models answer questions based on input videos and questions. The **CNN-LSTM** model [45] is a simple question-answering model. It employs a ResNet-50 [46] to extract frame-level features, averaging them across the time dimension. We encode questions using the final hidden state from an LSTM [39]. The visual features and question embedding are concatenated to make answer predictions with two fully-connected layers. **HCRN** [41] is a widely adopted model that hierarchically models visual and textual relationships. Both **CNN-LSTM** and **HCRN** were trained (or fine-tuned, if using pre-trained components like ResNet) on the training split until convergence on the validation set.

Visual Reasoning Models. The third category, visual reasoning models, includes **MAC** [42], which decomposes visual question answering into several attention-focused reasoning steps, making predictions based on the hidden output of the final step. In contrast, **ALOE** [43] capitalizes on transformers [47] and object-centric representation to deliver cutting-edge results on CLEVRER. We use MONet [48] to extract visual representation for **ALOE**. Similar to the VQA models, both **MAC** and **ALOE** were trained (or fine-tuned from general pre-trained weights where applicable) on the dataset's training split.

Large Vision Language Models. The final model category is large vision language models [9], [10], [44], which have been trained on massive vision-language data and shown excellent performance on both language understanding and visual question answering. For **ALPRO**, we fine-tune the model with ComPhy's training set until they achieve satisfactory results on the validation set. For **GPT-4V** and **Gemini**, we evenly sample a fixed number of frames from each target video as visual input, pairing them with corresponding questions and a carefully crafted text prompt to guide the model in generating formatted answers.

Baselines with Reference Videos. We also introduce variations of existing baseline models that utilize both the target video and reference videos as input. We enhance **CNN-LSTM**, **MAC**, and **ALOE** to create **CNN-LSTM (Ref)**, **MAC (Ref)**, and **ALOE (Ref)** by incorporating the features of both reference videos and the target video as visual input. We uniformly sample 25 frames from each target video and 10 frames from each reference video.

Training and Evaluation Fairness. To ensure fair comparison, all models that underwent training or fine-tuning (**Blind-LSTM**, **CNN-LSTM**, **HCRN**, **MAC**, **ALOE**, **ALPRO**, and the 'Ref' variants) were trained on the same training split. We employed consistent hyperparameter tuning strategies (where applicable) and evaluated all models under identical conditions on the validation/test splits using

Methods	Factual	Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.
Random	29.7	51.9	22.6	49.7	9.1
Frequent	30.9	56.2	25.7	50.3	8.7
Blind-LSTM	39.0	57.9	28.7	55.7	12.5
CNN-LSTM [40]	46.6	59.5	29.8	58.6	14.6
HCRN [41]	47.3	62.7	32.7	58.6	14.2
MAC [42]	**68.6**	60.2	32.2	60.2	16.0
ALOE [43]	54.3	**65.9**	35.2	**65.4**	**20.8**
CNN-LSTM (Ref) [40]	41.9	59.6	29.4	57.2	12.8
MAC (Ref) [42]	**65.8**	60.2	30.7	60.3	14.3
ALOE (Ref) [43]	57.7	**67.9**	**37.1**	**67.9**	**22.2**
ALPRO [44]	45.2	56.9	27.2	53.7	14.4
GPT-4V [9]	42.2	60.7	**47.1**	51.1	8.9
Gemini [10]	37.7	46.5	22.7	49.2	6.3
Human Performance	90.6	88.0	75.9	80.0	52.9

TABLE 3: Evaluation of physical reasoning on ComPhy. Human performance is based on sampled questions. See Section 4.2 for more details. **Red** text and **blue** text indicate the first and the second best results.

the specified metrics. The zero-shot evaluation of **GPT-4V** and **Gemini** is reported separately and interpreted in light of their lack of dataset-specific fine-tuning.

We employ the conventional accuracy metric to assess the performance of various methods. In the case of multiple-choice questions, we provide both per-option accuracy and per-question accuracy. A question is deemed correct if the model answers all of its options correctly.

4.2 Evaluation on physical reasoning

The question-answering results of various baseline models are shown in Table 3. Notably, there exist discrepancies in the relative performances of models across different kinds of questions, which suggests that diverse reasoning skills are necessitated by the questions in ComPhy.

Factual Reasoning. To address factual questions in ComPhy, models must identify visual attributes, analyze motion trajectories, and infer physical properties of objects. The results indicate that the "blind" models, namely **Random**, **Frequent**, and **Blind-LSTM**, perform significantly poorly on ComPhy compared to other models integrating visual context and linguistic information. Additionally, we observe that video question-answering models and pre-trained large vision language models exhibit lower performance compared to visual reasoning models like **MAC** and **ALOE**. We attribute this discrepancy to the fact that they are typically tailored for tasks such as object classification, action recognition, and activity understanding rather than understanding physical events in ComPhy. Among these, **MAC** outperforms the rest baselines when answering factual questions, underscoring the effectiveness of its compositional attention mechanism and iterative reasoning processes.

Dynamcis Reasoning. A notable feature of ComPhy is its demand for models to generate counterfactual and future dynamic predictions by leveraging their identified physical properties to address posed questions. Among all the baseline models, we have observed that **ALOE (Ref)** consistently attains the highest performance levels in tasks involving counterfactual and future reasoning. We posit that this superior performance is attributable to the utilization of self-attention mechanisms and self-supervised object masking techniques, enabling the model to effectively capture spatio-temporal visual context and imagine counterfactual scenarios for answering questions.

Reasoning with Large Vision-Language Models. We also evaluate the performance of the recent large vision-language models, **ALPRO**, **Gemini** and **GPT-4V** on ComPhy, which were pre-trained on massive image/video-text pairs from the internet. Despite their strong performance on traditional visual question-answering benchmarks such as GQA [49], VQAv2 [45], MSRVTT-QA [50] and MSVD-QA [50], all of them underperform on ComPhy. The inferior performance of large vision-language models (LVLMs) is attributed to a gap in their training. These models are pretrained on internet data, which primarily focuses on object categories and semantic relations, lacking emphasis on physical commonsense. ComPhy underscores its value in addressing the gap and complementing the missing physical commonsense in existing visual question-answering benchmarks.

Reasoning with Reference Videos. The results reveal that **CNN-LSTM (Ref)** and **MAC (Ref)** perform comparably or slightly worse than their original counterparts, **CNN-LSTM** and **MAC**. While **ALOE (Ref)** shows a modest improvement over **ALOE**, the variant models do not exhibit substantial improvements when incorporating the reference videos as supplementary visual input. This phenomenon is likely due to these models' primary training on extensive datasets comprising videos and question-answer pairs, hindering their adaptability to ComPhy's novel context, which necessitates discerning new compositional visible and hidden physical properties from a limited number of examples.

Human Performance. To evaluate human performance in ComPhy, 14 participants with a basic understanding of physics and proficiency in English were tasked. After an initial warm-up through a series of demonstration videos and questions to confirm their comprehension of events and physical properties, they were assigned to answer 25 diverse question samples from ComPhy. Their accuracy rates are as follows: 90.6% for factual questions, 88.0% for predictive questions per option, 80.0% for counterfactual questions per option, 75.9% for predictive questions per question, and 52.9% for counterfactual questions per question.

Reasoning in the Real World. We evaluated the performance of various models on our collected real-world dataset by fine-tuning each model on the dataset's training split and evaluating on the validation split (see Table 4). Results indicate that ALOE achieves the highest accuracy on factual questions (61.6%), consistent with its strong performance observed in simulated scenarios. Notably, MAC shows a balanced performance across all question types, particularly excelling in predictive questions (57.1% per question accuracy). Interestingly, state-of-the-art general-purpose vision-language models such as GPT-4o-mini and Gemini significantly underperform compared to specialized models, reflecting substantial limitations in their ability to reason about physical interactions in real-world contexts. The substantial gap between human performance (exceeding 88% across all categories) and the evaluated models underscores the complexity and challenge of physical reasoning tasks.

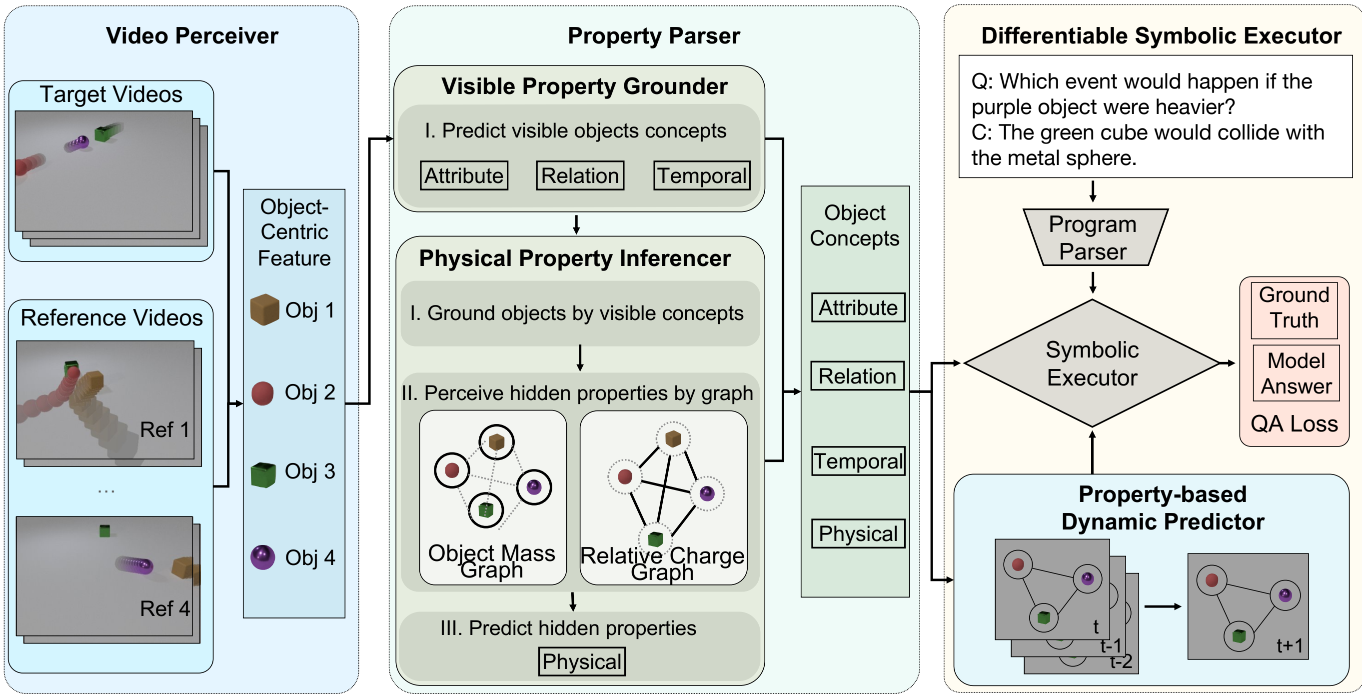

Fig. 4: The perception module detects objects' location and visual appearance attributes. The physical property learner learns objects' properties based on detected object trajectories. The dynamic predictor predicts objects' dynamics in the counterfactual scene based on objects' properties and locations. Finally, an execution engine runs the program parsed by the language parser on the predicted dynamic scene to answer the question.

Methods	Factual	Predictive per opt.	Predictive per ques.	Counterfactual per opt.	Counterfactual per ques.
Random	7.6	50.0	25.0	50.9	20.8
Frequent	41.7	53.6	28.7	50.0	23.9
Blind-LSTM	50.6	61.5	46.0	51.9	32.2
CNN-LSTM [40]	55.6	64.2	47.3	50.9	33.3
HCRN [41]	51.9	62.5	53.5	50.9	32.1
MAC [42]	58.9	60.9	57.1	52.8	35.8
ALOE [43]	60.8	60.6	42.4	47.1	28.7
CNN-LSTM (Ref) [40]	49.0	64.3	41.3	50.0	26.3
MAC (Ref) [42]	56.4	56.2	46.4	51.4	34.9
ALOE (Ref) [43]	61.6	61.4	42.8	51.6	32.1
ALPRO [44]	50.9	55.3	39.2	49.7	29.2
GPT-4o-mini [9]	42.6	49.6	23.2	47.5	26.0
Gemini [10]	32.5	57.7	23.1	52.1	29.8
Human Performance	90.0	95.0	90.0	94.4	88.9

TABLE 4: Evaluation of physical reasoning on the real video. Human performance is based on sampled questions.

5 Models

5.1 Model

In this section, we present Physical Concept Reasoner (PCR), a new physical reasoning model. It aims to comprehend objects' visible properties, infer hidden physical properties and events, and image corresponding physical dynamics by observing the videos and responding to the associated questions. Compared with our preliminary models [5], [8], it is able to infer hidden physical properties and predict corresponding property-based dynamics without explicit dense property annotations.

PCR can be factorized into different functional modules for physical reasoning in videos. As shown in Fig. 4, the model consists of five major modules: (1) video perceiver, (2) visible property grounder, (3) physical property inferencer, (4) property-based dynamic predictor, and (5) differentiable symbolic executor. When provided with a target video alongside four reference videos and a query, PCR employs a video perceiver to detect objects' spatial locations across frames and all videos. Subsequently, their trajectories are processed by the physical property inferencer to deduce their properties. Leveraging these data, the dynamic predictor forecasts object movements based on their physical traits. Lastly, a differentiable executor executes the program generated by a language parser [6], [47], utilizing the predicted object motions to answer the query. Note that the object-centric representation and outputs of various modules are maintained in a differentiable manner, enabling direct optimization of each module through backpropagation when answering video-related questions.

5.1.1 *Video Perceiver*

Object Tracking and Alignment. Given a target video and 4 reference videos, the video perceiver in PCR is responsible to track objects in every video and align them across different videos. The first step is to track objects in the videos. At the t-th frame, our model first applies a regional proposal network [51], [52] to detect all objects $\{b_i^t\}_{i=1}^{N_t}$, where N_t denote the object proposal number. The video perceiver then get a set of object trajectories $\{o_n\}_{n=1}^{N}$, where N is the number of object trajectories, $o_n = \{b^t\}_{t=1}^{T}$ and T is the number of frames. Similar to [5], [53], we first define the connection score $s_{cnn}(b_i^t, b_j^{t+1})$ between two proposals b_i^t and b_j^t in connective frames as

$$s_{cnn}(b_i^t, b_j^{t+1}) = s_c(b_i^t) + s_c(b_j^{t+1}) + \mathrm{IoU}(b_i^t, b_j^{t+1}), \quad (1)$$

where $s_c(b_i^t)$ is the confidence score predicted by the region proposal network and IoU denotes the interaction over union between two proposals. We define the connection score of a candidate object trajectory $o_n=\{b_n^t\}_{t=1}^T$ as $E(o_n) = \sum_{t=1}^{T-1} s_{cnn}(b_n^t, b_n^{t+1})$. We select the set of object trajectories $\{o_n\}_{n=1}^N$ with the highest connection scores and solve the problem with a linear sum assignment [54]. We then align objects in reference videos to the target videos with the predicted static visual attributes, color, shape, and material. Objects in reference videos are assigned to objects in the target video that has the most similar predicted labels.

Object-Centric Representation. We use a set of object-centric features to represent the videos for physical reasoning. Specifically, we compute the **averaged visual regional features** ($\mathbf{f}_n^v \in \mathbf{R}^{\mathbf{D_v}}$) from the faster-RCNN [52] for static visual appearance attributes like *shape*, *color* and *material*, where D_v equals to 512 and is the regional feature's dimension from ResNet-34. We adopt the **temporal trajectory features** ($\mathbf{f}_n^t \in \mathbf{R}^{\times \mathbf{D_t}}$) for predicting temporal concepts like *in* and *out*, where $D_s = T \times 4$ is the concatenation of the object location b_n^t across all T frames. Since we can only infer objects' physical property values from their movement and interaction, we use a set of **aligned trajectory features** for physical property inference. For the n-th object in the target video, we represent it with $\mathbf{p}_n$ and $\{\mathbf{p}_{n,r}\}_{r=1}^R$, where $\mathbf{p}_{n,r}^t$ and $\mathbf{p}_{n,r}$ are the concatenation of the object coordinates (x_n^t, y_n^t) along all T frames in the target video and the r-th reference video. R equals to 4 and is the number of the reference videos. We add all the objects without appearance in the specific reference videos with zero vectors.

We use the **interaction feature** ($f_{i,j,t}^{int} \in \mathbf{R}^{D_{int}}$) for prediction the the collision event between the i-th and the j-th objects at the t-th frame. we define $f_{i,j,t}^{int} = f_{i,j,t}^u || f_{i,j,t}^{loc}$, where $f_{i,j,t}^u$ is the ResNet feature of the union region of the i-th and j-th objects at the t-th frame and $f_{i,j,t}^{sp}$ is a spatial embedding for correlations between bounding box trajectories. We define $f_{i,j,t}^{sp} = \mathrm{IoU}(s_{i,t}, s_{j,t}) || (s_{i,t} - s_{j,t}) || (s_{i,t} \times s_{j,t})$, where $s_{i,t} = ||_{k=t-2}^{t+2} b_i^k$ is the concatenated segment of the i-th object centering at the t-th frame. It concatenates the intersection over union (IoU), difference $(-)$, and multiplication $(\times)$ of the normalized trajectory coordinates for the i-th and j-th objects centering at the t-th frame. For the collision event in the future and counterfactual scenes, we predict the collision event based on the $f_{i,j,t}$ only since there is no RGB image for extracting the $f_{i,j,t}^u$ feature.

5.1.2 Visible property grounder

The visible property grounder grounds objects' visible properties like *color*, *shape*, and *collision* onto the objects extracted by the video perceiver. PCR accomplishes this by aligning the representations of objects and events with learned concept embeddings in PCR. For example, to predict the n-th object is *red* or not, we use a confidence score s_n^{red}. We define $s_n^{red} = \left[\cos(c^{\mathrm{red}}, m_{sa}(f_n^v)) - \delta\right]/\lambda$, where c^{red} is a vector, representing the concept *red*, m_{sa} is a fully-connected layer, mapping the object feature f_n^v to the color space. cos calculates the cosine similarity between the two vectors. δ and λ_{sa} are constant scalars, controlling the value range of s_{red}. Similarly, we predict two objects collide at the t-th frame with $s_{i,j,t}^{cl}$, where $s_{i,j,t}^{cl}$ equals to $s_{i,j,t}^{cl} = \left[\cos(c^{\mathrm{cl}}, m_{cl}(f_{i,j,t}^{int})) - \delta\right]/\lambda$. c^{cl} represents the concept vector for the *collision* event and m_{cl} is a fully-connected layer transforming $f_{i,j,t}^{int}$ into the desired space.

5.1.3 Physical Property Inferencer

At the heart of our model, the Physical Property Inferencer (PPI) handles intricate and composite physical interactions by analyzing object motion trajectories extracted from both reference and target videos. The PPI utilizes a graph neural network [55] to predict mass and relative charge for each object pair, where node features capture object-centric properties (such as mass), and edge features encode pairwise properties (such as relative charge). The PPI employs a series of message-passing operations on the input trajectories of N objects within the video. The process is described by:

$$\begin{aligned} \mathbf{v}_n^0 &= f_{emb}(\mathbf{p}_n^t), \quad \mathbf{e}_{n_1,n_2}^l = f_{rel}^l(\mathbf{v}_{n_1}^l, \mathbf{v}_{n_2}^l), \\ \mathbf{v}_{n_1}^{l+1} &= f_{enc}^l\Big(\sum_{n_1 \neq n_2} \mathbf{e}_{n_1,n_2}^l\Big), \end{aligned} \tag{2}$$

Here, $f_{(\ldots)}$ are functions implemented by fully-connected layers. We then use two fully-connected layers to predict the output mass label $f_v^{pred}(\mathbf{v}_i^2)$ and edge charge label $f_e^{pred}(\mathbf{e}_{i,j}^1)$, respectively. Notably, the PPI is not trained in a fully-supervised manner but is optimized via leveraging the gradients from differentiable question answering. The complete physical property of a set of videos can be represented as a fully connected property graph, where each node corresponds to an object that appears in at least one video within the set. Meanwhile, each edge indicates whether the connected nodes possess the same, opposite, or no relative charge (i.e. it signifies whether one or both objects are charge-neutral). In Figure 4, we illustrate that the physical property inferencer (PPI) independently predicts the objects' properties in each reference video, covering only part of the property graph. To align predictions across different nodes and edges, we utilize the static attributes of objects identified by the video perceiver. By aggregating the sub-graphs generated from each video in the set through max-pooling over nodes and edge predictions, we obtain the complete object properties graph.

5.1.4 Property-based Dynamic Predictor

To predict objects' positions at the t + 1 frame, based on their full trajectories and properties (mass and charge) at the t-th frame, we employ a dynamic predictor implemented by graph neural networks. For the n-th object at the t-th frame, we represent it with $\mathbf{o}_n^{t,0} = ||_{t-3}^t(x_n^t, y_n^t, w_n^t, , h_n^t, m_n)$, using a concatenation of its object location (x_n^t, y_n^t), size (w_n^t, h_n^t) and the mass prediction (m_n) by the Physical Property inferener over a history window of 3. By incorporating a history of object locations rather than solely relying on the location at the t-th frame, we encode object velocity and accommodate for perception errors. Specifically, we have

$$\begin{aligned} \mathbf{h}_{n_1,n_2}^{t,0} &= \sum_k z_{n_1,n_2,k} g_{emb}^k(\mathbf{o}_{n_1}^{t,0}, \mathbf{o}_{n_2}^{t,0}), \\ \mathbf{o}_{n_2}^{t,l+1} &= \mathbf{o}_{n_2}^{t,l} + g_{rel}^l\left(\sum_{n_1}^{n_1 \neq n_2} (\mathbf{h}_{n_1,n_2}^{t,l})\right), \\ \mathbf{h}_{n_1,n_2}^{t,l+1} &= \sum_k z_{n_1,n_2,k} g_{enc}^{k,l}([\mathbf{o}_{n_1}^{t,l+1}, \mathbf{o}_{n_1}^{t,0}], [\mathbf{o}_{n_2}^{t,l+1}, \mathbf{o}_{n_2}^{t,0}]), \end{aligned} \tag{3}$$

where the variable $k \in 0, 1, 2$ represents whether the two connected nodes carry the same, opposite, or no relative charge. The k-th element of the one-hot indication vector $\mathbf{z}n_1, n_2$ is denoted as $\mathbf{z}_{n_1,n_2,k}$. The message-passing steps are indicated by $l \in [0, 1]$ and functions $g_{(...)}$ are implemented through fully-connected layers. For predicting object location and size in the $(t+1)$-th frame, we employ a function comprising a single fully-connected layer, $g_{pred}(\mathbf{o}^{t,2}_{n_2})$. To forecast future frames for predictive questions, we initialize the dynamic predictor with the last three frames of the target video and iteratively predict subsequent frames by feeding the generated predictions back into the model. For counterfactual queries, we use the first three frames of the target video as input, updating the predicted objects' mass labels (m_i) and the corresponding one-hot indicator vector $\mathbf{z}_{i,j}$ accordingly, to obtain physical predictions with counterfactual properties labels.

5.1.5 Differentiable Symbolic Executor

The differentiable symbolic executor first adopts a program parser [6], [56] to transform the input question into a series of program operations. The program parser is trained in a fully-supervised manner as in [5], [6]. The executor then executes the symbolic operations on the latent object-centric representation derived from the other modules and the output of the final operator serves as the solution to the question. We adopt a probabilistic approach, similar to the methodology proposed in [7], to represent the object states, events, and results of all operators during the training phase. This probabilistic representation allows for a differentiable execution process, considering the latent representations derived from both the observed and predicted scenes. As shown in the dotted lines of Figure 4, it becomes feasible to optimize the video perceiver, visible property grounder, physical property inferencer, and property-based dynamic predictor within the symbolic execution procedure.

5.2 Training Mechanisms

The proposed PCR features multiple functional modules, and optimizing these modules presents great challenges due to several factors: 1) the lack of dense property annotations for both visible concepts and hidden physical properties, 2) the complexity of physical properties and their interaction with other visible properties, and 3) fewer training examples compared to the previous physical reasoning dataset CLEVRER. To address these challenges, we propose two novel training mechanisms for model optimization: 1) Curriculum Learning for Physical Reasoning in Section 5.2.1, and 2) Learning by Imagination in Section 5.2.2.

5.2.1 Curriculum Learning for Physical Reasoning

We design a novel curriculum learning mechanism to optimize the PCR introduced in Section 5.1. We first train a program parser to parse the question and answers into executable programs with a sequence-to-sequence model [56]. In lesson 1, we filter out and select the factual questions without physical property description to learn an initial model to ground visible properties like *colors*, *shapes*, and *collisions*. In lesson 2, we include all the factual questions to teach the model to infer objects' physical properties with the physical property inferencer. During this lesson, we align the objects' dynamics in different videos and property predictions with the static visible property label prediction from lesson 1. In lesson 3, we utilize the property prediction results from the last lesson as pseudo labels to train a property-based dynamic predictor, which predicts objects' dynamics in the counterfactual and predictive scenes. Finally, we fine-tune all components in an end-to-end manner with all question-answer pairs from the training set.

5.2.2 Learning by Imagination

One key challenge for the PCR on the ComPhy dataset is the complexity of its video scenarios compared to previous datasets like CLEVRER [2], which has 152,572 question-answer pairs, while ComPhy has only 55,764 pairs but with more variances in physical property variances. To improve the training efficiency, we introduce a new training mechanism, named *Learning by Imagination*. Specifically, when a counterfactual question states "*Which event would happen if the purple object were heavier?*", it implicitly indicates that "*there is a purple object*" and "*the purple object is **not** heavy*". These implicit statements can be transformed into executable programs to enhance the learning of both the visible property grounder and the physical property inference introduced in Section 5.1.2. Note that the ability to learn and reason in counterfactual situations is a hallmark of human thought [57], [58].

5.3 Performance Analysis

Effectiveness of Physical Property Inference. We compare the proposed PCR with the previous neuro-symbolic method CPL [5] in table 5. We can see that the PCR performs better on all kinds of questions compared to the baseline methods in Table 3. This shows the effectiveness of neuro-symbolic models for physical reasoning. Second, although our PCR has no reliance on physical property labels and visual attribute labels during training, it can achieve comparable performance to the previous model CPL that requires dense annotation for videos on factual questions.

One distinguished advantage of PCR over end-to-end models [41], [43] is it enables step-by-step investigations and thorough analysis for physical concept learning in videos. We compare the model prediction from the PCR with the ground-truth labels and calculate the accuracy. Table 7 lists the result. We found that our model could effectively grasp visible concepts like "*colors*", "*moving*" and "*collisions*". We also notice that the physical property inferencer in PCR can achieve reasonable accuracy on physical concepts like "*mass*" and "*charge*". which shows PCR is able to learn physical properties from objects' trajectories and interactions. However, we also notice the performance gap between the hidden physical properties and the visible properties, which indicates that the bottleneck of the performance on factual questions lies in the hidden physical property inference.

Effectiveness of Dynamics Reasoning. We further compare our PCR with CPL and its variant CPL-DPI for dynamic reasoning in table 5. CPL-DPI follows the previous model NS-DR [6] to adopt dynamic particle interaction networks [59] (DPI) for dynamic prediction. Note that DPI adopts graph neural networks for dynamic prediction without considering the variance of physical properties. Compared CPL and PCR with CPL-DPI, we can see the importance of modeling

Methods	Factual	Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.
CPL-DPI [59]	-	73.3	50.8	61.1	16.6
CPL [8]	**80.5**	75.3	56.4	68.3	**29.1**
PCR	76.0	**80.0**	**62.0**	**70.0**	29.0

TABLE 5: Evaluation of PCR on the test set of ComPhy. The best performance is in boldface.

Methods	Factual	Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.
PCR w/o R	68.7	52.0	28.1	54.9	28.0
PCR w/o CI	70.3	51.2	24.4	54.0	28.0
PCR	**78.3**	75.0	56.5	70.5	50.2

TABLE 6: Ablation study of PCR on the validation set of ComPhy. The best performance is in boldface.

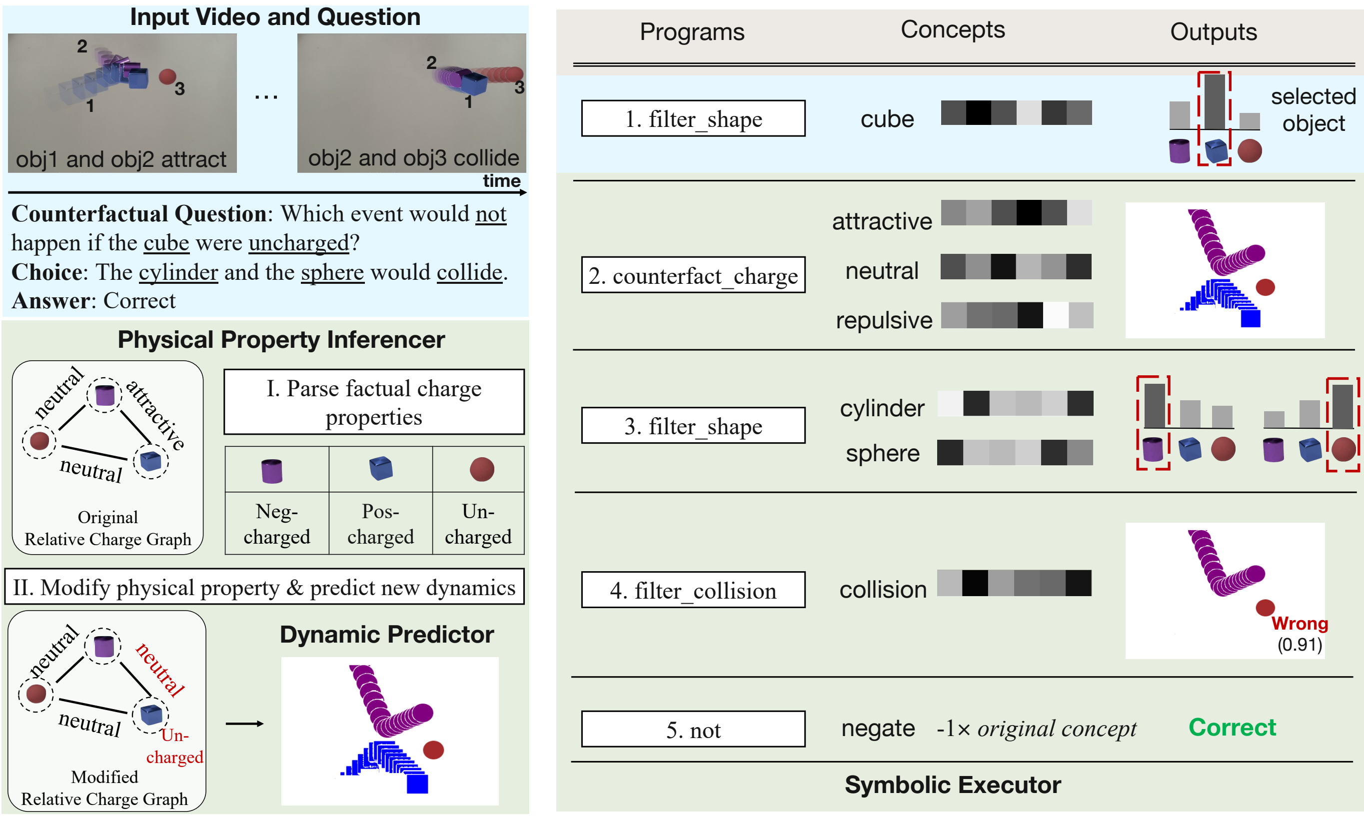

Fig. 5: A qualitative example of PCR on ComPhy. The left-up blue box shows the original video and a counterfactual question to answer. The right half table shows the executable program sequence parsed from the question with concepts related to it and outputs after execution. Specifically, the left-down chart illustrates the execution process of PCR for the program "counterfact charge": 1. PCR utilizes a PPI to parse factual charge properties of objects in the scene; 2. PCR modifies their properties according to the counterfactual concept and predict new dynamics using a dynamic predictor.

Methods	Static Attributes			Dynamic Attributes		Events			Physical Properties	
	Color	Shape	Material	Moving	Stationary	In	Out	Collision	Mass	Charge
PCR w/o R	91.0	91.8	92.8	83.3	85.2	85.6	81.8	86.8	79.5	45.0
PCR w/o CI	91.9	89.1	94.0	82.6	84.9	86.3	81.4	89.0	80.8	44.8
PCR	96.3	96.8	97.1	81.5	86.0	85.5	70.3	88.0	86.8	68.1

TABLE 7: Evaluation of video concept learning on the validation set.

mass and charges on nodes and edges of the graph neural networks for dynamic prediction. Moreover, compared with CPL, PCR achieves better performance on predictive questions and performs competitively on counterfactual questions, which shows the effectiveness of the differentiable executor for the optimization of property-based dynamic predictor, physical property inference, and visible property grounder. The performance of our approach surpasses that of the baselines listed in table 3, particularly in counterfactual and predictive questions. This achievement demonstrates the model's capability to predict the movements of objects in counterfactual and future scenarios, based on the identification of their underlying physical properties.

Furthermore, our evaluation in Section 4.2 highlights a noticeable disparity between the performance of our model, PCR, and human performance, particularly in the domain of counterfactual reasoning. We observed that PCR's dynamic predictor still exhibits limitations when it comes to long-term dynamic prediction. This indicates that further enhancements to the dynamic predictor could potentially yield even higher performance improvements for PCR.

Ablation Study. We conduct a series of ablation studies to prove the effectiveness of the PCR in table 6 and table 7. **PCR w/o R** denotes learning the property model without using reference videos. **PCR w/o CI** denotes the model without counterfactual imaging. We want to answer the following questions. We report the question-answering accuracy in table 6 and the concept classification accuracy in table 7. Comparing with **PCR w/o R** and **PCR**, we can see that reference videos provide important information for concept identification especially for the physical properties, *mass* and *charge* in table 7 and constantly improve question-answering performance in different kinds of questions. Comparing **PCR w/o CI** and **PCR**, we can see that the counterfactual imaging mechanism in Section 5.2.2 can improve the models' abilities in physical property identification in table 7, showing its effectiveness to learn physical reasoning.

More Diverse Physical Simulated Scenes. To better evaluate the model's performance on diverse physical scenes, we have simulated a diverse set of ComPhy dataset. The diverse set introduces 13 distinct object categories—including items such as mugs, pots, chairs, and more—in contrast to the primitive shapes used in the original benchmark. In addition, we incorporate 9 varied backgrounds with realistic textures and lighting conditions, and increase the total number of possible question-answer pairs to 175. The new objects span a wider range of shapes and material properties. These enhancements allow for a richer set of physical interactions, enabling the simulation of complex, compositional events.

We have also conducted new experiments on these new scenes, and the performance results can be seen in Table 8. From the table, we have the following observations. First, we can see that our model (PCR) still constantly outperforms the other baselines, showing the effectiveness of using neuro-symbolic models for physical reasoning. Second, we also observe that the average model performance is worse than their accuracy on the original data in Table 3 and Table 5. We believe that the reason is that the new physical scenes have provided more diverse physical interaction among the objects, making it more challenging for the AI models. We have also conducted a human study similar to the original ComPhy paper. The accuracy for different kinds of questions is 88.6 for factual questions, 73.7 for predictive questions, and 78.9 for counterfactual questions, much better than existing models in Table 8. This shows that although the scenes become more diverse, people can still handle these questions well. We provide more details on diverse simulated videos in the supplementary material.

Generalization to Real-World Scenes. We evaluated the performance of our new model, PCR, on the real-world dataset. It achieved 63.5% accuracy on factual questions, 70.4% on predictive questions (per option), 62.7% on predictive questions (per question), 54.6% on counterfactual questions (per option), and 36.5% on counterfactual questions (per question). From Table 4, PCR consistently outperforms the MAC model across all question types, demonstrating its enhanced effectiveness in physical reasoning.

Qualitative Case Study. As shown in Figure 5, PCR can transfer the question query into a series of executable operators, perceive objects' visible properties, infer objects' physical properties, and predict their corresponding dynamics to correct answer the question. Note that such step-by-step investigation is not possible in previous end-to-end models like **MAC** and **ALOE**, showing the transparency and interpretability of our PCR.

Methods	Factual	Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.
Random	1.8	50.1	22.9	48.1	24.0
Frequent	15.7	50.0	0.0	50.0	0.0
Blind-LSTM	43.2	50.3	25.0	49.2	23.2
CNN-LSTM [40]	49.6	52.8	29.9	55.7	29.7
HCRN [41]	51.5	**56.3**	**34.1**	51.9	30.1
MAC [42]	**51.7**	50.4	28.9	51.9	26.3
ALOE [43]	46.9	52.4	29.0	51.5	28.6
CNN-LSTM (Ref) [40]	49.7	51.4	23.3	55.6	30.5
MAC (Ref) [42]	50.6	51.9	33.3	50.8	25.2
ALOE (Ref) [43]	48.6	51.2	26.1	52.9	27.2
ALPRO [44]	47.1	51.8	28.9	52.6	28.4
GPT-4o-mini [9]	42.5	50.0	29.2	**58.8**	**30.7**
Gemini [10]	34.2	50.3	25.7	49.4	30.6
PCR (ours)	**68.4**	**58.3**	**34.9**	**60.3**	**32.8**
Human Performance	88.6	82.9	73.7	88.2	78.9

TABLE 8: Evaluation of physical reasoning on ComPhy-DIV. Human performance is based on sampled questions. See the text for more details. **Red** text and **blue** text indicate the first and second best results other than human performance.

5.4 Discussion on Intergrate PCR with LVLMs

Combining PCR with LVLMs offers a powerful paradigm for enhancing both robustness and flexibility. First, LVLMs can replace or augment the program parser in PCR via in-context learning, improving program synthesis for diverse linguistic formulations. Second, LVLMs' broad world knowledge can be invoked through a dedicated large language model-based module to handle commonsense reasoning tasks that lie outside PCR 's original training distribution. Finally, LVLMs can act as high-level controllers, orchestrating PCR 's neural modules alongside external modules to seamlessly tackle novel tasks. This integration leverages the precise, learned functionality of PCR and the generalist capabilities of LVLMs, yielding a more versatile and powerful system. We provide more experiments and analysis of integration of PCR and LVLMs in the supplementary material.

6 Conclusions

In this paper, we introduce the Compositional Physical Reasoning benchmarks, which challenge models to infer hidden physical properties such as mass and charge from limited video observations and leverage this information to predict dynamics and answer structured questions. Our evaluation of state-of-the-art models on ComPhy reveals substantial limitations in their ability to reason about these hidden attributes. We also propose a neuro-symbolic framework, PCR, that integrates object-centric representations with modular reasoning to jointly learn and infer both visible and hidden physical properties. We further present a real-world dataset to evaluate the generalization of physical reasoning models beyond simulation. Our findings highlight the critical role of hidden physical properties in dynamic scene understanding and expose the gap between current model capabilities and human-level reasoning, paving the way for more robust and generalizable physical reasoning in AI systems.

REFERENCES

[1] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick, "Phyre: A new benchmark for physical reasoning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. 1, 3

[2] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," in *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 5, 10

[3] F. Baradel, N. Neverova, J. Mille, G. Mori, and C. Wolf, "Cophy: Counterfactual learning of physical dynamics," in *International Conference on Learning Representations*, 2020. 1, 3

[4] T. Ates, 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*, 2020. 1, 3, 5

[5] Z. Chen, J. Mao, J. Wu, K.-Y. K. Wong, J. B. Tenenbaum, and C. Gan, "Grounding physical concepts of objects and events through dynamic visual reasoning," in *International Conference on Learning Representations*, 2021. 2, 3, 4, 8, 10

[6] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding," in *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 8, 10

[7] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision," in *International Conference on Learning Representations*, 2019. 2, 10

[8] Z. Chen, K. Yi, Y. Li, M. Ding, A. Torralba, J. B. Tenenbaum, and C. Gan, "Comphy: Compositional physical reasoning of objects and events from videos," in *International Conference on Learning Representations*. 2, 8, 11

[9] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv*, 2023. 2, 6, 7, 8, 12

[10] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv*, 2023. 2, 6, 7, 8, 12

[11] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017. 3, 5

[12] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[13] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[14] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvqa+: Spatio-temporal grounding for video question answering," in *Tech Report, arXiv*, 2019. 3

[15] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Agqa: A benchmark for compositional spatio-temporal reasoning," in *CVPR*, 2021. 3

[16] R. Riochet, 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*, 2018. 3, 4, 5

[17] N. F. Rajani, 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 *ACL*, 2020. 3

[18] D. M. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. F. Tung, R. Pramod, C. Holdaway, S. Tao, K. Smith, F.-Y. Sun *et al.*, "Physion: Evaluating physical prediction from vision in humans and machines," *arXiv*, 2021. 3

[19] H.-Y. Tung, M. Ding, Z. Chen, D. Bear, C. Gan, J. B. Tenenbaum, D. L. Yamins, J. E. Fan, and K. A. Smith, "Physion++: Evaluating physical scene understanding that requires online inference of different physical properties," *arXiv*, 2023. 3

[20] R. Girdhar and D. Ramanan, "Cater: A diagnostic dataset for compositional actions and temporal reasoning," in *ICLR*, 2020. 3

[21] M. Patel and T. Gokhale, "Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering," in *EMNLP*, 2022. 3

[22] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, pp. 18 327–18 332, 2013. 3

[23] J. B. Hamrick, P. W. Battaglia, T. L. Griffiths, and J. B. Tenenbaum, "Inferring mass in complex scenes by mental simulation," *Cognition*, vol. 157, pp. 61–76, 2016. 3

[24] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," *Advances in neural information processing systems*, vol. 28, pp. 127–135, 2015. 3

[25] A. Lerer, S. Gross, and R. Fergus, "Learning physical intuition of block towers by example," in *International conference on machine learning*. PMLR, 2016, pp. 430–438. 3

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017. 3

[27] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, and k. kavukcuoglu, "Interaction networks for learning about objects, relations and physics," in *Advances in Neural Information Processing Systems*, vol. 29, 2016. 3

[28] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum, "A compositional object-based approach to learning physical dynamics," *arXiv preprint arXiv:1612.00341*, 2016. 3

[29] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia, "Learning to simulate complex physics with graph networks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8459–8468. 3

[30] Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, and A. Torralba, "Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids," in *ICLR*, 2019. 3

[31] J. Mun, P. Hongsuck Seo, I. Jung, and B. Han, "Marioqa: Answering questions by watching gameplay videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3

[32] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," in *EMNLP*, 2018. 3

[33] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *NeurIPS*, 2016. 3

[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017. 3

[35] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018. 3

[36] C. Han, J. Mao, C. Gan, J. B. Tenenbaum, and J. Wu, "Visual Concept Metaconcept Learning," in *NeurIPS*, 2019. 3

[37] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021. 4

[38] B. O. Community, "Blender - a 3d modelling and rendering package," Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org 4

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 6

[40] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015. 6, 7, 8, 12

[41] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *CVPR*, 2020. 6, 7, 8, 10, 12

[42] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *ICLR*, 2018. 6, 7, 8, 12

[43] D. Ding, F. Hill, A. Santoro, and M. Botvinick, "Attention over learned object embeddings enables complex visual reasoning," *arXiv*, 2020. 6, 7, 8, 10, 12

[44] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963. 6, 7, 8, 12

[45] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015. 6, 7

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 6

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. 6, 8

[48] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, "Monet: Unsupervised scene decomposition and representation," *arXiv preprint arXiv:1901.11390*, 2019. 6

[49] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019. 7

[50] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1645–1653. 7

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *CVPR*, 2017. 8

[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015. 8, 9

[53] Z. Chen, L. Ma, W. Luo, and K.-Y. K. Wong, "Weakly-supervised spatio-temporally grounding natural sentence in video," in *ACL*, 2019. 8

[54] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, 1957. 9

[55] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2688–2697. 9

[56] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015. 10

[57] N. Van Hoeck, P. D. Watson, and A. K. Barbey, "Cognitive neuroscience of human counterfactual reasoning," *Frontiers in human neuroscience*, 2015. 10

[58] D. Buchsbaum, S. Bridgers, D. Skolnick Weisberg, and A. Gopnik, "The power of possibility: Causal learning, counterfactual reasoning, and pretend play," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2012. 10

[59] Y. Li, J. Wu, J.-Y. Zhu, J. B. Tenenbaum, A. Torralba, and R. Tedrake, "Propagation networks for model-based control under partial observation," in *ICRA*, 2019. 10, 11