PACS: A Dataset for Physical Audiovisual CommonSense Reasoning

Samuel Yu[1], Peter Wu[2], Paul Pu Liang[1], Ruslan Salakhutdinov[1], and Louis-Philippe Morency[1]

[1] Carnegie Mellon University
[2] University of California, Berkeley

Abstract. In order for AI to be safely deployed in real-world scenarios such as hospitals, schools, and the workplace, it must be able to robustly reason about the physical world. Fundamental to this reasoning is *physical common sense*: understanding the physical properties and affordances of available objects, how they can be manipulated, and how they interact with other objects. Physical commonsense reasoning is fundamentally a multi-sensory task, since physical properties are manifested through multiple modalities - two of them being vision and acoustics. Our paper takes a step towards real-world physical commonsense reasoning by contributing PACS: the first audiovisual benchmark annotated for physical commonsense attributes. PACS contains 13,400 question-answer pairs, involving 1,377 unique physical commonsense questions and 1,526 videos. Our dataset provides new opportunities to advance the research field of physical reasoning by bringing audio as a core component of this multimodal problem. Using PACS, we evaluate multiple state-of-the-art models on our new challenging task. While some models show promising results (70% accuracy), they all fall short of human performance (95% accuracy). We conclude the paper by demonstrating the importance of multimodal reasoning and providing possible avenues for future research.

1 Introduction

To safely interact with everyday objects in the real world, AI must utilize physical commonsense knowledge about everyday objects: including their physical properties, affordances, how they can be manipulated, and how they interact with other physical objects [6,29]. Humans use *physical commonsense reasoning* in all facets of day-to-day life, whether it is to infer properties of previously unseen objects ("the water bottle over there is made of plastic, not glass"), or to solve unique problems ("I can use a puffy jacket in place of my missing pillow") [6]. This type of general understanding of object interactions is necessary in building robust and complete AI systems that can be safely deployed in the real world (e.g., a package delivery robot needs to treat heavier or lighter objects differently).

Physical commonsense reasoning is fundamentally a multi-sensory task, as physical properties are manifested through multiple modalities, including vision and acoustics [13,40,65]. If two objects appear similar visually, audio can provide

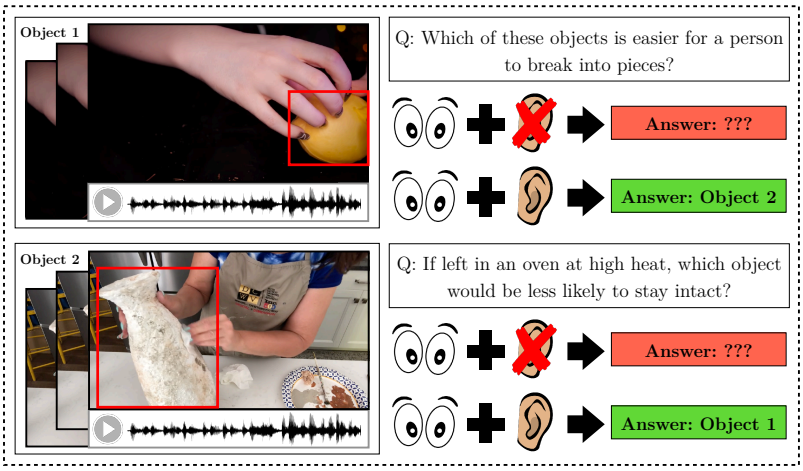


Fig. 1: PACS is the first audiovisual benchmark annotated for physical commonsense attributes, containing 13,400 question-answer pairs, 1,526 videos, and 1,377 unique questions. By benchmarking state-of-the-art unimodal and multimodal models to highlight *where* and *why* current models fail, PACS provides new opportunities to advance the research of physical reasoning through studying multimodal reasoning. This figure shows two example datapoints from PACS, with each datapoint containing a question and a pair of objects (in this figure, object 1 is a plastic lemon and object 2 is a ceramic vase). To view the video clips, please see the supplementary material.

valuable information to distinguish the physical properties between these objects. For example, in Figure 1, instead of plastic, object 1 could be mistaken for squishy foam, and instead of ceramic, object 2 could be mistaken for painted plastic, glass, or even paper. Without the necessary audio information, this could result in the erroneous answer that object 1 is easier to break than object 2. In the real world, this misunderstanding may lead to the damaging or mishandling of an object. Therefore, to enable physical commonsense reasoning in AI, it is essential for these models to reason across both audio and visual modalities.

Recent work has explored the use of vision and/or text to understand basic physical properties [30,32,35,50,61,67], or benchmark *physical commonsense* in language [6,16]. Our work complements these previous settings by adding the acoustic modality as part of the problem formulation. Furthermore, we include not only static frames but also temporal information using videos. In these directions, our paper takes a step towards real-world physical commonsense reasoning by contributing Physical Audiovisual CommonSense (PACS): the first *audiovisual* benchmark annotated for physical commonsense attributes. PACS contains a total of 13,400 question-answer pairs, involving 1,526 object-oriented videos that cover a diverse set of objects, and 1,377 unique physical commonsense questions involving a variety of physical properties.

In our paper, we first detail the construction of our new audiovisual benchmark of physical commonsense and establish the need for both the audio modality and

commonsense reasoning to succeed on our task. Using this benchmark, we evaluate the performance of multiple state-of-the-art unimodal and multimodal models in comparison with human performance. We also performed an analysis of *where* and *why* current models fail, highlighting the increased difficulty of reasoning about physical commonsense, the lack of fine-grained temporal information due to limitations in current models' video and audio processing, and the need for more advanced audiovisual models. We hope our work will elicit further research into building robust multimodal representations of the physical world. [3]

2 Related Work

We cover related work in commonsense reasoning, particularly on physical understanding, which has been studied in domains spanning psychology, language, vision, robotics, and multimodal machine learning.

Psychology: Physical commonsense was first studied in humans, with psychology experiments based on naive and intuitive physics [8,17,25,28,39]. In these experiments, humans are asked to predict object motion or the result of multi-object interactions. Further research has also been conducted in general physical modeling [7] and the multisensory perception of physical properties [13,40,65]. In particular, studies on human behavior indicate that the audio modality contains valuable information about the physical properties of objects [24,40,41,57].

Language: Related work has studied physical commonsense within the text modality [6,16,32,50,67]. To our knowledge, the generalizability of their findings to other modalities is still understudied. Our dataset extends these text-based knowledge graphs and language models to multimodal settings.

Vision: Methods utilizing physical commonsense have been applied to several visual commonsense tasks, including scene understanding [12,58], activity recognition [37], and cause-effect prediction [42]. We note that these methods focus solely on the visual modality, which may bring challenges in tasks with unknown or occluded objects. Including information from other modalities such as audio and language could help mitigate these challenges.

Audio provides valuable information for one's understanding of the world [24,57]. Currently, AI tasks studying physical properties through the lens of the audio modality include navigation [10], perception [65], and generative modeling [66,18]. We extend this research direction to higher-order reasoning through PACS.

Robotics: Comprehension of physical properties has been shown to be valuable for tool usage and object manipulation tasks [3,14,43,54,55]. Our paper provides a direction for generalizing physical commonsense reasoning utilizing both audio and visual modalities.

Multimodal: Recent work has introduced question-answering datasets with image and text inputs (e.g., VQA [5], NLVR [51], NLVR2 [52]), with some annotated for commonsense reasoning tasks (e.g., VCR [63,64], VisualCOMET [45]).

[3] For dataset download links, benchmarked models, and evaluation scripts, please visit https://github.com/samuelyu2002/PACS.

There has also been the use of multimodal answer choices, such as a combination of text and image regions in VCR [63] and VisualCOMET [45]. Other works have also introduced datasets with video and text inputs to test for temporal reasoning (e.g., MovieQA [53], MovieFIB [38], TVQA [36,64]). To our knowledge, none of these approaches have explored audio and video together for physical commonsense reasoning.

3 PACS Dataset

We introduce PACS, a benchmark dataset designed to help create and evaluate a new generation of AI algorithms able to reason about physical commonsense using both audio and visual modalities. The underlying task is binary question answering, where given a question $\boldsymbol{q}$ and objects $\boldsymbol{o_1}, \boldsymbol{o_2}$, the model must pick the more appropriate object to answer the question. Each object is represented by a video $\boldsymbol{v}$ showing a human interacting with the object, the corresponding audio $\boldsymbol{a}$, and a bounding-box $\boldsymbol{b}$ drawn around the object in the middlemost frame of $\boldsymbol{v}$.[4] Thus, each datapoint in PACS is a tuple of values $(\boldsymbol{q}, (\boldsymbol{b_1}, \boldsymbol{v_1}, \boldsymbol{a_1}), (\boldsymbol{b_2}, \boldsymbol{v_2}, \boldsymbol{a_2}), \boldsymbol{l})$, representing the question, two objects, and a binary label of which object is the correct answer (see Figure 1 for an example datapoint in our dataset).

In this section, we first outline various design principles used in the creation of our dataset. Then, we give an overview of PACS statistics (see Figure 2 for a complete overview), and finally discuss each component of our data collection and annotation process (see Figure 3 for our complete annotation pipeline). For a more detailed overview of our data collection pipeline, please refer to section A in the appendix.

3.1 Design Principles

Through synthesis of previous work, we divide physical commonsense into two main categories based on which we designed PACS. These categories were used as guidance for annotators when creating physical commonsense questions.

1. **Intuitive physics, and a functional world model:** This category is inspired by previous psychology and AI experiments relating to physical commonsense, such as predicting object motion [33,34,48,59], or how objects interact with each other [28]. Questions in this category focus on predicting the result of single or multi-object interactions. Easy questions involve a single object and action, such as: *"Which object will break after being dropped on the ground?" (a vase, a ball of paper)*. Harder questions involve multiple objects or actions, including interactions between the two objects, such as: *"Which object will become deformed if the other object is placed on top of it?" (a vase, a ball of paper)*.

[4] In our experiments, we usually represent the bounding box $\boldsymbol{b}$ as a red bounding box drawn directly on the middlemost frame of the video. Thus, we also interchangeably notate the bounding box as an image $\boldsymbol{i}$.

2. **Common real-world knowledge:** This category is inspired by previous commonsense datasets, which test for more concrete understandings of how and why humans or objects function in the real world [6,16,62,63]. Questions in this category ask about possible uses of an object in real-life scenarios. Importantly, these scenarios focus on less prototypical uses of an object, therefore reducing the possibility of abusing learned knowledge [6], such as *"Which object is better suited to clean up a watery mess" (an old t-shirt, a plastic box)*. Harder questions can introduce more complicated or uncommon scenarios involving multiple objects: *"If I were to stack the two objects, which would logically go on the bottom?" (an old t-shirt, a plastic box)*.

3.2 Dataset Statistics

This subsection presents the main object and question statistics of PACS. Each datapoint is the combination of a question, two objects, and the correct answer. Figure 2f shows the distribution of the number of questions relating to each object pair, with an average of 5.86 questions per pair.

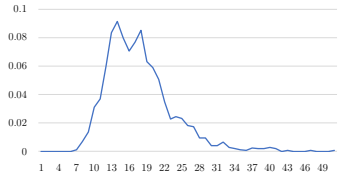
Object statistics: PACS contains a total of 1,526 objects, each represented by a unique video clip, with included audio and a bounding box in the middlemost frame of the video. Figure 2b shows a rough distribution of materials that the objects in our dataset are made of, as annotated in our video filtering step. Materials such as "Wax" or "Foam" occur more commonly in our dataset than in real life, due to our focus on creating a diverse set of objects. Figure 2e shows the length of each video. On average, videos in our dataset are 7.6 seconds long.

Question statistics: PACS contains a total of 1,377 unique questions each used multiple times across various pairs of objects. Figure 2d shows how many times each question was used, where on average, a question was distributed to 10.8 pairs of videos. Figure 2a shows the distribution of question length in terms of the number of words. On average, a question was 16.6 words long. Figure 2c shows the distribution of physical properties that our questions relate to. Figure 2g shows the most commonly occurring words in our dataset and is also color-coded by CLIP's accuracy on datapoints conditioned on the occurrence of a specific word. We can see a variety of action words (e.g., placed, dropped, thrown, roll, rubbed, pressed, blown), each associated with different physical properties. Furthermore, we see that AudioCLIP struggles with certain physical concepts, such as having low accuracy on heat-related words (e.g., hot, fire).

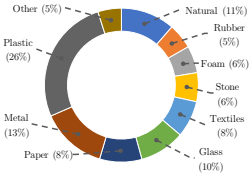
3.3 Dataset Creation

In this subsection, we outline the steps used to gather and label datapoints in PACS (see Figure 3 for a complete overview).

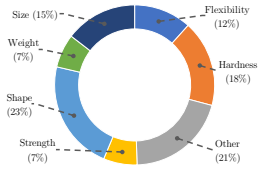
(a) Video collection: A broad set of ASMR videos were downloaded from YouTube. Specifically, we chose to use object-oriented ASMR videos , as they provide high-quality audio, and often incorporate objects that people less commonly interact with. We used a list of materials [1] to seed the search queries, which was later updated with more materials as we iterated through the first two data



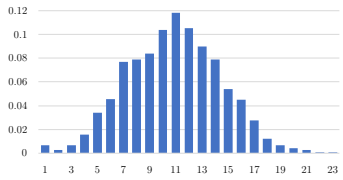
(a) Question length distribution (as tokenized by spaCy).



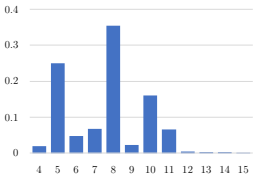
(b) Distribution of materials making up each object.



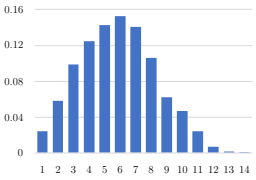
(c) Distribution of relevant physical properties.



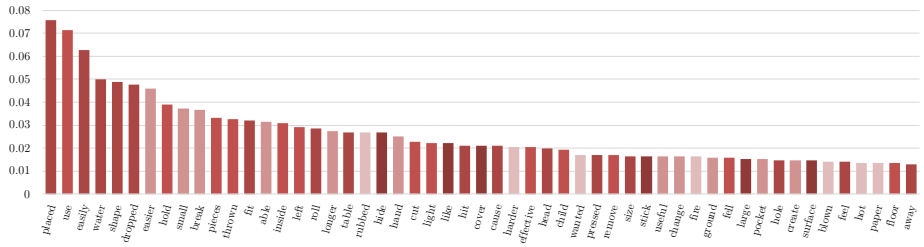
(d) Number of times each question was reassigned to new object pairs.



(e) Duration of videos (rounded to nearest second).



(f) Number of questions per pair of objects.



(g) Frequency of most common words in PACS. The top 4 words (object, item, likely, better) are excluded due to their high frequency being a result of our problem formulation. The bars are colored based on the accuracy that AudioCLIP [23] achieves on them, with darker being higher accuracy.

Fig. 2: Dataset statistics for PACS. Best viewed zoomed-in and with color. Figure 2b and Figure 2c show that the questions and objects in our dataset are diverse, involving different physical properties and materials. Figure 2g shows a variety of actions (e.g., placed, dropped, thrown, roll, rubbed, pressed, blown) covered in our diverse questions.

collection steps. For each video, we use a shot boundary detector [49] to split each video into separate scenes, and then further split each scene into roughly 5-10 second long clips. Finally, an audio classifier [20] was used to remove videos with background music, talking, or silence. The remaining clips were sparsely sampled to create the candidate set of clips.

(b) Video clip annotation and filtering: When analyzing the candidate set of clips, we noticed that a large number of objects that appeared in these clips were common household objects, resulting in many repeated objects. Furthermore, common objects do not require as much multimodal understanding, as a single image and a decent knowledge base may be enough to identify the object and extract necessary physical properties. Thus, as a heuristic for how common or obvious an object is, we test to see if annotators are able to classify the materials each object is made of. If annotators are able to correctly identify an object's

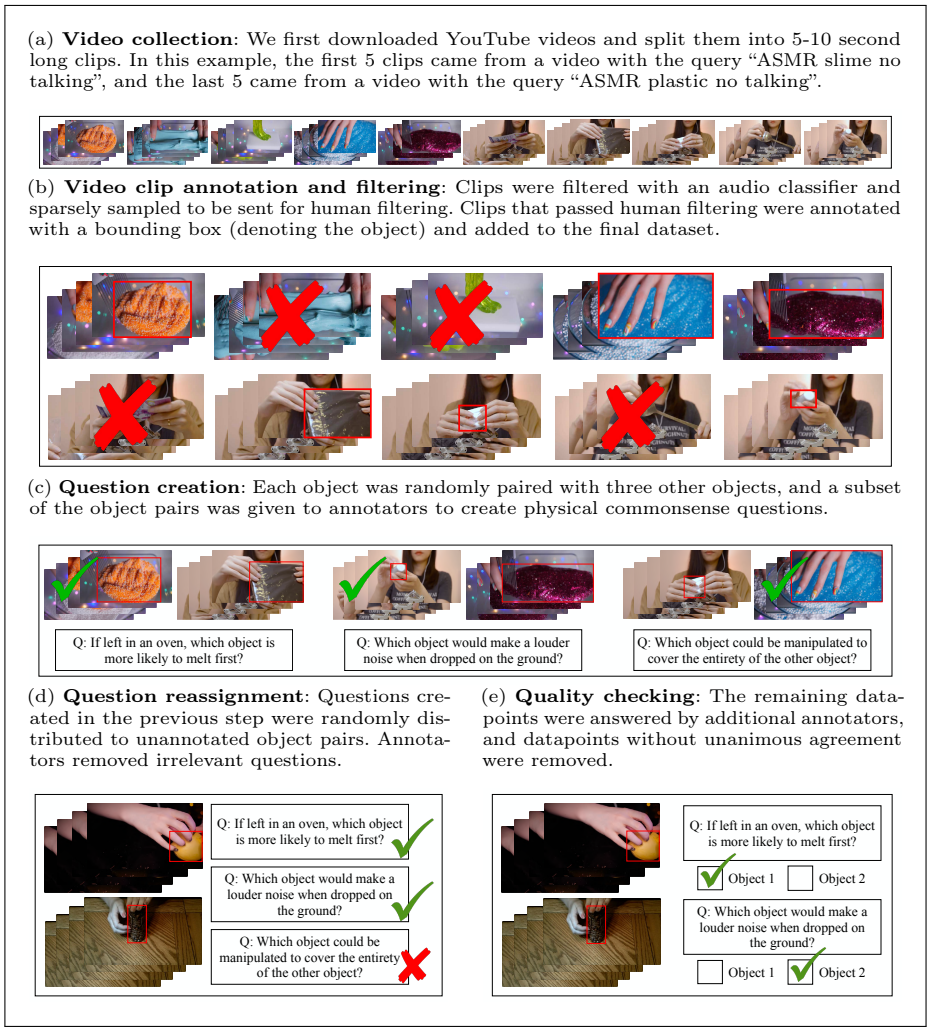


Fig. 3: Diagram of our data collection process, showing steps starting from gathering objects, to creating and checking datapoints. Best viewed zoomed-in and with color.

materials using just a single image, then this suggests that the object is likely common, and has physical properties that are easily distinguishable.

In this task, annotators were first given a single image from a candidate video clip and asked to draw a bounding box around the "object of focus", which we define as the object the person is touching in the video (if the guess is wrong, the clip is thrown away). Then, they were were asked to select the materials that make up the object from a list, and to provide a confidence score from $1-5$. Once they submitted their initial answer, annotators were then given access to

the whole video and audio and asked to redo the task. If their confidence did not increase and their answers did not change, then the clip was removed. Otherwise, the clip and the bounding box were added to the dataset as an object, with each clip containing exactly one bounding box annotation (one object).

The final set of 1,526 objects was partitioned into train, test, and validation of 1,224, 152, and 150 videos respectively. Then, each object was paired with three other objects in the same subset, resulting in 2,289 pairs of objects.

(c) Question creation: From the 2,289 object pairs gathered, 242 were randomly selected to be used in this step, while the other 2,047 pairs were used in the next step. In this step, annotators were asked to write questions that require physical commonsense knowledge to answer. Annotators were given two videos, and a frame from each video containing a bounding box that specified the object. The had the option to write one or two commonsense questions related to the pair of objects, and answer with "Object 1" or "Object 2". In total, 1,377 questions were created, with each pair of videos given to 5 separate annotators.

To facilitate the process of creating high-quality questions, we provided annotators with a more detailed version of the categorization developed in section 3.1 as guidance for what constitutes physical commonsense as instructions. They were also required to provide at least one relevant physical property for each question to encourage topical questions. Finally, questions were required to have a certain level of complexity, and were all quality-checked (e.g., questions that directly asked about a physical property such as *"Which object is more sticky"*, or *"Which object is larger?"* were forbidden).

(d) Question reassignment: We evenly redistribute the 1,377 questions created in the previous step to the remaining 2,047 object pairs. Reusing questions on new pairs of objects can create interesting scenarios, as it matches object pairs with questions that human annotators may not normally come up with [6]. The goal is to create matchings such as: *"If you absolutely needed to tie your hair up, which item would you use?" (a plastic straw, a piece of paper)*. In this example, the question and object pair are not normally associated with each other, but are still answerable by humans, who have the ability to draw new connections. This puts more of the challenge on drawing relationships between physical properties, rather than directly applying past knowledge.

Specifically, in this task, each unused object pair is assigned a list of 13 questions, which is then given to annotators. Then, annotators can either mark each object-question matching as "completely irrelevant", or choose to answer the question, thus creating a new datapoint.

(e) Quality checking: To ensure the quality of final datapoints, each candidate datapoint gathered from the Question Creation and Question Reassignment stages was given to additional annotators to double-check. Every candidate was answered three times between the question annotation stages and only kept in our dataset if there was unanimous agreement.

4 Experimental Setup

In this section, we first outline the setup for testing human performance. We then list the models for checking dataset biases, and several state-of-the-art models that we tested. Finally, we outline the creation of PACS-material, a material classification subtask on our dataset.[5] Our experiments were designed to answer the following research questions:

1. How difficult is our task, as measured by the performance of human annotators and state-of-the-art models? We evaluate open-source state-of-the-art models that have high performance on comparable datasets such as VCR [63], TVQA [36], and NLVR2 [11] (section 4.3), and compare these results to human performance on PACS (section 4.1).
2. Are there potential biases in our dataset? While the paired binary question answering format is designed to limit bias in the language modality (correlations between questions and correct vs incorrect answers) as opposed to standard QA datasets [2,4,31,62], we explore other sources of biases in language, video, and audio in PACS (section 4.2).
3. What is the importance of audio in our task, and what are the specific areas where audio is beneficial? We compare human and model performance with and without audio (with otherwise the same configurations) and analyze specific qualitative examples where including audio leads to better results (see section 4.1 and section 4.3 for how we set up human and model benchmarks).
4. How challenging is the level of reasoning required to capture physical commonsense? To establish this difficulty, we create an additional material classification task to compare with our physical commonsense task (section 4.4).

4.1 Human Performance

To test human performance with and without audio, we randomly sampled 243 datapoints from the dataset, and give them to 10 annotators to answer. The annotators were given half of the datapoints with audio and half without, such that each datapoint would be annotated with five answers with audio, and five answers without. Consistent with other works, we compute human accuracy as a majority vote [6,63], and also report 90% confidence intervals for the results.

4.2 Detecting Biases

We construct four different combinations of late-fusion models by combining state-of-the-art pre-trained image, audio, video, and text models. We used ViT [15] as the image model, AST [21] as the audio model, TDN [56] as the video model, and DeBERTa-V3 [26,27] as the text model. The specific configurations chosen for bias detection were inspired by past work studying bias on Visual Question Answering datasets [9,60,63]. We test for two main types of bias: *answer choice*

[5] For more details on experimental setups, refer to section B in the appendix.

bias (are there systematic biases in the answer choices that give away the correct answer without even seeing the question?), and *unimodal question-answerability* (is information from one modality enough to correctly answer the question?).
I + A + V: We study the predictability of our task given only information about the objects (no question is provided). This test demonstrates whether there is a pattern between the objects and the correct answer.
Q + I: Evaluates the usefulness of images (I) in predicting correct answers.
Q + V: Evaluates the usefulness of videos (V) in predicting correct answers.
Q + A: Evaluates the usefulness of audio (A) in predicting correct answers.

4.3 Baseline Models

Late Fusion [44]: We train a model using late fusion of all four input modalities as a simple baseline. We use SOTA image [15], audio [21], and video [56] models pretrained on large-scale classification datasets such as ImageNet21k [47], AudioSet [19], and Something-Something V2 [22], and the text [26] model is pretrained using replaced token detection. We concatenate the unimodal embeddings and use a linear layer to create multimodal embeddings for prediction.
CLIP [46] is a powerful image-text model pre-trained on a large set of images and text captions and can be used for a variety of zero-shot and finetuning tasks. CLIP embeds image and text into a shared vector space, where we can use *cosine similarity* to measure the similarity between image and text embeddings. We use CLIP to separately embed images of both objects and the question. The predicted object is the object with more similar embedding to the question embedding.
AudioCLIP [23] extends CLIP for audio inputs by training on AudioSet [19], which enables the embedding of audio inputs into the same vector space. Using this model, we extend the CLIP model mentioned above to include audio by concatenating the image and audio embedding, and using a linear layer to project them onto the same vector space as the text embedding.
UNITER [11] is an image and text model that is pre-trained using four different image-text tasks and achieves strong results on tasks such as NLVR2 [52]. We largely follow the procedure used to prepare and finetune UNITER on the NLVR2 dataset [52]. We split up both objects and generate two object-question embeddings, and finally concatenate them and use an MLP to classify the answer.
Merlot Reserve [64] uses image, audio, video, and text, achieving state-of-the-art results on VCR [63] and TVQA [36]. We follow the methods used to train Merlot Reserve on VCR and TVQA by constructing two multimodal sequences using all input modalities. Then, we separately generate confidence scores for both sequences and compare the two values as a classification output.

4.4 Material Classification

By comparing with the simpler task of classification, we can gain an understanding of the level of higher-order reasoning required in our task. In our main question-answering task, errors can come from multiple sources, either from misidentifying

Baseline Model	Accuracy (%)		
	With audio	Without audio	Δ
I + A + V [60,44]	51.9 ± 1.1	-	-
Q + I [62,44]	-	51.2 ± 0.8	-
Q + A [62,44]	50.9 ± 0.6	-	-
Q + V [62,44]	-	51.5 ± 0.9	-
Late Fusion [44]	55.0 ± 1.1	52.5 ± 1.6	2.5
CLIP/AudioCLIP [23,46]	60.0 ± 0.9	56.3 ± 0.7	3.7
UNITER (Large) [11]	-	60.6 ± 2.2	-
Merlot Reserve (Base) [64]	66.5 ± 1.4	64.0 ± 0.9	2.6
Merlot Reserve (Large) [64]	70.1 ± 1.0	68.4 ± 0.7	1.8
Majority	50.4	50.4	-
Human	96.3 ± 2.1	90.5 ± 3.1	5.9

Table 1: Results on PACS test set: baseline models are reported with the mean and standard deviation of 5 runs, while human accuracy is reported with a 90% confidence interval. There is a large gap between model and human performance, with the best performing model (Merlot Reserve) lagging behind by over 25%. Models with audio also consistently outperform the corresponding models without audio, demonstrating the need for information from all modalities to succeed in our task.

the properties of an object, or correctly identifying the objects, but failing to reason about the properties. Results from a material classification task using the same objects can give us an estimate on how much error stems from misidentifed objects, and how much comes from the failure to exhibit higher-order reasoning.

We create a material classification task (PACS-material) formulated identically to our dataset, where a pair of objects is accompanied by a comparison question (e.g., *"Which object is more likely to be made out of glass"*). The materials used are gathered from our data-collection stage (Figure 2b shows a distribution of material categories). We use the exact same object pairs as in the main task, and accompany each pair with comparison questions based on each object's material. In total, we created 3,460 training datapoints, 444 validation datapoints, and 445 testing datapoints. Each datapoint is a quadruplet $(\boldsymbol{o}^{(1)}, \boldsymbol{o}^{(2)}, \boldsymbol{q}, \boldsymbol{l})$, representing the two objects, the question, and the label.

5 Results and Discussion

In this section, we assess the whether audiovisual understanding and physical commonsense reasoning are required to succeed on our dataset, and look at where current models fail. For additional results, refer to section C in the appendix.

5.1 Human and Model Performance

A summary of all model performances is shown in Table 1. Notably, all methods struggle to achieve results close to human performance, with the gap in accuracy between the best model (Merlot Reserve) and human performance being over 25%.

This gap is much larger than the gap between SOTA and human performance on other datasets such as TVQA (3%) and VCR (14%) [64] , demonstrating the challenging nature of our dataset.

We believe that the gap in performance comes from (1) the inherent challenge of developing physical commonsense (section 5.4), and (2) the loss of information in each model. This includes the lack of video information in CLIP and UNITER, and the sparse sampling of video frames in the Merlot Reserve and Late Fusion models. Some physical information may require clear alignment between the actions displayed in the video and the audio signal to accurately understand the object, and thus require more fine-grained temporal information.

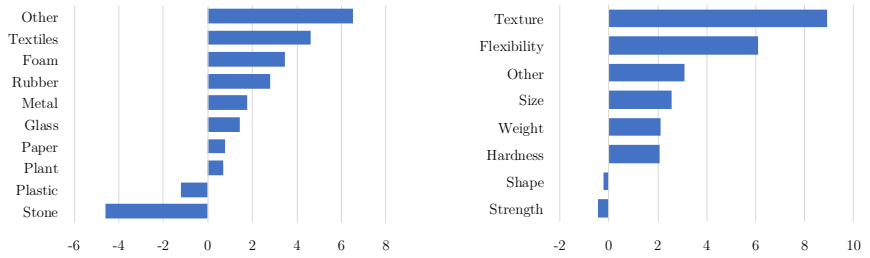
5.2 Checking for Biases in PACS

Table 1 shows the performance of our bias testing models, where we see that there is low performance among all configurations of models used. The I+A+V configuration tests for bias among the answer choices (objects), which achieves a low accuracy of 52%, demonstrating that the answer choices alone do not give away the answer. Furthermore, solely providing image, audio, or video information alongside the question yields poor performance, and it is only when all three modalities are combined that results solidly deviate from randomly guessing (55% accuracy). We believe the low results when provided with unimodal information are because all modalities play an important role. Only the image input specifies the object via a bounding box, thus making it difficult to succeed without the image. Additionally, since our dataset was curated to consist of complex objects that require video and/or audio to understand, removing such modalities also result in low performance.

5.3 Importance of Audio

In Table 1, we can see the benefit of including audio. Perhaps the most important experiment is how much audio helps humans, as the error rate decreases by more than half, with no overlap between the confidence intervals for the two values. When provided with audio, the models don't seem to improve as much. We theorize a few reasons for this: (1) for Merlot Reserve, the pretraining data is from a very different distribution, mostly consisting of human speech, and the input spectrograms may not be fine-grained enough to capture higher-pitched, sharper noises, such as tapping. (2) In contrast, AudioCLIP uses raw audio as an input, but the method of fusing audio and video through concatenation may be too simple.

Performance on the most "unique" objects: Using the material and physical property labels gathered in the annotation steps, we can also compare results conditioned on specific materials and properties. We calculate performance with respect to a specific material (e.g., metal) by only counting datapoints where at least one of the objects is made of metal. Similarly, we calculate performance with respect to a physical property (e.g., hardness) by only counting datapoints where the question is related to the property. In Figure 4a, we see that the



(a) Difference in accuracy with and without audio, conditioned on object materials.

(b) Difference in accuracy with and without audio, conditioned on physical properties.

Fig. 4: Comparison of results on Merlot Reserve when trained with and without audio. These results are conditioned on the material of the objects in the object pair, and on the physical properties relevant to the question (see section 5.3).

Baseline Model	Subset	**Accuracy (%)**		
		PACS-material	PACS	Δ
Late Fusion [44]	Val	67.8 ± 0.8	55.5 ± 0.3	12.3
	Test	67.4 ± 1.5	55.0 ± 1.1	12.4
AudioCLIP [23]	Val	81.9 ± 1.2	61.6 ± 0.9	18.8
	Test	75.9 ± 1.1	60.0 ± 0.9	15.0

Table 2: Comparison of PACS-material and PACS. Despite PACS-material being created from relatively noisy labels, we observe that it is a far easier task, with models performing $10-20\%$ better on it than on PACS. This suggests that our dataset requires a level of reasoning that goes beyond what is required in classification tasks.

biggest improvement in accuracy is on datapoints containing objects made of "Other" materials. Since our material labels cover the most common materials appearing in the dataset, this suggests that audio is especially important when reasoning about uncommon objects. From Figure 4b, we see that properties such as texture and flexibility show the most improvement, and no category's results suffer greatly with the addition of audio.

5.4 Difficulty of Reasoning

As seen in Table 2, the material classification task on our dataset is much easier than our main task, with models achieving 10-20% higher accuracy, despite being trained using fewer datapoints (11,044 vs 3,460). Since the only other difference between PACS and PACS-material lies in the content of the questions, we believe that this gap in performance is due to the added difficulty of physical commonsense reasoning. The remaining 20-30% of misclassified datapoints on PACS-material can be attributed to both noisy labels resulting in imperfect training and evaluation, and a true failure in understanding the objects' material makeup.

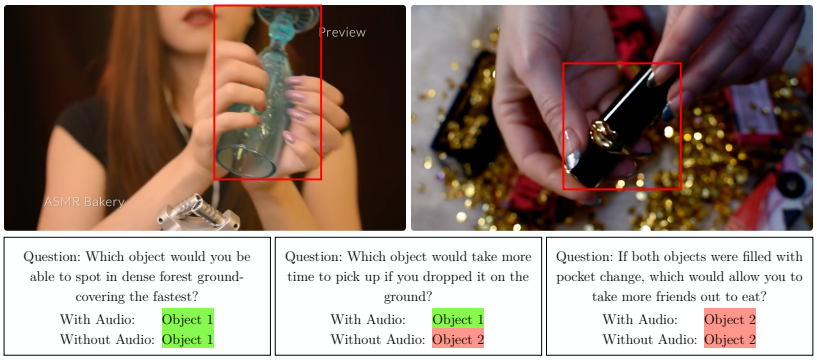


Fig. 5: Qualitative results showing predictions from Merlot Reserve models trained with and without audio. In this example, the first object could be mistaken as plastic and the second object could be made of plastic or metal. Thus, the model without audio doesn't realize that the glass object will shatter and takes longer to pick up off the ground. Furthermore, both models fail to answer the third question, which indirectly asks about the size and shape of both objects. This shows that models struggle on questions that are more complex, or require more implicit knowledge.

5.5 Example Predictions

Finally, we analyze some specific examples to see where audio is helpful, and where both models fail. Generally, audio is helpful when models are presented with visually ambiguous or uncommon objects. In these situations, audio is necessary to clarify the physical properties of the objects (e.g., question 2 in Figure 5). Furthermore, despite the presence of audio, both models may still fail when asked complex and/or uncommon questions that require the understanding of implicit information (e.g., question 3 in Figure 5).

6 Conclusion

We introduced PACS, a large-scale audiovisual dataset for physical commonsense reasoning. We find that the best models still struggle to (1) fully leverage multimodal information, and (2) develop a strong understanding physical commonsense. Through experiments, we evince the multimodal nature of PACS and its usefulness in benchmarking future work in multimodal commonsense reasoning. We also provide multiple promising directions for bridging the gap between human and AI performance, which we hope provides insight in progressing towards safe and robust multimodal representations of the physical world.

Acknowledgements This material is based upon work partially supported by the National Science Foundation (Awards #1722822 and #1750439) and National Institutes of Health (Awards #R01MH125740, #R01MH096951, and #U01MH116925). Additionally, we would also like to acknowledge NVIDIA's GPU support and Google's TPU support.

References

1. Standard list of material categories and types. `https://www.calrecycle.ca.gov/lgcentral/basics/standlst` (2018)
2. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1955–1960 (2016)
3. Agrawal, P., Nair, A.V., Abbeel, P., Malik, J., Levine, S.: Learning to poke by poking: Experiential learning of intuitive physics. Advances in neural information processing systems **29** (2016)
4. Anand, A., Belilovsky, E., Kastner, K., Larochelle, H., Courville, A.: Blindfold baselines for embodied qa. arXiv preprint arXiv:1811.05013 (2018)
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
6. Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al.: Piqa: Reasoning about physical commonsense in natural language. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 7432–7439 (2020)
7. Bliss, J.: Commonsense reasoning about the physical world. Studies in Science Education **44**(2), 123–155 (2008)
8. Bobrow, D.G.: Qualitative reasoning about physical systems: an introduction. Artificial intelligence **24**(1-3), 1–5 (1984)
9. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. Advances in Neural Information Processing Systems **32**, 841–852 (2019)
10. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: European Conference on Computer Vision. pp. 17–36. Springer (2020)
11. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020)
12. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ Scene Understanding: Single-View 3D Holistic Scene Parsing and Human Pose Estimation With Human-Object Interaction and Physical Commonsense. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
13. Corlett, P.R., Powers, A.R.: Conditioned hallucinations: historic insights and future directions. World Psychiatry **17**(3), 361 (2018)
14. Coumans, E., Bai, Y.: PyBullet, a Python Module for Physics Simulation for Games, Robotics and Machine Learning. `http://pybullet.org` (2016–2021)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
16. Forbes, M., Holtzman, A., Choi, Y.: Do neural language representations learn physical commonsense? In: CogSci (2019)
17. Forbus, K.D.: Qualitative process theory. Artificial intelligence **24**(1-3), 85–168 (1984)
18. Gao, R., Chang, Y.Y., Mall, S., Fei-Fei, L., Wu, J.: Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In: CoRL (2021)

19. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261
20. Giannakopoulos, T.: pyaudioanalysis: An open-source python library for audio signal analysis. PloS one **10**(12) (2015)
21. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer. In: Proc. Interspeech 2021. pp. 571–575 (2021). https://doi.org/10.21437/Interspeech.2021-698
22. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. pp. 5843–5851 (10 2017). https://doi.org/10.1109/ICCV.2017.622
23. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. arXiv preprint arXiv:2008.04838 (2020)
24. Handel, S.: Timbre perception and auditory object identification. Hearing **2**, 425–461 (1995)
25. Hayes, P., Nilsson, N.J.: Knowledge representation. Morgan Kaufman (1987)
26. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing (2021)
27. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=XPZIaotutsD`
28. Hespos, S.J., Ferry, A., Anderson, E., Hollenbeck, E., Rips, L.J.: Five-month-old infants have general knowledge of how nonsolid substances behave and interact. Psychological Science **27**, 244 – 256 (2016)
29. Hespos, S.J., Spelke, E.S.: Conceptual precursors to language. Nature **430**(6998), 453–456 (2004)
30. Hessel, J., Mimno, D., Lee, L.: Quantifying the visual concreteness of words and topics in multimodal datasets. In: Proceedings of NAACL-HLT. pp. 2194–2205 (2018)
31. Jabri, A., Joulin, A., Maaten, L.v.d.: Revisiting visual question answering baselines. In: European conference on computer vision. pp. 727–739. Springer (2016)
32. Jimenez, C.E.: Learning physical commonsense knowledge (2020)
33. Kaiser, M., Jonides, J., Alexander, J.: Intuitive reasoning about abstract and familiar physics problems. Memory & cognition **14**, 308–12 (08 1986). https://doi.org/10.3758/BF03202508
34. Kim, I.K., Spelke, E.S.: Perception and understanding of effects of gravity and inertia on object motion. Developmental Science **2**(3), 339–362 (1999). https://doi.org/https://doi.org/10.1111/1467-7687.00080, `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-7687.00080`
35. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
36. Lei, J., Yu, L., Bansal, M., Berg, T.: Tvqa: Localized, compositional video question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1369–1379 (2018)

37. Li, Y.L., Liu, X., Wu, X., Li, Y., Qiu, Z., Xu, L., Xu, Y., Fang, H.S., Lu, C.: Hake: A knowledge engine foundation for human activity understanding (2022)
38. Maharaj, T., Ballas, N., Rohrbach, A., Courville, A., Pal, C.: A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6884–6893 (2017)
39. McCloskey, M.: Intuitive physics. Scientific american **248**(4), 122–131 (1983)
40. Minsky, M.: Commonsense-based interfaces. Communications of the ACM **43**(8), 66–73 (2000)
41. Morrongiello, B.A., Fenwick, K.D., Chance, G.: Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. Infant Behavior and Development **21**(4), 543–553 (1998)
42. Mottaghi, R., Rastegari, M., Gupta, A., Farhadi, A.: "what happens if..." learning to predict the effect of forces in images. In: ECCV (2016)
43. Nair, L., Balloch, J., Chernova, S.: Tool macgyvering: Tool construction using geometric reasoning. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 5837–5843. IEEE (2019)
44. Pandeya, Y.R., Lee, J.: Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications **80**, 1–19 (01 2021). https://doi.org/10.1007/s11042-020-08836-3
45. Park, J.S., Bhagavatula, C., Mottaghi, R., Farhadi, A., Choi, Y.: Visualcomet: Reasoning about the dynamic context of a still image. In: European Conference on Computer Vision. pp. 508–524. Springer (2020)
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
47. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses (2021)
48. Smith, K.A., Battaglia, P.W., Vul, E.: Consistent physics underlying ballistic motion prediction. Cognitive Science **35** (2013)
49. Souček, T., Lokoč, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. arXiv preprint arXiv:2008.04838 (2020)
50. Storks, S., Gao, Q., Zhang, Y., Chai, J.Y.: Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In: EMNLP (2021)
51. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 217–223 (2017)
52. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6418–6428 (2019)
53. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
54. Toussaint, M.A., Allen, K.R., Smith, K.A., Tenenbaum, J.B.: Differentiable physics and stable modes for tool-use and manipulation planning (2018)
55. Tuli, S., Bansal, R., Paul, R., et al.: Tango: Commonsense generalization in predicting tool interactions for mobile manipulators. arXiv preprint arXiv:2105.04556 (2021)

56. Wang, L., Tong, Z., Ji, B., Wu, G.: Tdn: Temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1895–1904 (June 2021)
57. Wilcox, T., Woods, R., Tuggy, L., Napoli, R.: Shake, rattle, and... one or two objects? young infants' use of auditory information to individuate objects. Infancy **9**(1), 97–123 (2006)
58. Wu, J., Lu, E., Kohli, P., Freeman, B., Tenenbaum, J.: Learning to see physics via visual de-animation. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. Curran Associates, Inc.
59. Wu, J., Yildirim, I., Lim, J.J., Freeman, B., Tenenbaum, J.: Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper/2015/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf
60. Yang, J., Zhu, Y., Wang, Y., Yi, R., Zadeh, A., Morency, L.P.: What gives the answer away? question answering bias analysis on video qa datasets (2020)
61. Yatskar, M., Ordonez, V., Zettlemoyer, L., Farhadi, A.: Commonly uncommon: Semantic sparsity in situation recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7196–7205 (2017)
62. Zadeh, A., Chan, M., Liang, P.P., Tong, E., Morency, L.P.: Social-iq: A question answering benchmark for artificial social intelligence. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8799–8809 (2019). https://doi.org/10.1109/CVPR.2019.00901
63. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
64. Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., Choi, Y.: Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In: arxiv (2022)
65. Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J., Freeman, B.: Shape and material from sound. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/f4552671f8909587cf485ea990207f3b-Paper.pdf
66. Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J.H., Tenenbaum, J.B., Freeman, W.T.: Generative modeling of audible shapes for object perception. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
67. Zhao, Z., Papalexakis, E., Ma, X.: Learning Physical Common Sense as Knowledge Graph Completion via BERT Data Augmentation and Constrained Tucker Factorization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3293–3298. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.266, https://aclanthology.org/2020.emnlp-main.266