MCoMeT: Multimodal Fusion Transformer for Physical Audiovisual Commonsense Reasoning

Daoming Zong and Shiliang Sun*

School of Computer Science and Technology, East China Normal University, Shanghai, China
ecnuzdm@gmail.com, slsun@cs.ecnu.edu.cn

Abstract

Physical commonsense reasoning is essential for building reliable and interpretable AI systems, which involves a general understanding of the physical properties and affordances of everyday objects, how these objects can be manipulated, and how they interact with others. It is fundamentally a multisensory task, as physical properties are manifested through multiple modalities, including vision and acoustics. In this work, we present a unified framework, named Multimodal Commonsense Transformer (MCoMeT), for physical audiovisual commonsense reasoning. MCoMeT has two intriguing properties: *i*) it fully mines higher-ordered temporal relationships across modalities (*e.g.*, pairs, triplets, and quadruplets); and *ii*) it restricts the cross-modal flow through the feature *collection* and *propagation* mechanism with tight fusion bottlenecks, forcing the model to attend the most relevant parts in each modality and suppressing the dissemination of noisy information. We evaluate our model on a very recent public benchmark, PACS. Results show that MCoMeT significantly outperforms a variety of strong baselines, revealing powerful multi-modal commonsense reasoning capabilities.

Introduction

Humans use *physical commonsense reasoning* in all aspects of everyday life, whether to infer properties of previously unseen objects or to solve unique problems (*e.g.*, should I use salt or chocolate to clean my teeth when there is no toothpaste?). To interact with everyday objects in the real world, AI needs to acquire physical commonsense knowledge about everyday objects (Yu et al. 2022), such as their physical properties, affordances, how they are manipulated, and how they interact with other physical objects (Bisk et al. 2020; Forbes, Holtzman, and Choi 2019). The common understanding of object interactions plays an essential role in the cognitive abilities of robots (Zhao, Papalexakis, and Ma 2020a) (*e.g.*, when I apply eyeshadow without a brush, a robot should give me a cotton swab rather than a toothpick).

Previous works have pioneered the exploration of vision and text to understand fundamental physical properties (Hessel, Mimno, and Lee 2018; Krishna et al. 2017; Storks et al. 2021; Yatskar et al. 2017; Zhao, Papalexakis, and Ma 2020b). However, physical commonsense reasoning is naturally a multi-modal task because physical properties are manifested through multiple modalities, including vision and acoustics. If an object is visually ambiguous or rarely seen, audio can provide key information to identify its physical properties. As shown in Fig. 1, if only video without an audio track, `object1` could be mistaken for plastic or even paper, and `object2` could be mistaken for rubber or wax. For Q1 in Fig. 1, the absence of the necessary audio may easily lead to a wrong answer, *e.g.*, `object2` is more likely to break into pieces than `object1` if dropped from a height. Therefore, to empower AI with commonsense physical reasoning capabilities, models must learn to reason across both audio and visual modalities.

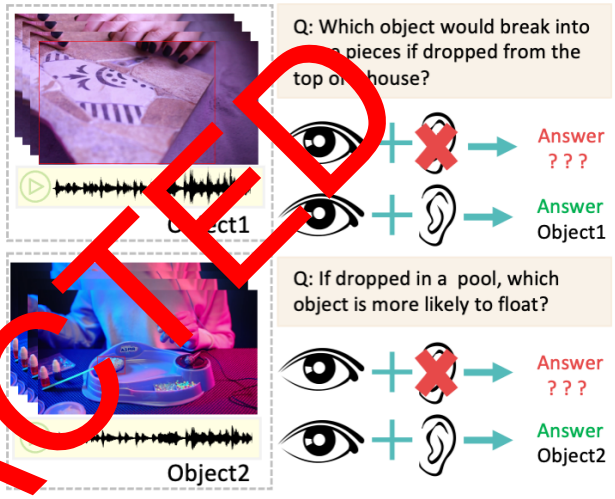

Figure 1: Illustration of two datapoints from PACS, with each datapoint containing a question and a pair of objects. Here `object1` is a ceramic tile and `object2` is foam.

In this work, we focus on physical commonsense reasoning under multimodal (audiovisual) settings. The additional acoustic information poses a great challenge to physical commonsense reasoning compared to tasks that focus solely on the visual or textual modality. On the one hand, we conjecture that the time-locked correlation allows audio to enrich the complementary visual information. On

*Corresponding Author.

the other hand, we are faced with: *i*) intrinsic differences between different modalities; *ii*) different noise topologies, where some modality streams contain more task-specific information than others (Nagrani et al. 2021). For example, as typical natural signals, both visual and audio features have heavy spatial-temporal redundancy and noisy information that is useless for other modalities; and *iii*) high computational complexity in integrating cross-modal information.

To alleviate the above issues, we first use two separate uni-modal encoders to capture the intra-modality correlations. Then, we utilize a multimodal encoder to capture the inter-modality relationship across modalities. The uni-modal encoder architecture is flexible enough to extract uni-modal patterns and can be aware of the distinctions in learning dynamics of each modality. Second, we propose to represent the video/audio as ordered tuples of varying frame numbers, which allows the matching of sub-sequences for object interactions at various speeds and temporal offsets. We consider the temporal relation of multiple frames sampled from a video or audio, as interactions (with objects) are typically shifting in appearance and are poorly represented by a single frame. The proposed temporal-relational module can effectively mitigate the temporal redundancy problem in both visual and audio streams (Yu et al. 2022). Third, inspired by (Nagrani et al. 2021), we restrict the flow of cross-modal information by introducing the feature collection and propagation mechanism with the tight fusion bottlenecks, dramatically reducing computational cost in computing audio-visual fusion. Such a paradigm imposes the model to *collect* and *compress* the most relevant parts across modalities and *broadcast* only the condensed features to each modality.

In summary, our main contributions are listed as follows: ***i***) We propose a novel Multimodal Commonsense Transformer (MCoMeT) for physical audiovisual commonsense reasoning. Our model can capture intra- and inter-modality relationships as well as high-ordered temporal relationships via two uni-modal encoders and a cross-modal encoder. Results show that MCoMeT outperforms a range of strong baselines. ***ii***) We explore how to efficiently fuse audio and visual streams to better discover the properties/affordances of objects as well as their interactions. We devise a temporal-relational module (TRM) and extend tight attention bottlenecks. The TRM constructs representations from ordered tuples of frames, enabling matching object interactions at various speeds and temporal shifts. And they work together to reduce the spatial-temporal redundancy and noise information in both visual and audio features. ***iii***) Experimental results on the benchmark have demonstrated the superiority of the proposed method. Abundant ablation studies are also conducted to validate the key ingredients of MCoMeT .

Related Work

Physical Commonsense Reasoning Natural language communication often requires reasoning about an object's affordances (*i.e.*, what actions are applicable to the object) from its properties (*e.g.*, whether an object is breakable, metal or wood). As a result, most of the previous work has concentrated on the physical commonsense within the text modality (Forbes, Holtzman, and Choi 2019; Bisk et al. 2020; Forbes and Choi 2017; Storks et al. 2021; Zhao, Papalexakis, and Ma 2020b). Specifically, (Forbes and Choi 2017) learned the physical attribute values for new words using dependency-based contextual information built on Wikipedia Corpus. (Forbes, Holtzman, and Choi 2019) investigated the extent to which neural language representations (*e.g.*, ELMo (Peters et al. 2018) and BERT (Kenton and Toutanova 2019)) can recover various facets of physical commonsense knowledge. (Bisk et al. 2020) developed a physical interaction question answering benchmark, named PIQA, for advancing physical commonsense understanding in natural language models.

MultiModal Fusion Multimodal fusion can be categorized into early fusion, late fusion and mid fusion (Baltrušaitis, Ahuja, and Morency 2018). Early fusion integrates features immediately after they are extracted (often achieved by simply concatenating their representations). Late fusion instead performs integration after each of the modalities has made a decision (*e.g.*, classification or regression) via various fusion schemes such as averaging, voting (Morvant, Habrard, and Ayache 2014), or a learned model (Mei, Bansal, and Walter 2016; Sun and Zhao 2020). In this work, we highlight the *mid fusion* for cross-modal information exchange in Transformer (Nagrani et al. 2021). The cross-attention provides an effective mechanism to capture the correlations and interactions across modalities, by *querying* the *keys* of another modality. In general, the input (tokens) to transformers is the output from pretrained feature extractors, such as convolutional neural networks (Akbari et al. 2021). Then different modality fusion strategies can be applied to the transformer. For example, Perceiver (Jaegle et al. 2021) introduces an iterative attention for early fusion by directly taking concatenated raw multimodal signals as inputs. In particular, MBT (Nagrani et al. 2021) proposes to restrict cross-modal attention flow between tokens within a layer via the bottleneck mid fusion.

MCoMeT

Notations and Definitions

Given a question $\boldsymbol{q}$ and objects $\boldsymbol{o_1}, \boldsymbol{o_2}$, the goal of our model is to select the more appropriate object to answer the question. Each object is represented by a video clip $\boldsymbol{v}$ depicting human interaction with the object, the corresponding audio $\boldsymbol{a}$, and a bounding-box $\boldsymbol{b}$ of the object drawn on the middlemost frame of the video. Formally, each datapoint we used is a quadruplet of $(\boldsymbol{q}, (\boldsymbol{b_1}, \boldsymbol{v_1}, \boldsymbol{a_1}), (\boldsymbol{b_2}, \boldsymbol{v_2}, \boldsymbol{a_2}), \boldsymbol{l})$, denoting the question, two objects, and a binary label of the answer.

Model Overview

Fig. 2 presents the overall architecture of our framework, which is built upon transformer encoder-decoder and consists of five core components, *i.e.*, temporal-relational module, uni-modal encoder, cross-modal encoder, query generator and query decoder. To begin with, we use three different pretrained feature extractors ($\boldsymbol{E_v}$, $\boldsymbol{E_a}$, and $\boldsymbol{E_t}$, see Sec. for details) equipped with the temporal-relational modules, to extract visual, audio, and textual features, respectively. Each datapoint can thereby be represented by three sets of feature

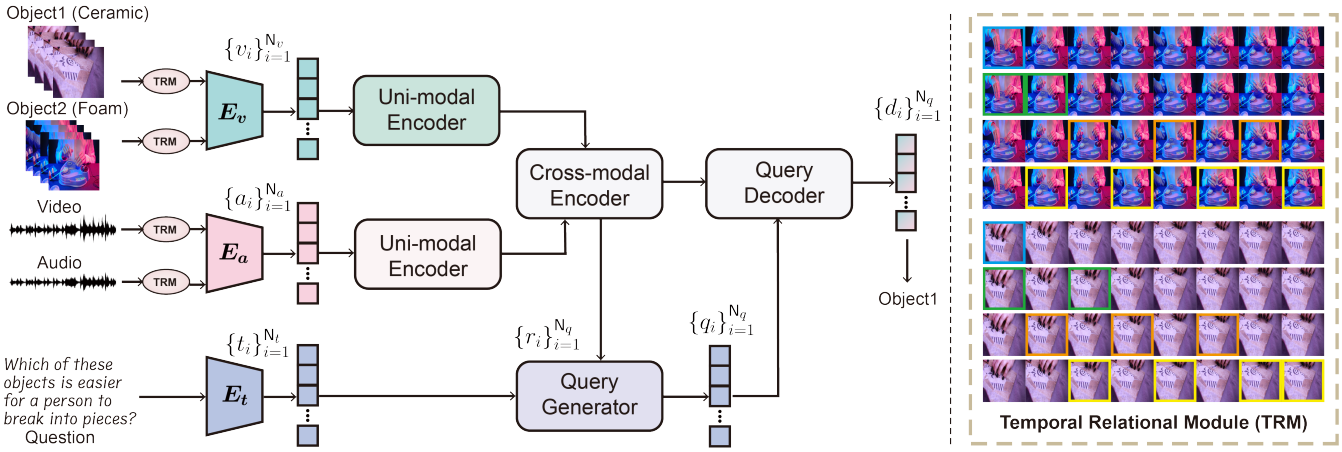

Figure 2: Overall architecture of the proposed MCoMeT . The corresponding uni-modal encoder and cross-modal encoder are deactivated when audio is unavailable. Illustrations of temporally-ordered tuples of frames sampled from the audio are omitted.

vectors, namely the visual features $\{v_i\}_{i=1}^{\mathsf{N}_v}$, audio features $\{a_i\}_{i=1}^{\mathsf{N}_a}$, and textual features $\{t_i\}_{i=1}^{\mathsf{N}_t}$. The visual and audio features are sent into separate uni-modal encoders to be contextualized under global receptive field, then be integrated by the cross-modal encoder for joint audiovisual representations $\{r_i\}_{i=1}^{\mathsf{N}_q}$. These joint representations, together with textual features, are used to generate clip-wise queries $\{q_i\}_{i=1}^{\mathsf{N}_q}$, which can be used to retrieve the most relevant parts from joint representations in the query decoder. After obtaining query-guided multimodal features $\{d_i\}_{i=1}^{\mathsf{N}_v}$, we adopt a sequence pooling and an MLP for prediction.

Temporal Relational Module

The temporal relational module (TRM) is identical to MCoMeT , which allows the tuple matching between a temporally-ordered sub-sequence of frames (sampled from a video) to the counterparts (sampled from an audio). The idea behind this is that interactions in videos are presented at various speeds and locations, and are poorly depicted by a single frame, *e.g.*, *moving a glass cup*. To this end, we explore high-order temporal relationships by aggregating multiple frames of representations. Given a pair of ordered frames from a video with indices $p=(p_1,p_2)$, where $1\le p_1<p_2\le \mathsf{F}$ and $V=\{v_1,\cdots,v_{\mathsf{F}}\}$ is a video with F uniformly sampled frames, we define the pair representation as follows:

$$V_p=[\Phi(v_{p_1})+\mathrm{PE}(p_1),\Phi(v_{p_2})+\mathrm{PE}(p_2)]\in\mathbb{R}^{2\times d}, \quad (1)$$

where $\Phi:\mathbb{R}^{\mathsf{H}\times\mathsf{W}\times 3}\mapsto\mathbb{R}^d$ is a convolutional network to obtain a d-dimensional embedding of an input frame, and $\mathrm{PE}(\cdot)$ is a positional encoding given a frame index (Vaswani et al. 2017). We define the set of all possible pairs as

$$\Pi=\{(n_1,n_2)\in\mathbb{N}^2:1\le n_1<n_2\le\mathsf{F}\}. \quad (2)$$

We then extend the pair of frames to a sub-sequence of ordered frames of any length. Denote by κ the length, or cardinality, of a tuple, *e.g.*, $\kappa=2$ for a pair, $\kappa=3$ for a triplet. All possible tuples for any κ are given by

$$\Pi^\kappa=\{(n_1,...,n_\kappa)\in\mathbb{N}^\kappa:\forall i(1\le n_i<n_{i+1}\le\mathsf{F})\}. \quad (3)$$

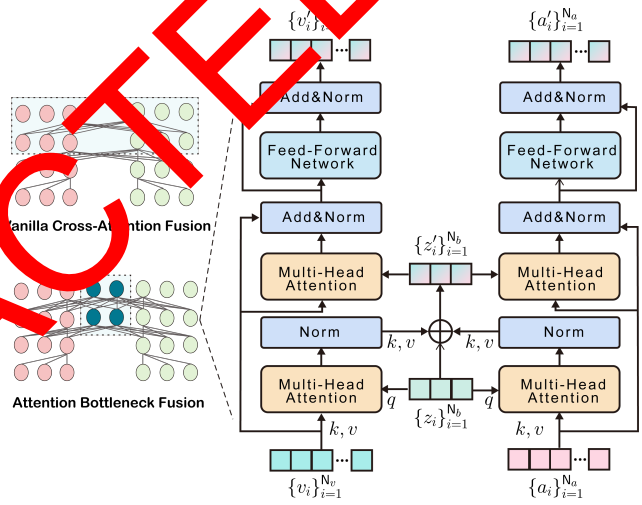

Figure 3: Schematic illustrations of the vanilla cross-attention (`CrossAtt`) fusion and our attention bottleneck fusion. The bottleneck tokens are designed for guiding the feature collection and propagation across modalities, greatly reducing computational cost compared to `CrossAtt`.

The associated video representation with respect to the tuple with indices $p=(p_1,...,p_\kappa)\in\Pi^\kappa$, generalizing the pair representation in Eq. 1, now becomes

$$V_p^\kappa=[\Phi(v_{p_1})+\mathrm{PE}(p_1),...,\Phi(v_{p_\kappa})+\mathrm{PE}(p_\kappa)]\in\mathbb{R}^{\kappa\times d}. \quad (4)$$

Let $\mathbb{C}$ denote the set of cardinalities. As an example shown in Fig. 2, pairs, triplets and quadruplets of frames are denoted by $\mathbb{C}=\{2,3,4\}$. This is accomplished similarly for the audio input. Note that, as the video frames are spaced so closely together in PACS, we take the surrounding 1.6 seconds of audio surrounding each video frame as an *audio frame* following (Yu et al. 2022).

Uni-modal Encoder

Understanding the physical properties and affordances of an object through a video requires an overall view of the global context (Zellers et al. 2022b). Therefore, to augment the features with a global context within each modality, we adopt a uni-modal encoder to mine the intra-modality correlations for video and audio. We stack several standard transformer encoder layers (Vaswani et al. 2017), each consisting of a multi-head self-attention block and a feed-forward network (FFN). In each attention head, self-attention for either of the visual or audio modality $x \in \{v, a\}$ can be computed as

$$x_i' = x_i + w_z \sum_{j=1} \text{Softmax}(\frac{(w_q x_i)(w_k x_j)^\top}{\sqrt{d}}) w_v x_j, \quad (5)$$

where x_i and x_i' are the input and output features of a video or audio clip i, followed by a two-layer FFN for feature transformation and non-linearity. d is the feature dimension, and $w_{\{q,k,v,z\}}$ are learnable parameters for the *query*, *key*, *value*, and output matrices, respectively.

Cross-modal Encoder

Previous studies (Zellers et al. 2022a; Nagrani et al. 2021) have pointed out that exploiting complementarity and redundancy across modalities helps to achieve a more general representation. Hence, we employ a cross-modal encoder to jointly capture the inter-modality correlations. Consider that as typical natural signals, visual and audio features have heavy spatial-temporal redundancy and noisy information that is useless for another modality. Besides, cross-modal attention is computationally prohibitive, *e.g.*, vanilla cross attention (Badamdorj et al. 2021) with quadratic complexity in computing pairwise attention with entire token sequence length. To mitigate the above problems, a recent work (Nagrani et al. 2021) attempts to restrict cross-modal communication to a small set of bottleneck tokens, showing promising results in reducing computation costs. Inspired by this, we extend the concept of bottleneck tokens and tailor a feature *collection* and *propagation* mechanism for better capturing correlations across modalities, as illustrated in Fig. 3.

Feature Collection Following (Nagrani et al. 2021), we introduce bottleneck tokens $\{z_i\}_{i=1}^{\mathsf{N}_b}$ to capture the inter-modality correlations. Here N_b is the number of bottleneck tokens and is set to be much smaller than the number of video clips N_v, in order to *condense* the most relevant parts across all modalities. The feature collection is achieved by several multi-head attentions between bottleneck tokens and the features from different modalities. Formally, we have

$$z_i' = z_i + w_z \sum_{j=1} \text{Softmax}(\frac{(w_q z_i)(w_k x_j)^\top}{\sqrt{d}}) w_v x_j, \quad (6)$$

where z_i and z_i' are input and output features of bottleneck tokens. Other notations are consistent with Eq. 5. The only difference between Eq. 5 and Eq. 6 lies in the query matrix z_i, aiming to aggregate features into bottleneck tokens. This feature collection operates on both visual and audio features so that multi-modal information is refined and *condensed* into bottleneck tokens.

Feature Propagation After collecting and condensing the multi-modal information, we propagate the *condensed* features into each modality using another multi-head attention by:

$$x_i' = x_i + w_z \sum_{j=1} \text{Softmax}(\frac{(w_q x_i)(w_k z_j)^\top}{\sqrt{d}}) w_v z_j, \quad (7)$$

where x_i' denotes the cross-modality enhanced features of a video or audio clip i. These features are then fed into FFNs for further transformation. Notice that in the first step we have obtained the refined features, our model only propagates information that is useful for the other modality in the second step, eliminating the diffusion of noisy information.

Query Generator

In line with (Carion et al. 2020; Meng et al. 2021), we argue that query embeddings should naturally guide the representation decoding process. Thus we design a query generator to adaptively generate clip-wise queries depending on the natural language input (**question**). We build this query generator via a multi-head attention layer, where joint audiovisual representations $\{r_i\}_{i=1}^{\mathsf{N}_q}$ serve as *query*, and textual features of the question are *key* and *value*. The joint audiovisual representations are output by the attention bottleneck layer, consisting of representations of two objects ($\boldsymbol{o_1}, \boldsymbol{o_2}$) with $\mathsf{N}_q = 2\mathsf{N}_b$. We hope that by computing attention weights between video clips and text queries, each clip can learn whether it contains the physical properties of objects involved in the problem and predict a query embedding.

Query Decoder and Prediction Heads

Query decoder takes the joint audio-visual representations $\{r_i\}_{i=1}^{\mathsf{N}_q}$ and text-guided query embeddings $\{q_i\}_{i=1}^{\mathsf{N}_q}$ as inputs, and decodes the joint multi-modal features to choose the most sensible answer. The output sequence of the query decoder has the same length as the encoder input. We adopt the sequence pooling (Hassani et al. 2021) to gather the outputs of the query decoder, followed by an MLP to predict the answer. During training, we apply the binary cross-entropy loss for optimization.

Experiments

Datasets and Evaluation Metric

We follow the well-established settings for physical commonsense reasoning (Yu et al. 2022). Concretely, we use the PACS dataset and benchmark MCOMET on two tasks, including a commonly-used proxy question-answer (QA) task and a PACS-material classification task. Errors in QA tasks can be attributed to two main factors, *i.e.*, either misidentifying an object's properties or correctly identifying an object but failing to reason about its properties. Material classification aims to distinguish the physical properties of a pair of objects by a comparison question, *e.g.*, *which object is more likely to be made out of glass?* Results of the material classification offer us an estimate of how many errors arise from the misidentified objects and how many from the failure of higher-order reasoning. Accuracy is adopted as the metric.

PACS (Yu et al. 2022) conceptualizes the datapoints. Each datapoint is a quadruplet $\langle \boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{q}, \boldsymbol{l} \rangle$, denoting the *two objects*, the *question* and the *answer*. PACS contains a total of 1,377 unique questions, each used multiple times in different pairs of objects, with an average of 5.86 questions per pair. PACS comprises a total of 1,526 objects, each represented by a unique video segment, with included audio and a bounding box in the middlemost frame of the video. Each video has an average length of 7.6 seconds. For material classification, we used the exact same object pairs as in the QA task, and accompany each pair with comparison questions based on each object's material. `Train/val/test` splits consist of 3,460/444/445 datapoints, respectively.

Human Performance. A random sampling set of 243 datapoints from the PACS (Yu et al. 2022) is given to 10 new annotators to answer. Each datapoint is annotated as five answers with audio and five answers without audio. Human accuracy is calculated as a majority vote (Bisk et al. 2020; Zellers et al. 2019) of the five datapoints and a 90% confidence interval is reported for the results.

Implementation Details

General Pipeline Given a datapoint $\boldsymbol{d} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{q}, \boldsymbol{l})$, where each object is represented by $\boldsymbol{o_n} = (\boldsymbol{i}_n, \boldsymbol{a}_n, \boldsymbol{v}_n)$, which are the image (containing a drawn-on bounding-box), audio, and video inputs. We first extract unimodal embeddings for both objects $(\boldsymbol{e_i}, \boldsymbol{e_a}, \boldsymbol{e_v})$ as well as the text embeddings of question $\boldsymbol{e}_q$. Specifically, we adopted a Vision Transformer (ViT) (Dosovitskiy et al. 2021) as the image encoder, Audio Spectogram Transformer (AST) (Gong, Chung, and Glass 2021) as the audio encoder, Temporal Difference Network (TDN) (Wang et al. 2021) as the video encoder, and DeBERTaV3 (He, Gao, and Chen 2021; He et al. 2020) as the question encoder. $(\boldsymbol{e_i}, \boldsymbol{e_a}, \boldsymbol{e_v})$ are concatenated and sent into MLP to generate an object embedding $\boldsymbol{e_o}$. Then, we separately concatenate the two object embeddings with the text (question) embedding $\boldsymbol{e}_q$ and use an MLP to generate two question-aware object embeddings. Finally, the two question-aware object embeddings are concatenated and sent to an MLP for answer prediction.

In particular, we considered the following baselines, and make them applicable for commonsense physics reasoning.

- **Late Fusion (Pandeya and Lee 2021)** The image, audio, and video encoders are pretrained on the ImageNet (Deng et al. 2009), AudioSet (Gemmeke et al. 2017), and Something-Something V2 (Goyal et al. 2017), respectively. We also use the pretrained DeBERTa (He et al. 2020)[1], outperforming the BERT (Kenton and Toutanova 2019) via disentangled attention and enhanced mask decoder. The way to fuse unimodal embeddings is the same as the general pipeline.
- **CLIP (Radford et al. 2021)** is a joint image-text model trained on million-scale image-text pairs via contrastive learning, which exhibits strong zero-shot learning ability. CLIP resorts to an image encoder and a text encoder to encode image and text respectively, where the similarity between image and text embeddings is measured by the cosine similarity: $\text{sim}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^\top \boldsymbol{b} / ||\boldsymbol{a}|| ||\boldsymbol{b}||$. Similar to CLIP, we separately embed images of both objects, and the question. We learn to maximize the cosine similarity between matched images and question embeddings, while minimizing the unmatched ones.
- **AudioCLIP (Guzhov et al. 2020)** extends CLIP for audio modality by introducing an audio head, which enables the embedding of audio inputs into the same vector space. Analogous to AudioCLIP, we extend the CLIP model mentioned above by including audio. We concatenate image and audio embeddings for both objects and use a linear layer to project them onto the same vector space as the question embedding. We predict the object that has an embedding closer to the question embedding.
- **UNITER (Chen et al. 2020)** is a large-scale pretrained model for joint image-text embedding. It first embeds image regions (visual features and bounding box features) and textual words (tokens and positions) into a common embedding space with an image embedder and a text embedder. Then it uses a transformer to learn sound contextualized embeddings via the four pre-training tasks. We finetuned UNITER on the NLVR2 dataset (Suhr et al. 2019). We split up both objects and generate two object-question embeddings, and finally concatenate them followed by a linear layer to classify the answer.
- **MERLOT (Zellers et al. 2021)** is a large-scale visual-language model that learns to match contextualized captions with their corresponding video frames. Beyond contrastive frame-transcript matching, it also includes a temporal reordering loss, *i.e.*, judging the order between two video frames. **MERLOT RESERVE (Zellers et al. 2022a)** extends MERLOT by including audio. We finetuned Merlot RESERVE on VCR (Zellers et al. 2019) and TVQA (Lei et al. 2018) by constructing two multimodal sequences using all input modalities. Like PACS, we also train MERLOT RESERVE w/ and w/o audio.

Comparison with State-of-the-Art Methods

PACS QA Table 1 shows the results of MCoMeT and its comparison approaches on PACS QA task. It can be seen that MCoMeT consistently outperforms existing methods across various setups. With multiple runs over all compared models to reduce randomness, MCoMeT achieves the best average performance across all setups, with a large margin of + 13.7% accuracy compared with the second-best method, namely Merlot Reserve. The strong performance demonstrates the superiority and robustness of the proposed model.

PACS Material Table 1 also lists the results on the PACS-material classification. It can be seen that, MCoMeT still outperforms all existing methods under all setups by even larger margins than that on PACS QA, as PACS-material classification is much easier than PACS QA. This can be potentially attributed to the effective exploitation of the temporal information and interactions both in intro-modality and inter-modality. Considering the only difference between PACS-material and PACS-QA lies in the content of the questions, we believe that the performance gap is due to the dif-

[1]The models and checkpoints are available at `https://huggingface.co/models?other=deberta-v3`

Baseline Model	PACS-QA Acc. (%) w/o audio	PACS-QA Acc. (%) w/ audio	PACS-QA Acc. (%) Δ	PACS-Material Acc. (%) w/o audio	PACS-Material Acc. (%) w/ audio	PACS-Material Acc. (%) Δ
I+A+V (Yang et al. 2020; Lei et al. 2018)	–	$51.9_{\pm 1.1}$	–	–	$51.9_{\pm 1.1}$	–
Q+I (Zadeh et al. 2019; Lei et al. 2018)	$51.2_{\pm 0.8}$	–	–	$51.2_{\pm 0.8}$	–	–
Q+A (Zadeh et al. 2019; Lei et al. 2018)	–	$50.9_{\pm 0.6}$	–	–	50.9±0.6	–
Q+V (Zadeh et al. 2019; Lei et al. 2018)	$51.5_{\pm 0.9}$	–	–	$51.5_{\pm 0.9}$	–	–
Late Fusion (Pandeya and Lee 2021)	$52.5_{\pm 1.6}$	$55.0_{\pm 1.1}$	2.5↑	$63.8_{\pm 1.3}$	$67.4_{\pm 1.5}$	3.6↑
CLIP (Radford et al. 2021)	$56.3_{\pm 0.7}$	–	–			
AudioCLIP (Guzhov et al. 2020)	–	$60.0_{\pm 0.9}$	–	–	$75.9_{\pm 1.1}$	–
UNITER (Large) (Chen et al. 2020)	$60.6_{\pm 2.2}$	–	–	$72.4_{\pm 1.8}$	–	–
MERLOT RESERVE (Base) (Zellers et al. 2022a)	$64.0_{\pm 0.9}$	$66.5_{\pm 1.4}$	2.6↑	$77.6_{\pm 2.3}$	$81.3_{\pm 1.7}$	3.7↑
MERLOT RESERVE (Large) (Zellers et al. 2022a)	$68.4_{\pm 0.7}$	$70.1_{\pm 1.0}$	1.8↑	$79.8_{\pm 2.4}$	$83.6_{\pm 1.3}$	3.8↑
MCoMeT ($\mathbb{C}=\{1\}$)	**71.6**$_{\pm 1.1}$	**75.9**$_{\pm 0.6}$	**4.3**↑	**81.1**$_{\pm 1.8}$	**87.9**$_{\pm 1.6}$	**6.8**↑
MCoMeT ($\mathbb{C}=\{2,3\}$)	**75.2**$_{\pm 0.9}$	**79.7**$_{\pm 1.1}$	**4.5**↑	**86.2**$_{\pm 1.3}$	**92.2**$_{\pm 1.4}$	**6.0**↑
Human Performance	$90.5_{\pm 3.1}$	$96.3_{\pm 2.1}$	5.9↑	*	*	*

Table 1: Performance comparison on PACS test set: all compared models are reported with the mean and standard deviation of 5 runs, while human accuracy is reported with a 90% confidence interval. Δ represents the performance gap between models with and without audio. ∗ indicates results are not available.

Model	V $\mathbb{C}=\{1\}$	V $\mathbb{C}=\{2,3\}$	A+V Co-Att	A+V Bo-Att	PACS-QA Acc. (%)
MCoMeT	✓				$71.6_{\pm 1.1}$
		✓			$75.2_{\pm 0.9}$
			✓		$78.3_{\pm 0.7}$
				✓	$79.7_{\pm 1.1}$

Table 2: Ablation study to evaluate the effectiveness of key design choices in MCoMeT, including visual-only modality and different audio-visual fusion manners.

Cardinality	Num. of Tuples	PACS-QA w/o audio	PACS-QA w/audio
$\mathbb{C}=\{1\}$	-	71.6	75.9
$\mathbb{C}=\{2\}$	28	74.4	77.9
$\mathbb{C}=\{3\}$	56	74.9	79.3
$\mathbb{C}=\{4\}$	70	73.9	79.4
$\mathbb{C}=\{2,3\}$	84	**75.2**	**79.7**
$\mathbb{C}=\{2,4\}$	98	73.8	78.8
$\mathbb{C}=\{3,4\}$	126	74.6	79.7
$\mathbb{C}=\{2,3,4\}$	154	74.6	79.4

Table 3: Performance comparison (Accuracy, %) between different sets of cardinalities. The number of tuples for each model variant is defined by $\sum_{\kappa\in\mathbb{C}}|\Pi^{\kappa}|$.

ficulty of reasoning about physical commonsense in the QA task. The remaining 10%-15% of misidentified datapoints on PACS-material may stem from a true failure in understanding the material composition of the object.

Ablation Studies

We conduct comprehensive ablation studies to verify the effectiveness of our design choices. All results are averaged over 5 runs on the same train/val/test split of PACS.

Effect of Fusion Strategy We first ablate the fusion strategy, where the `CrossAtt` (illustrated in Fig. 3) denotes the unrestricted pairwise attention with all other latent units from both modalities. In Table 2, we can see that using temporal-relational representations from ordered tuples of frames works better than using single-frame representations. The benefit of including audio is also significant, with an accuracy gain of (at least) +3.1 compared to the visual-only modality. Moreover, we observe that `Bottleneck` performs marginally better than the vanilla `CrossAtt`, revealing the effectiveness of our attention bottleneck fusion module in facilitating multimodal (audio-visual) learning.

Temporal-Relational Module Table 1 reports results using $\mathbb{C}=\{2,3\}$, which indicates pair and triplet ordered tuple of frames shown in Fig. 2. We now evaluate each cardinality of $\mathbb{C}\in\{1,2,3,4\}$ independently and their combinations in Table 3. First of all, results show a considerable improvement of (+3.9%) in MCoMeT moving from single frame to pair comparisons. The performance of the audio-visual modality is further improved for triplets (+1.7%) and is only marginally improved for quadruplets (+0.1%). Using all cardinalities $\mathbb{C}=\{2,3,4\}$ slightly degrades the performance. Overall, $\mathbb{C}=\{2,3\}$ performs best and is used by default.

Number of Sampled Frames We study the impact of the number of sampled audio/video frames on MCoMeT . In PACS, the average length of each video is 7.6 seconds. We uniformly select a range of $[4,11]$ video frames to determine the optimal sampled frames for MCoMeT . Correspondingly, the number of sampled audio frames is the same as the video. Concretely, we take the 1.6 seconds of audio surrounding each video frame. This leads to an overlap of audio in the input sequence, but this is necessary because 1.6 sec-

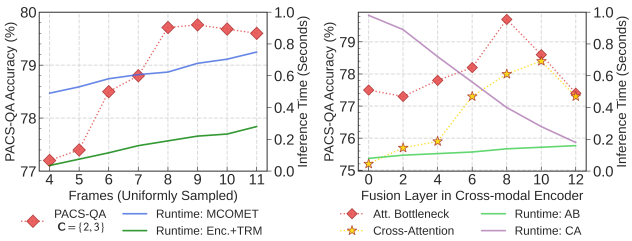

Figure 4: *Left*: Ablations on sampling different number of video/audio frames. *Right*: Comparison of fusion strategies using the vanilla cross-attention (CA) and the proposed attention bottleneck (AB) on different fusion layers L_f.

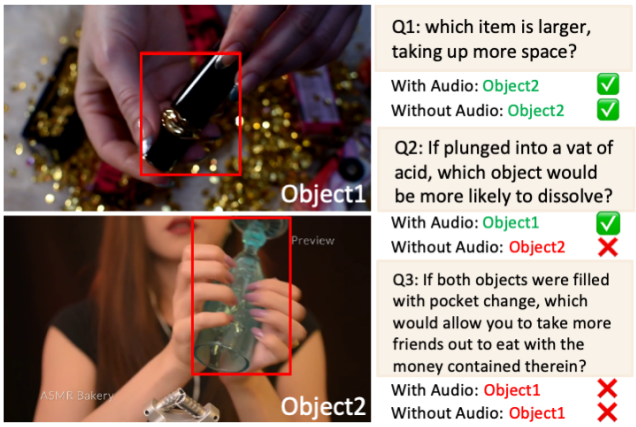

Figure 5: Demonstration of the predictions from the two MCoMeT variants (trained with and without audio), where `object1` is a metal tube and `object2` is a glass cup. In a visual-only scenario, object1 could be mistaken for ceramic, while object2 could be mistaken for plastic. In the absence of audio, the model is hard to recognize the materials of both objects, thus giving a wrong answer to Q2. The failure of Q3 may be attributed to implicit inquiring about the size/shape of the two objects.

onds is the minimum length allowed for audio clips (Yu et al. 2022). Fig. 4 (*right pane*) plots the results of MCoMeT with $\mathbb{C} = \{2, 3\}$ varying with the number of video/audio frames. As illustrated, moving frames from 4 to 8 increases the performance. Frames greater than 8 would not result in significant gains and even cause a slight performance drop. In addition, we see that the runtime of our model scales linearly with the number of frames. The two separate encoders equipped with the temporal-relational module account for approximately $1/3$ of the total running time.

Impact of Fusion Strategy Fig. 4 (*right pane*) unveils the influence of varying the fusion layer L_f for two fusion strategies, *i.e.*, vanilla cross-attention (`CrossAtt`) and attention bottleneck (`Bottleneck`) fusion. The number of bottleneck tokens is fixed to 4 across all layers. As shown in Fig. 4, `Bottleneck` consistently outperforms `CrossAtt` on all layers, achieving the optimal performance at $L_f = 8$. This suggests that the model benefits from restricting cross-modal attention via a small set of bottleneck tokens, allowing earlier layers to focus on learning unimodal features while deep layers still capture cross-modal interactions. It is also seen that, for audio-visual modality fusion in transformer, a later layer fusion (within a certain range) is often a better choice. However, layers that are overly late, *e.g.*, the last layer, may suffer performance degradation due to the inability to interact effectively with low-level features.

Fig. 4 also compares the inference time between the two fusion strategies. It is clear that using a small number of bottleneck tokens, *e.g.*, $B = 4$ for our experiments, brings marginal computation complexity, with inference speed remaining largely constant with varying fusion layer L_f. This is in contrast to the vanilla cross-attention fusion, which has a non-negligible computational cost for every layer it is applied to. We note that for mid fusion with $L_f = 8$, `Bottleneck` outperforms `CrossAtt` by a substantial margin, with almost half the running time.

Case Study

To gain a better understanding of our model, we analyze some examples from PACS test set. Generally, when the physical properties of an object can be directly identified visually, reasoning can be done using visual modality alone. As is seen in Fig. 5 Q1, determining which object takes up more space only requires a general understanding of the size/shape. When models need to deal with visually ambiguous objects, audio is necessary to clarify the physical properties of an object, *e.g.*, in Fig. 5 Q2, the materials of object1 and object2 are key to answering the question. It is hard to tell by vision whether object2 is made of plastic or glass. If audio is added, object2 can easily be identified as glass, which is known to be insoluble in acid. However, despite the presence of audio, our model may still fail when the problem itself involves more implicit physical properties or requires *multi-hop reasoning* (*e.g.*, see Fig. 5 Q3). This enlightened us to leverage graph structures (Ding et al. 2019) to design fine-grained reasoning paths and rules in the future.

Conclusion

This paper presents a unified framework for physical audio-visual commonsense reasoning, namely MCoMeT . It has two attractive points: *i*) it fully explores higher-ordered temporal relationships between different modalities (*e.g.*, pairs, triplets, and quadruplets); and *ii*) it limits the cross-modal flow via the feature *collection* and *propagation* mechanism with tight fusion bottlenecks, largely reducing computation cost and suppressing the noisy information. Experimental results on the PACS benchmark demonstrate the superiority and effectiveness of MCoMeT . We hope that this work will shed light on multimodal physical commonsense reasoning.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Project 62076096, Shanghai Municipal Project 20511100900, and Shanghai Knowledge Service Platform Project (No. ZF1213). Prof. Shiliang Sun is

the corresponding author of this paper.

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, volume 34, 24206–24221.

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2021. Joint visual and audio learning for video highlight detection. In *ICCV*, 8127–8137.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2): 423–443.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, volume 34, 7432–7439.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Ding, M.; Zhou, C.; Chen, Q.; Yang, H.; and Tang, J. 2019. Cognitive graph for multi-hop reading comprehension at scale. *arXiv preprint arXiv:1905.05460*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Forbes, M.; and Choi, Y. 2017. Verb Physics: Relative Physical Knowledge of Actions and Objects. In *ACL*, 266–276.

Forbes, M.; Holtzman, A.; and Choi, Y. 2019. Do Neural Language Representations Learn Physical Commonsense? In *CogSci*.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 776–780.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, 571–575.

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.; and Memisevic, R. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. 5843–5851.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2020. AudioCLIP: Extending CLIP to Image, Text and Audio. *arXiv preprint arXiv:2008.04838*.

Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; and Shi, H. 2021. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hessel, J.; Mimno, D.; and Lee, L. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL*, 2194–2205.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. In *ICML*, 4651–4664. PMLR.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

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

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*, 1369–1379.

Mei, H.; Bansal, M.; and Walter, M. R. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *AAAI*.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *ICCV*, 3651–3660.

Morvant, E.; Habrard, A.; and Ayache, S. 2014. Majority vote of diverse classifiers for late fusion. In *SSPR*, 153–162. Springer.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. In *NeurIPS*, volume 34, 14200–14213.

Pandeya, Y. R.; and Lee, J. 2021. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80: 1–19.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Storks, S.; Gao, Q.; Zhang, Y.; and Chai, J. Y. 2021. Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. In *EMNLP*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *ACL*, 6418–6428.

Sun, S.; and Zhao, J. 2020. *Pattern Recognition and Machine Learning*. Beijing: Tsinghua University Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, volume 30.

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021. TDN: Temporal Difference Networks for Efficient Action Recognition. In *CVPR*, 1895–1904.

Yang, J.; Zhu, Y.; Wang, Y.; Yi, R.; Zadeh, A.; and Morency, L.-P. 2020. What Gives the Answer Away? Question Answering Bias Analysis on Video QA Datasets. arXiv:2007.03626.

Yatskar, M.; Ordonez, V.; Zettlemoyer, L.; and Farhadi, A. 2017. Commonly uncommon: Semantic sparsity in situation recognition. In *CVPR*, 7196–7205.

Yu, S.; Wu, P.; Liang, P. P.; Salakhutdinov, R.; and Morency, L.-P. 2022. PACS: A Dataset for Physical Audiovisual CommonSense Reasoning. *arXiv preprint arXiv:2203.11130*.

Zadeh, A.; Chan, M.; Liang, P. P.; Tong, E.; and Morency, L.-P. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, 8807–8817.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 6720–6731.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022a. MERLOT Reserve: Multimodal Neural Script Knowledge through Vision and Language and Sound. In *arxiv*.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022b. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 16375–16387.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 34: 23634–23651.

Zhao, Z.; Papalexakis, E.; and Ma, X. 2020a. Learning Physical Common Sense as Knowledge Graph Completion via BERT Data Augmentation and Constrained Tucker Factorization. In *EMNLP*.

Zhao, Z.; Papalexakis, E.; and Ma, X. 2020b. Learning Physical Common Sense as Knowledge Graph Completion via BERT Data Augmentation and Constrained Tucker Factorization. In *EMNLP*.