Counterfactual Debiasing for Physical Audiovisual Commonsense Reasoning

Daoming Zong[1*], Chaoyue Ding[2*], Kaitao Chen[2], Yinsheng Li[2†], Shuaiyu Wang[2]

[1]SenseTime Research
[2]School of Computer Science, Fudan University, Shanghai, China
{ecnuzdm,cydingcs}@gmail.com, chenkaitao@pjlab.org.cn, liys@fudan.edu.cn

Abstract

Physical commonsense is an essential aspect of human cognition, involving an *intuitive understanding of the physical properties and interactions of everyday objects and materials*. Though physical commonsense reasoning should inherently be a multisensory task, integrating both video and audio signals, existing physical audiovisual commonsense reasoning (PACR) models predominantly rely on visual information. This reliance leads to spurious correlations and undermines the models' reasoning and generalization abilities. To counteract this, we introduce a *model-agnostic* Counterfactual Physical Audiovisual Commonsense Reasoning (`CF-PACR`) framework aimed at mitigating visual bias-induced spurious effects. Specifically, we construct a traditional PACR model using both audio and visual information as the factual reasoning model. Subsequently, in the counterfactual reasoning model, we isolate visual information to estimate direct effects. Finally, we subtract the direct effects from the total effects across modalities to derive indirect effects, thereby mitigating visual biases. Extensive experiments validate the effectiveness and generalizability of `CF-PACR` in alleviating the *spurious correlations* between visual modality and model predictions.

Introduction

To facilitate the safe deployment of artificial intelligence (AI) systems across a myriad of real-world scenarios, it is imperative that these systems acquire an understanding of the *physical properties, affordances, and interactions of everyday objects, as well as how to manipulate them* (Krishna et al. 2017; Yatskar et al. 2017; Hessel, Mimno, and Lee 2018; Forbes, Holtzman, and Choi 2019; Yu et al. 2022). This general understanding of physical attributes and interactions of objects is crucial for the development of secure and trustworthy AI systems (Bisk et al. 2020). For instance, "*in the absence of a brush for makeup application, the expectation would be for a robot to proffer a cotton swab rather than a toothpick*". Previous research has delved into harnessing visual and textual data to explore physical commonsense (Jimenez 2020; Storks et al. 2021). Nonetheless, reasoning about physical commonsense is inherently a cross-modal task, as physical attributes can manifest across various modalities including, but not limited to, *visual and auditory cues*. In scenarios where an object appears visually ambiguous or rare, auditory information may provide essential cues for identifying its physical characteristics.

However, existing physical audiovisual commonsense reasoning (PACR) methods heavily rely on the correlation between visual features and class labels for prediction. This dependency appears particularly fragile when dealing with unknown or rare testing samples (Sun et al. 2021, 2022). Specifically, when objects in test samples are presented in ways unseen or rare to models, they tend to infer based on the distribution of visual clues in prior learned experience, leading to prediction errors. For instance, in the test datapoint depicted in Fig. 1, the ground truth labels of the involved objects (`Object1`, `Object2`) are `[foam]` and `[ceramic]`, respectively. In the training set, the number of samples with the same object combination (*i.e.* `[foam]+[ceramic]`) is only 2, while those annotated with (`[foam]+[wood]`) and (`[foam]+[plastic]`) are 22 and 62, respectively. Due to the shortcut effect (Ye and Kovashka 2021), the model is prone to misidentifying `Object2` as *wood*, as it fails to adapt to this rare pattern of visual object combination, thereby diminishing the model's generalization ability in long-tail scenarios. This correlation between visual features and predictive outcomes is also referred to as the *spurious effect* in causal inference (Niu et al. 2021; Ma et al. 2022), highlighting the adverse effects of over-reliance on the visual modality for `PACR`.

To address the issue of spurious correlations, one intuitive approach is the collection of unbiased datasets. However, the prevalence of the long-tail distribution of the visual world (Kang et al. 2019; Cao et al. 2019; Liu et al. 2019) limits our ability to offer a sufficient number of matched training samples for all test samples. An alternative is to develop advanced cross-modal fusion methods (Lu et al. 2019; Xiao et al. 2020; Zellers et al. 2021), such as those based on attention mechanisms (Nagrani et al. 2021) or pretraining surrogate tasks (Georgescu et al. 2023). These sophisticated fusion strategies enable models to transcend reliance on solitary visual features by leveraging auditory information to enhance visual perception and improve the interpretation of complex scenes. Nonetheless, these methods do not adequately eliminate the inherent risk of visual biases.

[*]Equal contribution.
[†]Corresponding author.

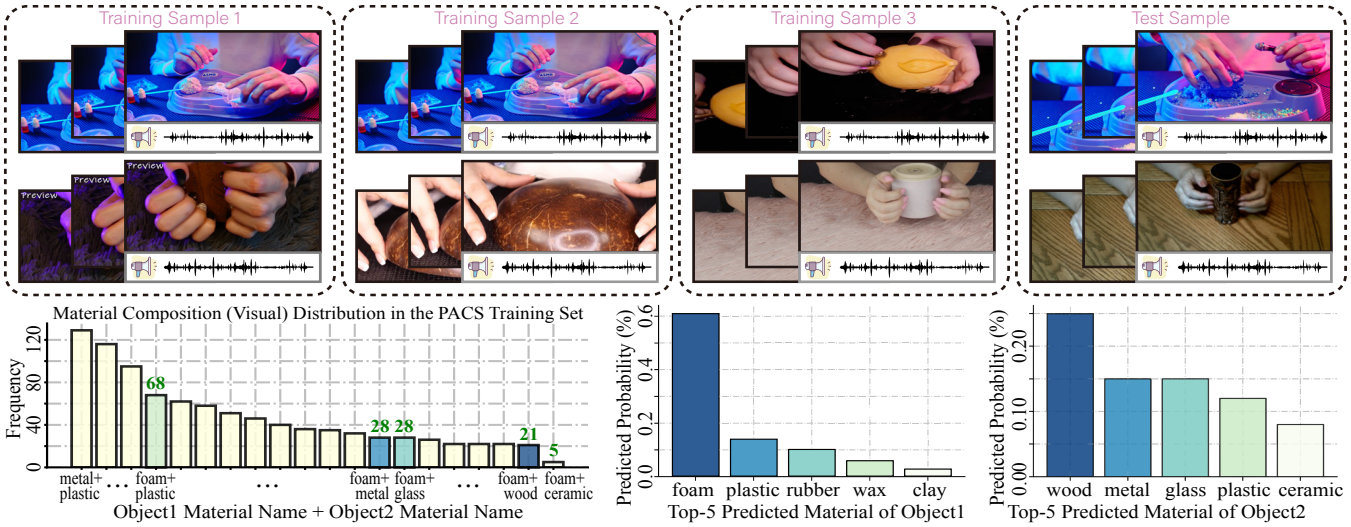

Figure 1: Visual Bias in PACS Task. The observed visual bias results from an imbalanced distribution of object combinations in the training data, causing the model to learn spurious correlations. As shown in the test case, AudioCLIP (Guzhov et al. 2020) incorrectly classifies '`ceramic`' as '`wood`', demonstrating how the model's overexposure to the '`foam`'+'`wood`' combination during training leads to erroneous associations.

To this end, we introduce a general framework—the Counterfactual Physical Audiovisual Commonsense Reasoning (`CF-PACR`) framework—to mitigate the adverse effects of visual biases on physical commonsense reasoning tasks. This framework is motivated by an analysis of the role of the visual modality in model predictions, revealing two key influences: *direct and indirect effects*. Direct effects refer to predictions relying directly on visual object combinations, often manifesting as spurious correlations in the visual modality, while indirect effects, through the multimodal fusion of visual and audio, extract reliable visual cues, showing the benefits of visual information. `CF-PACR` achieves debiased reasoning by separating direct effects from the total effect, preserving the pure indirect effects.

In the training phase of `CF-PACR`, we conduct factual reasoning using video, audio, and their fusion information; in the testing phase, we endow `CF-PACR` with counterfactual analysis capabilities. Counterfactual reasoning scenarios explore such a hypothetical situation: *how would the model identify the material category of an object or answer commonsense questions if it only had access to visual information?* By contrasting the results of factual and counterfactual reasoning, the model achieves more accurate classification scores. The highlights of this research are summarized as follows: (1) We identify the *spurious correlations* between visual concept combinations and model predictions in the PACS dataset, revealing their limitations on model reasoning and generalization. (2) The proposed model-agnostic `CF-PACR` framework leverages causal graph models to distinguish visual biases learned within visual features (captured by direct effects of the visual modality) from effective visual cues learned within multimodal fusion (captured by indirect effects of the fused modality). (3) Extensive experiments validate the effectiveness and generalizability of `CF-PACR`, demonstrating considerable improvements over traditional PACR models using counterfactual inference.

Related Work

Physical commonsense refers to a general understanding of the *physical attributes* and *affordances* of common objects in everyday life (Forbes, Holtzman, and Choi 2019; Bisk et al. 2020; Jimenez 2020; Zhao, Papalexakis, and Ma 2020; Storks et al. 2021). Physics commonsense reasoning is fundamentally a multimodal task, as physical properties can be presented through multiple modalities including visual and audio (Yu et al. 2022; Lv et al. 2024). Recent research has explored physics commonsense in various visual tasks, including scene understanding (Chen et al. 2019), visual deanimation (Wu et al. 2017), activity recognition (Li et al. 2023), and *cause-and-effect* relationship prediction (Mottaghi et al. 2016). However, these approaches focus solely on the visual modality, posing significant challenges for tasks involving unseen or occluded objects. If two objects are visually similar, audio can provide valuable information to distinguish their physical properties. To this end, Yu et al. established the first physics audiovisual commonsense reasoning (PACS) benchmark. Lv et al. introduced a disentangling counterfactual learning approach for physics audiovisual commonsense reasoning. In contrast to these methods, `CF-PACR` is tailored to mitigate the impact of *visual biases* in PACS tasks. The motivation for `CF-PACR` stems from an analysis of the complexity of visual information, wherein visual cues simultaneously contain beneficial clues for identifying physical attributes of objects and detrimental biases that induce prediction errors.

Method

Task Definition of PACS

PACS (Yu et al. 2022) is a video-based audiovisual benchmark designed to evaluate the model's ability to reason about physical commonsense using audio and visual modalities.

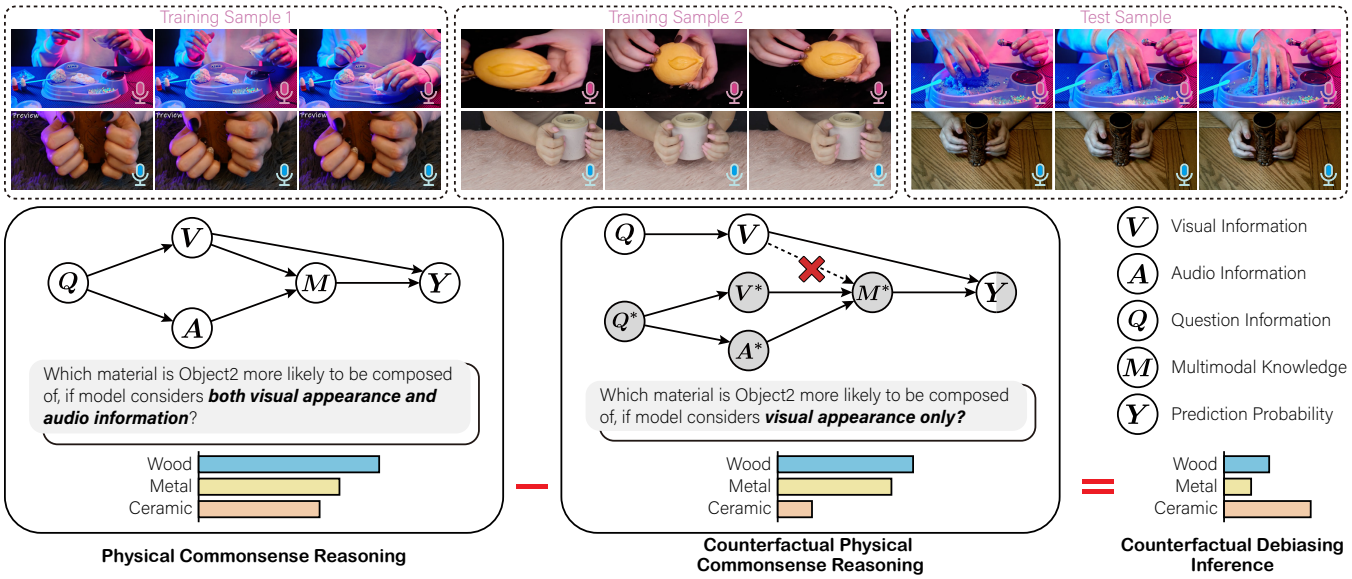

Figure 2: Illustration of the causal graph in the physical audiovisual commonsense reasoning (PACR) task.

Its core tasks include (multi-class) object material classification (i.e. PACS-Material) and binary question answering (i.e. PACS-QA). Given a question q and two objects o_1 and o_2, the model is tasked with selecting the material category for a query object or picking up the object that best fits the question. Each object o is represented by a video v, an audio a, and a bounding box b drawn around the object in the video frames. Hence, each datapoint in the PACS task is a tuple $(q, (b_1, v_1, a_1), (b_2, v_2, a_2), \ell)$, where ℓ denotes the label indicating the material category of the query object (PACS-Material) or which object is the correct answer (PACS-QA).

Causal Graph in PACS Tasks

Fig. 2 presents the causal graph constructed specifically for the PACS tasks, comprising elements such as question $\mathcal{Q}$, video $\mathcal{V}$, audio $\mathcal{A}$, cross-modal representation $\mathcal{M}$, and model prediction $\mathcal{Y}$. This causal graph is designed with a broad universality, unrestricted by specific implementation details.

Node $\mathcal{Q}$ (Question): Utilizes a text feature extractor to derive features from questions, generating queries related to visual or auditory features. For a given question q, this node outputs utterance-level question representation $\mathcal{Q}$: Input: $\{q\} \rightarrow$ Output: $\{Q\}$.

Node $\mathcal{V}$ (Video): Contains an object visual encoder for learning the visual representation of objects. Given a pair of object videos (v_1, v_2), and a question q, this node outputs object-level visual representation $\mathcal{V}$: Input: $\{q, v_1, v_2\} \rightarrow$ Output: $\{V\}$, where $\mathcal{V}$ carries the visual information of objects related to the question, including beneficial context information and visual biases that may mislead the model.

Node $\mathcal{A}$ (Audio): Incorporates an object audio encoder to extract audio features of objects. Given audio clips (a_1, a_2), and a question q, this node outputs object-level audio representation $\mathcal{A}$: Input: $\{q, a_1, a_2\} \rightarrow$ Output: $\{A\}$, where $\mathcal{A}$ includes audio features of objects related to the question, providing crucial auditory clues for accurate material classification and physical property reasoning.

Link $(\mathcal{V}, \mathcal{A}) \rightarrow \mathcal{M}$ (Cross-modal Fusion): Combines video and audio information to generate a better joint audiovisual representation $\mathcal{M}$.

Node $\mathcal{M}$ (Cross-modal Representation): Contains a cross-modal fusion module for generating an audiovisual representation of each object $\mathcal{M}$: Input: $\{V, A\} \rightarrow$ Output: $\{M\}$, embedding more robust and comprehensive information than single-modal representations.

Link $\{\mathcal{V}, \mathcal{A}, \mathcal{M}\} \rightarrow \mathcal{Y}$ (Classifier): Formalized as:

Input: $\{V\} \rightarrow$ Output: $\{Y_v\}$, Input: $\{M\} \rightarrow$ Output: $\{Y_m\}$, where Y_v and Y_m correspond to classification scores associated with the visual representation $\mathcal{V}$ and the multimodal representation $\mathcal{M}$, respectively.

Node $\mathcal{Y}$ (Final Classification Result): Combines all scores $\{Y_v, Y_m\}$ through a fusion function h, yielding the final prediction score $Y_{v,m}$: Input: $\{Y_v, Y_m\} \rightarrow$ Output: $\{Y_{v,m}\}$. Three fusion functions were explored, including *Naive Sum*, *Log-sigmoid Sum*, and *Harmonic Mean*, to investigate their impact on prediction performance.

CF-PACR Framework

CF-PACR is designed to reduce visual bias in physical audiovisual commonsense reasoning tasks through counterfactual debiasing. This framework mitigates adverse impacts by blocking the direct pathway from visual object combinations to model predictions ($V \rightarrow Y$), which reflects the detrimental influence stemming from spurious correlations within the visual modality. We formalize this process by introducing a visual model E_v, and an audiovisual model E_m, as follows:

$$\{Y_v, f_v\} = \mathsf{E}_v(q, v), \quad \{Y_m\} = \mathsf{E}_m(q, v, a), \qquad (1)$$

where q, $v = \{v_1, v_2\}$, and $a = \{a_1, a_2\}$ denote a question, a pair of videos, and a pair of audios input to the model, respectively. As illustrated in Fig. 3b, the final score is derived

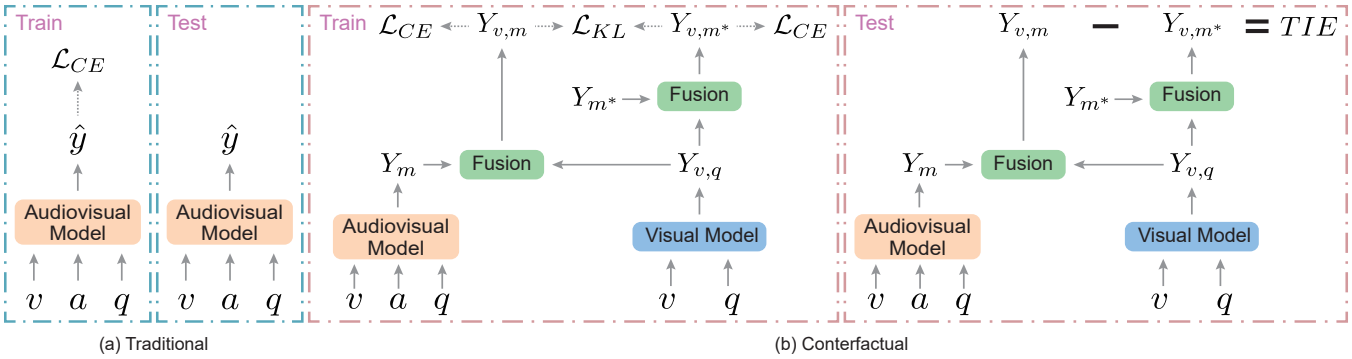

Figure 3: Illustration of traditional PACS models and the counterfactual debiasing inference of our CF-PACR framework.

by aggregating the two activation paths directly connected to Y through a fusion function h:

$$Y_{v,m} = h(Y_v, Y_m). \tag{2}$$

The *non-treatment condition* is represented by superscript $*$. For instance, giving the observed videos is denoted as $V = v$, and not offering the video is denoted as $V = v^* = \varnothing$. To accurately capture the spurious correlation between the visual modality and model predictions, we focus on the causal effect of the direct path $V \rightarrow Y$ while blocking other pathway activations. Given the challenge of inputting null values into neural networks, we introduce a learnable vector d to represent model output under the no-treatment condition. Our framework can be further refined as:

$$Y_v = \begin{cases} y_v = \mathsf{E}_v(q, v) & \text{if } V = v, \\ y_v^* = c & \text{if } V = v^*, \end{cases} \tag{3}$$

$$Y_m = \begin{cases} y_m = \mathsf{E}_m(q, v, a) & \text{if } V = v \text{ and } A = a, \\ y_m^* = d & \text{if } V = v^* \text{ or } A = a^*. \end{cases} \tag{4}$$

The total effect (TE) reflects the combined impact of the visual object $V = v$ and auditory information $A = a$ on the classification result Y, decomposable into the sum of direct and indirect effects:

$$TE = Y_{v,m} - Y_{v^*,m^*}, \tag{5}$$

where $Y_{v,m}$ is the observed outcome based on $V = v$ and $A = a$, and Y_{v^*,m^*} represents the potential outcome under the non-treatment condition. The proposed CF-PACR framework aims to block the direct effect $V \rightarrow Y$ while preserving the indirect effect $V \rightarrow M \rightarrow Y$. The natural direct effect (NDE) encapsulates the direct impact of visual information, calculated as:

$$NDE = Y_{v,m^*} - Y_{v^*,m^*}. \tag{6}$$

By subtracting the NDE from the TE (see Fig. 3b), we eliminate visual bias to obtain a more rational and accurate classification outcome, namely the total indirect effect (TIE):

$$TIE = TE - NDE = Y_{v,m} - Y_{v,m^*}. \tag{7}$$

In practice, the hyperparameter τ regulates the proportion of the NDE removed from the TE, thus controlling the computation of the indirect effect:

$$TIE = Y_{v,m} - \tau \cdot Y_{v,m^*}. \tag{8}$$

We select the classification result that maximizes TIE, unlike traditional posterior probability-based methods.

Implementation of CF-PACR

We implement the CF-PACR framework based on the causal graph depicted in Fig. 2.

Model Architecture The CF-PACR pipeline, depicted in Fig. 3, begins with computing Y_v and Y_m using a visual model E_v and an audiovisual model E_m, respectively, as defined in Eq. 1. For late fusion (Eq. 2), we evaluate three functions h: 1) *Naive Sum*: $Y_{v,m} = Y_v + Y_m$ 2) *Log-sigmoid Sum*: $Y_{v,m} = \log\left(\sigma\left(Y_v + Y_m\right)\right)$ 3) *Harmonic Mean*: $Y_{v,m} = \log \frac{\sigma(Y_v)\cdot\sigma(Y_m)}{1+\sigma(Y_v)\cdot\sigma(Y_m)}$, where $\sigma(\cdot)$ is the sigmoid function, and Y_v, Y_m are predicted probabilities from distinct PACS models. The visual model E_v processes queries by filtering relevant objects and extracting visual features. For example, in PACS-QA, CLIP's image and text encoders generate embeddings for keyframes and questions. Then we leverage the similarity between image and question embeddings to determine the probability of each object being the correct answer. The audiovisual model E_m supports multimodal inputs (video, audio, text) using frameworks like AudioCLIP, MERLOT, or MERLOT Reserve. The fused probabilities $Y_{v,m}$ are re-normalized to sum to 1.

Training To train CF-PACR, we adopt the cross-entropy (CE) loss for $Y_{t,m}$ and Y_{t,m^*} as follows:

$$\mathcal{L}_{CE} = \alpha * \mathrm{CE}(Y_{v,m}, y) + \beta * \mathrm{CE}(Y_{v,m^*}, y), \tag{9}$$

where α and β are the weights balancing the two loss functions, and y denotes the label for the PACS tasks. For instance, in PACS-Material, $y \in \mathbb{R}^{\mathsf{C}}$, where C represents the categories of materials, while in PACS-QA, $y \in \{0, 1\}$ indicating which object is the correct answer. Note that, apart from optimizing TE using $Y_{v,m}$, we introduce CE loss to encourage counterfactual NDE predictions Y_{v,m^*}, where $Y_{v,m^*} = h(Y_v, Y_{m^*})$, and Y_{m^*} represents the potential outcomes of the audiovisual multimodal model without video or audio inputs. Theoretically, we should utilize randomly initialized vectors or all-zero vectors to represent v^* and a^*, then input them into the audiovisual multimodal model E_m to obtain m^*. As only v^* and a^* affect the prediction of Y_{m^*}, and Y_{m^*} is unaffected by other features, we directly utilize a randomly initialized vector d to represent the final result Y_{m^*}, i.e., $Y_{m^*} = d$. Note that the dummy vector d is trainable during training, and all samples share it.

Inappropriate Y_{m^*} may cause discrepancies in the scales of TE and NDE, thus TIE will be dominated by one of them (Sun et al. 2022; Niu et al. 2021). Hence, we utilize the Kullback-Leibler (KL) divergence to minimize the difference between Y_{v,m^*} and $Y_{v,m}$ to estimate Y_{m^*}:

$$\mathcal{L}_{KL} = \text{KL}(Y_{v,m^*}, Y_{v,m}). \tag{10}$$

In summary, the final training loss is given by:

$$\mathcal{L} = \sum_{(q_i,v_i,a_i,y_i)\in\mathcal{D}} \mathcal{L}_{CE} + \gamma * \mathcal{L}_{KL}, \tag{11}$$

where γ is the hyperparameter controlling the effect of $\mathcal{L}_{KL}$. We optimize `CF-PACR` end-to-end.

Inference To perform inference using TIE, the final prediction of `CF-PACR` can be obtained as follows:

$$\text{TIE} = Y_{v,m} - \tau * Y_{v,m^*} = h(Y_v, Y_m) - \tau * h(Y_v, Y_{m^*}),$$

where τ controls the extent of subtracting the shortcut between the visual modality and model predictions.

Experiment

Dataset and Metric

Each datapoint in the PACS benchmark is represented as a quadruplet $\langle o_1, o_2, q, \ell \rangle$, consisting of a pair of objects, a question, and its associated label. The dataset comprises 1,377 unique questions, each reused across different object pairs, averaging 5.86 questions per object pair. PACS encompasses a total of 1,526 objects, each represented by a unique video segment v, along with an audio a and a bounding box b of the object in the middlemost frame of v. The average length of each video is 7.6 seconds. PACS dataset covers two tasks: a binary question answering task (i.e. PACS-QA) and an object material classification task (i.e. PACS-Material[1]), the latter containing 18 different material types such as wood, plastic, and metal. The training, validation, and test sets for PACS-QA consist of 11,044, 1,192, and 1,164 samples respectively. For PACS-Material, the training, validation, and test sets comprise 3,460, 444, and 445 samples respectively. Accuracy serves as the evaluation metric.

Baseline Models and Bias Detection

We employ pretrained models for processing image, audio, video, and text, constructing five different fusion models. Specifically, we adopt the Vision Transformer (ViT) (Dosovitskiy et al. 2021) as the image encoder, the Audio Spectrogram Transformer (AST) (Gong, Chung, and Glass 2021) as the audio encoder, the Temporal Difference Network (TDN) (Wang et al. 2021) as the video encoder, and DeBERTa-V3 (He et al. 2021) as the text encoder. To detect visual biases, we consider a simple **late fusion** baseline. This baseline concatenates single-modal embeddings obtained from image, video, and audio encoders to form multimodal embeddings of objects. Then, it concatenates the multimodal embeddings of two objects with question embeddings, utilizing an MLP to generate two question-object embeddings $e_{qo}^{(1)}$ and $e_{qo}^{(2)}$. Finally, it concatenates the two question-object embeddings and employs an MLP to generate multi-class or binary classification outputs. We focus on three types of biases:

Q1: Answer Selection Bias. Is there a systemic bias in answer selection, meaning providing correct answers without seeing the questions?

Q2: Unimodal Question Answerability. Is information from a single modality sufficient to correctly answer questions?

Q3: Visual Bias. Does the imbalance in visual object combinations in the training set affect model predictions?

In particular, we design the following experiments:

- I+A+V: Assessing the predictability of tasks considering only object information (without providing questions) to examine patterns between objects and correct answers.
- Q+I: Evaluating the impact of images on predictions.
- Q+V: Evaluating the impact of videos on predictions.
- Q+A: Evaluating the impact of audios on predictions.
- **Late Fusion (Q+I+V+A)**: Assessing the impact of multimodal composition on model predictions.
- **Late Fusion ‡**: Rebalancing the distribution of visual object combinations in the original training set by recombining object pairs and questions while keeping the test set unchanged, and using this to examine the influence of *visual bias* on model predictions.

Then, we apply `CF-PACR` to two traditional PACR baselines, namely AudioCLIP and MERLOT Reserve, to validate the effectiveness and generalization of the proposed `CF-PACR`. AudioCLIP extends CLIP to the audio modality by introducing an audio head, embedding audio inputs into the same vector space. MERLOT Reserve extends MERLOT by incorporating audio into pretraining, aimed at learning to match contextualized captions with video frames.

Implementation Details All hyperparameters for the comparison algorithms were optimized via grid search on the PACS validation set to maximize prediction performance. For the `CF-PACR` framework, hyperparameters α, β, γ, and τ were tuned within [0, 1] at 0.1 intervals. The CF-PACR framework, being model-agnostic, was integrated with different visual and audio-visual models to create three variants: (1) `Late Fusion+CF-PACR`, combining Q+I+V (visual) and Q+I+A+V (audio-visual); (2) `AudioCLIP+CF-PACR`, using CLIP (visual) and AudioCLIP (audio-visual); and (3) `MER-Res+CF-PACR`, employing MERLOT (visual) and MERLOT Reserve (audio-visual). All variants were trained on four NVIDIA Tesla V100 GPUs with a batch size of 16, 30 epochs, a weight decay of $1e{-}4$, and an initial learning rate of $1e{-}3$. Each video was sampled with 8 frames, including the middle frame annotated with object bounding boxes.

[1]Note that our PACS-Material is different from the PACS-material task in the original PACS (Yu et al. 2022), which is a binary classification problem designed to identify *which object is more likely to be the specified material in the question.*

Baseline Model	Accuracy (%) PACS-QA	PACS-Material
Detecting Bias		
I+V+A	$51.9_{\pm 1.1}$	$58.5_{\pm 1.1}$
Q+I	$51.2_{\pm 0.8}$	$57.4_{\pm 1.2}$
Q+V	$51.5_{\pm 0.9}$	$58.0_{\pm 1.0}$
Q+A	$50.9_{\pm 0.6}$	$56.7_{\pm 1.1}$
Late Fusion	$55.0_{\pm 1.1}$	$60.4_{\pm 1.2}$
Late Fusion‡	$59.3_{\pm 0.9}$	$65.9_{\pm 1.3}$
Late Fusion + CF-PACR	$57.8_{\pm 0.9}$	$63.8_{\pm 1.3}$
CLIP (Radford et al. 2021)	$56.3_{\pm 0.7}$	$63.3_{\pm 0.9}$
UNITER (L) (Chen et al. 2020)	$60.6_{\pm 2.2}$	$67.2_{\pm 1.5}$
MERLOT (Zellers et al. 2021)	$61.4_{\pm 1.6}$	$68.4_{\pm 1.2}$
AudioCLIP (Guzhov et al. 2020)	$60.0_{\pm 0.9}$	$66.2_{\pm 1.0}$
AudioCLIP (Guzhov et al. 2020) + CF-PACR	$63.6_{\pm 0.8}$	$69.4_{\pm 0.9}$
MER-Res (B) (Zellers et al. 2022)	$66.5_{\pm 1.4}$	$72.2_{\pm 1.2}$
MER-Res (B) (Zellers et al. 2022) + CF-PACR	$68.7_{\pm 0.8}$	$74.7_{\pm 0.9}$
MER-Res (L) (Zellers et al. 2022)	$70.1_{\pm 0.7}$	$74.2_{\pm 0.9}$
MER-Res (L) (Zellers et al. 2022) + CF-PACR	$72.8_{\pm 1.1}$	$76.5_{\pm 1.0}$

Table 1: Performance comparison of different baselines.

Model Variants	Accuracy (%) PACS-QA	PACS-Material
Visual Model (Radford et al. 2021)	$56.3_{\pm 0.7}$	$63.3_{\pm 0.9}$
Audiovisual Model (Guzhov et al. 2020)	$60.0_{\pm 0.9}$	$66.2_{\pm 1.0}$
CF-PACR (SUM) w/o CF	$60.7_{\pm 0.8}$	$66.6_{\pm 1.1}$
CF-PACR (HM) w/o CF	$60.9_{\pm 0.7}$	$66.8_{\pm 0.9}$
CF-PACR (LogSUM) w/o CF	$61.4_{\pm 1.1}$	$67.1_{\pm 1.0}$
CF-PACR (SUM)	$61.9_{\pm 0.8}$	$67.5_{\pm 0.9}$
CF-PACR (HM)	$62.8_{\pm 0.9}$	$68.7_{\pm 1.1}$
CF-PACR (LogSUM)	$63.6_{\pm 0.8}$	$69.4_{\pm 0.9}$

Table 2: Ablations on the impact of the fusion function h.

Bias Analysis in PACS

Table 1 presents the model performance for bias testing in PACS. Upon examining the PACS-QA (binary) task, we observe that the model's prediction of whether a question is correct significantly depends on the specific query content. This can be inferred from the almost random guessing accuracy (50%) of the I+A+V model. Conversely, for the PACS-Material (multiclass) task, the performance of the model is scarcely affected by query content. However, the introduction of query content aids the model in learning more relevant and accurate visual features, as evidenced by the performance contrast between the Q+I+A+V and I+A+V models (60.4% vs. 58.5%). Furthermore, providing only image (I), video (V), or audio (A) information alongside a question (Q) does not improve model accuracy. Only when combining all three modalities do the prediction results significantly deviate from random guessing, such as 55.0% for PACS-QA and 60.4% for PACS-Material. Finally, after training on rebalanced training sets, we can witness a noticeable improvement in model performance, with increases of 4.3 percentage points for PACS-QA and 4.5 percentage points for PACS-Material. These results strongly imply the presence of significant visual biases in the original PACS dataset.

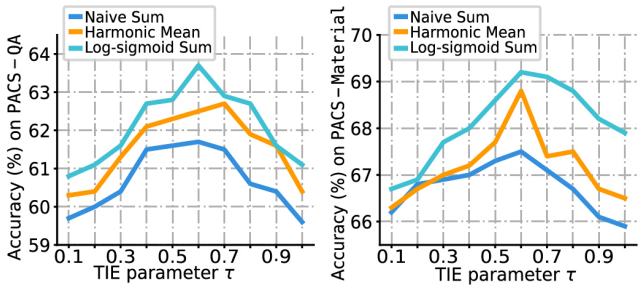

Figure 4: Sensitivity analysis of TIE parameters.

Main Results

Table 1 illustrates the performance comparison between some baseline models and models integrated with the CF-PACR framework. Firstly, we observe that models with audio inputs generally outperform those without audio inputs. For instance, in PACS-QA, the accuracy comparison between the AudioCLIP and CLIP models is 60.0% and 56.3% respectively. Similarly, in PACS-Material, the accuracy comparison between MERLOT Reserve (B) and the MERLOT model is 72.2% and 68.4% respectively. Furthermore, we observe a considerable improvement in the performance of baseline models after integrating our counterfactual debiasing framework. On average, there is an increase of approximately 2.82 percentage points in PACS-QA and approximately 2.85 percentage points in PACS-Material. This indicates that our CF-PACR framework can make minimal modifications during the testing phase to alleviate visual bias while retaining valuable cues from visual information. Overall, the CF-PACR framework can be seamlessly applied to existing multimodal PACS models and significantly enhance their prediction accuracy via counterfactual debiasing.

Ablation Study

Impact of Fusion Functions In CF-PACR, fusion functions play a pivotal role as both factual and counterfactual outcomes are derived from them. As depicted in Fig. 3, before counterfactual debiasing, the fusion function is indispensable for merging two prediction scores. Table 2 lists the results of using three different fusion functions across the two PACS tasks. Here, our visual model is CLIP, and the audiovisual model is AudioCLIP. The results indicate: i) Counterfactual debiasing contributes the most to CF-PACR; ii) Among the three candidates, CF-PACR shows a preference for employing the Log-sigmoid Sum fusion function.

Impact of the TIE Parameter The parameter τ significantly impacts the computation of TIE, governing the extent to which NDE is subtracted from the TE (see Eq. 8). A higher τ implies a lower reliance of the model on visual information. As τ approaches 0, the TIE of the classification result tends toward the total effect, equivalent to the results obtained from traditional posterior probability-based inference strategies. As depicted in Fig. 4, as τ gradually

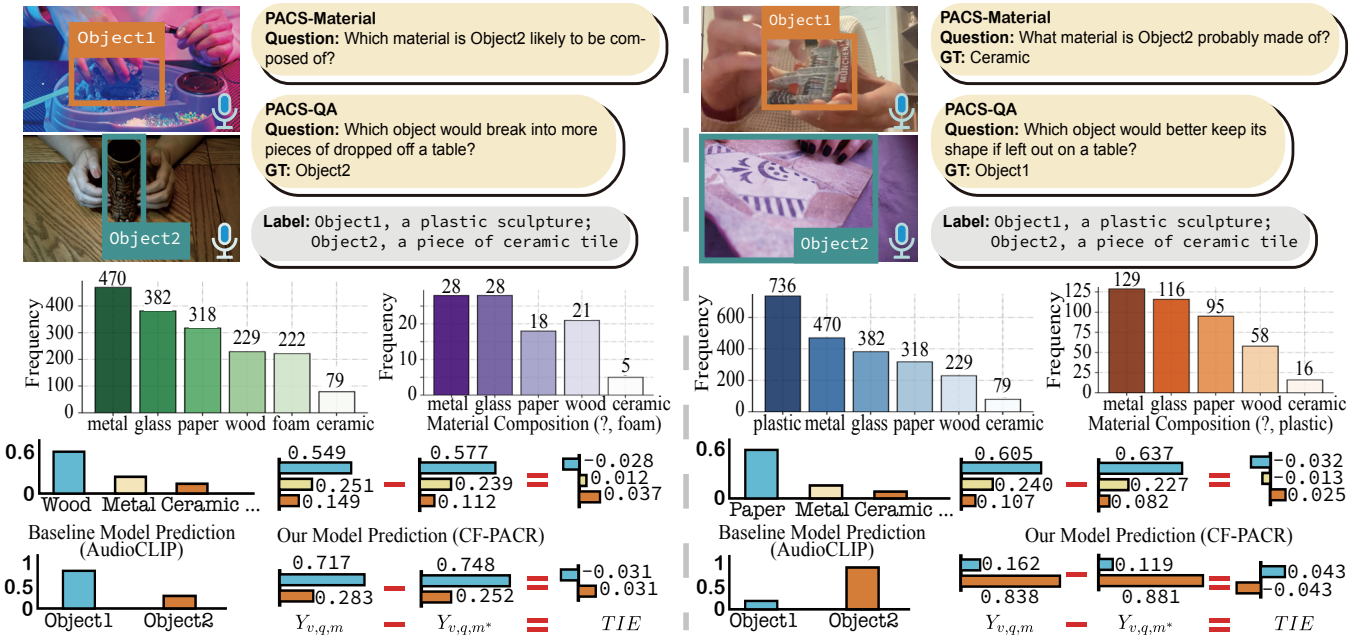

Figure 5: Demonstration of model predictions of four questions (covering both the PACS-QA and PACS-Material tasks) derived from AudioCLIP and CF-PACR (with CLIP as the visual model and AudioCLIP as the audiovisual model) respectively.

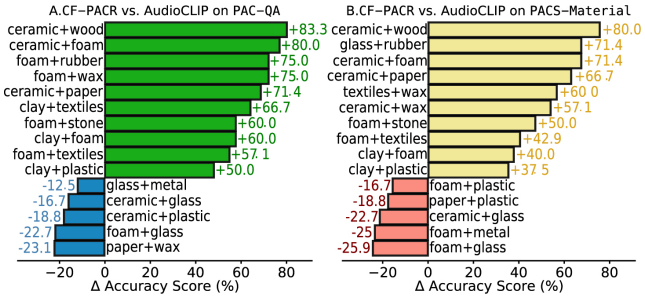

Figure 6: The top-10 visual object combinations with increased accuracy and top-5 visual object combinations with decreased accuracy before and after applying CF-PACR.

increases from 0.1 to 1, the model's performance shows a trend of initially rising and then declining. Specifically, we opted for the Log-sigmoid Sum fusion function and set τ to 0.6. An appropriate τ value enables CF-PACR to mitigate the bad effects of visual bias in predictions while simultaneously leveraging useful cues from visual information.

Category Analysis Fig. 6 illustrates the accuracy improvements of various visual object combinations in both PACS-QA and PACS-Material tasks before and after applying CF-PACR to AudioCLIP. Notably, visual object combinations that have few training samples and those closely related to visual information, such as [ceramic+wood] and [ceramic+foam], benefit significantly from our counterfactual debiasing. Additionally, we observe a slight decline in accuracy for some visually dominant object combinations, which may be attributed to CF-PACR eliminating excessive visual cues when computing TIE.

Case Study

Fig. 5 compares the performance of AudioCLIP and CF-PACR (CLIP for vision, AudioCLIP for audio-visual tasks) on PACS-QA and PACS-Material tasks. In Case 1 (c.f. left pane), AudioCLIP misclassifies Object2 as *wood* and incorrectly predicts Object1 as more fragile. While both visual and audiovisual models of CF-PACR initially favor the *wood* class for Object2, counterfactual debiasing through TIE computation enables correct *ceramic* classification. Fig. 5 (the middlemost column) reveals a training set bias, with [foam+wood] occurring 21 times versus only 5 instances of [foam+ceramic]. This imbalance leads traditional models to learn spurious correlations, causing classification errors. CF-PACR addresses this by employing counterfactual debiasing to eliminate biased priors, enabling accurate predictions without relying on shortcut correlations. Case 2 (right pane) demonstrates similar improvements.

Conclusion

This study identifies the spurious correlation between visual object compositions and model predictions in the PACS tasks, severely impeding the model's physical commonsense reasoning capability. To address this, we propose a novel model-agnostic counterfactual debiasing inference framework to alleviate the model's reliance on visual spurious correlations. Combining factual learning and counterfactual reasoning, we can capture the natural direct effect of pure visual object compositions on classification scores, and eliminate the biased direct effects from the total effect by computing the total indirect effect. We demonstrate the effectiveness of CF-PACR by integrating it with multiple baselines on the PACS tasks, validating its superiority and generalizability.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2023YFC3304800). The computations in this research were performed using the CFFF platform of Fudan University.

References

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

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32.

Chen, Y.; Huang, S.; Yuan, T.; Qi, S.; Zhu, Y.; and Zhu, S.-C. 2019. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *ICCV*, 8648–8657.

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

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. In *ICLR*.

Forbes, M.; Holtzman, A.; and Choi, Y. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Georgescu, M.-I.; Fonseca, E.; Ionescu, R. T.; Lucic, M.; Schmid, C.; and Arnab, A. 2023. Audiovisual masked autoencoders. In *ICCV*, 16144–16154.

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

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2020. Audioclip: Extending clip to image, text and audio. *arXiv preprint arXiv:2008.04838*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

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

Jimenez, C. E. 2020. Learning physical commonsense knowledge.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *ICLR*.

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.

Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; Qiu, Z.; Xu, L.; Xu, Y.; Fang, H.-S.; and Lu, C. 2023. HAKE: A Knowledge Engine Foundation for Human Activity Understanding. *IEEE TPAMI*, 45(7): 8494–8506.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *CVPR*, 2537–2546.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32.

Lv, C.; Zhang, S.; Tian, Y.; Qi, M.; and Ma, H. 2024. Disentangled counterfactual learning for physical audiovisual commonsense reasoning. In *NeurIPS*, 12476–12488.

Ma, J.; Guo, R.; Mishra, S.; Zhang, A.; and Li, J. 2022. Clear: Generative counterfactual explanations on graphs. *NeurIPS*, 35: 25895–25907.

Mottaghi, R.; Rastegari, M.; Gupta, A.; and Farhadi, A. 2016. "What happens if..." learning to predict the effect of forces in images. In *ECCV*, 269–285.

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

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual VQA: A cause-effect look at language bias. In *CVPR*, 12700–12710.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Storks, S.; Gao, Q.; Zhang, Y.; and Chai, J. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of EMNLP*, 4902–4918.

Sun, P.; Wu, B.; Li, X.; Li, W.; Duan, L.; and Gan, C. 2021. Counterfactual debiasing inference for compositional action recognition. In *ACM MM*, 3220–3228.

Sun, T.; Wang, W.; Jing, L.; Cui, Y.; Song, X.; and Nie, L. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *ACM MM*, 15–23.

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021. TDN: Temporal difference networks for efficient action recognition. In *CVPR*, 1895–1904.

Wu, J.; Lu, E.; Kohli, P.; Freeman, B.; and Tenenbaum, J. 2017. Learning to see physics via visual de-animation. In *NeurIPS*, 152–163.

Xiao, F.; Lee, Y. J.; Grauman, K.; Malik, J.; and Feichtenhofer, C. 2020. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*.

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

Ye, K.; and Kovashka, A. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *AAAI*, 4, 3181–3189.

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*.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. 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. In *NeurIPS*, 23634–23651.

Zhao, Z.; Papalexakis, E.; and Ma, X. 2020. Learning physical common sense as knowledge graph completion via BERT data augmentation and constrained tucker factorization. In *EMNLP*.