Disentangled Counterfactual Learning for Physical Audiovisual Commonsense Reasoning

Changsheng Lv[1,2*], **Shuai Zhang**[1,2*], **Yapeng Tian**[3], **Mengshi Qi**[1,2✉†], **and Huadong Ma**[1,2]
[1]Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia
[2]Beijing University of Posts and Telecommunications
[3]Department of Computer Science, The University of Texas at Dallas
{lvchangsheng, zshuai, qms, mhd}@bupt.edu.cn, yapeng.tian@utdallas.edu

Abstract

In this paper, we propose a Disentangled Counterfactual Learning (DCL) approach for physical audiovisual commonsense reasoning. The task aims to infer objects' physics commonsense based on both video and audio input, with the main challenge is how to imitate the reasoning ability of humans. Most of the current methods fail to take full advantage of different characteristics in multi-modal data, and lacking causal reasoning ability in models impedes the progress of implicit physical knowledge inferring. To address these issues, our proposed DCL method decouples videos into static (time-invariant) and dynamic (time-varying) factors in the latent space by the disentangled sequential encoder, which adopts a variational autoencoder (VAE) to maximize the mutual information with a contrastive loss function. Furthermore, we introduce a counterfactual learning module to augment the model's reasoning ability by modeling physical knowledge relationships among different objects under counterfactual intervention. Our proposed method is a plug-and-play module that can be incorporated into any baseline. In experiments, we show that our proposed method improves baseline methods and achieves state-of-the-art performance. Our source code is available at https://github.com/Andy20178/DCL.

1 Introduction

Humans acquire physical commonsense knowledge through integrating multi-modal information and have the ability to reason the physical properties of previously unseen objects in the real world [1].

Whether inferring the physical properties of unseen objects ("this object is probably made of wood") or solving practical problems ("which object would make a bigger mess if dropped") [2], such reasoning abilities represent a sophisticated challenge in the realm of machine intelligence, and play the important role in robust robot navigation [3] as well as AR/VR [4]. In this work, we utilize audiovisual question answering as a proxy task to advance the machine's capacity for physical commonsense reasoning.

The major challenge in physical commonsense reasoning is how to develop methods that can extract and reason the implicit physical knowledge from abundant multi-modal data, especially video inputs. This requires the ability to process complex video information, identify the category and the corresponding physical property of each object, and understand the causal relationships among objects. These cognitive processes are similar to those that humans use to know, learn, and reason about the physical world. Current existing methods [5, 6, 7, 8] often extract generic visual features from videos that contain human-object interactions, resulting in a mixed feature representation that

[*] Equal contribution
[†] Corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

https://doi.org/10.52202/075280-0546

does not distinguish between object and action information. However, in physical commonsense reasoning, it is critical to clearly figure out the attributes and physical properties of objects.

To mitigate the challenge, our idea is to meticulously disentangle the whole video content into static parts and dynamic ones, which means they remain constant features over time, and the features vary with time, respectively. Another driving force behind our work is to establish relationships of physical knowledge among various objects within both video and audio modalities. We enhance the optimization of the results by taking into account the relevance of multiple samples and incorporating causal learning, by applying such relationships as confounders. The application of counterfactual interventions further bolsters the model's explainability capacity.

In this paper, we propose a new ***D****isentangled* ***C****ounterfactual* ***L****earning* (**DCL**) framework for physical audiovisual commonsense reasoning, which explicitly extracts static and dynamic features from video and uses causal learning to mine implicit physical knowledge relationships among various objects. Specifically, we design a Disentangled Sequential Encoder (DSE), which is based on a sequential variational autoencoder to clearly separate the static and dynamic factors of the input videos through self-supervision. Meanwhile, we incorporate the contrastive estimation method to enhance the mutual information (MI) between the input data and two latent factors, while also discouraging the MI between the static and dynamic factors. Furthermore, we propose a novel Counterfactual Learning Module (CLM) to capture physical knowledge relationships from a wide array of data samples and employ counterfactual intervention with causal effect tools during the reasoning process. We further refine the training objectives of the model by maximizing the probability likelihood in DSE and the Total Indirect Effect value in CLM. Extensive experimental results on PACS dataset [2] validate that our proposed DCL approach achieves superior results over previous state-of-the-art methods, demonstrating our model can effectively improve the reasoning ability and enhance the explainability.

Our main contributions can be summarized as follows:

(1) We propose a new Disentangled Counterfactual Learning (DCL) approach for physical audio-visual commonsense reasoning, which disentangles video inputs into static and dynamic factors via a Sequential Variational Autoencoder. To the best of our knowledge, we are the first to adopt disentanglement learning into this specific task.

(2) We introduce a novel Counterfactual Learning Module to model physical knowledge relationships among various objects, and then apply them as counterfactual interventions to enhance the ability of causal reasoning.

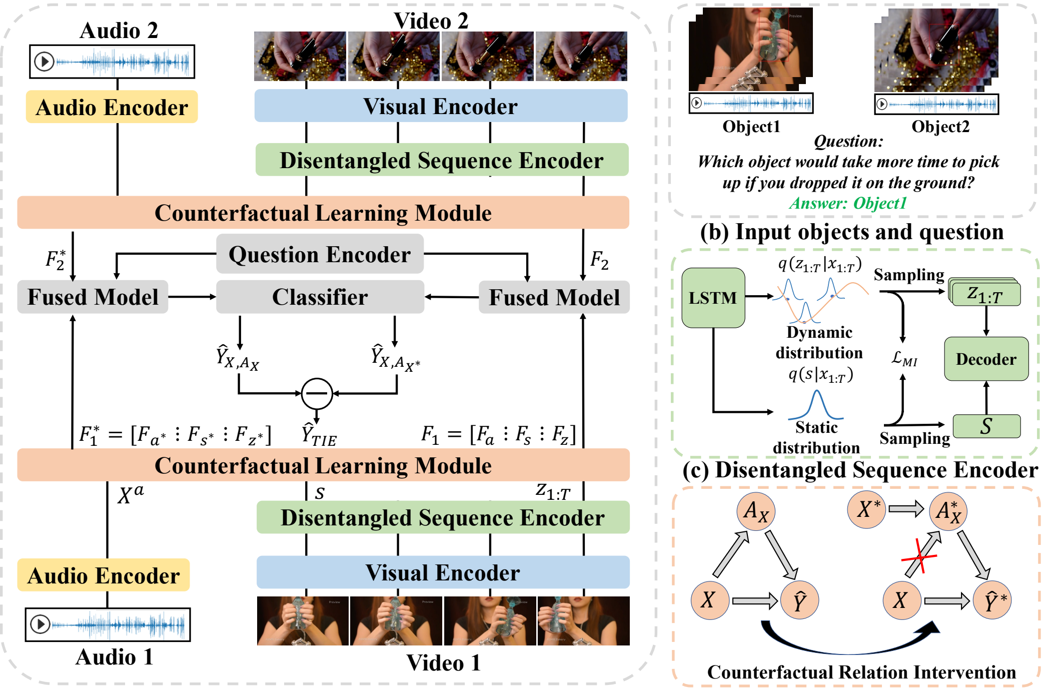
(3) Experimental results demonstrate our proposed DCL is a plug-and-play module that can be incorporated into any baseline, and show DCL can achieve 3.2% absolute improvement over the state-of-the-art methods.

2 Related Work

Physical Commonsense Reasoning. Naive and intuitive physics was first studied in human psychology experiments [9, 10, 11]. Further research has also been conducted in the multi-sensory perception of physical properties [12]. Nowadays, visual and language are adopted to learn and utilize physical commonsense in different tasks [13, 14, 15, 16, 17], which only focus on single-modality knowledge and disregard multi-modal fusion. In embodied audio-visual AI tasks, existing research utilizes multi-modal information to address inverse physics problems [5, 6, 7, 8]. Different from them, we introduce a disentangled counterfactual learning approach to reason the physical commonsense from both video and audio inputs.

Disentangled Representation Learning. Disentangled representation learning aims to learn various hidden explanatory factors behind observable data [18], which has already been widely applied in computer vision and natural language processing [19, 20, 21, 22, 23], such as image-to-image translation [19], image generation [20] and text generation [21, 21]. However, such methods are limited in the Visual Question Answering (VQA) task. In this paper, we propose a VAE-based disentangled learning method that disentangles visual information into static and dynamic aspects and improves the interpretability of the model.

Causal learning. Traditional VQA datasets suffer from "language prior" [24] or "visual priming bias" [25], leading current methods focus on features' inherent bias of either language or vision and



(a) Overview of our method

(b) Input objects and question

(c) Disentangled Sequence Encoder

(d) Counterfactual Learning Module

Figure 1: The illustration of our proposed DCL model: (a) shows the overall architecture: the input videos with audios (b) are first encoded by the corresponding visual and audio encoders, then the Disentangled Sequence Encoder (c) is used to decouple video features into static and dynamic factors by LSTM-based VAE. Through the Counterfactual Learning Module (d), we construct the affinity matrix A as a confounder and obtain the prediction $\hat{Y}_{X,A_X}$ and counterfactual result $\hat{Y}_{X,A_X^*}$. Finally, we achieve $\hat{Y}_{TIE}$ by subtracting these two results and optimizing the model by cross-entropy loss.

then output inaccurate answers. In recent years, counterfactual thinking and causal reasoning have been widely applied in visual explanations [26, 27, 28, 29], scene graph generation [30, 31, 32, 33, 34], video understanding [35, 36, 37, 38, 39] and image recognition [31]. Recently, there has been some work on audio-visual video question answering [40, 41, 42], but they generally focus on cross-modal modeling. Our approach, in contrast, focuses on constructing physical knowledge relationships among different samples and using them as confounders in causal reasoning.

3 Proposed Approach

3.1 Overview

Figure 1(a) illustrates the overall framework of our proposed DCL. First, the model extracts features of input videos and the corresponding audio by visual and audio encoders, respectively. Then, a newly designed Disentangled Sequence Encoder (Sec. 3.2) is used to decouple the video into static and dynamic factors. Subsequently, the raw and intervened multi-modal features are obtained through the Counterfactual Learning Module (Sec. 3.3), and then fused with the question feature. Finally, the prediction result is obtained by optimizing the classification results based on the raw and intervened features (Sec. 3.4).

Problem Formulation. As shown in Figure 1(b), given a pair of videos (*i.e.,* $< v_1, v_2 >$) with the corresponding audios (*i.e.,* $< a_1, a_2 >$), where each video contains an object, and our task is to select a more appropriate object from the video inputs to answer the questions (*i.e.,* q) about the physical commonsense. We define the estimated answer as $\hat{Y}$ and the ground-truth answer as Y in the dataset. During the pre-processing, we denoted the extracted features of audio, video, and question text as X^a, X^v, and X^t, respectively, where $X^a, X^v, X^t \in \mathbb{R}^d$, d denotes the feature dimension. Especially,

https://doi.org/10.52202/075280-0546

we capture the audio feature and text feature as non-sequence while extracting the video feature as sequence data as $X^v = \{X_1^v, X_2^v, \cdots, X_T^v\}$, where T refers to the number of video frames.

3.2 Disentangled Sequential Encoder

The proposed Disentangled Sequential Encoder, as shown in Figure 1(c), can separate the static and dynamic factors for the multi-modal data (we take the video data as the example in the following), which extends the traditional sequential variational autoencoder by incorporating the mutual information term.

Specifically, we assume the latent representations of the input video's feature $X_{1:T}^v$ can divide into the static factor s and the dynamic factors $z_{1:T}$, where z_t is the latent dynamic representation at time step t. Following [43, 44], we assume such two factors are independent of each other, i.e., $p(s, z_{1:T}) = p(s)p(z_{1:T})$, where $p(\cdot)$ represents the probability distribution. And z_i depends on $z_{<i} = \{z_0, z_1, .., z_{i-1}\}$, where $z_0 = 0$ and the reconstruction of ${x_i}$[3] is independent of other frames conditioned on z_i and s. Hence we need to learn a posterior distribution $q(z_{1:T}, s|x_{1:T})$ where the two factors are disentangled formulated as:

$$q(z_{1:T}, s|x_{1:T}) = q(z_{1:T}|x_{1:T})q(s|x_{1:T}) = q(s|x_{1:T}) \prod_{i=1}^{T} q(z_i|z_{<i}, x_{\leq i}). \tag{1}$$

We model the posterior distribution by Bi-LSTM [45], where $q(z_i|z_{<i}, x_{\leq i})$ is conditioned on the entire time series by taking the hidden states as the input, and $q(s|x_{1:T})$ can be computed by input $x_{1:T}$. Then we sample the two decoupled factors s and $z_{1:T}$ by utilizing the distribution $q(s|x_{1:T})$ and $q(z_i|z_{<i}, x_{\leq i})$ via the reparameterization trick [46]. Afterward, we use the obtained two disentangled factors to reconstruct the $x_{1:T}$ based on a VAE-based decoder [44]. The prior of static factor s and dynamic factor z_i can be defined as the Gaussian distribution with $\mathcal{N}(0, I)$ and $\mathcal{N}(\mu(z_{<i}), \sigma^2(z_{<i}))$, respectively, where $\mu(\cdot)$ and $\sigma(\cdot)$ are modeled by Bi-LSTM. The reconstruction process can be formalized by the factorization:

$$p(x_{1:T}, s, z_{1:T}) = p(s) \prod_{i=1}^{T} p(z_i|z_{<i})p(x_i|z_i, s). \tag{2}$$

Furthermore, we introduce the mutual information to encourage the disentangled two factors mutually exclusive and incorporate non-parameter contrastive estimation into the standard loss function for learning latent representations [43, 47], which can be formulated as:

$$\mathcal{C}(z_{1:T}) = \mathbb{E}_{p_D} log \frac{\phi(z_{1:T}, x_{1:T}^+)}{\phi(z_{1:T}, x_{1:T}^+) + \sum_{j=1}^{n} \phi(z_{1:T}, x_{1:T}^j)}, \tag{3}$$

where x^+ means the video has the same object as "positive example", while $x^j, j = \{1, 2, ..., n\}$ is a set of n "negative" videos that have different objects. In order to alleviate the high dimensionality curse [48, 49], we use $\phi(z_{1:T}, x_{1:T}^+) = exp(sim(z_{1:T}, x_{1:T}^+)/\tau)$, where $sim(\cdot, \cdot)$ means cosine similarity function and $\tau = 0.5$ is a temperature parameter. $\mathcal{C}(s)$ can be calculated in the similar way.

Following [44], we use content augmentation by randomly shuffling the time steps of the video, and performing motion augmentation via the Gaussian blur [49]. The results can be denoted as $\mathcal{C}(z_{1:T}^m)$ and $\mathcal{C}(s^m)$, where $z_{1:T}^m$ and s^m are the augmented data of $z_{1:T}$ and s, respectively. The MI term can be formulated as follows:

$$MI(z_{1:T}; x_{1:T}) \approx \frac{1}{2}(\mathcal{C}(z_{1:T}) + \mathcal{C}(z_{1:T}^m)), \qquad MI(s; x_{1:T}) \approx \frac{1}{2}(\mathcal{C}(s) + \mathcal{C}(s^m)). \tag{4}$$

The objective function can be formulated by adding MI terms to the vanilla ELBO as follows:

$$\begin{aligned} \mathcal{L}_{DSE} = \mathbb{E}_{q(z_{1:T}, s|x_{1:T})}[-\sum_{t=1}^{T} \log p(x_t|s, z_t)] + \gamma \cdot [KL(q(s|x_{1:T})||p(s)) \\ + \sum_{t=1}^{T} KL(q(z_t|x_{\leq t})||p(z_t|z_{<t}))] - \alpha \cdot MI(z_{1:T}; x_{1:T}) - \beta \cdot MI(s; x_{1:T}) + \theta \cdot MI(z_{1:T}; s), \end{aligned} \tag{5}$$

[3] For the sake of simplicity, in the following content, we will use x_i to represent X_i^v.

where γ, α, β and θ are hyper-parameters. The full proof can be found in [44].

3.3 Counterfactual Learning Module

In this module, we obtain the decoupled static and dynamic factors s and $z_{1:T}$ after performing the Disentangled Sequential Encoder. Then, these factors with the corresponding audio features are sent to construct physical knowledge relationships. Meanwhile, we utilize the counterfactual relation intervention to improve knowledge learning.

Physical Knowledge Relationship. Inspired by Knowledge Graph [50, 51], the physical knowledge contained in different samples may have a certain correlation. Therefore, we propose to model such implicit relationships as a graph structure and draw the affinity matrix A to represent this physical knowledge relationship among various objects. Similarly, we build an affinity matrix for audio features or other modal data. With the well-constructed affinity matrix A, we can enhance the static, dynamic, and audio features denoted as X^v_s, X^v_z, and X^a through passing the message and transferring across various samples, formulated as:

$$X = [\ X^a \;\vdots\; X^v_s \;\vdots\; X^v_z\ ], \qquad A_X = [\ A_{X^a} \;\vdots\; A_{X^v_s} \;\vdots\; A_{X^v_z}\ ], \qquad F = A_X X^\top, \tag{6}$$

where A_X represents the augmented matrix composed of three affinity matrices, while F refers to the transferred features, and $\top$ denotes the transpose of a matrix. To achieve A, we take $A_{X^v_s}$ as an example: we first compute the similarity matrix based on the static features and filter out the noisy relationships by defining a near neighbor chosen function as $\mathcal{T}(\cdot, k)$ to retain the top-k values in each row, and then adopt the Laplacian matrix D^{-1} of $\mathcal{S}'$ to normalize the total affinities as the following:

$$\mathcal{S}^{i,j} = \exp(\frac{sim(x_i, x_j)}{\tau}), x_i, x_j \in X_s \qquad \mathcal{S}' = \mathcal{T}(\mathcal{S}, k), \qquad A_{X_s} = D^{-1} \cdot \mathcal{S}', \tag{7}$$

where $sim(\cdot, \cdot)$ is the cosine similarity. The calculation of A^v_z and A^a is similar to A^v_s.

Counterfactual Relation Intervention. To add additional supervision of the affinities A_X, we present to highlight the role of the object's physical knowledge relationship during the optimization. First, we express our method as a Structural Causal Model (SCM) [52, 53] as shown in Figure 1(d), and then introduce causal inference into our method. $\hat{Y}$ represents the final classification output of the model, which is obtained by feeding forward F input into the fusion model and classifier:

$$\hat{Y}_{X,A_X} = CLS(Fusion(F_1, F_2, X^t)), \tag{8}$$

where F_1 and F_2 denote the fused feature of input video pair v_1 and v_2, and X^t denotes the feature of the question text. 'CLS' and '$Fusion$' represent the classifier and fusion model, respectively. Obviously, the process of deriving the output $\hat{Y}$ from input X can be regarded as two types of effects: one is the direct effect $X \rightarrow \hat{Y}$, and the other is an indirect type $X \rightarrow A_X \rightarrow \hat{Y}$. Our final loss function is to maximum likelihood estimation, which will have an end-to-end effect on both types of effects, resulting in insufficient enhancement of A in the indirect effects path. Thus we leverage the Total Indirect Effect (TIE) [54] here to underline the A_X:

$$\hat{Y}_{TIE} = \hat{Y}_{X,A_X} - \mathbb{E}_{X^*}[\hat{Y}_{X,A_{X^*}}], \tag{9}$$

where $\hat{Y}_{X,A_{X^*}}$ refers to calculating the results by replacing the original affinity A_X to a intervened one A_{X^*}, and X^* is the given intervened inputs. Note that $\hat{Y}_{X,A_{X^*}}$ cannot happen in the real world because features X and affinities A_{X^*} come from different X and X^*, which is called counterfactual intervention. Therefore, we modify $\hat{Y}_{X,A_X}$ to $\hat{Y}_{X,A_{X^*}}$ is equal to keeping all features fixed but only changing the affinity A, which can highlight the critical effect of A. We compute the expectation of that effect to get the more stable one, and the intervened input features X^* are sampled by a Gaussian distribution:

$$X^* = X_\sigma \cdot W + X_\mu, \tag{10}$$

where W is the standard random vector whose dimension is as same as X, and the mean X_μ and standard deviation X_σ are both learned by the re-parameterization trick in the end-to-end learning.

3.4 Fusion Model and Optimization

As we have stated, our approach is a plug-and-play module that can be directly integrated into various multimodal fusion methods. We take Late fusion [55] as an example to illustrate the structure of fusion models, as well as the optimization function of our DCL. The fusion model can be expressed in the following formula:

$$Fusion(F_1, F_2, X^t) = MLP(Concat(MLP(Concat(F_1, F_2)), X^t)), \tag{11}$$

where $Concat$ represents row-wise concatenation, and MLP represents multi-layer perception.

Finally, our final optimization goal can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{DSE} + \mathcal{L}_{TIE}, \tag{12}$$

where $\mathcal{L}_{DSE}$ denote the optimization function we described in Eq. (5) and $\mathcal{L}_{TIE}$ denotes the cross-entropy loss of $\hat{Y}_{X,A_X}$ in Eq. (9) given the ground-truth Y.

4 Experiments

4.1 Experimental Setup

Dataset. The Physical Audiovisual CommonSense Reasoning Dataset (PACS) [2] is a collection of 13,400 question-answer pairs designed for testing physical commonsense reasoning capabilities. The dataset contains 1,377 unique physical commonsense questions spanning various physical properties and is accompanied by 1,526 videos and audio clips downloaded from YouTube. The total number of data points in the PACS dataset is 13,400, and for PACS-Material it is 4,349. Following [2], We divide PACS into 11,044/1,192/1,164 as train/val/test set, each of which contains 1,224/150/152 objects respectively. We partitioned the PACS-Material subset into 3,460/444/445 for train/val/test under the same object distribution as PACS. To ensure a fair evaluation of models, we evaluate our method on both the entire dataset and a subset focused on material-related problems, reporting the results separately for each subset during testing.

Evaluation Metric. We follow the prior work [2] and use the accuracy to evaluate both PACS and PACS-material subsets. To evaluate the effectiveness of our static factor via disentanglement learning, we measure the accuracy of the PACS-material subset using only the static factor. In our experiments, we report results by running the experiments five times and taking the average following [2].

Implementation. We implement our proposed model with PyTorch on two NVIDIA RTX 3090 GPUs. Specifically, we downsampled each video to $T = 8$ frames during pre-processing and set the feature dimension as $d = 256$. In the Disentangled Sequence Encoder, we used a hidden layer size of 256 for Bi-LSTM. During optimization, we set the batch size as 64, which consisted of 64 video pairs and the corresponding questions. In the Counterfactual Learning Module, $\tau = 2$ and $k = 5$ were used when calculating similarities and constructing the physical knowledge relationships. More details about hyper-parameters analysis can be found in our supplementary.

Compared Baseline Methods. To demonstrate the effectiveness of our method across multiple models, we apply it to the following baselines: 1) Late fusion [55], which adopts independent text, image, audio, and video encoders to obtain features, and the unimodal embeddings are concatenated and passed through a linear layer to create multimodal embeddings for prediction. 2) CLIP/AudioCLIP [56, 57], which embed video, text, and audio data into a shared vector space using CLIP and AudioCLIP, and a linear layer is used to create multimodal embeddings for prediction. Notes that CLIP is unable to extract audio features, so we did not use audio in our experiments with CLIP. 3) UNITER [58], which is a pre-trained image and text model for four different image-text tasks and performs well on tasks such as NLVR2 [59]. As for all of the above-mentioned baseline methods, we adopt the reported parameters following their corresponding paper. Note that another state-of-the-art Merlot Reserve [60] is a large-scale multimodal pretraining model that learns joint representations of audio, visual, and language by guiding the new pretraining objectives trained on TPU-v3-512. However, due to the restriction of experimental resources, we are unable to test the performance of DCL with this model and leave it as future work.

https://doi.org/10.52202/075280-0546

Baseline Model	Accuracy (%)	
	PACS	PACS-Material
Late Fusion [55]	55.0 ± 1.1	67.4 ± 1.5
Late Fusion w/ DCL	57.7 ± 0.9	69.7 ± 1.2
CLIP [56]	56.3 ± 0.7	72.4 ± 1.1
CLIP w/ DCL	58.4 ± 0.8	75.4 ± 1.2
AudioCLIP [57]	60.0 ± 0.9	75.9 ± 1.1
AudioCLIP w/ DCL	**63.2** ± **0.8**	**76.2** ± **1.4**
UNITER(Large) [58]	60.6 ± 2.2	75.0 ± 2.8
UNITER w/ DCL*	62.0 ± 2.4	75.7 ± 2.8

(a) Comparison of our DCL with different baselines.

Baseline Model	Accuracy (%)	
	PACS	PACS-Material
Late Fusion [55]	55.0 ± 1.1	67.4 ± 1.5
Late Fusion w/ MLP	54.9 ± 0.9	67.7 ± 1.1
Late Fusion w/ DSE	56.2 ± 0.8	68.5 ± 1.2
CLIP[56]	56.3 ± 0.7	72.4 ± 1.1
CLIP w/ MLP	56.5 ± 0.5	72.6 ± 1.2
CLIP w/ DSE	57.0 ± 0.6	73.2 ± 1.1
AudioCLIP[57]	60.0 ± 0.9	75.9 ± 1.1
AudioCLIP w/ MLP	60.3 ± 0.8	76.2 ± 1.3
AudioCLIP w/ DSE	**61.1** ± **0.8**	**76.5** ± **1.0**

(b) Ablation study of DSE.

Baseline Model	Accuracy (%)	
	PACS	PACS-Material
Late Fusion [55] w/ DSE	56.2 ± 0.8	68.5 ± 1.2
Late Fusion w/ DSE,A	57.0 ± 1.1	68.9 ± 1.3
Late Fusion w/ DSE,A,C	57.7 ± 0.9	69.7 ± 1.2
CLIP [56] w/ DSE	57.0 ± 0.6	73.2 ± 1.1
CLIP w/ DSE,A	57.8 ± 0.8	74.5 ± 1.1
CLIP w/ DSE,A,C	58.4 ± 0.8	75.4 ± 1.2
AudioCLIP [57] w/ DSE	61.1 ± 0.8	76.0 ± 1.0
AudioCLIP w/ DSE,A	61.9 ± 0.9	75.8 ± 1.1
AudioCLIP w/ DSE,A,C	**63.2** ± **0.8**	**76.2** ± **1.4**

(c) Ablation study of CLM

Baseline Model	Accuracy (%)	
	PACS-Material	Δ
Late Fusion [55] w/ dynamic	60.4 ± 1.3	-
Late Fusion w/ static	66.0 ± 0.6	5.6
Late Fusion w/ static, dynamic	69.7 ± 1.2	3.7
CLIP [56] w/ dynamic	63.0 ± 0.7	-
CLIP w/ static	71.1 ± 0.9	8.1
CLIP w/ static, dynamic	75.4 ± 1.2	4.3
AudioCLIP [56] w/ dynamic	67.1 ± 0.6	-
AudioCLIP w/ static	72.5 ± 1.0	5.4
AudioCLIP w/ static, dynamic	**76.2** ± **1.4**	3.7

(d) Ablation analysis of the static factors.

Table 1: Experimental results and analysis in terms of the accuracy achieved on PACS and PACS-Material. 'DSE' refers to Disentangled Sequential Encoder, while 'CLM' refers to Counterfactual Learning Module. 'A' represents physical knowledge relationship, and 'C' stands for Counterfactual Relation Intervention. Δ represents the performance improvement compared with the previous row.

4.2 Comparison to Baselines

Quantitative Results. We report the quantitative performance comparisons using the PACS dataset in Table 1(a). The results show that with the addition of our proposed DCL, both the Late Fusion and UNITER models register absolute improvements of 2.7% and 1.4%, respectively. Similarly, CLIP and AudioCLIP models, which align the image, audio, and text into a shared space, achieve absolute improvements of 2.1% and 3.2%, respectively. These results highlight the strong reasoning and generalization capabilities of our DCL method.

Interestingly, even with the DCL, CLIP's performance trails that of AudioCLIP, emphasizing the critical role of audio information. When comparing CLIP and AudioCLIP enhanced with our DCL, it is evident that the inclusion of audio information leads to an absolute improvement of 4.8%. The same conclusion applies to the PACS-Material subset, indicating that our method, serving as a plug-and-play module, can consistently enhance material reasoning performance and be flexibly incorporated across various models. It is noteworthy that all objects in the test set were absent from the training and validation sets, attesting to the zero-shot reasoning ability of our model.

Qualitative Results. In order to provide a comprehensive qualitative perspective on our results, we present the visualized results of different models for the same questions in Figure 2. In Figure 2(a), the two objects exhibit significant visual and auditory differences, facilitating accurate inference by both the baseline model and our DCL. Conversely, the two objects in Figure 2(b) share similar auditory and static information but differ significantly in dynamic information. Our method accurately infers the answers to both questions. This is especially notable in the case of the second question, which requires dynamic information for reasoning, thus highlighting the efficacy of our DCL in the decoupling of static and dynamic elements.

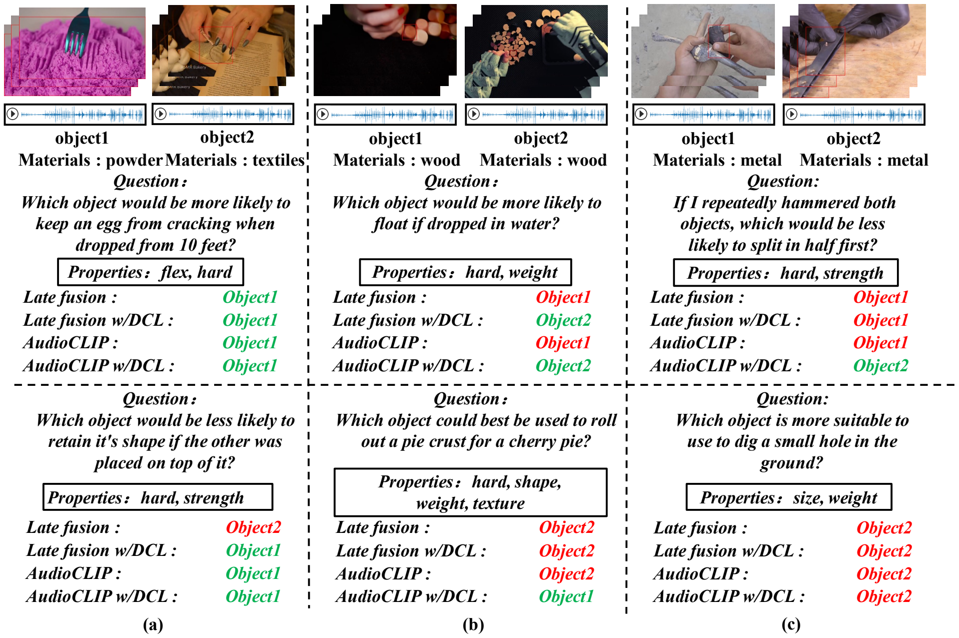

object1
object2
Materials : powder
Materials : textiles
Question :
Which object would be more likely to keep an egg from cracking when dropped from 10 feet?
Properties : flex, hard
Late fusion : Object1
Late fusion w/DCL : Object1
AudioCLIP : Object1
AudioCLIP w/DCL : Object1
Question :
Which object would be less likely to retain it's shape if the other was placed on top of it?
Properties : hard, strength
Late fusion : Object2
Late fusion w/DCL : Object1
AudioCLIP : Object1
AudioCLIP w/DCL : Object1
(a)
object1
object2
Materials : wood
Materials : wood
Question :
Which object would be more likely to float if dropped in water?
Properties : hard, weight
Late fusion : Object1
Late fusion w/DCL : Object2
AudioCLIP : Object1
AudioCLIP w/DCL : Object2
Question :
Which object could best be used to roll out a pie crust for a cherry pie?
Properties : hard, shape, weight, texture
Late fusion : Object2
Late fusion w/DCL : Object2
AudioCLIP : Object2
AudioCLIP w/DCL : Object1
(b)
object1
object2
Materials : metal
Materials : metal
Question:
If I repeatedly hammered both objects, which would be less likely to split in half first?
Properties : hard, strength
Late fusion : Object1
Late fusion w/DCL : Object1
AudioCLIP : Object1
AudioCLIP w/DCL : Object2
Question:
Which object is more suitable to use to dig a small hole in the ground?
Properties : size, weight
Late fusion : Object2
Late fusion w/DCL : Object2
AudioCLIP : Object2
AudioCLIP w/DCL : Object2
(c)

Figure 2: Qualitative Results of baseline w/ and w/o our DCL, where "Material" refers to the material of the object, and "Properties" refers to the physical properties that the question focuses on. The correct answers are depicted in green while the incorrect ones are depicted in red.

In Figure 2(c), only AudioCLIP enhanced with our DCL accurately answers the first question, while the other models fail. This discrepancy arises from inaccurate labeling within the dataset, where two objects appear in the same video, and the bounding box appears to have been assigned to the wrong object. By integrating a more robust object detection model and fine-grained annotations, our proposed DCL can circumvent such issues.

4.3 Ablation Study

In this section, we perform the ablation studies to examine the effectiveness of each of our proposed modules, including the Disentangled Sequential Encoder, physical knowledge relationship matrix, and counterfactual relation intervention in the Counterfactual Learning Module.

Disentangled Sequential Encoder (DSE). Table 1(b) shows the performance comparison of our proposed DSE with different baselines. And we can see AudioCLIP can achieve the best results. After incorporating DSE, all baselines are able to achieve significant improvements in PACS and PACS-Material. However, adding an MLP module with the same parameters as DSE did not yield such improvement, indicating that the proposed DSE can effectively enhance the ability of physical characteristic expression in video features.

Counterfactual Learning Module (CLM). Table 1(c) presents the ablative analysis of two designed parts in the Counterfactual Learning Module (CLM). After adding the physical knowledge relationship denoted as affinity matrix A, all baselines showed an improvement in performance. Similarly, the same result can be clearly found when adding the Counterfactual Relation Intervention denoted as 'C'. Additionally, the same conclusion can be drawn from the PACS-Material, demonstrating that both parts of our Counterfactual Learning Module can independently improve results and achieve optimal performance when used simultaneously.

4.4 The Analysis of Disentanglement Learning

In this section, we carry out additional experiments and analysis focusing on the decoupled static and dynamic factors obtained through the proposed Disentangled Sequential Encoder. Specifically, we utilize separated individual factors to predict the object's material information, excluding multi-modal

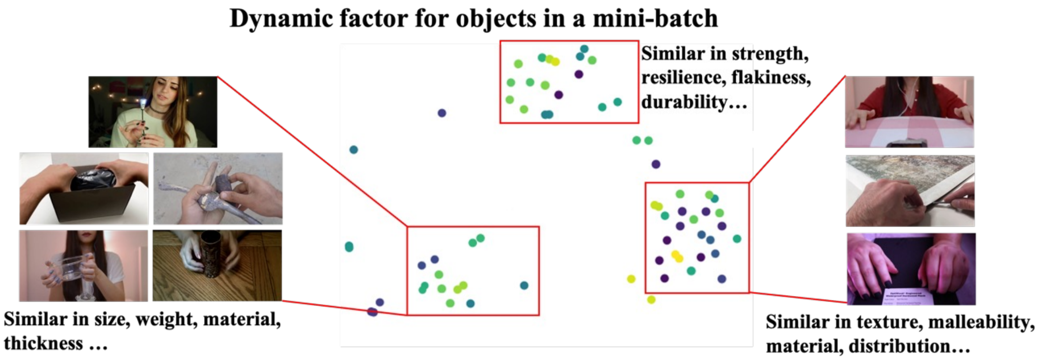


Figure 3: The t-SNE visualization of the obtained dynamic factors in the latent space of video samples in the test set. The points located close to each other as a cluster exhibit similar or identical actions, such as rubbing with the hand (left) and friction with the hand (right), while such two clusters are pulled away showing they have completely different dynamic information.

feature fusion. For the captured dynamic factors, we present the t-SNE visualization of various video samples to illustrate the distribution of these dynamic factors within the latent space.

Analysis of static factor. The material classification results are shown in Table 1(d), where “w/ static” and “w/ dynamic” respectively signify reasoning using only decoupled static factors and dynamic factors, while “w/ static, dynamic” denotes using both static and dynamic factors simultaneously. It can be observed that combining the static and dynamic factors significantly improves the accuracy of reasoning for material problems, compared to using only dynamic factors. This indicates the crucial importance of static factors in material property inference. When predicting using only static factors, there is no significant reduction in accuracy, confirming the adequacy of static factors in providing sufficient material-related information that is not captured by dynamic factors. This also implies that the static and dynamic factors contain little overlapping information and validate the effectiveness of our comparative estimation of mutual information (MI).

Analysis of dynamic factor. For dynamic factors, we decoupled the test set videos and showed the t-SNE visualization of the results, as shown in Figure 3. It can be observed that there is a certain clustering phenomenon in the latent representation of dynamic factors. By displaying the corresponding videos of the data points, we found that the points on the left panel of the figure mostly represent actions involving rubbing with objects by hand. In this case, the visual features are often insufficient to fully represent the characteristics of the object due to its thickness or weight, and the audio information is often the main basis for human reasoning. On the other hand, the points on the right side of the figure correspond to actions involving the friction of objects. Such actions can reveal more dynamic properties of the object, such as malleability. In this case, the dynamic information in the video is often the main basis for human reasoning. Through the above analysis, we can see that for different types of dynamic information, humans will also focus on different modalities of information. This is helpful for future improvements of our reasoning models.

5 Conclusion

In this paper, we introduced a method named Disentangled Counterfactual Learning (DCL) for physical audiovisual commonsense reasoning, in which a Disentangled Sequential Encoder was designed to decouple the video into static and dynamic factors represented as time-invariant and time-varied information related to actions, respectively. Additionally, we model the physical knowledge relationship among different target objects as an affinity matrix and apply counterfactual relation intervention to guide the model to emphasize the physical commonalities among different objects. Designed as a plug-and-play component, our method can be readily incorporated and has demonstrated its potential to significantly enhance multiple baselines.

https://doi.org/10.52202/075280-0546

Acknowledgement

This work was partly supported by the Funds for Creative Research Groups of China under Grant 61921003, the National Natural Science Foundation of China under Grant 62202063, Beijing Natural Science Foundation (No.L223002), the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2021QNRC001, the 111 Project under Grant B18008, the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2022C01).

References

[1] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[2] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Pacs: A dataset for physical audiovisual commonsense reasoning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 292–309. Springer, 2022.

[3] Christopher J Conti, Aparna S Varde, and Weitian Wang. Robot action planning by commonsense knowledge in human-robot collaborative tasks. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–7. IEEE, 2020.

[4] Jean-Luc Lugrin and Marc Cavazza. Making sense of virtual environments: action representation, grounding and common sense. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 225–234, 2007.

[5] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192, 2021.

[6] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.

[7] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020.

[8] Chuang Gan, Yi Gu, Siyuan Zhou, Jeremy Schwartz, Seth Alter, James Traer, Dan Gutfreund, Joshua B Tenenbaum, Josh H McDermott, and Antonio Torralba. Finding fallen objects via asynchronous audio-visual integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10523–10533, 2022.

[9] Daniel G Bobrow. Qualitative reasoning about physical systems: an introduction. *Artificial intelligence*, 24(1-3):1–5, 1984.

[10] Kenneth D Forbus. Qualitative process theory. *Artificial intelligence*, 24(1-3):85–168, 1984.

[11] Susan J Hespos, Alissa L Ferry, Erin M Anderson, Emily N Hollenbeck, and Lance J Rips. Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological Science*, 27(2):244–256, 2016.

[12] Joan Bliss. Commonsense reasoning about the physical world. *Studies in Science Education*, 44(2):123–155, 2008.

[13] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[14] Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*, 2019.

https://doi.org/10.52202/075280-0546

[15] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. 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*, pages 8648–8657, 2019.

[16] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: a knowledge engine foundation for human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[17] Teresa Wilcox, Rebecca Woods, Lisa Tuggy, and Roman Napoli. Shake, rattle, and... one or two objects? young infants' use of auditory information to individuate objects. *Infancy*, 9(1): 97–123, 2006.

[18] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[20] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems*, 31, 2018.

[21] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, 2017.

[22] Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1602. URL `https://aclanthology.org/P19-1602`.

[23] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. Curriculum disentangled recommendation with noisy multi-feedback. *Advances in Neural Information Processing Systems*, 34:26924–26936, 2021.

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[25] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[26] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

[27] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 264–279, 2018.

[28] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020.

[29] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020.

https://doi.org/10.52202/075280-0546

[30] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[31] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4613–4623, 2019.

[32] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2019.

[33] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5237–5246, 2019.

[34] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Stc-gan: Spatio-temporally coupled generative adversarial networks for predictive scene parsing. *IEEE Transactions on Image Processing*, 29:5420–5430, 2020.

[35] Mengshi Qi, Jie Qin, Xiantong Zhen, Di Huang, Yi Yang, and Jiebo Luo. Few-shot ensemble learning for video classification with slowfast memory networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3007–3015, 2020.

[36] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633, 2019.

[37] Mengshi Qi, Jie Qin, Yi Yang, Yunhong Wang, and Jiebo Luo. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing*, 30: 2989–3004, 2021.

[38] Mengshi Qi, Jie Qin, Yu Wu, and Yi Yang. Imitative non-autoregressive modeling for trajectory forecasting and imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2020.

[39] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.

[40] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021.

[41] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.

[42] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. *arXiv preprint arXiv:2308.05421*, 2023.

[43] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6538–6547, 2020.

[44] Junwen Bai, Weiran Wang, and Carla P Gomes. Contrastively disentangled sequential variational autoencoder. *Advances in Neural Information Processing Systems*, 34:10105–10118, 2021.

[45] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[46] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

https://doi.org/10.52202/075280-0546

[47] Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pages 5670–5679. PMLR, 2018.

[48] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[49] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[50] Shuwen Liu, Bernardo Grau, Ian Horrocks, and Egor Kostylev. Indigo: Gnn-based inductive knowledge graph completion using pair-wise encoding. *Advances in Neural Information Processing Systems*, 34:2034–2045, 2021.

[51] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2022.

[52] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[53] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[54] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

[55] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80: 2887–2905, 2021.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[57] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[58] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020.

[59] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

[60] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.