

# Supplementary Material of Toward Explainable Physical Audiovisual Commonsense Reasoning

Anonymous Authors

## ABSTRACT

In this Supplementary Material, we present details of the instruction dataset used for fine-tuning of the PAVC-Reasoner as well as the collection principles of PACS-Reason (A), details of the implementation of PAVC-Reasoner (B), example demonstrations of the predictions generated by PAVC-Reasoner variants on PACS-Reason and PACS-Reason+ benchmarks (C), visualizations of the effects of perceivers in PAVC-Reasoner (D), and an analysis of the model-generated hallucination rationales (E).

## CCS CONCEPTS

• Computing methodologies → Knowledge representation and reasoning.

## KEYWORDS

multimodal commonsense reasoning, commonsense reasoning, explainable reasoning, physical audiovisual commonsense reasoning

## A INSTRUCTION TUNING

PAVC-Reasoner empowers frozen LLMs with the capability of understanding both visual and auditory signals in videos as well as following human instructions by independently training the three perceivers. Specifically, for the *vision-language* perceiver, a two-stage training scheme is employed. In Stage I, cross-modal pretraining is conducted using two large-scale vision-caption datasets. This stage primarily focuses on achieving vision-language alignment. In Stage II, the perceiver is fine-tuned with several image-language instruction-following datasets. The training in this stage aims to enhance the model’s instruction-following ability. Regarding the *audio-language* and *audiovisual-language* perceivers, due to the absence of large-scale aligned audio-text pairs, we exclusively perform instruction fine-tuning for these two perceivers. This is accomplished using our self-constructed audio-text and video-audio-text instruction-tuning datasets. Both stages of training employ an autoregressive language modeling task, wherein the model learns to generate subsequent prediction text tokens based on the preceding context, aiming to maximize the log-likelihood of the generated text tokens.

**Unpublished working draft. Not for distribution.**

Permission to make digital or hard copies of all or part of this work for personal or professional use, by individuals or small businesses, is granted by ACM, provided that the copies are not made for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

## A.1 Vision Instruction Tuning

Based on the autoregressive training objective, we prompt the LLM to sequentially generate prediction tokens. Let  $\mathbf{x}_v$  and  $\mathbf{x}_a$  denote a video clip and its corresponding audio. Let  $\mathbf{x}_q$  and  $\mathbf{x}_r$  represent the instructions (questions) related to the audible video and the ground-truth responses (answers) respectively, which can be represented by a sequence of discrete text tokens. Then, for a sequence of length  $L$ , we can calculate the probability of generating the target answer as follows:

$$p(\mathbf{x}_r|\mathbf{x}_v, \mathbf{x}_q) = \prod_{i=1}^L p_{\theta}(x_i|\mathbf{x}_v, \mathbf{x}_q, \mathbf{x}_{r,<i}), \quad (1)$$

where  $\theta$  denote the trainable parameters. And  $\mathbf{x}_q$  denote the question (instruction) tokens.  $\mathbf{x}_{r,<i}$  are the answer (response) tokens before the current prediction token  $x_i$ , respectively. To pre-train the visual-language perceiver, we perform two distinct tasks: video-caption generation and image-caption generation. Here, images are considered as single-frame videos. For this purpose, we employ Webvid-2M [4], an extensive short video dataset that encompasses text descriptions collected from stock footage websites. Additionally, we utilize LLaVA-CC3M-Pretrain-595K [32], which constitutes a subset of the CC3M [40] image-text dataset, filtered to achieve a more balanced distribution of conceptual coverage. Specifically, we feed images/videos into the visual-language perceiver to obtain video prompt embeddings  $\tilde{\mathbf{Q}}_v$ , which are subsequently passed into the LLM to generate textual descriptions of the videos/images. Original video captions and image captions serve as the ground truth for prediction. During training, we maintain the weights of the visual encoder and LLM frozen, allowing only the trainable parameters of the Video Q-Former, the positional embedding layer, and the linear projection layer. This pretraining process ensures that the visual embeddings  $\tilde{\mathbf{Q}}_v$  align with the pretrained LLM text embeddings. And this phase can be conceptualized as training a compatible visual sensor for the frozen LLM, thereby enhancing the visual perception capabilities of the LLM.

After the pretraining phase, our model has acquired the fundamental capability to comprehend video/image content. Subsequently, the focus shifts to enhancing the LLM’s ability to express video details and follow multi-modal instructions. Therefore, in the second phase, we consider fine-tuning the model using a series of high-quality multi-modal instruction datasets. One such dataset is the LLaVA-Instruct-150K [32], generated through interactions with the COCO [30] dataset and GPT-4 comprising a total of 158K distinct language-image instruction-following samples with 58K in conversations, 23K in detailed description, and 77k in complex reasoning respectively. Another image-centric instruction dataset is the CC-Sub-Align [57] dataset, a high-quality collection of 3,500 image-text pairs collected by MiniGPT-4 [57] and structured in the instruction format. We also consider the VideoChat-Instruct-11K

dataset, a video-centric instruction dataset constructed by Video-Chat [27] based on WebVid-10M [4], encompassing 7K detailed video descriptions and 4K multi-run video conversations.

### A.3 Audio Instruction Tuning

Multimodal instruction datasets with a primary focus on images have seen a significant surge in development due to the abundance of image-text pairs [32]. However, datasets that center around audio in multimodal instructions are relatively scarce, mainly due to the limited availability of audio-text data [27]. It is worth noting that recent research has resorted to employing ASR datasets to train speech interfaces in order to align the output of the speech interface with the LLM. Nevertheless, given that our task is related to physical commonsense attributes, we aspire for the *audio-language* perceiver to endow PAVC-Reasoner with the ability to comprehend the natural audio semantics, rather than being overly oriented towards the interpretation of human speech.

To this end, we have considered three audio datasets aimed at characterizing natural environments, namely the Audio-Visual Event (AVE) [42] dataset, the VGG-Sound [7], and the ESC-50 [37] dataset. Notably, **AVE** comprises 4,143 videos spanning 28 event categories, with each video being temporally labeled for audio-visual events along with a background label. **VGG-Sound** encompasses over 200,000 video clips containing both visual frames and audio sound, encompassing 310 audio classes that span diverse and challenging acoustic environments and real-world noise characteristics. **ESC-50** constitutes a labeled collection of 2,000 environmental audio recordings, with each audio recording lasting 5 seconds and being categorized into one of 50 semantic classes. Given an audio segment and its associated semantic label ( $\mathbf{x}_a, \mathbf{x}_\ell$ ), it is inherent to generate a set of instructions  $\mathbf{X}_q$  aimed at prompting the LLM to articulate the audio content. We prompt GPT-4 [1] to curate such a list of instructions (see Fig. 1 for details) and randomly sample a datapoint  $\mathbf{x}_q$ , e.g., “*briefly describe the audio*”. We extend an audio-label pair to its instruction-following version by the following prompt: <System Message> ###Human:  $\mathbf{x}_a$  Question:  $\mathbf{x}_q$  <EOC> ###Assistant: This audio describes the sound of  $\mathbf{x}_\ell$  <EOC>. Here,  $\mathbf{x}_\ell$  is the semantic label of an audio clip, e.g., “*door knock*” or “*glass breaking*”. And <EOC> is the special token indicating end of chunk. The list of instructions used to describe audio content is illustrated in Fig. 16.

### A.4 Joint Instruction Tuning

The probability of generating text tokens for training the audiovisual language perceiver can be defined as follows:

$$p(\mathbf{x}_r | \mathbf{x}_a, \mathbf{x}_v, \mathbf{x}_q) = \prod_{i=1}^L p_\phi(x_i | \mathbf{x}_a, \mathbf{x}_v, \mathbf{x}_{q,<i>, \mathbf{x}_{r,<i>}), \quad (2)$$

where the objective of the learnable parameters  $\phi$  in the audiovisual language perceiver is to align the joint outputs of the frozen video encoder and audio encoder with the embedding space of LLM, enabling the model to possess joint audio-visual perception capabilities. We generate three types of instruction-following data based on PACS [50] and GPT-4: *detailed description*, *complex reasoning*, and *video conversation*.

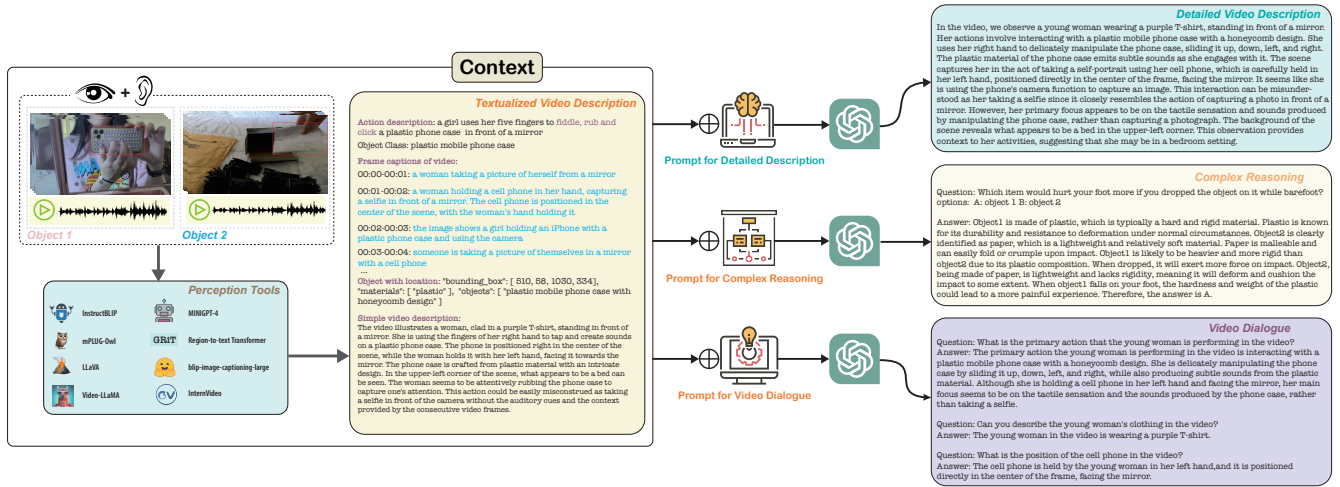
Fig. 1 presents an example of each type. For each type, we initially manually designed some examples, which served as unique human annotations during the data collection and were used as demonstrations to elicit GPT-4 in-context learning. We collected a total of 158K unique video-language instruction-following samples, including 23K video description samples, 77K complex reasoning samples, and 58K dialogue samples. Below, we elaborate on the construction process of this instruction-following data:

**Detailed Description** In order to provide rich and detailed descriptions of the joint audio-visual content in the videos, we created a series of questions with this intent. We prompted GPT-4 and filtered to obtain a list of questions, as shown in Table 1 in the Appendix. For each pair of videos  $\mathbf{o}_1$  and  $\mathbf{o}_2$ , we randomly selected a question from the list, querying GPT-4 to generate a detailed description. Fig. 17 shows the instructions list for asking model to describe the video content in detail.

**Complex Reasoning.** The instruction data of detailed descriptions primarily focuses on strengthening the model’s comprehension of audio-visual content in videos. Building upon this foundation, we further evaluate the model’s ability in the realm of physical commonsense reasoning. To achieve this, we have curated a series of physical commonsense questions, which encompass binary questions akin to those found in the original PACS [50] dataset and its derivative, PACS-Material [50]. In the given scenario involving a question  $\mathbf{x}_q$ , a video containing object  $\mathbf{o}_1$ , and a video containing object  $\mathbf{o}_2$ , the model is tasked with selecting the more suitable object to answer the question and provide a detailed rationale. Additionally, we have leveraged the capabilities of GPT-4 to create comparative audio-visual commonsense reasoning questions, aligned with the PACS style. Concretely, our design principles are categorized into two classes. The first class of questions pertains to intuitive physics and functional world models, such as predicting the *motion of objects*, or *interactions between objects*. These questions revolve around forecasting the effects of interactions between one or more objects. Simple questions involve individual objects and behaviors, e.g., “*Which object (a rubber ball or a ceramic tile) is more likely to shatter upon impact?*” More challenging questions entail multiple objects or behaviors, encompassing interactions between two objects, e.g., “*Which object would you put on top of the other (a plastic phone case and a piece of paper) to keep both from blowing away?*”

The second class of questions delves into common real-world knowledge, inspired by previous commonsense datasets, which examine human or object’s more specific understanding of how the real world operates [5, 15, 51]. These questions inquire about the potential applications of objects in real-life scenarios. It is of significance that these scenarios focus on non-standard uses of objects, thereby reducing the likelihood of misusing existing knowledge, e.g., “*In the absence of an eye shadow brush, which object should I use (a cotton swab or a toothpick)?*” More challenging questions may introduce more complex or uncommon scenarios involving multiple objects, e.g., “*If I want to stack these two objects together, which one should be placed underneath?*” We have amalgamated the formidable capabilities of GPT-4 with the meticulous review of multiple human experts to generate high-quality answers pertaining to both categories of questions. These answers typically adhere to rigorous logical reasoning and undergo a step-by-step reasoning process. The prompt used to ask GPT-4 to generate *complex*





**Figure 1: Pipeline of the GPT-assisted video-audio-text instruction-following data generation.** We generate three types of instruction-following data based on GPT-4 [57] and PACS [50]: detailed description, complex reasoning, and video conversation. We utilize various vision models for the transformation of visual data extracted from videos into textual video descriptions. Following this, employing diverse prompts and using human-annotated demonstrations as context, we elicit responses from GPT-4, inducing it to generate specific questions and corresponding rationales (inclusive of answers) aligned with our specified queries. The generated data will serve as our video-audio-text instruction-following dataset for model finetuning.

reasoning instruction data is shown in Fig. 18. And the prompt for the construction of the PACS-Reason+ using GPT-4 [1] to query all object pairs in PACS [50] is illustrated in Fig. 20.

**Video Conversation.** This design aims to boost the instruction-following and audiovisual commonsense reasoning capabilities of our model. We orchestrate a dialogue in which PAVC-Reasoner would act as an agent, engaging in conversation with a person who poses questions about the pair of videos. PAVC-Reasoner performs as if it is actively watching the videos and providing answers. Based on the audiovisual content of the videos, we use GPT-4 to formulate a diverse set of physical commonsense questions and answers, encompassing aspects such as *material composition of objects*, *physical properties and affordances of objects*, *effects of interactions between objects*, and *how to manipulate them*, as well as *potential real-world applications of objects*. We exclusively consider questions that have explicit answers. The prompt used to ask GPT-4 to generate video conversation instruction data is shown in Fig. 19.

## B IMPLEMENTATION DETAILS

### B.1 Baseline Multimodal Models

Specifically, we use state-of-the-art encoders to obtain unimodal outputs. For instance, we take into account the following pretrained encoders:

**Image Encoder:** Using the ViT/B-16 model [13] pre-trained on ImageNet-21k [12] to derive the image embedding for each object. **Audio Encoder:** Utilizing the AST (Audio Spectrogram Transformer) model [20], featuring a time and frequency stride of 10 and weight averaging, pre-trained on the full AudioSet [22] as the audio encoder for generating the object's audio embeddings.

**Video Encoder:** Employing the TDN (Temporal Difference Network) model [44], pre-trained on Something-Something-V2 [21] with a ResNet101 [23] backbone and 8-frame inputs, as the video encoder to obtain the object's video embeddings.

**Question Encoder:** Employing a pre-trained DeBERTa-V3-Large model [24] as the text encoder to derive the question embeddings of a pair of objects. We do not use any text augmentations, and we pre-extract the text embeddings from the [CLS] token of the output layer (pre-pooler) of the text model to optimize training efficiency. We amalgamate unimodal embeddings from *image*, *audio*, *video* and *question* encoders to construct five late fusion models as follows:

**I+A+V:** Concatenating the unimodal embeddings of image, audio, and video for both objects, followed by an MLP to derive two object embeddings,  $e_o^{(1)}$  and  $e_o^{(2)}$ . Finally, the two object embeddings are concatenated and fed into the MLP to generate binary classification output.

**Q+I:** Concatenating the separate image embeddings of the two objects with the question embedding, and then utilizing an MLP to produce two question-object embeddings,  $e_{qo}^{(1)}$  and  $e_{qo}^{(2)}$ . Finally, these question-object embeddings are input to an MLP to generate binary classification output.

**Q+A:** Concatenating the individual audio embeddings of the two objects with the question embeddings, followed by an MLP to create two question-object embeddings,  $e_{qo}^{(1)}$  and  $e_{qo}^{(2)}$ . Consequently, these embeddings are then fed into an MLP to derive binary classification output.

**Q+V:** Concatenating the individual video embeddings of the two objects with the question embeddings, followed by using an MLP to generate two question-object embeddings,  $e_{qo}^{(1)}$  and  $e_{qo}^{(2)}$ . Finally, an

MLP processes these question-object embeddings to obtain binary classification output.

**Late Fusion (Q+I+V+A)** Concatenating the image, audio, and video embeddings of both objects, using an MLP to generate two object embeddings,  $e_o^{(1)}$  and  $e_o^{(2)}$ . These object embeddings are concatenated with the question embeddings, followed by an MLP to yield two question-object embeddings,  $e_{qo}^{(1)}$  and  $e_{qo}^{(2)}$ . Finally, these question-object embeddings are concatenated and sent to an MLP to yield binary classification output. In addition, we also consider the following advanced methods:

- **CLIP** [38] (contrastive language–image pre-training), known as a powerful zero-shot transfer, pre-trains an image encoder and a text encoder to predict which images were paired with which texts by maximizing the cosine similarity of their embeddings. Employing CLIP, we encode two images and a question for a pair of objects, aiming to predict the object whose image embedding is closer to the question embedding.
- **AudioCLIP** [22] extends CLIP by integrating an audio encoder, resulting in a tri-modal hybrid architecture. Employing the audio encoder and image encoder from AudioCLIP, we acquire audio embeddings and visual embeddings of an object. These are concatenated and projected onto the same embedding space as the question embedding via a linear layer. The objective is to predict the object whose embedding aligns more closely with the question embedding.
- **UNITER** [8] is a large-scale pre-trained model for joint multimodal embedding. It initially employs image encoders and text encoders to encode image regions (containing visual features and bounding box features) and textual words (tokens and positions) into a shared embedding space. Subsequently, a transformer module is applied through four meticulously designed pre-training tasks to learn generalizable contextualized embeddings for each region and word. Employing UNITER, we input image-question pairs to generate image embeddings and question embeddings, then concatenate these embeddings and feed them into an MLP for binary answer classification.
- **MERLOT Reserve** [52] extends MERLOT [53] by including the audio modality. It acquires a multimodal neural script representation of videos, jointly reasoning over video frames, text, and audio via a unified vision-audio-language Transformer. We obtain the question-object embeddings by feeding the sequence-level representations of video frames, question words, and audio segments to the joint encoder. This is achieved by pooling the output hidden representations of the last layer. We then feed these question-object embeddings into a linear layer to generate binary answer predictions.

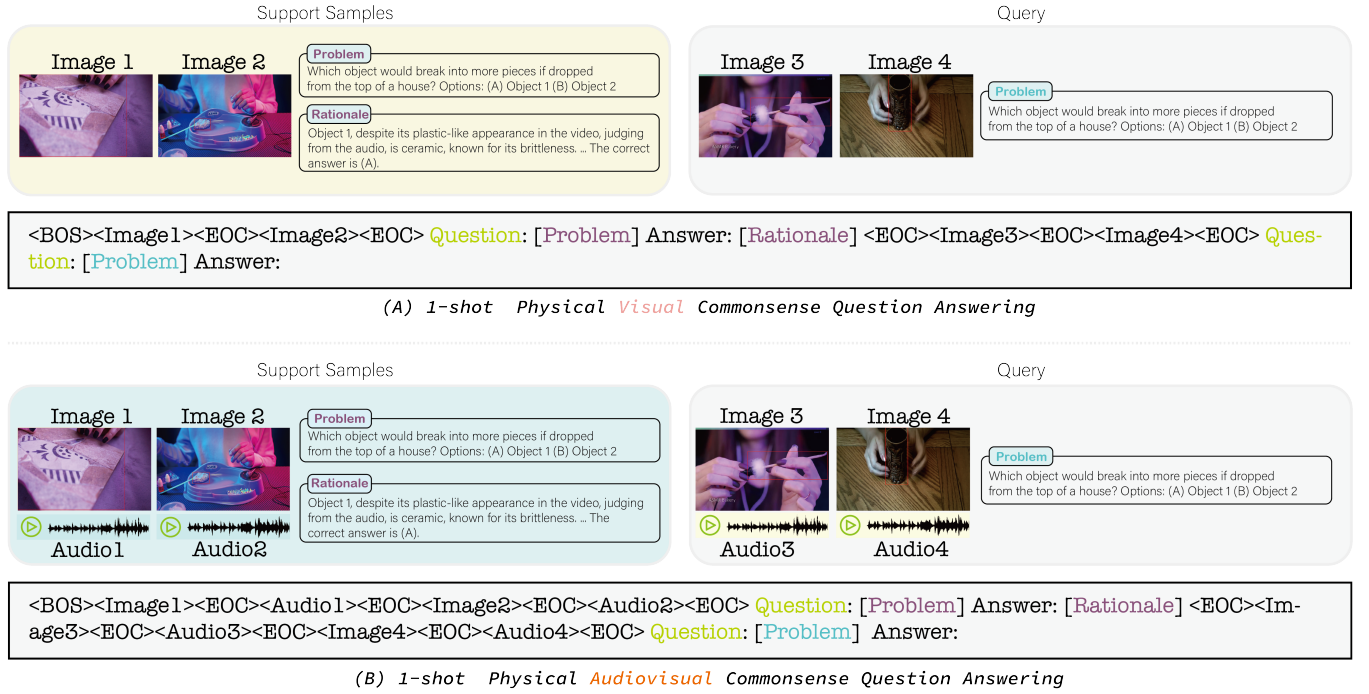
## B.2 Baseline Multimodal Large Language Models (MLLMs)

We also consider various popular open-source multimodal large language models (MLLMs) as baseline models. These MLLMs have shown excellent performance on comprehensive multimodal evaluation benchmarks such as MME [16] and MMBench [33] which include multiple dimensions or tasks evaluating reasoning abilities

such as *physical attribute reasoning*, *functional reasoning*, and *commonsense reasoning*. In essence, these MLLMs exhibit a certain degree of perceptual and cognitive capabilities, rendering them robust baseline models for our explainable physical audiovisual commonsense reasoning (EPACR) task. Specifically, we evaluate the following MLLMs: FrozenBiLM [47], MiniGPT-4 [57], InstructBLIP [11], OpenFlamingo [3], MM-GPT [19], LLaVA [32], mPLUG-Owl [48], VideoChat [27], LLaMA Adapter [55], Video-ChatGPT [36] and Video-LLaVA [29].

Table 2 records the parameter counts of these MLLM models and the specific configurations. Since most MLLMs are only equipped to handle visual and textual inputs, we only test their physical visual commonsense reasoning capabilities. We did not perform expensive fine-tuning on these MLLMs on instruction-following data (i.e. PACS-Reason and PACS-Reason+). Instead, we employed a multimodal few-shot in-context learning approach similar to the methods used in GPT-3 [6] and Flamingo [2] for PACR task adaptation. Specifically, we receive a set of *support examples* in the form of  $\langle \text{video}, \text{text} \rangle$  or  $\langle \text{video}, \text{audio}, \text{text} \rangle$ , where *video* is regarded as a sequence of images, and *text* includes the expected response and relevant task information such as questions or instructions. At the end of these demonstrations, we append a single *query* requesting the model to make a prediction. Fig. 2 illustrates how we construct multimodal prompts by concatenating support examples and queries. Due to resource constraints, we use a one-shot evaluation. In addition, we consider two MLLMs capable of processing audio, Video-LLaMA [54] and PandaGPT [41]. We fine-tune these two MLLMs on the PACS-Reason and PACS-Reason+ instruction datasets, using the same instruction fine-tuning strategy as PAVC-Reasoner (combined with LoRA). Below, we briefly introduce each MLLM used in our experiment:

- **FrozenBiLM** [47] builds on *frozen bidirectional* language models (BiLM) and show that such an approach provides a stronger and cheaper alternative for zero-shot VideoQA. In particular, it combine visual inputs with the frozen BiLM using light trainable modules, and trains such modules using Web-scraped multi-modal data, and finally performs zero-shot VideoQA inference through masked language modeling, where the masked text is the answer to a given question.
- **MiniGPT-4** 🦜 [57] consists of a vision encoder with a pre-trained ViT and Q-Former, a single linear projection layer, and an frozen LLM. MiniGPT-4 only requires training the linear layer to align the visual features with the Vicuna [9] large language model.
- **InstructBLIP** 🦜 [11] Mirroring the architecture of BLIP-2 [26], InstructBLIP encompasses a frozen image encoder, a trainable Q-Former, and a frozen Language Model (LLM). Specifically, the Q-Former is used to extract instruction-aware visual features from the output embeddings of the frozen image encoder. These visual features are then presented as soft prompts to the frozen LLM. Employing language modeling loss, the model undergoes instruction tuning to generate responses.
- **OpenFlamingo** 🦜 [3] is an open-source implementation of the vision-language model, Flamingo [2]. It integrates



**Figure 2: Conceptualization of Few-shot Multimodal Prompt Generation for PACR.** Given a few task-specific few-shot examples (referred to as support samples, as illustrated in the figure with a 1-shot demonstration) and a query for which MLLMs must make a prediction, we construct prompts by interleaving video and audio preceding each corresponding text. To facilitate this, we introduce a structured format tailored for physical visual or audiovisual commonsense reasoning, exemplified as follows: “Question: {problem} Answer: {rationale}”. Zoom in for the best view.

powerful pre-trained visual models with pure language models by introducing novel cross-attention layers within each transformer block of the language model (LM). The keys and values of these layers are derived from visual features when utilizing language queries. Subsequently, dense feed forward layers are inserted between the existing pre-trained and frozen LM layers to fine-tune the language model based on visual input. These layers are gated to ensure the integrity of the LM during initialization, thereby enhancing stability and performance. The model adeptly handles arbitrary interleaved sequences of visual and textual data, including the seamless ingestion of images or videos as input.

- **MM-GPT** [19] excels at following various human directives, such as generating detailed instructions, counting objects of interest, and answering general questions from users. **MM-GPT** is parameter-efficiently fine-tuned on OpenFlamingo [3], where low-rank adapters (LoRA [25]) have been incorporated into both the cross-attention and self-attention components for parameter-efficient fine-tuning. It leverages language-only instruction-following data alongside visual-language instructions for joint training, enabling the model to comprehend and adhere to human instructions.
- **LLaMA Adapter** [55] is a lightweight adaption method for instruction fine-tuning a multi-modal LLaMA [43] model. It inserts lightweight adapters with learnable prompts into

the transformer layers of LLaMA and adopts zero initialized attention with gating mechanisms for stable training in early stages. In our experiments, we used LLaMA-Adapter (V2) [17] for evaluation.

- **LLaVA** [32] is an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language understanding, achieving impressive chat capabilities mimicking spirits of the multimodal GPT-4. We adopted the LLaVA-v1.5-13b [31] for evaluation, which is a stronger version of LLaVA.
- **mPLUG-Owl** [48] is a novel training paradigm that equips LLMs with multi-modal abilities via modularized learning of LLM, a vision foundation model to encode visual knowledge, and a visual abstractor module. It involves a two-stage method for aligning image and text, which learns visual knowledge with the assistance of LLM, while maintaining and even improving the text generation abilities of LLM.
- **mPLUG-Owl2** [49] utilizes a modularized network design, with the modality-adaptive language decoder acting as a universal interface for managing different modalities. mPLUG-Owl2 incorporates shared functional modules to facilitate modality collaboration and introduces a modality-adaptive module that preserves modality-specific features.
- **VideoChat** [27] is an end-to-end chat-centric video understanding system. It integrates video foundation models, namely InternVideo [45], and large language models via a



**Table 1: Performance comparisons between PAVC-Reasoner and diverse baseline models on the newly-collected PACS-Reason+ benchmark. Results are reported as the mean and standard deviation over five runs. For the case of MLLMs, evaluations were conducted using the 1-shot multimodal in-context learning or fine-tuning on the PACS-Reason+ training set. The symbol "-" indicates that the current model lacks support for rationales generation or audio input, thus rendering results unattainable.**

Baseline Model	Accuracy (%)		Helpfulness (%)		Relevance (%)		Granularity (%)		HumanEval (%)	
	+audio	-audio	+audio	-audio	+audio	-audio	+audio	-audio	+audio	-audio
<i>Simple Late Fusion Baselines</i>										
I+A+V [13, 20, 44]	51.6±0.8	-	-	-	-	-	-	-	-	-
Q+I [13, 24]	-	49.1±0.7	-	-	-	-	-	-	-	-
Q+A [20, 24]	48.3±1.1	-	-	-	-	-	-	-	-	-
Q+V [24, 44]	-	49.2±0.9	-	-	-	-	-	-	-	-
Q+I+A+V [50]	51.4±0.8	48.8±1.0	-	-	-	-	-	-	-	-
<i>Multimodal Models</i>										
CLIP [38]	-	55.6±0.8	-	-	-	-	-	-	-	-
AudioCLIP [22]	58.1±1.2	-	-	-	-	-	-	-	-	-
UNITER (Large) [8]	-	59.9±1.7	-	-	-	-	-	-	-	-
Merlot Reserve (B) [52]	64.8±1.3	62.2±0.8	-	-	-	-	-	-	-	-
Merlot Reserve (L) [52]	68.4±0.9	66.3±0.5	-	-	-	-	-	-	-	-
<i>Multimodal Large Language Models (MLLMs) with 1-shot Multimodal Prompt</i>										
FrozenBiLM [47]	-	57.2±1.3	-	64.7±1.4	-	69.0±1.2	-	66.6±1.3	-	57.7±1.6
MiniGPT-4 [57]	-	59.2±1.5	-	65.8±1.4	-	70.2±2.0	-	67.2±0.6	-	58.5±2.2
InstructBLIP [11]	-	61.4±1.8	-	67.6±1.2	-	71.0±0.5	-	68.7±1.4	-	61.3±0.3
MM-GPT [19]	-	63.3±1.6	-	68.5±0.4	-	72.6±2.0	-	69.2±1.9	-	62.3±1.5
OpenFlamingo [3]	-	64.7±0.7	-	71.5±1.4	-	74.2±1.8	-	71.6±1.8	-	63.7±1.7
LLaMA Adapter [55]	-	62.9±1.6	-	68.4±1.3	-	72.6±1.6	-	70.2±0.8	-	59.6±2.1
LLaVA [32]	-	63.5±2.0	-	67.1±1.5	-	70.5±1.3	-	69.0±2.0	-	59.5±1.0
mPLUG-Owl [48]	-	65.1±0.6	-	69.7±2.1	-	75.8±0.6	-	71.0±2.1	-	60.5±2.2
mPLUG-Owl2 [49]	-	66.2±1.3	-	70.4±1.2	-	75.4±0.8	-	71.4±1.7	-	62.4±0.7
VideoChat [27]	-	61.7±1.1	-	68.5±1.3	-	71.5±1.1	-	70.1±0.6	-	60.5±1.5
Video-ChatGPT [36]	-	63.1±1.1	-	70.0±0.9	-	71.7±1.9	-	46.8±1.4	-	60.0±0.5
Video-LLaVA [29]	-	64.1±1.0	-	69.0±1.7	-	73.0±1.7	-	48.4±1.9	-	62.1±1.7
<i>Finetuned Multimodal Large Language Models (MLLMs) on PACS-Reason+</i>										
Video-LLaMA [54]	61.4±1.8	59.3±1.3	71.5±1.8	66.4±1.6	74.1±1.3	72.6±1.2	73.2±1.8	70.2±1.4	62.6±1.9	60.5±2.1
Video-LLaMA-2 [54]	62.3±0.5	60.0±0.8	72.6±1.4	66.9±2.0	75.3±0.8	73.6±1.5	74.1±1.4	70.8±0.7	64.0±2.2	61.4±2.0
PandaGPT [41]	62.2±1.6	62.0±1.2	73.9±1.5	68.8±1.7	76.1±1.9	73.9±1.3	74.6±1.3	72.0±1.8	64.8±2.4	61.6±2.0
PAVC-Reasoner-V2	73.5±0.7	71.8±0.5	74.7±1.6	71.6±1.1	77.6±0.4	76.9±0.9	74.7±1.5	71.8±1.1	67.6±1.4	63.8±1.5
PAVC-Reasoner-V4	76.7±1.3	74.3±1.0	79.0±1.7	75.7±0.1	86.9±0.8	85.1±1.4	83.1±0.4	78.6±2.0	68.9±1.8	64.9±1.3

learnable neural interface, excelling in spatiotemporal reasoning, event localization, and causal relationship inference. **VideoChat2** [28] is a robust video multimodal LLM baseline, achieving progressive multimodal training with diverse instruction-tuning data. The progressive multimodal training encompasses three distinct phases: Phase 1 aligns the visual encoder UMT-L with QFormer, effectually compressing extensive visual input. Stage 2 extends this association to incorporate LLM, while Stage 3 concentrates on effective instruction tuning to enable the model to follow human instructions.

- **Video-ChatGPT** [36] leverages the CLIP-L/14 [38] visual encoder to extract both spatial and temporal video feature, and feeds the resulted spatiotemporal features into a learnable linear layer, which then projects them into the LLMs input space.
- **Video-LLaVA** [29] extracts visual features using LanguageBind [56] encoders and aligns images and videos before projection, allowing LLM to learn from a unified visual

representation and endowing LLM with the ability to understand both images and videos simultaneously.

- **Video-LLaMA** [54] is a multimodal framework that empowers LLMs with the capability of understanding both visual and auditory content in the video. It derives a Video Q-former to assemble the pre-trained image encoder into the video encoder and introduce a video-to-text generation task to learn video-language correspondence, and introduces an Audio Q-former on top of ImageBind [18] to learn reasonable auditory query embeddings for the LLM module. Video-LLaMA is trained on massive video/image-caption pairs as well as visual-instruction-tuning datasets of moderate amount but higher quality.
- **PandaGPT** [41] combines ImageBind's multimodal encoders and a Vicuna [9] LLM to achieve outstanding capabilities across six modalities (text, image/video, audio, depth, thermal, and IMU). With a modest set of trainable parameters added (a linear projection layer and LoRA weights on the attention modules of Vicuna), PandaGPT is trained with

aligned image-text pairs (160K image-language instruction-following data released by LLaVa and MiniGPT-4).

### B.3 Model Settings

For the frozen image encoder, we investigated two state-of-the-art pre-trained visual transformer models: (1) ViT-L/14 from CLIP [38], which has 24 layers with hidden dimension set as 1024, number of attention heads as 16 and patch size set as 14; and (2) ViT-G/14 from EVA-CLIP [14]. As for the frozen audio encoder, we explored audio encoders from ImageBind [18] and LanguageBind [56]. Regarding the frozen language models, we considered two language model series: (1) FlanT5 [10], which is an encoder-decoder language model fine-tuned from T5 [39], and (2) Vicuna [9], which is a decoder-only language model finetuned from LLaMA [43]. For the *visual-language* and *audio-language* perceivers, we use the six-layer, 16-head transformer blocks, comprising 64 learnable queries, each with a dimension of 768. For the *audiovisual-language* perceiver, the number of learnable queries is set to 64, with each query having a dimension of 768. Notice that in the cross-modal training stage, only the three perceivers, the position embedding layer, and the linear projection layer are trainable. In the instruction fine-tuning stage, besides the modules mentioned above, we also update the parameters of the LoRA module.

### B.4 Data Preprocessing and Training Details

Due to the inclusion of a video pair in PACS-Reason and PACS-Reason+, we conduct temporal sampling by extracting 4 frames per second (fps) for each video. For every sampled image, we employ UniDetector [46], a versatile object detector, to detect the target object within each sampled image. Subsequently, we apply crop and resize operations to the region containing the target object, resulting in a size of  $224 \times 224$  for each processed image. In the case of other videos, (such as Webvid-2M and VideoChat), we uniformly sample 8 frames and resize each frame to  $224 \times 224$ . Additionally, we perform image augmentations, including random resized cropping and horizontal flipping. Regarding audio input, we uniformly extract 4 segments of 2-second short audio clips from the video. To encode the audio, we transform each 2-second audio segment into spectrograms using 128 mel-spectrogram bins. Since the spectrogram shares similarities with a 2D image signal, we utilize the same patch embedding technique with a patch size of 16 and a stride of 10, as described in [20]. All variants of PAVC-Reasoner are conducted on four 80GB A100 GPUs. Consistent pre-training hyperparameters are employed across all models. We use the AdamW optimizer [35] for optimization with  $\beta = (0.9, 0.98)$  and a weight decay set at 0.05. We use a cosine learning rate decay with a peak learning rate of  $1e-4$  and a linear warmup of 2k steps. The minimum learning rate in the second stage is set to  $1e-6$ . As shown in Table 3, we record the hyperparameters of the training PAVC-Reasoner at the pretraining stage and the instruction fine-tuning stage in detail. Through grid search on the validation subset of the training set (or on the available official or standard validation sets), we opt for certain hyperparameters for each task. These hyperparameters encompass the learning rate (ranging from  $1e-8$  to  $1e-3$ ), decay schedule (exponential decay by a factor of 10), training steps, batch size (256 or 512), and the utilization of visual data augmentation (including

color augmentation and random horizontal flipping and random resized cropping etc.)

### B.5 GPT-4-Assisted Choice Extraction

Due to the limited instruction-following capabilities of various MLLMs, despite our questions being in the form of clear multiple-choice questions with well-structured options (e.g., "*Which object is softer to the touch? Options: A. object 1 B. object 2*"), many MLLMs still produce answers in free-form text (e.g., the model's direct output could be "*The correct answer is object 1,*" but not explicitly 'A'). Extracting choices from free-form predictions is easy for humans but challenging for rule-based matching. Therefore, inspired by MMBench [33], we devise a universal evaluation strategy for all MLLMs with varying instruction-tracking abilities. Specifically, we employed GPT-4 as a choice extractor. We provided GPT-4 with the question, options, and model predictions, prompting it to align the prediction with one of the given options and generate the corresponding option label. We then extract the predicted option (e.g., 'A') from the response generated by GPT-4. For most questions, GPT-4 is able to return a single character (e.g., 'A', 'B' or 'X') as the model's label prediction. For each input, we compare the model's label prediction with the ground-truth true label. If the prediction matches the true label, the test sample is considered correct. The prompt for choice extraction is illustrated in Fig. 3.

## C PREDICTION DEMONSTRATION

In this section, we present additional exemplars of the commonsense reasoning proficiency of PAVC-Reasoner on the PACS-Reason+ (as illustrated in Figures 7, 8, 9, 10, 12, and 14) and PACS-Reason++ benchmarks (as illustrated in Figures 11, 13, and 15). We illustrate the model's capacity to generate corresponding rationales for various questions, i.e., intermediate reasoning chains serving as the rationale. These visualizations comprehensively showcase our model's capacity in generating step-by-step rationales for diverse physical commonsense questions involving a variety of physical properties—namely, intermediate reasoning chains that elucidate the logical steps taken to derive the answer.

## D VISUAL PREFERENCE V.S. AUDIOVISUAL PREFERENCE

We aim to intuitively analyze the impact of various perceivers on PAVC-Reasoner. To achieve this, in Fig. 5, we present two questions with distinct preferences. Question A, "*which object could hold water?*" is relatively straightforward, involving an examination of object shape and size. In contrast, Question B, querying *which object is useful in starting a fire if lost in the cold and dark woods*, entails a certain level of complexity, involving an assessment of the non-prototype uses of two objects in daily life. This necessitates the model's accurate understanding of the material composition, physical properties, and availability of the objects. In Fig. 4, we visualize the attention heatmaps in the V-L branch and A-L branch for both questions.

From the comparison of attention maps between visual tokens and audio tokens in the 0th and 1st layers in Fig. 4, we observe that the model predominantly focuses more on visual content when generating responses to Question A. Conversely, the attention maps

**Table 2: All parameter counts and detailed model configurations for the multimodal large-scale language models used in our experiments.**

Methods	Large Language Model	Vision Encoder	Audio Encoder	Trainable LLM	Trainable Visual Encoder
FrozenBiLM [47]	Deberta-V2-XLarge (890M)	CLIP-ViT-L/14 (304M)	✗	🔥	❄️
MiniGPT-4 [57]	Vicuna-V0-7B	CLIP-ViT-L/14 (304M)	✗	❄️	❄️
InstructBLIP [11]	Vicuna-V1.1-7B	CLIP-ViT-L/14 (304M)	✗	❄️	❄️
OpenFlamingo [3]	mpt-1b-redpajama-200b	CLIP-ViT-L/14 (304M)	✗	🔥	❄️
MultiModal-GPT [19]	LLaMA-7B	CLIP-ViT-L/14 (304M)	✗	🔥	❄️
LLaMA Adapter V2 [55]	LLaMA-7B	CLIP-ViT-L/14 (304M)	✗	🔥	❄️
LLaVA-1.5 [32]	Vicuna-V1.5-13B	CLIP-ViT-L-336px (304M)	✗	🔥	❄️
mPLUG-Owl [48]	LLaMA-7B	CLIP-ViT-L/14 (304M)	✗	🔥	🔥
mPLUG-Owl2 [49]	LLaMA-7B	CLIP-ViT-L/14 (304M)	✗	🔥	🔥
VideoChat [27]	Vicuna-7b-delta-v0	EVA-ViT-G/14 (1.0B)	✗	🔥	❄️
VideoChat2 [28]	Vicuna-7b-delta-v0	UMT-L (304M)	✗	🔥	🔥
Video-ChatGPT [29]	Vicuna-v1.1-7B	CLIP-ViT-L/14 (304M)	✗	❄️	❄️
Video-LLaVA [29]	Vicuna-v1.5-7B	CLIP-ViT-L/14 (304M)	✗	🔥	❄️
Video-LLaMA [54]	Vicuna-7B	EVA-ViT-G/14 (1.0B)	ImageBind-H (307M)	❄️	❄️
Video-LLaMA-2 [54]	LLaMA-2-7B-Chat	EVA-ViT-G/14 (1.0B)	ImageBind-H (307M)	❄️	❄️
PandaGPT [41]	Vicuna-7B	ImageBind-H (307M)	ImageBind-H (307M)	❄️	❄️
PAVC-Reasoner-V2	FlanT5 <sub>XL</sub> -3B	EVA-ViT-G/14 (1.0B)	ImageBind-H (307M)	🔥	❄️
PAVC-Reasoner-V4	Vicuna-7B	EVA-ViT-G/14 (1.0B)	ImageBind-H (307M)	🔥	❄️



### Choice Extraction Template

```

messages = [ {"role": "system", "content": f'"" Suppose you operate as an AI assistant
to help me matching an answer with several options of a multiple choice question.
You are provided with a question, several options, and an answer, and you need to
find which option is most similar to the answer.
If the meaning of all options are significantly different from the answer, output X.
Your should output a single uppercase character in A or B (if they are valid
options), and X.
Example 1:
Question: Which object would be less painful to rest your head on while napping?
Options: A. object 1 B. object 2
Answer: object 2
Your output: B
Example 2:
Question: Which item would hurt your foot more if you dropped the object on it
while barefoot? Options: A. object 1 B. object 2
Answer: The answer is C
Your output: X
Example 3:
Question: {question}? Options: {options}
Answer: {prediction}
Your output:
"" } ]
for sample in fewshot_samples:
    messages.append({ "role": "user", "content": sample["context"] })
    messages.append({ "role": "assistant", "content": sample["response"] })
    messages.append({ "role": "user", "content": "\n".join(query) })

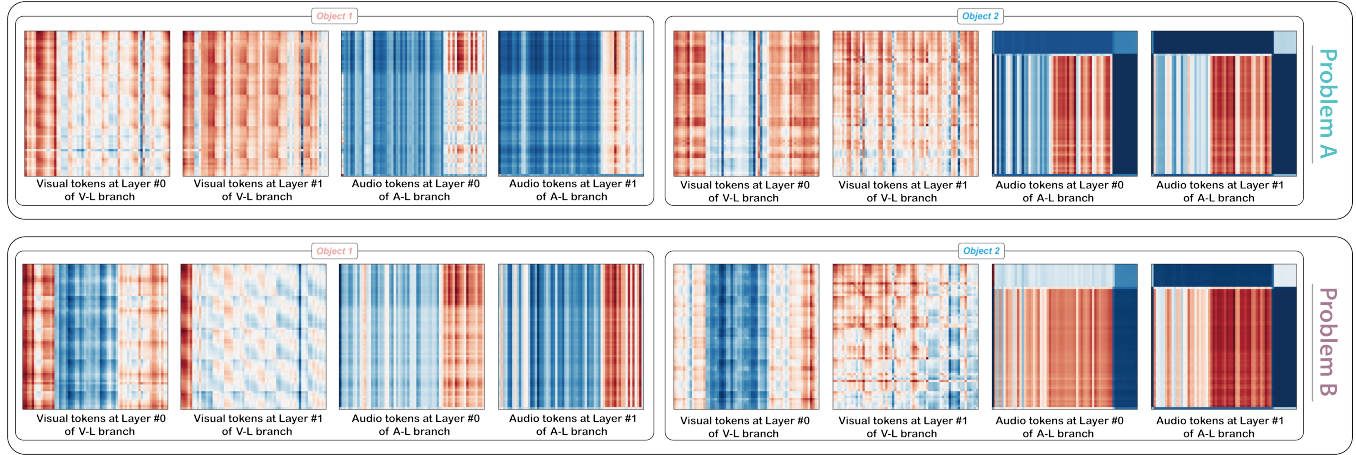
```

**Figure 3: The prompt used to generate the binary answer candidate based on GPT-4 [1].**

in the second row indicate that the model, in generating responses to Question B, simultaneously attends to both visual and audio content. This may be attributed to the valuable information that audio can provide to distinguish the physical properties between the two objects, such as object 2 being easily misidentified as a wooden cup in the absence of audio (actually a ceramic cup). In comparison to Question A, when addressing Question B, the model elevates

the contribution of audio features to correctly identify the physical properties of the object. This signifies that our perceptual branches can adaptively capture modal features of varying degrees based on the nature of the question, indirectly reflecting their effectiveness.





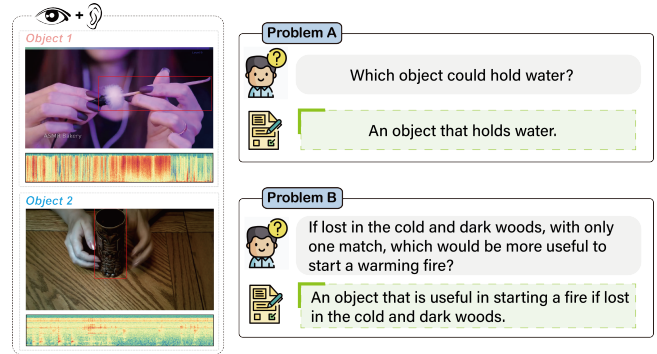
**Figure 4: Illustrations of attention maps for visual tokens and audio tokens in layers 0 and 1 about various questions. Blocks closer to the red spectrum indicate higher attention values, while those closer to the blue spectrum signify lower attention values. The range of all attention maps is normalized to a fixed interval for enhanced contrast in differences.**

**Table 3: Training hyper-parameters of PAVC-Reasoner.**

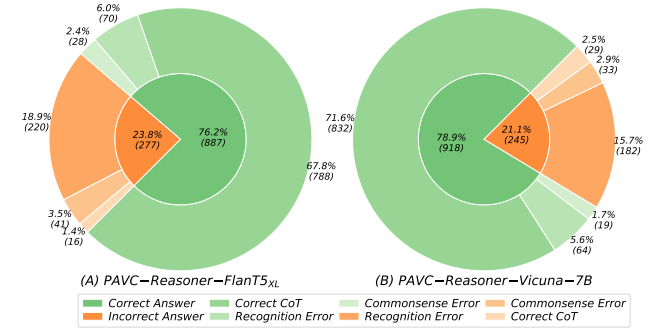
Configuration	Pretraining Stage	Instruction Tuning
Vision encoder init.	ViT-L or ViT-G [26]	Pretraining Stage
Audio encoder init.	ImageBind-Huge [18]	
Audiovisual encoder init.	Random	Random
LLM init.	FlanT5 <sub>XL</sub> [10] or Vicuna-7B [9]	
Image Resolution	224×224	224×224
ViT sequence length	256	1024
LLM sequence length	256	2048
Learnable query numbers	64	64
Optimizer	AdamW [35]	
Optimizer momentum	$\beta_1, \beta_2 = (0.9, 0.999)$	
Peak learning rate	1e-4	2e-5
Minimum learning rate	1e-6	1e-7
Layer-wise lr decay	0.75	0.85
Drop path rate	0	0
Learning rate schedule	cosine decay [34]	
Weight decay	0.05	0
Gradient clip	1.5	
Training steps	250,000	80,000
Warmup steps	2000	250
Global batch size	4096	256
Gradient Acc.	16	
Numerical precision	bfloat16	
Optimizer sharding	✓	
Activation checkpointing	✓	

## E HALLUCINATION RATIONALE ANALYSIS

To delve deeper into the quality of rationales generated by our model, we classified the model's predictions on the PACS-Reason test set. Fig. 6 summarizes the classification results produced by PACS-Reasoner. Firstly, we observe that even samples classified as correct (*i.e.*, accurate answers) contain a certain number of erroneous rationales (CoTs). These erroneous rationales primarily fall into two categories: one arising from model misidentifications of object materials and physical properties, constituting errors in



**Figure 5: Example of two questions with different modality preferences.**



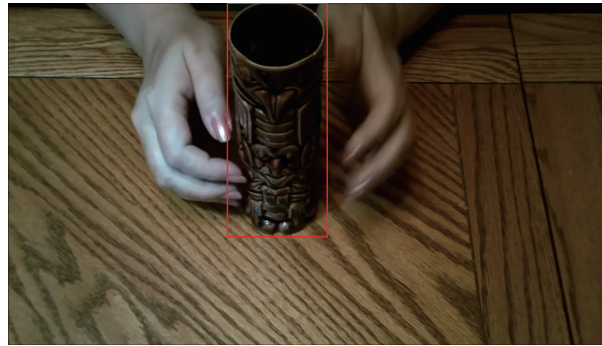
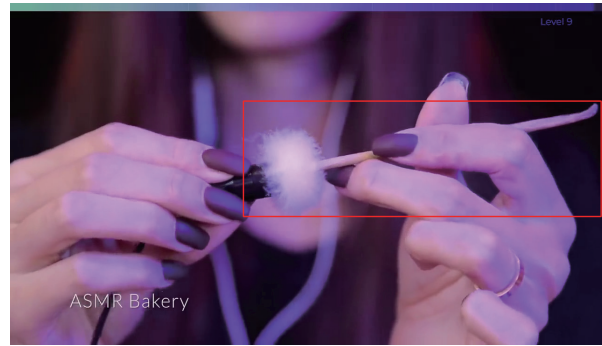
**Figure 6: Categorization error analysis of PAVC-Reasoner.**

CoTs, and the other involving generated CoTs containing commonsense errors (*e.g.*, "... smaller objects generally experience more air resistance relative to their size. . ."). Notably, PAVC-Reasoner (with Vicuna-7B [9]) demonstrates a reduction in the generation of CoTs with commonsense errors compared to PAVC-Reasoner

(with FlanT5<sub>XL</sub> [10]), decreasing from 2.4% to 1.9%. This implies that a more powerful LLM contributes to mitigating commonsense error issues. On the other hand, despite the incorrect rationales generated by the model, the final answers (*i.e.*, choosing A or B) were correct. This suggests that our model exhibits a degree of robustness, capable of predicting the correct answers by disregarding incorrect rationales.

For incorrect samples (*i.e.*, incorrect answers), misidentification errors in the CoTs dominate, constituting 18.9% and 15.7% of errors among two model predictions, respectively. When answering questions involving judgments about the physical properties and material categories of objects, such as understanding what material the cup in the video is made of (see Fig. 5), the model often makes misidentification errors. For instance, *object 2* might be misidentified as a wooden cup instead of a ceramic one. Similarly, we found that the more powerful visual perception model ViT-G/14 [14], as compared to ViT-L/14 [38], can reduce the occurrence of misidentification errors, lowering the error rate from 18.9% to 15.7%. Nevertheless, limitations persist in our model’s understanding of visual and auditory content in videos. This motivates our exploration of *more efficient multimodal cooperative and modality-separate network modules to capture both modality-shared and modality-specific features*. Additionally, we are actively constructing high-quality video-audio-text alignment datasets to further enhance the model’s perception capabilities. Within the category of incorrect CoT, another type of error is commonsense errors, where CoTs contain inaccuracies in commonsense knowledge. Moreover, in some cases where the answer is incorrect, the CoT is correct (with ratios of 1.4% and 2.5%), but may not necessarily be relevant to the answers.

1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160



PACS-Reason



#### System Message

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.



Which object could hold water?  
Options: (A) Object 1 (B) Object 2



If hiding in a building, which object would you toss to best distract and lure away your pursuers?  
Options: (A) Object 1 (B) Object 2



If lost in the cold and dark woods, with only one match, which would be more useful to start a warming fire?  
Options: (A) Object 1 (B) Object 2



PAVC-Reasoner  
w/o audio

Object 1 appears to be a wooden stick with fuzz. It is unlikely that this object could hold water due to its narrow and non-hollow structure. Object 2 is likely a wooden cup with painted carvings. Its cup-like shape and the material that it is made of wood make it a suitable vessel for holding liquids. Therefore, the correct answer is (B) Object 2.



PAVC-Reasoner  
w/o audio

Object 1 is likely a wooden stick with fuzz on it, while Object 2 appears to be a painted wood cup with carved patterns. Object 2, being a painted wood cup with carved patterns, has a visually interesting and potentially eye-catching appearance. When thrown, it could draw the attention of pursuers due to its unique design and colors, providing a strong visual distraction that may lure them away from the person hiding. Therefore, the better choice would be to toss Object 2.



PAVC-Reasoner  
w/o audio

Object 1, being likely a small wooden stick with fuzz, tends to have better ignition capabilities compared to Object 2. The fuzz on Object 1 can act as tinder, which is highly flammable and can catch fire easily when exposed to a flame. On the other hand, Object 2, being likely a painted wood cup with carved patterns, may not be as suitable for starting a fire due to its solid structure and lack of easily ignitable components. Therefore, the better choice would be Object 1.



PAVC-Reasoner  
w/ audio (full)

Object 1 is likely a wooden stick with fluff and Object 2 appears to be an engraved ceramic cup. Object 1, being a wooden stick, is unsuitable for holding water. Even the fluff may be absorbent, it may soak up water, but the wooden stick itself may not retain water effectively. Object 2 (the ceramic cup) is more likely to be able to hold water effectively. The waterproof nature of ceramics and the cup's design make it suitable for containing liquids, including water. Therefore, the correct answer is Object 2. ✓



PAVC-Reasoner  
w/ audio (full)

Object 1 is likely a small wooden stick with fuzz on it, while Object 2 is possibly a long ceramic cup. A ceramic cup, if thrown or dropped, is likely to produce a distinct sound. The noise generated by breaking ceramic could attract attention. The wooden stick with fluff may not produce a loud noise when tossed, and the sound may not be very distinctive. Therefore, the better choice would be to toss Object 2. ✓

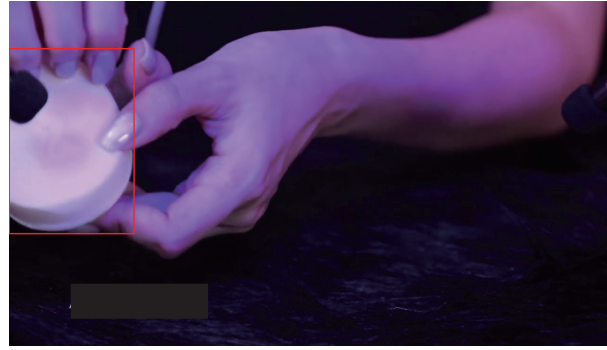
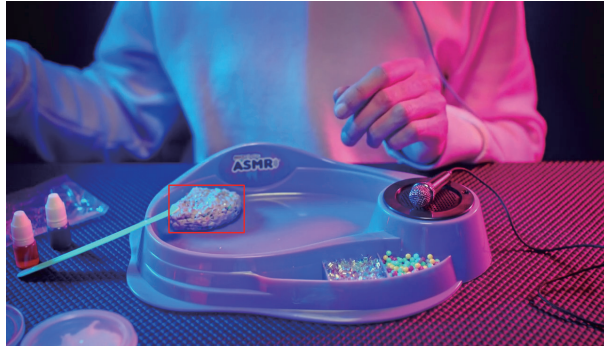


PAVC-Reasoner  
w/ audio (full)

Object 1 seems to be a wooden stick with fluff and Object 2 is possibly a long ceramic cup. The fluff on the wooden stick could serve as potential tinder, aiding in the ignition of the stick. Wood is generally flammable, and the fluff on the stick might catch fire relatively easily. Ceramic is not combustible and does not serve as fuel for a fire. The long tube shape of the ceramic cup does not lend itself well to serving as tinder or kindling. Therefore, the better choice would be Object 1. ✓

**Figure 7: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. In the leftmost scenario, Object2 is erroneously identified as a wooden cup by the model without audio, rather than a ceramic cup. Nevertheless, due to the explicit inquiry in the leftmost question, which pertains only to the physical attributes such as size, shape and prototype usage of the object, it manages to generate the correct answer. The middle question and response depict a situation where the model without audio fails to recognize that the ceramic object, when dropped, will shatter and produce a louder sound. Despite this, through an incorrect chain-of-thought (CoT), it fortuitously deduces the correct answer. In contrast, the full model gives both the correct answer and accurate CoT as the rationale. In the rightmost scenario, the question explicitly examines the material composition of the object and its non-prototype usage. While both models can predict the correct answer, the model without audio offers an incorrect CoT. Conversely, the full model simultaneously yields the correct answer and accurate CoT.





PACS-Reason



#### System Message

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.



Which object could be packed into a container smaller than its original shape?  
Options: (A) Object 1 (B) Object 2



Which object would have its shape be more easily changed by pulling it from both of its ends?  
Options: (A) Object 1 (B) Object 2



Which object would take more time to pick up if you dropped it on the ground?  
Options: (A) Object 1 (B) Object 2



PAVC-Reasoner  
w/o audio

Object 1 is likely a clay-like lump and Object 2 is likely a plastic lid-like item. Clay is known for its malleability and ability to be shaped. Object 1, being a clay-like lump, may not be easily reshaped or compressed to fit into a smaller container without losing its original form. In contrast, plastic is known for its flexibility and ability to deform without breaking. It can be easily compressed or bent to fit into tight spaces. Therefore, the correct answer is (B) Object 2. ❌



PAVC-Reasoner  
w/o audio

Object 1 appears to be a clay-like lump, which suggests that it is malleable and can be easily reshaped. When pulled from both ends, the plastic container may provide some support to the object, but the object itself is still able to deform and change shape. In contrast, Object 2 appears to be a plastic-like mouse, which is less malleable than clay. When pulled from both ends, the mouse may provide some resistance to deformation, making it more difficult to change its shape. Therefore, the correct answer is (A) Object 1. 🟢



PAVC-Reasoner  
w/o audio

Object 1 would take more time to pick up if you dropped it on the ground. This is because it is a clay-like lump, which is a small and delicate object that can easily get damaged if not handled carefully. In contrast, Object 2 is a larger and more robust object, such as a mouse, which is less likely to get damaged when dropped on the ground. Therefore, the correct answer is (A) Object 1. 🟢



PAVC-Reasoner  
w/ audio (full)

Object 1 seems to be a ball of foam and Object 2 appears to be a plastic lid based on the visual and audio cues. Foam is compressible and can be easily squished or compressed. Being a foam-like material, Object 1 can likely be compressed or deformed to fit into a smaller container. Plastic lids are generally rigid and less compressible. The lid-like shape of Object 2 may not allow for easy compression or deformation to fit into a smaller container. Therefore, the correct answer is (A) Object 1. 🟢



PAVC-Reasoner  
w/ audio (full)

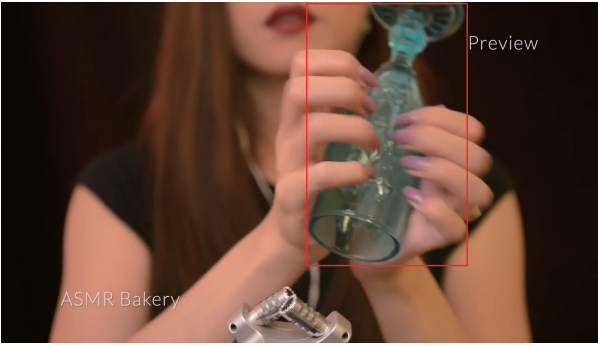
Object 1 is identified as a foam-like material and Object 2 is identified as a plastic material based on both sound and visual cues. Foam-like materials are generally more flexible and deformable compared to rigid plastics. Foam can compress and change shape more easily when force is applied. Plastic materials, especially if they are in a block or lid-like shape, tend to be more rigid and less deformable than foam. Consequently, Object 1 is more likely to have its shape easily changed when pulled from both ends. The flexible and compressible nature of foam allows it to deform more readily under applied force. Therefore, the correct answer is (A) Object 1. 🟢



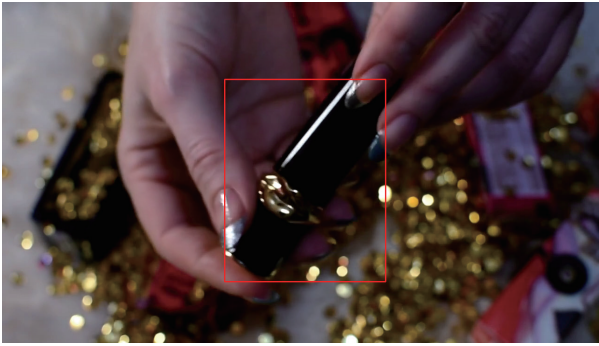
PAVC-Reasoner  
w/ audio (full)


Object 1 is a foam-like material and Object 2 appears to be a plastic block. Object 1 is smaller in size than Object 2, picking it up might be more challenging, potentially requiring more effort and time compared to the larger Object 2. Smaller objects generally experience more air resistance relative to their size. Object 1, being smaller, might experience more air resistance than Object 2. Higher air resistance tends to slow down the fall of an object. Consequently, Object 1 would likely take more time to reach the ground compared to Object 2. Therefore, the correct answer is (A) Object 1. 🟢

**Figure 8: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. The leftmost question and response illustrate a scenario where object1, being uncommon, is susceptible to being misconstrued as made of plastic or wax rather than foam. Object2 appears visually ambiguous, resembling a plastic mouse or a plastic lid. Object1 is mistakenly identified as a clay-like lump by the model without audio. Due to the leftmost question's focus on the objects' malleability and softness, the model without audio produces incorrect CoT and an erroneous answer. In contrast, the full model successfully recognizes the physical properties of both objects, providing accurate thought chains as rationales and the correct answer. The middle question and response depict a scenario where PAVC-Reasoner-Vision struggles to comprehend the flexibility or elasticity of these two objects in the absence of audio. Although both models yield correct predictions, the intermediate reasoning path generated by the model without audio is unreasonable. The rightmost question and response portray a scenario where both models successfully grasp the implicit query about the physical properties, such as shape and material, of the objects. Both models provide accurate answers.












ASMR Bakery



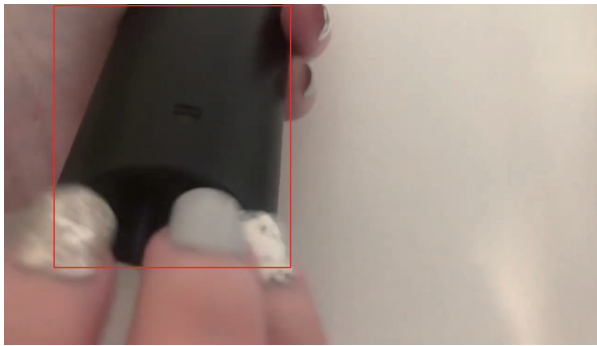
**PACS-Reason**


**System Message**

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.

 <p><b>Which object would you be able to spot in dense forest ground-covering the fastest?</b> Options: (A) Object 1 (B) Object 2</p>	 <p><b>Which object would take more time to pick up if you dropped it on the ground?</b> Options: (A) Object 1 (B) Object 2</p>	 <p><b>If both objects were filled with pocket change, which would allow you to take more friends out to eat with the money contained therein?</b> Options: (A) Object 1 (B) Object 2</p>
<p><b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 is likely a plastic cup and Object 2 is likely a metal cylindrical object. While Object 2 has a smaller visual shape compared to Object 1, it is still relatively small, which could make it difficult to spot in a dense forest ground-covering. Object 1, being visually larger due to its cup shape, would be easier to spot. Therefore, the correct answer is (A) Object 1. </p>	<p><b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 is likely a plastic cup, which is visually larger in size compared to Object 2, which may be a metal cylindrical object. Larger objects generally require less precision to grasp, making them quicker to pick up compared to smaller, more intricate objects. Object 2, being a metal cylindrical object, may have a narrower profile, potentially making it more difficult and time-consuming to pick up from the ground. Therefore, the correct answer is (B) Object 2. </p>	<p><b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 is likely a plastic cup and Object 2 is probably a small metal tube. Object 1 has a larger size visually compared to Object 2, indicating that Object 1 may be able to hold more pocket change. Object 2, due to its tubular shape, may have a larger capacity to hold pocket change compared to a glass cup. Therefore, the correct answer is (B) Object 2. </p>
<p><b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is possibly a glass cup, and Object 2 might be a metal tube. The glass cup is visually larger, which means it would likely have a higher contrast against the forest floor. Larger objects are generally easier to spot, especially if they have a distinctive shape like a cup. Object 2 is likely a smaller metal tube, and its shape is visually smaller compared to Object 1. In a dense forest, a smaller and more cylindrical object might be harder to spot among the foliage. Therefore, the correct answer is (A) Object 1. </p>	<p><b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is probably a glass cup and Object 2 is potentially a metal tube. Glass is generally more fragile and can break upon impact. And metal is typically more durable and less prone to breakage. Due to its larger size and fragility, Object 1 might experience a more significant impact upon hitting the ground, possibly leading to a more scattered arrangement of broken pieces. Object 2, being smaller and more durable, may experience less impact damage. Retrieving a shattered glass may take more time and effort due to the need to gather broken pieces carefully. Therefore, the correct answer is (A) Object 1. </p>	<p><b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is possibly a glass cup, and Object 2 is likely a metal tube. While Object 1 may be larger in size, the volume it can hold is limited by its shape. The volume of the glass cup may not hold a substantial amount of pocket change. However, when it comes to pocket change, the shape of the coins is a critical factor. The shape of coins may also affect the number of coins that can fit into the objects. Since Object 2 is a tube, it has the potential to hold a larger amount of pocket change within a smaller physical volume. Therefore, the correct answer is (B) Object 2. </p>

**Figure 9: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. The *leftmost* question and response depict a scenario where the object1 could be mistaken as plastic, and the object2 could be made of plastic or metal. Although the model without audio misidentifies the first object as plastic, it still generates the correct answer due to the explicit nature of the question regarding the size of the objects. *Middle*: The model without audio fails to recognize that the glass object will shatter and take longer to pick up off the ground. In contrast, the full model produces accurate answer and correct chain-of-thought (CoT) as rationale. The *rightmost* question and response describe a scenario where both models successfully analyze the third question, which implicitly inquires about the size and shape of the objects. However, both models generate incorrect answers and CoTs.



#### System Message

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.



Which of the objects could be manipulated to increase its length?  
Options: (A) Object 1 (B) Object 2



If you dropped your keys down a drainpipe, which item would you attach some sticky gum to and lower down to attempt a retrieval?  
Options: (A) Object 1 (B) Object 2



If you accidentally fell asleep at your desk with one of these objects under your cheek, which would be less likely to cause an impression?



PAVC-Reasoner  
w/o audio

Object 1 seems to be a plastic TV remote, and Object 2 is probably a plastic snake toy. TV remotes typically have a fixed and non-expandable structure, while wooden snake toys are designed to be manipulated, allowing users to increase their length by unfolding or extending the toy. Video 2 also shows that Object 2 can be deformed. Therefore, the correct answer is (B) Object 2.



PAVC-Reasoner  
w/o audio

Object 1 appears to be a plastic TV remote control and Object 2 may be a plastic snake. The plastic snake toy has the ability to bend, fold, extend, and retract. This means that it can be manipulated to navigate through the drainpipe and reach the keys more effectively than Object 1, the plastic TV remote, which does not have the same flexibility or maneuverability. Therefore, the correct answer is (B) Object 2.



PAVC-Reasoner  
w/o audio

Object 1 may be a plastic TV remote, and Object 2 may be a plastic snake toy. Object 1 has a flat rectangular shape. When pressure is applied to a flat surface, it is more likely to create a distinct and potentially deeper impression due to the uniformity of the surface. Object 2 is composed of many small square blocks, creating an irregular surface. When pressure is applied to such an irregular surface, it is less likely to leave a distinct and deep impression on the skin because the pressure is distributed across multiple points of contact. Therefore, the correct answer is (B). ❌



PAVC-Reasoner  
w/ audio (full)

Object 1 is likely a plastic TV remote, and Object 2 is likely a wooden snake toy. Object 1, being a plastic TV remote, is typically rigid and not designed to be easily manipulated or extended. It is unlikely that this object can be increased in length. Object 2, being a wooden snake toy, is specifically mentioned to be capable of folding, extending, and retracting. This suggests that it has a flexible design and can be manipulated to increase its length. Therefore, the correct answer is (B) Object 2. ✅



PAVC-Reasoner  
w/ audio (full)

Object 1 is likely a plastic TV remote, and Object 2 is likely a wooden snake toy. TV remotes typically have a fixed and non-expandable structure. Snake toys are often designed with joints, allowing them to be bent, folded, and extended. The flexibility and manipulability of the wooden snake toy make it a better choice for adapting to the confined space of a drainpipe and potentially reaching the keys. Therefore, the correct answer is (B) Object 2. ✅



PAVC-Reasoner  
w/ audio (full)

Object 1 might be a plastic TV remote, and Object 2 might be a wooden snake toy. According to the videos, the TV remote has a flat and elongated shape. Its surface is likely to be relatively uniform and smooth. The wooden snake toy is composed of multiple small cubes, forming a more textured and irregular surface. The flat and smooth surface of the plastic remote is less likely to leave irregular and deep marks on your skin. Therefore, the correct answer is (A) Object 1. ✅

**Figure 10: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. In this illustration, the object on the left (object1) is a plastic television remote control, while the object on the right (object2) is a wooden snake toy.






PACS-Reason++



**System Message**

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.

 <p>If both objects were submerged in water, which one do you think would float? Options: (A) Object 1 (B) Object 2</p>	 <p>If you were to use one of these objects to press a button that requires a bit of force, which one do you think would be more effective? Options: (A) Object 1 (B) Object 2</p>	 <p>In a situation where you needed to stack multiple objects on top of each other, which one would be more stable and less likely to topple over? Options: (A) Object 1 (B) Object 2</p>
 <p><b>PAVC-Reasoner w/o audio</b></p> <div style="border: 1px solid #FFDAB9; padding: 10px; margin-top: 10px;"> <p>Object 1 is likely be a plastic TV remote, and Object 2 is likely a plastic snake toy. The object with a lower volume-to-mass ratio would float, and the object with a lower volume-to-mass ratio would have a lower density. Object 1 would likely have a higher density than Object 2. This is because the TV remote is visually smaller and denser than the snake toy, which may have hollow spaces within the cubes itself. Therefore, the correct answer is (B) Object 2.</p> </div>	 <p><b>PAVC-Reasoner w/o audio</b></p> <div style="border: 1px solid #FFDAB9; padding: 10px; margin-top: 10px;"> <p>Object 1 seems to be a plastic TV remote, and Object 2 appears to be a plastic snake toy. TV remotes are typically made of sturdy, rigid plastic to withstand daily use and handling. They are designed to be able to endure the force applied when pressing buttons without deforming or breaking easily. Object 2 is made of interconnected small square blocks, which suggests that it is likely to be more flexible and prone to deformation when force is applied. Therefore, the correct answer is (B) Object 2.</p> </div>	 <p><b>PAVC-Reasoner w/o audio</b></p> <div style="border: 1px solid #FFDAB9; padding: 10px; margin-top: 10px;"> <p>Object 1 may be a plastic TV remote, and Object 2 appears to be a plastic snake toy. Object 1 has a flat and slim profile with a semi-ellipsoidal shape. The surface of Object 1 is visually smooth. Object 2 is composed of interconnected square blocks. These blocks can be folded into the shape of a cube or rectangular prism, indicating that they could interlock and form a stable structure. It can be inferred that Object 2 would be more stable and less likely to topple over when stacked. Therefore, the correct answer is (B) Object 2.</p> </div>
 <p><b>PAVC-Reasoner w/ audio (full)</b></p> <div style="border: 1px solid #FFDAB9; padding: 10px; margin-top: 10px;"> <p>Object 1 might be a plastic TV remote, and Object 2 might be a wooden snake toy. Although plastic is generally denser than water, the flat and elongated shape of Object 1 may not displace enough water to generate sufficient buoyant force. Wood is typically less dense than water. The multiple small cubes and the ability to be compressed or expanded may allow for better displacement of water. Therefore, the correct answer is (B) Object 2.</p> </div>	 <p><b>PAVC-Reasoner w/ audio (full)</b></p> <div style="border: 1px solid #FFDAB9; padding: 10px; margin-top: 10px;"> <p>Object 1 may be a plastic TV remote, and Object 2 may be a wooden snake toy. Being made up of many small square blocks hinged together, the ability of Object 2 to be compressed or expanded may make it less stable for applying direct force. Object 1, being a TV remote, the rigidity of the plastic material and the flat shape provide stability, allowing for a more controlled application of force without significant deformation. Therefore, the correct answer is (B) Object 2.</p> </div>	 <p><b>PAVC-Reasoner w/ audio (full)</b></p> <div style="border: 1px solid #FFDAB9; padding: 10px; margin-top: 10px;"> <p>Object 1 may be a plastic TV remote, and Object 2 may be a wooden snake toy. The flat and elongated shape of Object 1 may make it less stable for stacking, especially when trying to create a tall stack. The smooth surface of Object 1 largely reduce friction between stacked objects, making them more prone to slipping. Object 2, composed of multiple cubes and having the ability to be folded into different shapes, provides a more stable structure for stacking. Therefore, the correct answer is (B) Object 2.</p> </div>

**Figure 11: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason++ benchmark for diverse questions. In this illustration, the object on the left (object1) is a plastic television remote control, while the object on the right (object2) is a wooden snake toy.




PACS-Reason



**System Message**

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.

 <p>Which object could be woven into a pattern? Options: (A) Object 1 (B) Object 2</p>	 <p>Which object would be more likely to be colonized by mold if it was made wet and put into a closet without drying? Options: (A) Object 1 (B) Object 2</p>	 <p>Which object would you be able to throw farther through a wind tunnel? Options: (A) Object 1 (B) Object 2</p>
 <p><b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 may be a plastic brush, and Object 2 may be a ball of foam. The plastic brush is typically rigid and not conducive to weaving. However, foam is generally flexible and can be manipulated. This means that Object 2 can be easily woven or manipulated into patterns due to its flexibility. Therefore, the correct answer is (B) Object 2. ✓</p>	 <p><b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 may be a plastic brush, and Object 2 may be a ball of shining foam. A Plastic brush is generally non-porous and does not absorb moisture easily. It is also easy to clean and dry, making it less likely to support mold growth even if it is wet and put in a closet without drying. Object 2, as soft and stick foam, is porous and can absorb water easily, which creates a favorable environment for mold growth. it is more likely to retain moisture and promote mold growth than Object 1. Therefore, the correct answer is (B) Object 2. ✓</p>	 <p><b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 is likely a plastic brush, and Object 2 is likely a ball of shining foam. Object 1 is made of plastic, which is generally rigid and may not easily adapt to airflow. While Object 2, as a foam, is likely less dense and may have better aerodynamic properties, allowing it to travel a greater distance compared to Object 1. Therefore, the correct answer is (B) Object 2. ✓</p>
 <p><b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 seems to be a plastic brush, and Object 2 seems to be a ball of glittering soft foam. Object 1 has bristles that are arranged in a specific pattern or layout. However, the bristles of a brush are usually fixed in place and do not have the flexibility or interlocking capability required for weaving. Object 2, being a ball of foam, the softness and stickiness of the foam material allow for easy manipulation and interlocking, enabling the creation of intricate woven patterns. Therefore, the correct answer is (B) Object 2. ✓</p>	 <p><b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 appears to be a plastic brush, and Object 2 looks like a ball of soft foam. Plastic is generally non-porous and less conducive to mold growth. The smooth and rigid surface of plastic brushes provides fewer crevices for mold to establish and thrive. While Object 2, being a Soft and sticky foam, has a porous structure, providing more favorable conditions for mold growth. The soft and sticky nature of foam can retain moisture, creating an environment conducive to mold colonization. Therefore, the correct answer is (B) Object 2. ✓</p>	 <p><b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is identified as a plastic brush, and Object 2 is identified as a soft foam. Object 1 has a solid and rigid structure. When thrown through a wind tunnel, the aerodynamic properties of the brush may not be optimized for efficient airflow. Object 2, on the other hand, being a soft foam, is generally lightweight. The lightweight nature of foam allows it to be propelled more easily by the wind, potentially resulting in a longer distance traveled. Therefore, the correct answer is (B) Object 2. ✓</p>

**Figure 12: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full)), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. In this illustration, Object1 (i.e., the object on the left) is commonly recognized as a plastic brush, whereas Object2 (i.e., the object on the right) is relatively uncommon. However, through audiovisual analysis, it can be discerned that Object2 is composed of a soft and viscous foam material.




PACS-Reason++



**System Message**

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.

<p> If you were to use one of these objects as a makeshift hammer, which one would likely be more effective? Options: (A) Object 1 (B) Object 2</p> <p> <b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 looks like a plastic brush and Object 2 looks like a foam ball. Object 1 would likely be more effective as a makeshift hammer. While plastic may not be as hard as some other materials, it is generally harder than foam. Additionally, the plastic brush may offer more rigidity, allowing for better force transfer when used as a hammer. The weight of the plastic brush may also contribute to its effectiveness in delivering impact. Therefore, the correct answer is (A) Object 1. ✓</p>	<p> In a situation where you needed a material for sound insulation, which object is a better choice? Options: (A) Object 1 (B) Object 2</p> <p> <b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 seems to be a plastic brush and Object 2 seems to be a foam ball. Object 2, being a soft foam, the lower density and increased flexibility make it more effective at absorbing and dampening sound waves. It provides better sound insulation properties compared to the rigid plastic material of Object 1. Therefore, the correct answer is (B) Object 2. ✓</p>	<p> If you had to design an object that needed to absorb impact, which object would be more suitable for this purpose? Options: (A) Object 1 (B) Object 2</p> <p> <b>PAVC-Reasoner w/o audio</b></p> <p>Object 1 seems to be a plastic brush and Object 2 seems to be a foam ball. Object 2, being a soft foam, is more suitable for absorbing impact. The flexibility and compressibility of foam make it effective in absorbing and dispersing impact energy, providing better shock absorption properties compared to the rigid plastic material of Object 1. Therefore, the correct answer is (B) Object 2. ✓</p>
<p> <b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is identified as a plastic brush, and Object 2 is identified as a soft foam. Object 1 would likely be more effective as a makeshift hammer. While plastic may not be as hard as some materials used for hammers, it is likely to provide better impact resistance and weight for tasks that require some force. The foam's soft and lightweight nature might make it less efficient for hammering, as it lacks the necessary hardness and weight to transmit impactful force. Therefore, the correct answer is (A) Object 1. ✓</p>	<p> <b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is identified as a plastic brush, and Object 2 is identified as a soft foam. Object 1 is generally rigid. Its smooth surface may reflect sound rather than absorb it. While Object 2, being a soft foam, is known for its sound-absorbing qualities. The porous and soft nature of foam allows it to trap and absorb sound waves, reducing reverberation. Therefore, the correct answer is (B) Object 2. ✓</p>	<p> <b>PAVC-Reasoner w/ audio (full)</b></p> <p>Object 1 is likely as a plastic brush, and Object 2 is probably a ball of soft foam. Object 1 is generally rigid, and is less effective at absorbing and dispersing impact energy, as it tends to transmit the force rather than absorb it. While Object 2, as a soft foam, is well-suited for absorbing impact. Its flexibility allows it to compress and deform, dissipating impact energy and reducing the force transmitted. Therefore, the correct answer is (B) Object 2. ✓</p>

**Figure 13: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 (full), presenting responses (i.e., rationales) generated on the PACS-Reason++ benchmark for diverse questions. In this illustration, Object 1 (i.e., the object on the left) is commonly recognized as a plastic brush, whereas Object 2 (i.e., the object on the right) is relatively uncommon. However, through audiovisual analysis, it can be discerned that Object 2 is composed of a soft and viscous foam material.





**System Message**

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.

**Question 1:** Which object is more likely to survive being tapped with a hammer?  
Options: (A) Object 1 (B) Object 2

**PAVC-Reasoner w/o audio**

Object 1 looks like a plastic card, and Object 2 looks like a metal bell. Object 2, a metal copper bell, is more likely to survive being tapped with a hammer. Metals, including copper, are known for their durability and ability to absorb and distribute impact energy without breaking or shattering, unlike the potentially brittle ceramic material of Object 1. Therefore, the correct answer is (B) Object 2.

**Question 2:** Which object would hurt more if you stepped on it?  
Options: (A) Object 1 (B) Object 2

**PAVC-Reasoner w/o audio**

Object 1 appears to be a plastic card, and Object 2 appears to be a metal bell. Object 1, possibly a plastic item, would likely hurt more if stepped on. From videos, Object 1 has some hardness and edges that can cause discomfort. Object 2, being metal and smoother, is less likely to have sharp edges and would be more forgiving underfoot. Therefore, the correct answer is (A) Object 1.

**Question 3:** Which object would be worse at rolling cookie dough?  
Options: (A) Object 1 (B) Object 2

**PAVC-Reasoner w/o audio**

Object 1 appears to be a small plastic card, and Object 2 appears to be a metal bell. Object 1 is visually rigid and thin. It could be easily break into pieces if used to roll dough. Object 2 has a smoother metal surface compared to Object 1. Therefore, the correct answer is (A) Object 1.

**Question 4:** Which object is more likely to be a ceramic small decoration, while Object 2 seems to be a metal copper bell. Object 1, being made up of ceramic, is generally brittle and can be prone to cracking or breaking under force. Object 2, being a durable and resilient material, is better suited to withstand impact compared to ceramic, which is more prone to breakage. Therefore, the correct answer is (B) Object 2.

**PAVC-Reasoner w/ audio (full)**

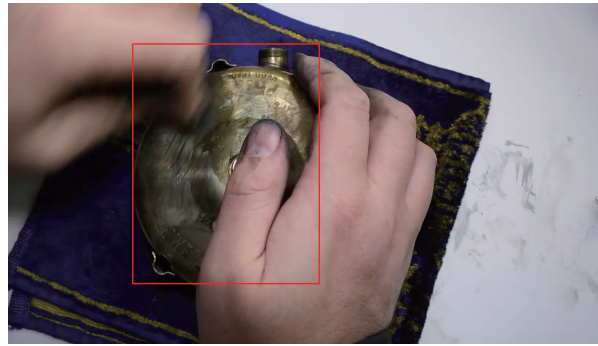
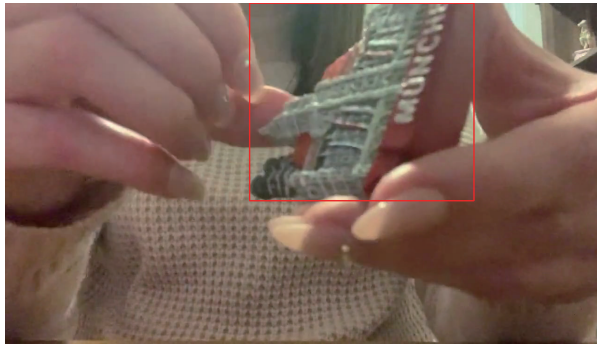
**Question 5:** Object 1 is possibly a ceramic small building souvenir, while Object 2 is possibly a metal copper bell. Object 1, being made up of ceramic, is generally fragile and prone to break into sharp pieces when stepping on it. While Object 2, being a metal copper bell, has a smoother surface and is less prone to breaking into sharp pieces compared to Object 1. Therefore, the correct answer is (A) Object 1.

**PAVC-Reasoner w/ audio (full)**

**Question 6:** Object 1 is likely a slim ceramic building souvenir, while Object 2 is likely a metal copper bell. Ceramic is generally a rigid and inflexible material, making it difficult to roll cookie dough evenly. Additionally, the small size and pointy shape of Object 1 may make it less effective at smoothing out the dough and achieving a uniform thickness. Object 2 has a smooth surface and is harder than Object 1. Therefore, the correct answer is (A) Object 1.

**PAVC-Reasoner w/ audio (full)**

**Figure 14: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. In this illustration, Object1 (i.e., the object on the left) poses a visual challenge in determining its material composition, as it may be misconstrued as a plastic product rather than crafted from ceramic. Object2 (i.e., the object on the right) is composed of metal. The absence of crucial audio cues could potentially lead the model without audio to misconstrue the material composition of these objects, consequently misinterpreting their strength.



PACS-Reason+



#### System Message

You are a PAVC-Reasoner, a large physical audiovisual commonsense reasoning model trained with video-instruction data. Please answer the question directly and provide a detailed explanation step-by-step or reasonable intermediate reasoning steps for inferring the answer.



Which object would be more prone to breaking if accidentally dropped on a hard surface?  
Options: (A) Object 1 (B) Object 2



Which object is more likely to produce a clear ringing sound when struck with a mallet?  
Options: (A) Object 1 (B) Object 2



If you needed a paperweight on a windy day, which object would be more effective at keeping papers in place?  
Options: (A) Object 1 (B) Object 2



PAVC-Reasoner  
w/o audio

Object 1 looks like a small plastic item, while Object 2 looks like a metal bell. Object 1, a small plastic object, is more fragile and susceptible to breaking when dropped on a hard surface, while Object 2, a metal copper bell, would be more durable and less likely to break in such a scenario. Therefore, the correct answer is (A) Object 1.



PAVC-Reasoner  
w/o audio

Object 1 is likely a slim plastic building souvenir, while Object 2 is likely a metal bell. Plastic is generally not known for producing clear and resonant sounds when struck. It tends to have a duller and less vibrant acoustic quality compared to materials like metal. Additionally, Object 1 is like a small and solid plastic item, it may lack the hollow space or resonance necessary for generating a distinct ringing sound. Therefore, the correct answer is (B) Object 2.



PAVC-Reasoner  
w/o audio

Object 1 resembles an elongated plastic building souvenir, while Object 2 resembles a metal bell. Object 1, being a small plastic item, is visually lightweight and may not provide enough weight to hold down papers effectively, especially in windy conditions. Object 2, the metal copper bell, would be expected to provide better weight and stability. Therefore, the more effective answer is (B) Object 2.



PAVC-Reasoner  
w/ audio (full)

Object 1 is likely a slim ceramic building souvenir, while Object 2 is possibly a metal bell. Ceramic is generally a less durable material compared to metal. It is more prone to cracking or shattering when subjected to impact or pressure. Object 1 would be more prone to breaking if accidentally dropped on a hard surface. The brittleness of ceramic, combined with its sharp edges, makes it more susceptible to shattering upon impact compared to the more durable metal surface of Object 2. Therefore, the correct answer is (A) Object 1.



PAVC-Reasoner  
w/ audio (full)

Object 1 is identified as a thin ceramic building souvenir, while Object 2 is identified as a metal copper bell. Both Object 1 and Object 2 are capable of producing clear ringing sounds when struck. However, metal, especially copper bells, is renowned for its distinct and resonant tones when struck with a mallet. The specific design and material properties of a metal bell often contribute to a more pronounced and sustained ringing sound. Therefore, the correct answer is (B) Object 2.



PAVC-Reasoner  
w/ audio (full)

Object 1 is identified as a thin and small ceramic souvenir, while Object 2 is identified as a metal bell. Object 2, possibly a metal copper bell, is more likely to be effective as a paperweight on a windy day. Metal, being heavier and more stable, is generally better suited for this purpose than Object 1, a ceramic item, which can be more fragile and may have less mass and friction. Therefore, the more effective answer is (B) Object 2.

**Figure 15: Demonstration of predictions by PAVC-Reasoner.** We showcase two model variants, namely PAVC-Reasoner-Vision (without audio) and PAVC-Reasoner-V4 ((full), presenting responses (i.e., rationales) generated on the PACS-Reason+ benchmark for diverse questions. In this illustration, Object1 (i.e., the object on the left) poses a visual challenge in determining its material composition, as it may be misconstrued as a plastic product rather than crafted from ceramic. Object2 (i.e., the object on the right) is composed of metal. The absence of crucial audio cues could potentially lead the model without audio to misconstrue the material composition of these objects, consequently misinterpreting their strength.

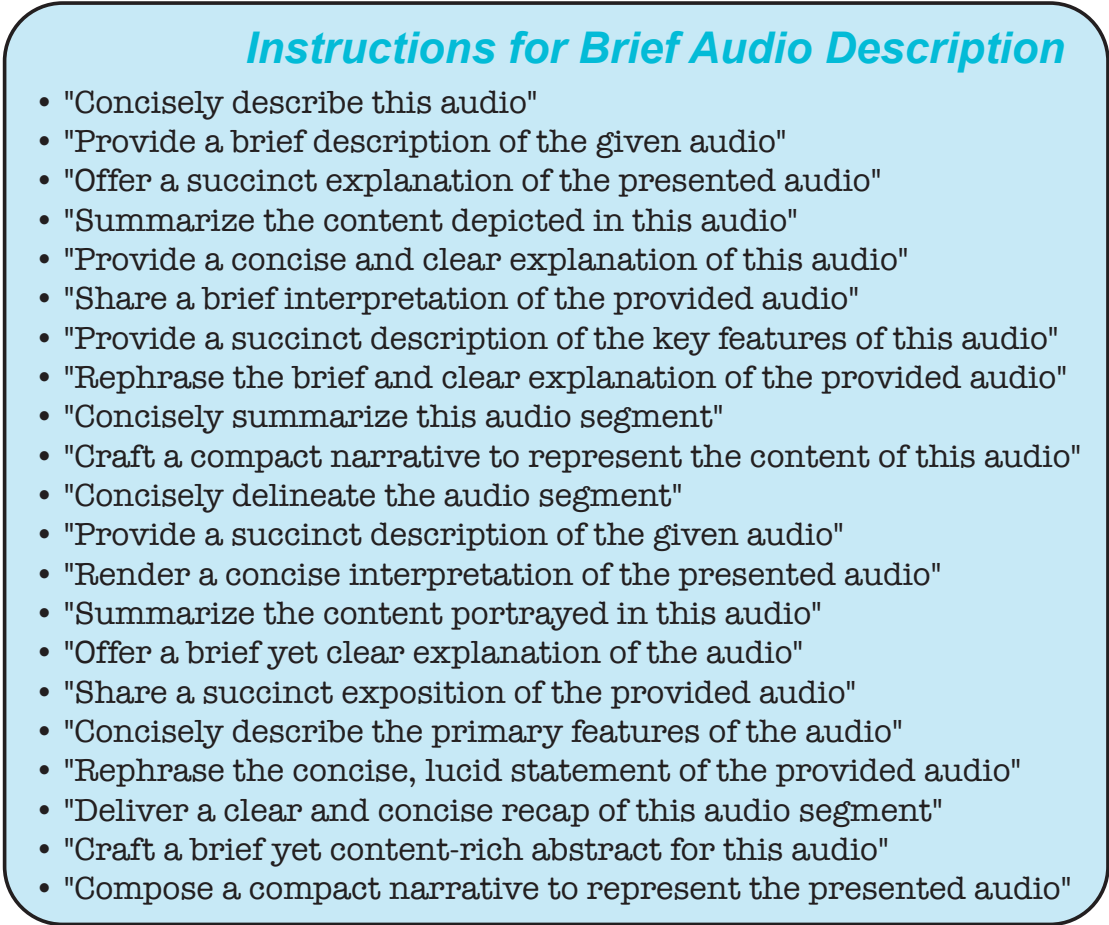


Figure 16: The list of instructions for brief audio description.




## Instructions for Detailed Video Description

- "Provide a comprehensive description of the video segment"
- "Furnish a detailed account of the given video"
- "Elaborate on the aspects of the video you have observed"
- "Offer a comprehensive overview of the presented video"
- "Provide an in-depth analysis of various facets of the displayed video"
- "Thoroughly expound the content of the showcased video"
- "Characterize the video using intricate descriptions"
- "Deconstruct the elements of the video in a certain manner"
- "Delve into the essential details of the imagery"
- "Portray the video in a rich, descriptive narrative"
- "Precisely articulate the content of the video"
- "Conduct a comprehensive and detailed analysis of the video"
- "Illustrate the video through descriptive explanations"
- "Examine the video closely and share its intricacies"
- "Offer a meticulous description of the given video"
- "Carefully examine the image and share its intricacies"
- "Conduct a comprehensive description of the given image"
- "Elaborate on the specifics of this video segment"
- "Provide a comprehensive description of the given video"
- "Offer a detailed explanation of the visual content observed"
- "Present a thorough overview of the video being showcased"
- "Conduct a comprehensive analysis of the displayed video"
- "Provide an intricate elucidation of various facets of the video"
- "Clearly expound upon the content presented in the video"
- "Characterize the video using detailed descriptions"
- "Deconstruct the elements of the video in a systematic manner"
- "Articulate the crucial details of the imagery with precision"
- "Portray the video in a rich, descriptive narrative"
- "Narrate the content of the video with accuracy"
- "Analyze the video thoroughly and in detail"
- "Illuminate the video through descriptive exposition"
- "Scrutinize the video carefully and share its intricate details"
- "Provide an exhaustive description of the given video"

Figure 17: The prompt for generating *detailed video description* instruction-following data.

### Instructions for Complex Reasoning

`messages` = [ {"role": "system", "content": f"" You are an AI vision assistant with the capability to analyze videos and an expert in physical common-sense reasoning, you will receive the following components:

1. Video1 Description : This provides a detailed natural language description of Video 1.

2. Video2 Description : This offers an elaborate natural language description of Video 2.

3. A set of physical common-sense reasoning questions related to the objects in the aforementioned videos. Each question is accompanied by a correct answer and follows the template structure as shown below:

```
"object1_object2": {
  "question1": {
    "label": 1,
    "text": ""
  },
  "question2": {
    "label": 0,
    "text": ""
  },
  ... }
```

Please note that "object1" exists solely in Video 1, while "object2" is exclusively present in Video 2. Your task now is to craft a detailed and step-by-step rationale for each physical common-sense reasoning question. This rationale should include, but is not limited to:

(a) Identifying the basis for determining the categorization of object1 and object2 from Video 1 and Video 2—whether the identification of object1 and object2 can be solely visual or requires the incorporation of auditory signals for accurate determination.

(b) Identifying which specific physical property the current question is testing for, such as **Texture**, **Flexibility**, **Shape**, **Hardness**, **Weight**, **Size**, **Strength**, **Porosity**, and so on, for both object1 and object2. Include relevant physical common-sense knowledge about these objects, covering their physical characteristics, availability, everyday functions, how they are operated, and how they interact with other objects.

The rationale you design should be as concise, succinct, and focused as possible, highlighting the core points. Your responses should be presented in a professional tone befitting an expert in common-sense reasoning. The final output should be structured in JSON format, resembling the example provided below:

```
"object1_object2": {
  "question1": {
    "label": 1,
    "text": "",
    "rationale": ""
  },
  "question2": {
    "label": 0,
    "text": "",
    "rationale": ""
  },
  ... }
```

Please proceed to provide your responses accordingly. "" ]

for sample in fewshot\_samples:

```
    messages.append({"role": "user", "content": sample['context']})
```

```
    messages.append({"role": "assistant", "content": sample['response']})
```

```
messages.append({"role": "user", "content": '\n'.join(query)})
```

Figure 18: The prompt for generating complex reasoning instruction-following data.

### Instructions for Detailed Video Description

`messages` = [ {"role": "system", "content": f"" You serve as an AI visual assistant with the capability to analyze video content across various categories:

- **Action Description:** This category provides a description of the action categories observed in the video, including object descriptions associated with the actions.
- **Object Class:** It pertains to specific objects within the video with which humans interact. Descriptions include the object category, physical properties, and their usability, focusing on objects of primary interest in the video. Attributes may encompass shape, mass, material, and usability or ease of operation.
- **Frame Captions of Video:** These frame captions segment the video into multiple frames, each frame accompanied by one or more sentences describing the visual content.
- **Object with Location:** This category specifies the precise position within the intermediate frames of specific objects of interest in the video. Descriptions include the material composition, category, and detailed coordinates of the object within the image, represented by a bounding box format: (x1, y1, x2, y2), corresponding to the upper-left x, upper-left y, lower-right x, and lower-right y.
- **Brief Video Description:** This description provides a concise overview of the entire video scene, encompassing object categories, material composition, physical attributes, usability, object positions, and interactions with humans.

Utilizing these five categories of information, you are tasked with providing a comprehensive depiction of the video scene. This should include detailed information about object categories, material composition, physical attributes, usability, object positions, and specifics regarding human interaction. Please present the narrative in natural language, elucidating the video scene without directly mentioning bounding box coordinates or video timestamps. Your responses should read as if directly describing the video content observed by a person, refraining from explicitly citing the sources of information from the provided categories. "" ]

for `sample` in `fewshot_samples`:


```
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": '\n'.join(query)})
```


Figure 19: The prompt for generating *video conversation* instruction-following data.



## Instructions for Extending Question-Answer-Rationale

`messages = [ {"role": "system", "content": f"" Suppose you operate as an AI video analysis assistant with expertise in physical common-sense reasoning. You will receive the following components:`

1. Video1 Description : A comprehensive natural language depiction of Video 1.

2. Video2 Description : A detailed natural language description of Video 2.

3. Object1 Description: Elaboration of the category of Object1 in Video 1.

4. Object2 Description: Description of the category of Object2 in Video 2.

5. A set of physical common-sense reasoning questions associated with Object1 and Object2 in the videos. Each question has a correct answer and follows a template structure, as indicated below:

```
"object1_object2": {
  "question1": {
    "label": 1,
    "text": ""
  },
  "question2": {
    "label": 0,
    "text": ""
  },
  ...
}
```

Note that Object1 is present solely in Video 1, and Object2 exclusively exists in Video 2.

Now, your task is to design ten new physical common-sense reasoning questions in a similar style to the provided questions. Additionally, provide the correct answer and a detailed rationale, elucidating the specific reasoning pathway leading to the answer.

Your design principles should encompass the following:

(a) Intuitive physics and functional world models: Inspired by previous psychological and AI experiments related to common-sense knowledge, these questions focus on predicting the outcome of single or multiple object interactions. Simple questions involve a single object and action, such as "Which object will break if dropped?" (A vase, a paper ball). More complex questions include multiple objects or actions, considering interactions between two objects, for instance, "If I place another object over which object will deform?" (A vase, a paper ball).

(b) Common real-world knowledge: Based on prior common-sense datasets, these questions inquire about the possible uses of objects in real-life scenarios. The scenarios emphasize atypical uses of objects, reducing the possibility of misusing learned knowledge, like "Which object is more suitable for cleaning water stains?" (An old T-shirt, a plastic box). More challenging questions may introduce complex or uncommon scenarios involving multiple objects, such as, "If I stack two objects together, logically, which one will be at the bottom?" (An old T-shirt, a plastic box).

Subsequently, for each new physical common-sense reasoning question you design, provide an answer (0 for object1, 1 for object2), and a detailed rationale. The rationale should include but is not limited to:

(a) Identifying the basis for categorizing object1 and object2 from Videos 1 and 2 (whether visual cues alone or a combination of audio and visual inputs are necessary for accurate identification).

(b) Highlighting the specific physical property tested in the current question (e.g., Texture, Flexibility, Shape, Hardness, Weight, Size, Strength, Porosity, etc.). Additionally, include common-sense knowledge about the material composition and the physical characteristics, functionalities, operations, and interactions of these objects with others.

The rationales you create should be concise, succinct, and focused on core points. Your responses should mirror the tone of a professional expert in common-sense reasoning. The final output should be structured in JSON format, similar to the example below:

```
"object1_object2": {
  "question_new_1": {
    "label": "",
    "text": "",
    "rationale": ""
  },
  "question_new_2": {
    "label": "",
    "text": "",
    "rationale": ""
  },
  ...
}
```

Please proceed to provide your responses accordingly. "" ]

for sample in fewshot\_samples:

```
    messages.append({"role": "user", "content": sample["context"]})
    messages.append({"role": "assistant", "content": sample["response"]})
    messages.append({"role": "user", "content": "\n".join(query)})
```

Figure 20: The prompt for generating instruction-following data for PACS-Reason++.

## REFERENCES

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*. 23716–23736.
- [3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).
- [4] Max Bain, Arsha Nagraani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*. 1728–1738.
- [5] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about physical commonsense in natural language. In *AAAI*. 7432–7439.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*. 1877–1901.
- [7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vg-sound: A large-scale audio-visual dataset. In *ICASSP*. 721–725.
- [8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*. 104–120.
- [9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/>
- [10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instruct-BLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*. 49250–49267.
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. 248–255.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
- [14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*. 19358–19369.
- [15] Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899* (2019).
- [16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiaowu Zheng, et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394* (2023).
- [17] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyue Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).
- [18] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*. 15180–15190.
- [19] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).
- [20] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Interspeech*. 571–575.
- [21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianiilos, Moritz Mueller-Freitag, et al. 2017. The "something something" video database for learning and evaluating visual common sense. In *ICCV*. 5842–5850.
- [22] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP*. 976–980.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
- [24] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *ICLR*.
- [25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-rank adaptation of large language models. In *ICLR*.
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. 19730–19742.
- [27] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).
- [28] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2023. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *arXiv preprint arXiv:2311.17005* (2023).
- [29] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. 740–755.
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).
- [32] Haotian Liu, Chunyuan Li, Qingyuan Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *NeurIPS*. 34892–34916.
- [33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281* (2023).
- [34] Ilya Loshchilov and Frank Hutter. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*.
- [35] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).
- [37] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *ACM MM*. 1015–1018.
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.
- [39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* 21, 1 (2020), 5485–5551.
- [40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*. 2556–2565.
- [41] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355* (2023).
- [42] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*. 247–263.
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [44] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2021. TDN: Temporal difference networks for efficient action recognition. In *CVPR*. 1895–1904.
- [45] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191* (2022).
- [46] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. 2023. Detecting everything in the open world: Towards universal object detection. In *CVPR*. 11433–11443.
- [47] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*. 124–141.
- [48] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).
- [49] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257* (2023).
- [50] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. PACS: A dataset for physical audiovisual commonSense reasoning. In *ECCV*. 292–309.

- [51] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*. 6720–6731.
- [52] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohamadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*. 16375–16387.
- [53] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *NeurIPS*. 23634–23651.
- [54] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* (2023).
- [55] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199* (2023).
- [56] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*.
- [57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*.