Contra4: Evaluating Contrastive Cross-Modal Reasoning in Audio, Video, Image, and 3D

Artemis Panagopoulou[∇*] Le Xue[□] Honglu Zhou[□]
Silvio Savarese[□] Ran Xu[□] Ciaming Xiong[□]
Chris Callison-Burch[∇] Mark Yatskar[∇] Juan Carlos Niebles[□]
□ Salesforce AI Reseach ∇ University of Pennsylvania

Project Page: https://artemisp.github.io/contra4-web

Abstract

Real-world decision-making often begins with identifying which modality contains the most relevant information for a given query. While recent multimodal models have made impressive progress in processing diverse inputs, it remains unclear whether they can reason *contrastively* across multiple modalities to select the one that best satisfies a natural language prompt. We argue this capability is foundational, especially in retrieval-augmented and decision-time contexts, where systems must evaluate multiple signals and identify which one conveys the relevant information. To evaluate this skill, we introduce **Contra4**, a dataset for contrastive cross-modal reasoning across four modalities: image, audio, video, and 3D. Each example presents a natural language question alongside multiple candidate modality instances, and the model must select the one that semantically aligns with the prompt. Contra4 combines human-annotated captions with a mixture-of-models round-trip-consistency filter to ensure high-quality supervision, resulting in 174k training examples and a manually verified test set of 2.3k samples. While task-specific fine-tuning helps improve performance by 56% *relative* to baseline, state-of-the-art models still achieve only an *absolute* of 56% accuracy overall and 42% in four-modality settings, underscoring a significant limitation in current multimodal models.

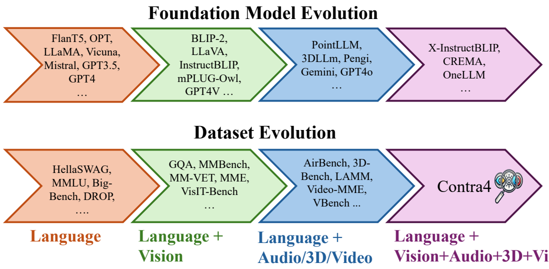

Foundation Model Evolution
FlanT5, OPT, LLaMA, Vicuna, Mistral, GPT3.5, GPT4 …
BLIP-2, LLaVA, InstructBLIP, mPLUG-Owl, GPT4V …
PointLLM, 3DLLm, Pengi, Gemini, GPT4o …
X-InstructBLIP, CREMA, OneLLM …
Dataset Evolution
HellaSWAG, MMLU, Big-Bench, DROP, ….
GQA, MMBench, MM-VET, MME, VisIT-Bench …
AirBench, 3D-Bench, LAMM, Video-MME, VBench ...
Contra4
Language
Language + Vision
Language + Audio/3D/Video
Language + Vision+Audio+3D+Video

Figure 1: Evolution of Foundation Models and Benchmarks. Contra4 evaluates models on multiple modalities concurrently (image, video, audio, 3D, and language).

1 Introduction

In many real-world situations, systems are presented with multiple streams of sensory data, images, audio recordings, video clips, or 3D scans, and must determine which one best answers a specific question. For example, a security monitoring system might receive a video feed, a 3D scan, and ambient audio and be asked: "Which modality suggests someone is approaching stealthily?". A clinician reviewing diagnostic inputs might ask: "Which of these captures evidence of respiratory distress?", where the answer could lie in an X-ray image, a stethoscope audio clip, or a patient video. In design review, a team might ask: "Which representation best conveys the product's motion?", comparing a static sketch, a 3D render, and an animation. These tasks do not require combining signals, but rather understanding each modality in isolation and selecting the one that aligns best with a given prompt. This process, *cross-modal semantic comparison*, is foundational to many multimodal applications, yet underexplored in current benchmarks.

Recent advancements, such as OpenAI's GPT-4o[1] and Google's Gemini (Team et al., 2023), highlight the growing emphasis on models capable of comprehensively integrating diverse modalities, mirroring the multi-sensory nature of human perception. While these models promise broad multimodal capabilities, currently there is limited access to them: OpenAI's API currently offers limited access beyond image processing, and Gemini only allows for audio, video, and image inputs. Nevertheless, developing robust benchmarks remains essential to assess their performance across modalities as these features become widely available.

Despite the growing interest in cross-modal mod-

[*] Work done during internship at Salesforce.

[1] https://openai.com/index/hello-gpt-4o/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25351–25365
November 4-9, 2025 ©2025 Association for Computational Linguistics

els,[2] there remains a significant gap in the benchmarks available to evaluate their proficiency in handling inputs across multiple modalities simultaneously. Table 3 in the Appendix provides an overview of the major multimodal benchmarks, underscoring this deficiency. The DisCRn benchmark (Panagopoulou et al., 2023) stands out as the only dataset that integrates inputs from all four modalities. However, it lacks probing for examples that contain more than two modalities.

To address this gap, we introduce **Contra4**, a benchmark for *contrastive cross-modal reasoning* across up to four modalities: image, audio, video, and 3D. Each sample presents a natural language question alongside multiple candidate modality inputs, only one of which semantically satisfies the prompt. The dataset is intentionally designed to *evaluate a critical precursor to information integration across modalities, the ability to determine which modality is most relevant to a given query.*

Beyond increasing the number of concurrent modalities, Contra4 includes a training set and a human-annotated test set to evaluate model performance both out-of-the-box and after fine-tuning. We implement two negative sampling strategies, high-similarity and random, to test model robustness, and apply a mixture-of-models round-trip-consistency filtering step with option permutation to ensure data quality.

In summary, our contributions are the following:
(i) We introduce Contra4, a dataset requiring reasoning on up to four modalities simultaneously.
(ii) We leverage captions and a mixture-of-models round-trip-consistency strategy (MoM-RTC) for multiple-modality data generation.
(iii) We benchmark cross-modal models and show the task's difficulty, even under fine-tuning setups.

2 Related Work

Advancements in vision-language tasks have paved the way for models capable of reasoning across multiple non-linguistic inputs, such as multiple images (Bansal et al., 2020; Li et al., 2022b; Tanaka et al., 2023; Wang et al., 2024c) or cross-modal reasoning involving images and tables (Li et al., 2022b). Despite their complexity, these tasks predominantly focus on image-text modalities. While cross-modal benchmarks exist, primarily evaluating models on joint audio-video reasoning (Alamri et al., 2018; Li et al., 2022a), there remains a gap in assessing models' capabilities for comparative cross-modal reasoning. Even comprehensive multimodal benchmarks like MultiBench (Liang et al., 2021) and OmniXR (Chen et al., 2025) primarily operate with single-modality inputs or, at most, video with corresponding audio. The medical domain follows a similar trend; for instance, M3 (Huang et al., 2021) evaluates models using only corresponding X-ray images, audio, and textual input. To address this gap, we introduce `Contra4`, a dataset to evaluate contrastive reasoning by differentiating between cross-modal inputs.

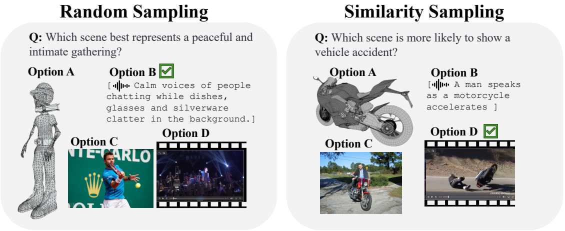

Random Sampling
Q: Which scene best represents a peaceful and intimate gathering?
Option A
Option B ☑
[Calm voices of people chatting while dishes, glasses and silverware clatter in the background.]
Option C
Option D
Similarity Sampling
Q: Which scene is more likely to show a vehicle accident?
Option A
Option B
[A man speaks as a motorcycle accelerates]
Option C
Option D ☑

Figure 2: Examples from Contra4. Additional examples are found in Figure J in the Appendix.

The rise of high-performing LLMs has enabled automated data annotation, with most datasets relying on huge proprietary models like GPT-4 (OpenAI, 2023). Initially used for text (Dai et al., 2023; He et al., 2023), these methods now extend to images (Changpinyo et al., 2021; Bitton et al., 2024; Xue et al., 2024), video (Muhammad Maaz and Khan, 2023), audio (XinhaoMei, 2023; Yang et al., 2024), and 3D (Wu et al., 2015; Zhang et al., 2024), leveraging multimodal models for alignment and LLMs for annotation. Our work differs by focusing on synthetic datasets for tasks involving 3+ modalities, where we show cross-modal reasoning remains a challenge.

3 Contra4: Task Definition

Let $\mathbf{x} = \{x_M^i\}_{i=1}^N$ be a set of N multimodal inputs, where each x_M^i is drawn from a specific modality M and is paired with a text query q, as shown in Figure 9. The function $T(\cdot)$ is used for tokenizing and embedding any textual elements, while $P_M(\cdot)$ projects an input from the modality M into the model's linguistic embedding space. In addition, each input x_M^i, encoded with a modality-specific encoder $Enc(\cdot)$, has an associated enumeration prefix E_i. To form the final input to the MLLM, we concatenate the tokenized prefix $T(E_i)$ with the projected multimodal representation $P_M(Enc(x_M^i))$ for all $i = 1, \ldots, N$, and then

[2] Cross-modal models involve 3+ modalities (Panagopoulou et al., 2023).

further concatenate the tokenized query $T(q)$. Symbolically, this can be written as: $MLLM(\mathbf{x}, q) = MLLM\Big(\bigoplus_{i=1}^{N} [T(E_i) \oplus P_M(Enc(x_M^i))] \oplus T(q)\Big)$, where $\oplus$ denotes the concatenation operation in the embedding space. The model's task is to correctly identify which enumeration prefix E_i corresponds to the correct answer for the query q.

4 Dataset

Data generation: Our method leverages textual descriptions as a *universal connector* across modalities to build a dataset that enables querying across diverse modalities *without* requiring an additional multimodal linking model. Figure 3 illustrates our process: given a set of single-modality M datasets with associated captions, $D_M = \{(x_M, c_M)\}$, we apply a four-stage data augmentation method to generate contrastive cross-modal reasoning data.

Step 1. Negative Sampling Selection: We employ two negative selection strategies: *high [caption] similarity* and *random* to enhance the evaluation potential of the dataset. This process results in tuples of two, three, and four modalities denoted as $D_{\hat{M}}$, where $\hat{M}$ denotes the subselected modalities.

Step 2. Question Generation: After generating tuples in Step 1, we use an LLM with four in-context examples to generate a contrastive question about the multimodal inputs. Questions focusing on textual qualities of the captions are filtered out, ensuring relevance to the multimodal scene depiction.

Step 3. Answer-Explanation Generation: Conditioned on the captions in the original dataset and the questions refined in Step 2, we prompt the same LLM to answer and explain its reasoning.

Step 4. Mixture-of-Models Round-Trip-Consistency (MoM-RTC): We validate dataset quality by running a round-trip-consistency check on an ensemble of distinct models, prompting each LLM to answer and explain the contrastive questions based on their captions. We keep only samples that pass certain filtering criteria, **Majority Filter (MF)**, **Unanimous Filter (UF)**, **Permute Majority Filter (PMF)**, and **Permute Unanimous Filter (PUF)**, which we compare in Table 1. In particular, MF requires that a majority of models agree with the original answer; UF requires unanimous agreement; PMF extends MF and PUF extends UF under all permutations of the cross-modal options. Pseudocode for the procedure is presented in Algorithm 1 in the Appendix for clarity.

Dataset Statistics: Using the above pipeline, we produce 174k automatically annotated samples for training and release a test set of 2.3k manually-annotated examples. Answer distribution is balanced post-hoc. See details in Appendix F.

Filter	Human Acc.	N/A	O/A	GPU (hrs)	Aggregated		
					Rand	Sim	All
None*	46.7	18.3	18.3	0	254k	261k	515k
MF	60.0	18.3	17.5	40	190k	188k	378k
UF	60.0	16.7	13.3		130k	126k	256k
PMF	68.3	13.3	15.0	120	147k	131k	278k
PUF	**83.3**	**6.7**	**5.8**		91k	83k	174k

Table 1: Human inspection of Round-Trip-Consistency checks on training data. N/A is the fraction of questions not applicable to any choice and O/A to more than one choice.
*Some rule-based word filtering is applied; see Appendix C.

5 Experiments

Implementation Details: For Step 2 and Step 3 we employ LLaMA-3.1-8B-Instruct (Dubey et al., 2024) . For Step 4 we also use mistralai/Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and microsoft/Phi-3-medium-128k-instruct (Abdin et al., 2024). For the permutation checks we consider all possible permutations of the answer choices. For the text similarity we use all-MiniLM-L6-v2 encodings via `sentence-transformers`. Single run accuracy is reported. The datasets used to generate `Contra4` are summarized in Table 4 in the Appendix with additional implementation details in Appendix G.

Models: To assess task difficulty and position this dataset as a community challenge, we evaluate several state-of-the-art (SOTA) models capable of handling all four modalities. Two models, **X-InstructBLIP** (Panagopoulou et al., 2023) and **CREMA** (Yu et al., 2024), use a frozen LLM with separate modality encoders. They differ in that CREMA uses a fused Q-Former for modality alignment, requiring additional RGB input for 3D, whereas X-InstructBLIP maintains separate Q-Formers. We also evaluate **OneLLM** (Han et al., 2023), which unifies modalities into a common space, connecting a fused modality encoder to the LLM, and trains the entire architecture, including the LLM. We also report performance of OneLLM finetuned on subsets sampled from each filtering pool in Step 4. We chose OneLLM for fine-tuning because it is the top-performing model that natively processes 3D point clouds, allowing us to measure the direct impact of finetuning on the core cross-modal reasoning task without confounding factors from input signal conversion. We further baseline Gemini using `gemini-2.0-flash-exp` on exam-

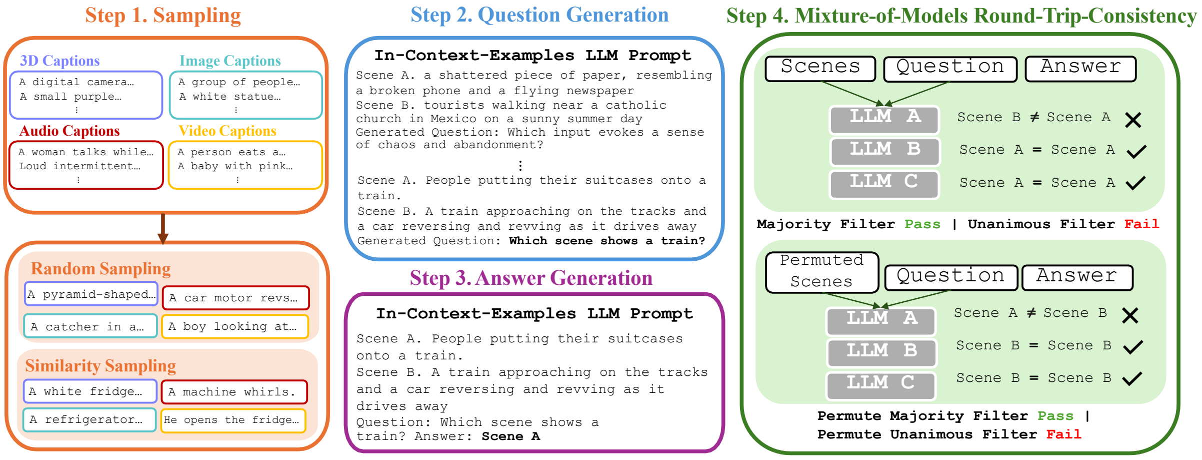

Step 1. Sampling
3D Captions
A digital camera…
A small purple…
Image Captions
A group of people…
A white statue…
Audio Captions
A woman talks while…
Loud intermittent…
Video Captions
A person eats a…
A baby with pink…
Random Sampling
A pyramid-shaped…
A car motor revs…
A catcher in a…
A boy looking at…
Similarity Sampling
A white fridge…
A machine whirls.
A refrigerator…
He opens the fridge…
Step 2. Question Generation
In-Context-Examples LLM Prompt
Scene A. a shattered piece of paper, resembling a broken phone and a flying newspaper
Scene B. tourists walking near a catholic church in Mexico on a sunny summer day
Generated Question: Which input evokes a sense of chaos and abandonment?
Scene A. People putting their suitcases onto a train.
Scene B. A train approaching on the tracks and a car reversing and revving as it drives away
Generated Question: Which scene shows a train?
Step 3. Answer Generation
In-Context-Examples LLM Prompt
Scene A. People putting their suitcases onto a train.
Scene B. A train approaching on the tracks and a car reversing and revving as it drives away
Question: Which scene shows a train? Answer: Scene A
Step 4. Mixture-of-Models Round-Trip-Consistency
Scenes
Question
Answer
LLM A
LLM B
LLM C
Scene B ≠ Scene A ✗
Scene A = Scene A ✓
Scene A = Scene A ✓
Majority Filter Pass | Unanimous Filter Fail
Permuted Scenes
Question
Answer
LLM A
LLM B
LLM C
Scene A ≠ Scene B ✗
Scene B = Scene B ✓
Scene B = Scene B ✓
Permute Majority Filter Pass |
Permute Unanimous Filter Fail

Figure 3: Data Generation Pipeline. In Step 1, candidate choices are sampled either randomly or by selecting those with high text similarity. Step 2 employs in-context learning to generate a question based on the captions, which is answered in Step 3. Step 4 utilizes a mixture-of-models round-trip-consistency (MoM-RTC) check to eliminate incorrect samples.

	2 Modalities			3 Modalities			4 Modalities			All		
Model	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All
CREMA*	**0.71**	**0.64**	**0.68**	**0.61**	**0.55**	**0.58**	**0.45**	**0.39**	**0.42**	**0.60**	**0.53**	**0.56**
X-InstructBLIP	0.47	0.48	0.47	0.30	0.27	0.29	0.13	0.22	0.18	0.31	0.33	0.32
OneLLM	0.52	0.52	0.52	0.16	0.22	0.19	0.24	0.27	0.25	0.31	0.34	0.32
Gemini-2.0†	0.24	0.21	0.23	0.10	0.14	0.13	×	×	×	0.23	0.20	0.22
Caption Baseline	0.52	0.46	0.49	0.33	0.33	0.33	0.26	0.27	0.26	0.38	0.36	0.37

Table 2: Zero-Shot Evaluations on Contra4 Test Set.

† Proprietary LLM. Samples with 3D are excluded due to incompatibility.

* RGB rendering signal used for 3D point clouds.

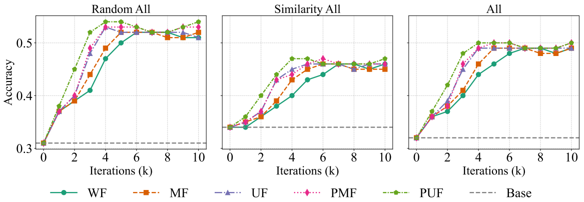


Figure 4: Finetune OneLLM on different MoM-RTC data.

ples that do not contain 3D since it is not supported. Lastly, our **Caption Baseline** replaces multimodal scenes with predicted captions for an LLM-only approach (details in Appendix H).

6 Discussion

How does MoM-RTC affect dataset quality? We conduct a human inspection of 120 randomly selected dataset samples, evenly split by negative sampling (high-similarity vs. random) and input modality choices, to validate dataset quality and our MoM-RTC procedure. Table 1 presents these results using the interface in Figure 10. PUF though highly selective, produces superior quality samples without relying on costly, closed-source APIs, mitigating selection bias (Pezeshkpour and Hruschka, 2023; Balepur et al., 2024; Wang et al., 2024b). By admitting only examples that remain correct under choice permutations, we counteract LLM biases, improving overall correctness. While permutation-based RTC methods require three times the GPU hours of non-permutation approaches, they improve human-perceived precision by over 20 points.

How do SOTA models perform on Contra4? Table 2 reports performance of SOTA models on the task, showing that caption-based baselines outperform most approaches as the number of modalities increases. The top performer, CREMA, relies on external RGB rendering for point clouds, though resource-intensive, it significantly boosts performance across 3D, Image, and Video (Figure 5). While the test set is manually curated, we provided a sample of 50 examples from the test set to two independent annotators who yielded an average performance of 92%, making CREMA's 56% accuracy a significant gap. Architecturally, CREMA employs distinct modules for cross-modal token extraction, similar to X-InstructBLIP, but fuses them before aligning with the base LLM, aiming at a more uniform modality representation. OneLLM, in contrast, uses a fused X-modal token extraction module that appears less effective. Surprisingly, Gemini achieves the lowest score, likely due to lack of fine-tuning for this task; in many responses, it fails to recognize all three inputs, and instead resolves to captioning only the last input provided, ignoring the question. Overall, these findings suggest that systems that fuse modalities after independent extraction, followed by LLM fine-tuning, are most effective for the task.

How does fine-tuning affect task performance? To further validate our findings, we fine-tuned OneLLM on MoM-RTC data, resulting in a noticeable performance boost from 32% to

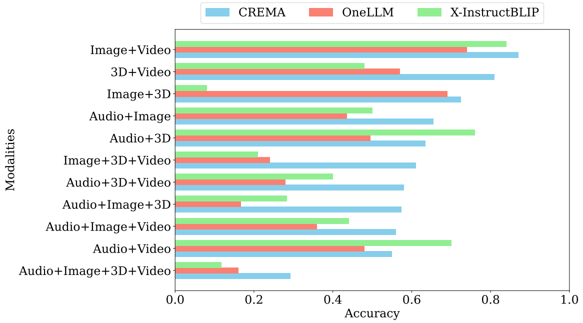


Figure 5: Performance breakdown by input modalities.

50%. However, overall accuracy remained low, indicating that fine-tuning alone is insufficient and alternative approaches are needed. Interestingly, despite lower human-perceived quality, all data filtering methods ultimately achieve similar performance given enough training iterations. Notably, PUF converges with the least data, followed by PMF and UF, aligning with their human-inspected accuracy rankings.

Overall, our results reveal that even state-of-the-art multimodal models struggle with the task introduced in this work, underscoring the need for more robust cross-modal understanding.

7 Limitations

A key limitation of our work lies in the simplified nature of the task formulation. While real-world applications often require models to determine which modality best supports a decision or query, Contra4 reduces this to a controlled, contrastive selection problem. **This abstraction is intentional**: it allows us to isolate a core cognitive capability, identifying the most relevant modality for a given prompt, without the confounds of task-specific logic or procedural complexity. However, we acknowledge that this setup does not fully capture the rich, dynamic nature of real-world multimodal reasoning.

Additionally, Contra4 emphasizes commonsense semantic matching (e.g., identifying peaceful scenes or emotional tones) rather than testing deeper inferential or task-specific reasoning. Our goal is not to exhaustively probe all aspects of multimodal intelligence, but to provide a focused diagnostic benchmark for a foundational yet underexplored skill. That said, future work should explore benchmarks that integrate higher-level decision-making and temporal inference across modalities, and complement evaluations like Contra4 with information synthesis settings to push the boundaries of model understanding.

Lastly, while our dataset provides a rigorous benchmark for cross-modal reasoning, performance evaluations depend on current state-of-the-art models, which may not yet be fully optimized for this task. As multimodal architectures evolve, future benchmarks should adapt accordingly to reflect their growing capabilities.

8 Ethics Statement

In conducting this research, we acknowledge the significant limitations and potential dangers associated with the use of Large Language Models (LLMs). One of the primary concerns is the presence of inherent biases within LLMs, which are a direct consequence of the data on which they are trained. These biases can inadvertently perpetuate harmful stereotypes and lead to discriminatory outcomes, particularly in sensitive applications. Additionally, LLMs, especially those with large parameter counts, may generate outputs that are factually incorrect or misleading, posing a risk in contexts that demand high levels of accuracy and reliability. To mitigate these risks we inspected the test samples of the dataset and used multimodal sources that would limit the potential of generation of such harmful questions. However, we emphasize the importance of ongoing vigilance and the need for responsible use of these models and our dataset to prevent unintended negative consequences.

Note on AI Assistants: AI assistants were used for grammar checks and sentence level rephrasing to improve paper flow. Coding assistants were also used to streamline development.

Acknowledgments

We are grateful to the reviewers for their thorough review and insightful feedback on our manuscript. This research was developed in part with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE*

International Conference on Computer Vision, pages 8948–8957.

Huda Alamri, Chiori Hori, Tim K Marks, Dhruv Batra, and Devi Parikh. 2018. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*, volume 2.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question? *arXiv preprint arXiv:2402.12483*.

Ankan Bansal, Yuting Zhang, and Rama Chellappa. 2020. Visual question answering on image sets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 51–67. Springer.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2024. Visit-bench: A dynamic benchmark for evaluating instruction-following vision-and-language models. *Advances in Neural Information Processing Systems*, 36.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, YANDONG LI, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, and Boqing Gong. 2025. Omnixr: Evaluating omni-modality language models on reasoning across modalities. In *The Thirteenth International Conference on Learning Representations*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL-HLT*, pages 2368–2378.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700*.

Xingwei He, Zhenghao Lin, Yeyun Gong, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yong Huang, Edgar Mariano Marroquin, and Volodymyr Kuleshov. 2021. A multi-modal and multitask benchmark in the clinical domain.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022b. MMCoQA: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231, Dublin, Ireland. Association for Computational Linguistics.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, Russ Salakhutdinov, and Louis-Philippe Morency. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Yike Yuan, Wangbo Zhao, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023. Mmbench: Is your multi-modal model an all-around player?

Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, et al. 2024. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6890–6894. IEEE.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. 2022. Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024c. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, Bangkok, Thailand. Association for Computational Linguistics.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920.

XinhaoMei. 2023. Wavcaps. `https://github.com/XinhaoMei/WavCaps`. Accessed: 2023-07-1.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Fanqing Meng, Siyuan Huang, Meng Lei, Ping Luo, and Yu Qiao. 2023a. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2023b. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. 2024. Lamm: Language-assisted multimodal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. 2024. Crema: Multimodal compositional video reasoning via efficient modular adaptation and fusion. *arXiv preprint arXiv:2402.05889*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multidiscipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Junjie Zhang, Tianci Hu, Xiaoshui Huang, Yongshun Gong, and Dan Zeng. 2024. 3dbench: A scalable 3d benchmark and instruction-tuning dataset. *arXiv preprint arXiv:2404.14678*.

A Data Format

The dataset is stored in an easy-to-use `json` format. Each entry in the dataset consists of various fields including a unique identifier, selection type, question type, examples from various modalities, and the associated question and answer.

A.1 Structure

- **id**: A unique identifier for the dataset entry.
- **selection_type**: The method used for selecting negative examples.
- **q_type**: The question type indicating the number of choices.
- **examples**: A list of examples, each containing:
  - **source**: The dataset from which the example is taken.
  - **id**: A unique identifier for the example within its source.
  - **caption**: A description of the content or scene depicted in the example.
- **modalities**: A list of modalities corresponding to each example.
- **questions**: The question presented to the model.
- **answers**: The correct answer or ground truth.
- **category**: The category of the question, used for organizing the dataset.

B Benchmark Comparisons

Table 3 provides a succinct comparison across multimodal benchmarks,[3] showing that `Contra4` is unique in its incorporation of up to four distinct modalities in a single example.

C Data Generation Details

Step 1: Negative Sampling Selection We employ two negative selection strategies: *random* and *high similarity* to reduce the likelihood that the task can be solved using unimodal heuristics. For the high similarity negative samples, we first encode all captions across all modalities using `all-MiniLM-L6-v2` embeddings via sentence-transformers. Subsequently, we anchor one modality randomly as the basis for selection. From this anchored modality, we identify and select a negative sample from among the thirty most similar instances across the different modalities, as ranked by the cosine similarity of their text captions. For the random setup, we perform the same procedure but sample randomly instead.

Step 2 Question Generation: Upon generating tuples in Step 1, we employ `meta-llama/Llama-3.1-8B-Instruct` to generate contrastive questions. For each tuple, we provide the LLM with four in-context examples to facilitate the generation of a question which is then considered for inclusion in the final dataset. The prompt for question generation is the following:

> <s>You are given some scenes described in text. Each scene is represented by a short caption. Your task is to generate a question that compares the scenes based on their content. The generated question should be relevant to the context of the scenes and should require a comparison between them. There should be only one correct answer. Here are some examples to guide you:
>
> Scene A. "a shattered piece of paper, resembling a broken phone and a flying newspaper"
> Scene B. "tourists walking near a catholic church in Mexico on a sunny summer day"
> Generated Question: Which scene evokes a sense of chaos and abandonment?
>
> Scene A. "Someone is using a rip saw in a carpenter's workshop"
> Scene B. "An elegant bathroom featuring a tub, sink, mirror, and decorations"
> Generated Question: Which scene is more likely to involve louder noises?
>
> Scene A. "The night sky showcasing the Milky Way"
> Scene B. "A bustling city street at midday"
> Scene C. "A serene mountain landscape in the morning"
> Generated Question: Which scene is different from the other two?
>
> Scene A. "A painting depicting a stormy sea"
> Scene B. "A photograph of a calm beach at sunset"
> Scene C. "A digital illustration of a bustling space station"

[3] We do not include vision benchmarks such as GQA (Hudson and Manning, 2019), VizWiz (Bigham et al., 2010), and NoCaps (Agrawal et al., 2019) since they appear as subsets of other benchmarks included in the table such as LVLM-eHUB (Xu et al., 2023a).

Dataset	Image	Audio	Video	3D	Max Modalities per Sample
DROP (Dua et al., 2019)	×	×	×	×	1
MMLU (Hendrycks et al., 2020)	×	×	×	×	1
MULTIBench (Liang et al., 2021)	✓	✓	✓	×	2
BigBench (Srivastava et al., 2023)	×	×	×	×	1
LVLM-eHUB (Xu et al., 2023a)	✓	×	×	×	1
SEED (v1) (Li et al., 2023b)	✓	×	✓	×	1
SEED (v2) (Li et al., 2023a)	✓	×	✓	×	2
MM-BENCH (Liu et al., 2023)	✓	×	×	×	1
VisIT-Bench (Bitton et al., 2024)	✓	×	×	×	1
MM-VET (Yu et al., 2023)	✓	×	×	×	1
MMMU (Yue et al., 2023)	✓	×	×	×	1
LAMM (Yin et al., 2024)	✓	×	×	✓	1
AV-Superb (Tseng et al., 2024)	✓	✓	✓	×	1
HEAR (Turian et al., 2022)	×	✓	×	×	1
Dynamic Superb (Tseng et al., 2024)	×	✓	×	×	1
AIR-Bench (Yang et al., 2024)	×	✓	×	×	1
Video-Bench (Ning et al., 2023)	×	✓	✓	×	2
3D-Bench (Zhang et al., 2024)	×	×	×	✓	1
OmniXR (Chen et al., 2025)	✓	✓	✓	×	1
DisCRn (Panagopoulou et al., 2023)	✓	✓	✓	✓	2
Contra4	✓	✓	✓	✓	4

Table 3: **Comparison of Multimodal Challenge Datasets.** Columns *Image*, *Audio*, *Video*, and *3D* specify whether the dataset includes that modality (✓) or not (×).

Scene D. "A sculpture of a tranquil garden"
Generated Question: Which scene is most different from the other three?

Scene A. "A team of firefighters putting out a blaze in a city"
Scene B. "A family enjoying a picnic in a peaceful park"
Generated Question: Which scene involves a greater sense of danger and urgency?

Scene A. "A snowy mountain peak illuminated by the golden light of sunrise"
Scene B. "A tropical beach with crystal-clear water and palm trees swaying in the breeze"
Scene C. "A bustling city park filled with people enjoying outdoor activities"
Scene D. "A vast desert under a blazing sun with sand dunes stretching to the horizon"
Generated Question: Which scene represents a colder and more remote environment?

We implement a filtering process to exclude questions that focus on textual or difficult to measure qualities. This excludes questions containing terms (and derivatives) such as *'word', 'text', 'verb', 'noun', 'describe', 'question', 'sentence', 'detail', 'visual', 'image', 'video', 'audio', 'sound', 'heard', '3d', 'point cloud', 'caption', 'more elements', 'most elements', 'more objects', 'more people', 'most objects', 'more colors', 'most colors', 'more than one', 'similar', 'rating', 'score'*.

Step 3: Answer-Explanation Generation Building on the captions in the original dataset and the questions refined in Step 2, we require the same LLM to answer and explain its reasoning using the following prompt:

<s>You are given some scenes described in text as well as a question about them. Each scene is represented by a short caption. Your task is to provide a clear and concise answer that explains the reasoning behind the correct choice. Here are some examples to guide you:

Scene A. "a shattered piece of paper, resembling a broken phone and a flying newspaper"
Scene B. "tourists walking near a catholic church in Mexico on a sunny summer day"
Question: Which scene evokes a sense of chaos and abandonment?
Answer: Scene A. Scene A evokes feelings of chaos and abandonment, contrasting sharply with the joy and vibrancy of Scene B.

Scene A. "Someone is using a rip saw in a carpenter's workshop"
Scene B. "An elegant bathroom featuring a tub, sink, mirror, and decorations"
Question: Which scene is more likely to involve louder noises?
Answer: Scene A. Scene A is characterized by the noise and activity of craftsmanship, whereas Scene B offers a serene and luxurious ambiance for relaxation.

Scene A. "The night sky showcasing the Milky Way"
Scene B. "A bustling city street at midday"
Scene C. "A serene mountain landscape in the morning"
Question: Which scene is different from the other two?
Answer: Scene B. Scene B, with its bustling city life, differs in its dynamic and urban setting from the tranquil and natural settings of Scenes A and C.

Scene A. "A painting depicting a stormy sea"
Scene B. "A photograph of a calm beach at sunset"
Scene C. "A digital illustration of a bustling space station"
Scene D. "A sculpture of a tranquil garden"
Question: Which scene is most different from the other three?
Answer: Scene C. Scene C, a digital illustration of a bustling space station, diverges in its futuristic and technological theme from the natural and serene subjects of the other inputs.

Scene A. "A team of firefighters putting out a blaze in a city"
Scene B. "A family enjoying a picnic in a peaceful park"

Question: Which scene involves a greater sense of danger and urgency?
Answer: Scene A. Scene A, with firefighters responding to a blaze, conveys a strong sense of danger and urgency compared to the calm and leisurely atmosphere of Scene B.

Scene A. "A snowy mountain peak illuminated by the golden light of sunrise"
Scene B. "A tropical beach with crystal-clear water and palm trees swaying in the breeze"
Scene C. "A bustling city park filled with people enjoying outdoor activities"
Scene D. "A vast desert under a blazing sun with sand dunes stretching to the horizon"
Question: Which scene represents a colder and more remote environment?
Answer: Scene A. Scene A, featuring a snowy mountain peak, exemplifies a cold and remote environment in contrast to the other settings, which are warmer or more populated.

Step 4: Mixture-of-Models Round-Trip-Consistency (MoM-RTC): This step verifies the answers of Step 3, via querying multiple models under all possible permutations of the inputs. For clarity, we present a pseudo-algorithm for the MoM-RTC procedure in Algorithm 1. Each of the three LLMs in this procedure is prompted as follows:

Select which of the scenes best answers the question. Respond with brevity, and only include your choice in the response.
Question: {question}
Choices:
Scene A. {first modality caption}
Scene B. {second modality caption}
and so on...
Answer:

D Category Distribution

To analyze the breadth of the dataset we automatically extract instance categories by employing an LLM which are then grouped based on keyword matching. In particular, we use `meta-llama/Llama-3.1-8B-Instruct` served via VLLM (Kwon et al., 2023) and prompt it to predict the topic of each question using the following prompt:

You are tasked with categorizing a question that compares or evaluates inputs based on a specific property (e.g., which input is more positive, has more action, etc.).

Example Questions and Outputs:
Question: "Which input is more positive in tone?"
Category: Sentiment Analysis
Reasoning: The question explicitly asks about emotional tone, a sentiment-related property.

Question: "Which video has more action?"
Category: Activity Level
Reasoning: The question focuses on the level of dynamism or activity in the input videos.

Question: "Which object is larger?"
Category: Size Comparison
Reasoning: The question compares a specific property, size, between inputs.

Question: "Which scene is more likely to involve human presence?"
Category: Human Presence
Reasoning: The question asks about the likelihood of human presence

Question: "Which scene involves more unpredictable or sudden changes?"
Category: Dynamic Changes
Reasoning: The question asks about the level of unpredictability or sudden changes in the scene.

Question: {question}
Category:

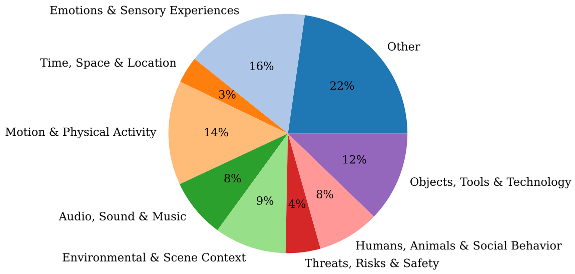


Figure 6: Category distribution in human annotated set

Figure 6 illustrates the resulting category distribution on the annotated test set.

E Caption Datasets

In Table 4 we provide details on the captioning datasets used to connect the separate modalities in `Contra4`.

Modality	Dataset	Train Split	Test Split	Captions License	Data License
Image	MSCOCO (Changpinyo et al., 2021)	Train2017	Val2017	CC by 4.0	CC by 4.0
Video	MSRVTT (Xu et al., 2016)	Train	Test	MIT License	MIT License
3D	PointLLM (Xu et al., 2023b)	train	test	ODC-By 1.0	CC-by-4.0
Audio	AudioCaps (Kim et al., 2019)	Train	Validation	MIT License	CC by 4.0
	Clotho (Drossos et al., 2020)	Development	Evaluation(v1) + Validation(v2)	Non-Commercial	Non-Commercial

Table 4: Datasets used to generate `Contra4`

F Additional Dataset Statistics

Using the MoM-RTC pipeline, we produce 174k automatically annotated samples for training and release a test set of 2.3k human-annotated examples. Table 5 shows a more detailed breakdown on the types of data maintained across different MoM-RTC methods. Fig. 7 shows the distribution of different modalities in the train and test data.

Filter	Human Acc.	Recall	N/A	O/A	GPU (hrs)	2 Modalities			3 Modalities			4 Modalities			Aggregated		
						Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All
None*	39.5	**89.1**	13.6	10.2	0	143k	145k	287k	90k	94k	183k	21k	23k	44k	254k	261k	515k
MF	65.5	85.7	7.9	7.1	40	115k	113k	227k	63k	62k	125k	13k	13k	26k	190k	188k	378k
UF	71.5	79.7	4.4	10.9		83k	80k	163k	40k	38k	78k	8k	7k	15k	130k	126k	256k
PMF	72.8	76.7	6.6	7.3	120	102k	93k	196k	39k	34k	73k	5k	4k	9k	147k	131k	278k
PUF	**83.3**	74.4	**0.0**	**2.5**		69k	64k	133k	20k	18k	38k	2k	1k	3k	91k	83k	174k

Table 5: Human Inspection of Different Round-Trip-Consistency Checks on Train Data. N/A corresponds to the percentage of wrong examples that are wrong due to lack of applicability to any choice, and O/A to the percentage of wrong examples due to the question applying to more than one choice. * some rule based word filtering is applied, see Appendix C.

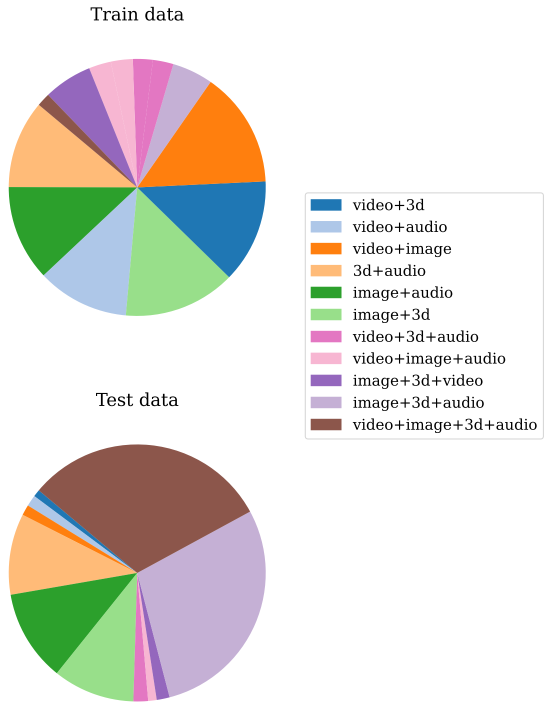


Figure 7: Modality Combination Distribution

G Implementation Details

All models are served via VLLM (Kwon et al., 2023) on 4 A100 40GB GPUs. LLMs are always queried using nucleus sampling with top_p=0.9. For Step 2 meta-llama/Llama-3.1-8B-Instruct is queried with temperature equal to 1.05 to encourage diverse questions, and 0.3 for Step 3 and Step 4. All cross-modal LLMs are benchmarked using the default parameter settings in their corresponding repositories and API. For fine-tuning OneLLM we employ LoRA (Hu et al., 2021) with batch size 8, weight decay 0.02, learning rate 1e-7, and a gradient clipping norm of 2.0 for 10k iterations. We ensure that modalities are balanced in training via appropriate sampling of all modalities equally.

H Caption Baseline Details

The caption baseline employs OpenGVLab/InternVL2-8B (Chen et al., 2024) to generate captions for images, Qwen/Qwen2-VL-7B-Instruct (Wang et al., 2024a) for videos, Qwen/Qwen2-Audio-7B-Instruct (Chu et al., 2024) for audio, and X-InstructBLIP (Panagopoulou et al., 2023) for 3D point clouds. With the exception of X-InstructBLIP where we use the official implementation, all other models are queried via VLLM. All models are queried with the default hyperparameters. We use the following prompts: 'Describe the [image/audio/3d model]' and 'Describe this set of frames. Consider the frames to be a part of the same video.'. Table 6 shows the captioning performance on each modality for the validation subset of Contra4. The reasoning LLM underlying this baseline is `meta-llama/Llama-3.1-8B-Instruct`.

	Image	Video	Audio	3D
METEOR	0.21	0.15	0.20	0.17

Table 6: Predicted caption performance (METEOR)

I Detailed OneLLM fine-tuning Results

Figure 8 shows a breakdown of fine-tuning performance across different question types. We find the trend to be similar across all subsets, with lower performance on examples sampled with high similarity.

J Dataset Examples

Figure 9 displays data examples from the test split of Contra4 for each of the categories identified in Appendix D.

K Human Annotation

In evaluating the effectiveness of mixture-of-models round-trip-consistency for synthetic data generation, we develop a user interface presented in Figure 10. These volunteers were not offered monetary compensation and participated primarily out of academic interest and willingness to contribute to ongoing research as they are all graduate students in computer science in an American university. All annotators provided informed consent and were briefed on the nature of the task prior to participation. For each example, we present the question and the corresponding modality choices, with the option to select 'None of the above.' Note that human annotation was not part of the data generation process itself, as we relied on automatic filtering methods. However, for the test set, all samples were manually reviewed by the authors to ensure clarity, correctness, and overall quality.

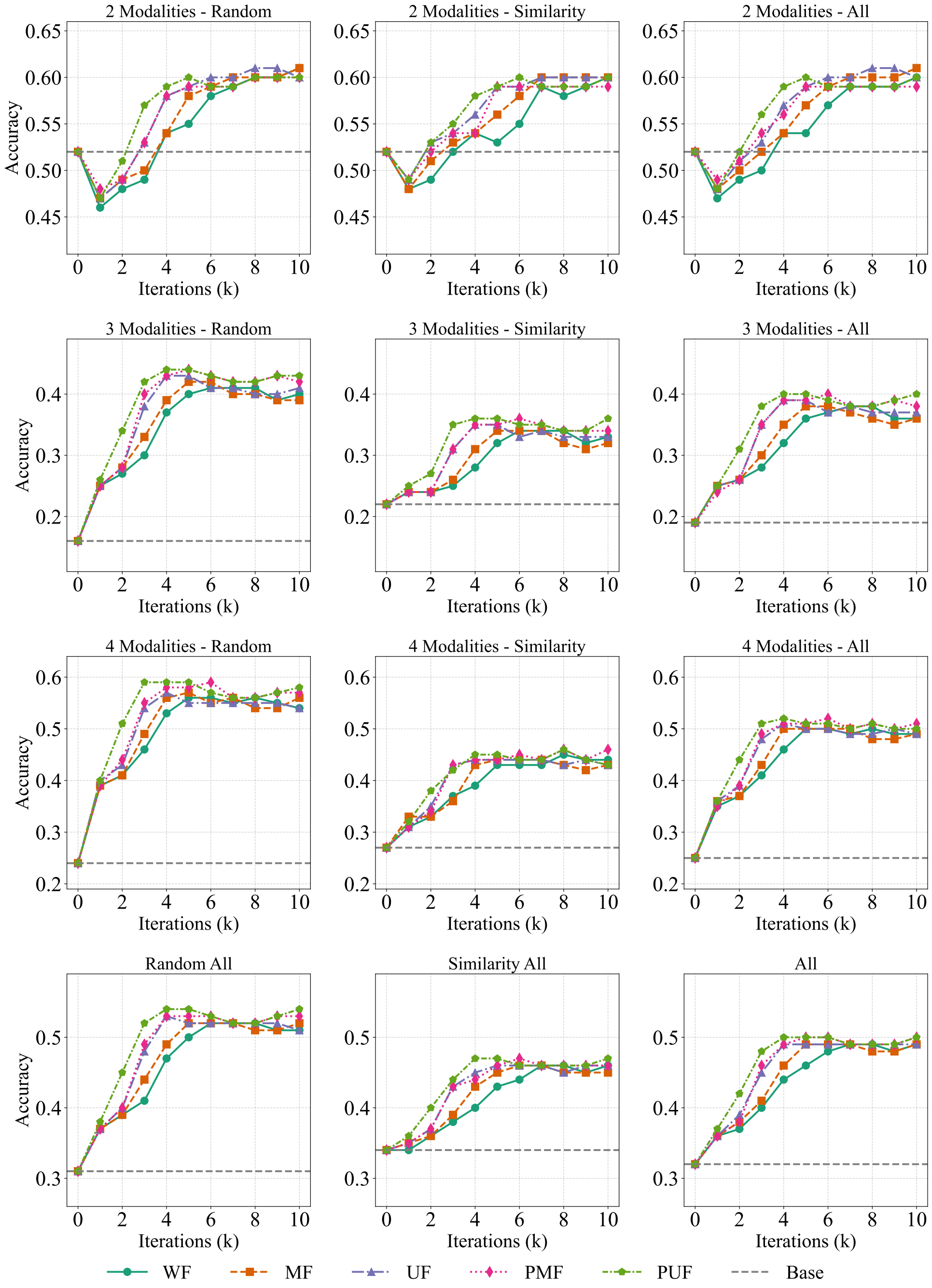


Figure 8: Detailed break down of MoM-RTC data effectiveness for fine-tuning OneLLM

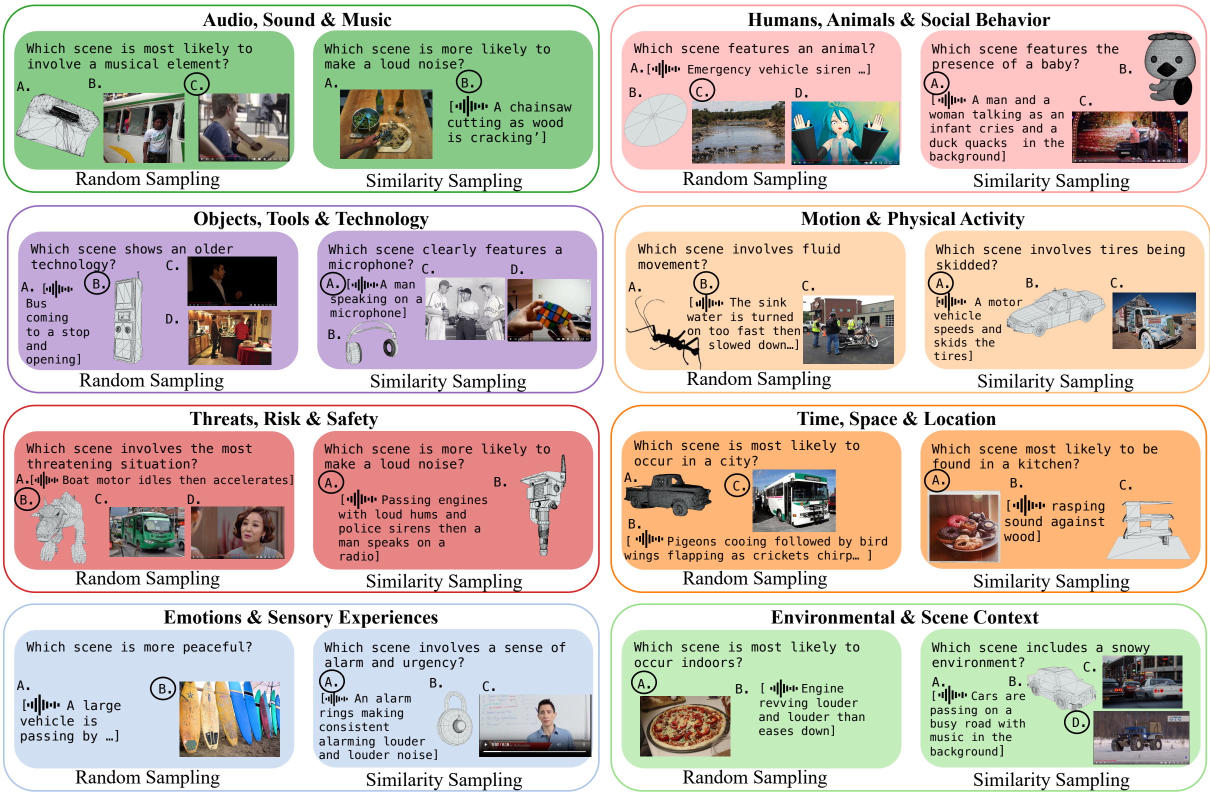

Audio, Sound & Music
Which scene is most likely to involve a musical element?
A.
B.
C.
Random Sampling
Which scene is more likely to make a loud noise?
A.
B.
[A chainsaw cutting as wood is cracking']
Similarity Sampling
Humans, Animals & Social Behavior
Which scene features an animal?
A.[Emergency vehicle siren …]
B.
C.
D.
Random Sampling
Which scene features the presence of a baby?
A.
[A man and a woman talking as an infant cries and a duck quacks in the background]
B.
C.
Similarity Sampling
Objects, Tools & Technology
Which scene shows an older technology?
A. [Bus coming to a stop and opening]
B.
C.
D.
Random Sampling
Which scene clearly features a microphone?
A.[A man speaking on a microphone]
B.
C.
D.
Similarity Sampling
Motion & Physical Activity
Which scene involves fluid movement?
A.
B.
[The sink water is turned on too fast then slowed down…]
C.
Random Sampling
Which scene involves tires being skidded?
A.
[A motor vehicle speeds and skids the tires]
B.
C.
Similarity Sampling
Threats, Risk & Safety
Which scene involves the most threatening situation?
A.[Boat motor idles then accelerates]
B.
C.
D.
Random Sampling
Which scene is more likely to make a loud noise?
A.
[Passing engines with loud hums and police sirens then a man speaks on a radio]
B.
Similarity Sampling
Time, Space & Location
Which scene is most likely to occur in a city?
A.
B.
[Pigeons cooing followed by bird wings flapping as crickets chirp…]
C.
Random Sampling
Which scene most likely to be found in a kitchen?
A.
B.
[rasping sound against wood]
C.
Similarity Sampling
Emotions & Sensory Experiences
Which scene is more peaceful?
A.
[A large vehicle is passing by …]
B.
Random Sampling
Which scene involves a sense of alarm and urgency?
A.
[An alarm rings making consistent alarming louder and louder noise]
B.
C.
Similarity Sampling
Environmental & Scene Context
Which scene is most likely to occur indoors?
A.
B. [Engine revving louder than and louder than eases down]
Random Sampling
Which scene includes a snowy environment?
A.
[Cars are passing on a busy road with music in the background]
B.
C.
D.
Similarity Sampling

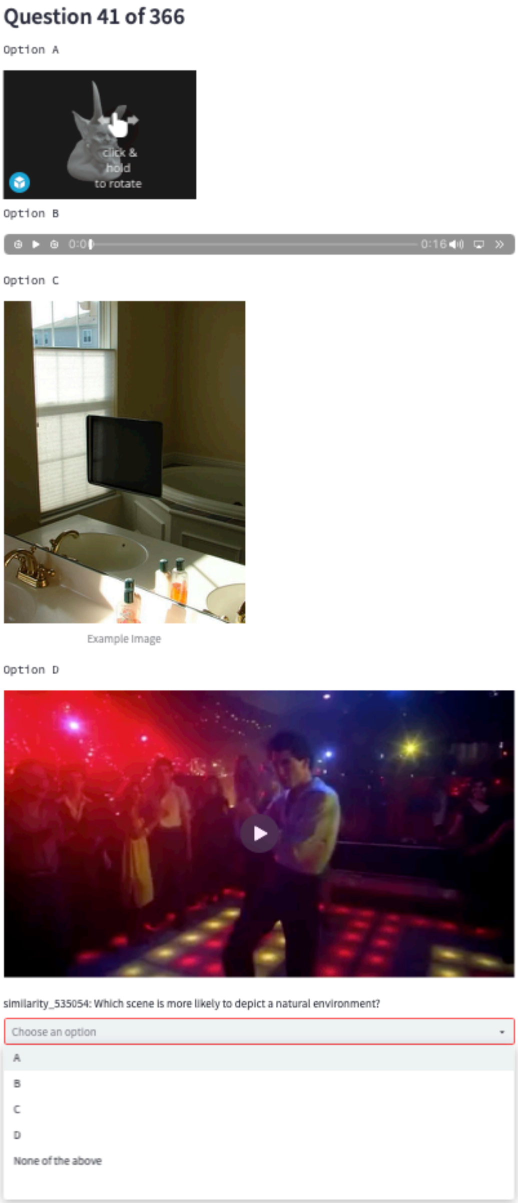
Figure 9: Dataset examples for each category.

Filter	2 Modalities			3 Modalities			4 Modalities			Aggregated		
Correctness												
	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All
WF	75.0	60.0	67.5	30.0	30.0	30.0	55.0	30.0	42.5	53.3	40.0	46.7
MF	90.0	70.0	80.0	55.0	50.0	52.5	55.0	40.0	47.5	66.7	53.3	60.0
UF	60.0	90.0	75.0	50.0	55.0	52.5	75.0	30.0	52.5	61.7	58.3	60.0
PMF	75.0	65.0	70.0	60.0	80.0	70.0	65.0	65.0	65.0	66.7	70.0	68.3
PUF	85.0	70.0	77.5	75.0	90.0	82.5	95.0	85.0	90.0	85.0	81.7	83.3
Over-Applies (OA)												
Filter	**2 Modalities**			**3 Modalities**			**4 Modalities**			**Aggregated**		
	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All
WF	5.0	10.0	7.5	30.0	20.0	25.0	10.0	35.0	22.5	15.0	21.7	18.3
MF	0.0	20.0	10.0	5.0	15.0	10.0	20.0	45.0	32.5	8.3	26.7	17.5
UF	0.0	5.0	2.5	30.0	15.0	22.5	0.0	30.0	15.0	10.0	16.7	13.3
PMF	15.0	20.0	17.5	20.0	5.0	12.5	15.0	15.0	15.0	16.7	13.3	15.0
PUF	0.0	0.0	0.0	5.0	15.0	10.0	10.0	5.0	7.5	5.0	6.7	5.8
None-Applies (NA)												
Filter	**2 Modalities**			**3 Modalities**			**4 Modalities**			**Aggregated**		
	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All	Rand.	Sim.	All
WF	10.0	20.0	15.0	30.0	25.0	27.5	10.0	15.0	12.5	16.7	20.0	18.3
MF	10.0	10.0	10.0	40.0	25.0	32.5	20.0	5.0	12.5	23.3	13.3	18.3
UF	30.0	5.0	17.5	20.0	15.0	17.5	20.0	10.0	15.0	23.3	10.0	16.7
PMF	5.0	10.0	7.5	20.0	15.0	17.5	15.0	15.0	15.0	13.3	13.3	13.3
PUF	10.0	10.0	10.0	5.0	10.0	7.5	5.0	0.0	2.5	6.7	6.7	6.7

Table 7: Detailed results of human inspection. We report percentages for each of the metrics on the corresponding data subsets.

Algorithm 1 Mixture-of-Models Round-Trip-Consistency (MoM-RTC)

```
Require: data: List of samples, each with {question, cross-modal info, original_answer};
models: Ensemble of LLMs/classifiers;
permute_strategy $\in$ {NONE, RANDOM, ALL};
filtering_criteria $\in$ {MF, UF, PMF, PUF};
Ensure: filtered_data: Subset of samples passing the consistency check
function GENERATE_PERMUTATIONS(sample, strategy)    ▷ Returns a list of permuted versions of sample
end function
function GET_MODEL_PREDICTION(model, sample)    ▷ Prompts the model on the given sample and returns a predicted answer
end function
function MAJORITY_VOTE(answers)    ▷ Returns the most frequent answer in answers; or handle ties as needed
end function
function UNANIMOUS_VOTE(answers)    ▷ Returns the unique answer if all are identical, else “no unanimous consensus”
end function
filtered_data $\leftarrow$ [ ]
for all sample $\in$ data do
    original_answer $\leftarrow$ sample.original_answer
                                                    ▷ 1. Generate permutations of the sample
    permutations $\leftarrow$ GENERATE_PERMUTATIONS(sample, permute_strategy)
                                                    ▷ 2. Query each model on each permutation
    predictions_by_perm $\leftarrow$ [ ]
    for all perm $\in$ permutations do
        model_preds $\leftarrow$ [ ]
        for all model $\in$ models do
            pred $\leftarrow$ GET_MODEL_PREDICTION(model, perm)
            model_preds.append(pred)
        end for
        predictions_by_perm.append(model_preds)
    end for
                                                    ▷ 3. Check the consistency criteria
    if filtering_criteria $\in$ {MF, UF} then    ▷ Single (unpermuted) scenario; use the first permutation’s predictions
        model_preds $\leftarrow$ predictions_by_perm[0]
        if filtering_criteria = MF then
            voted_answer $\leftarrow$ MAJORITY_VOTE(model_preds)
            if voted_answer = original_answer then
                filtered_data.append(sample)
            end if
        else if filtering_criteria = UF then
            unanimous_answer $\leftarrow$ UNANIMOUS_VOTE(model_preds)
            if unanimous_answer $\neq$ "no unanimous consensus" and unanimous_answer = original_answer then
                filtered_data.append(sample)
            end if
        end if
    else                                            ▷ PMF or PUF: multiple permutations
        consistent_across_all $\leftarrow$ True
        for all model_preds $\in$ predictions_by_perm do
            if filtering_criteria = PMF then
                voted_answer $\leftarrow$ MAJORITY_VOTE(model_preds)
                if voted_answer $\neq$ original_answer then
                    consistent_across_all $\leftarrow$ False
                    break
                end if
            else if filtering_criteria = PUF then
                unanimous_answer $\leftarrow$ UNANIMOUS_VOTE(model_preds)
                if (unanimous_answer = "no unanimous consensus") $\vee$ (unanimous_answer $\neq$ original_answer) then
                    consistent_across_all $\leftarrow$ False
                    break
                end if
            end if
        end for
        if consistent_across_all then
            filtered_data.append(sample)
        end if
    end if
end for
return filtered_data
```



Question 41 of 366

Option A

Option B

Option C

Example image

Option D

similarity_535054: Which scene is more likely to depict a natural environment?

Choose an option

A

B

C

D

None of the above

Figure 10: Interface for Human Inspection