AVTrustBench: Assessing and Enhancing Reliability and Robustness in Audio-Visual LLMs Supplementary

The supplementary is organised as follows:

1. More Details About the Data

1.1. Exclusion of single modality questions.

In the original AVQA [17], MUSIC-AVQA [9] a subset of the questions were agnostic either of visual or the audio modality, which can be answered with only one modality. However, while forming the QA pairs, we perform a careful inspection to eliminate such samples. To ensure the validity of the AVTrustBench benchmark, we carefully excluded these questions. we removed $\sim$ 10% of samples from MUSIC-AVQA for the Adversarial attack and $\sim$ 50% for the Modality-specific dependency respectively. For Compositional reasoning we carefully choose the samples that encompass both the modalities from the AudioSet dataset following a semi-automated strategy. Nearly 30% of the samples are synthetically generated.

1.2. Construction of AVTrustBench

Tab. 1 contains the task-wise question and instruction templates for each task. We carefully construct up to $\sim$ 5 different prompts for each task type. The collected samples have audio durations of 10–20s and video resolutions of 240p–1080p. Next, we elaborate on the data preparation strategy for each task.

Adversarial attack. For Adversarial attack we consider the AVQA [17] and MUSIC-AVQA dataset [9]. We retain the original labels from the MUSIC-AVQA dataset ('Existential', 'Localization', etc.) and annotate samples from AVQA with one of the 'Existential', 'Temporal', 'Localisation' and 'World Knowledge' categories depending on the QA pair. For AVQA, we prepare two sets that act as look-up tables while forming the options in the below-mentioned cases. The first one (**T1**) contains a mapping between a given sounding object class of interest and other classes which are not associated with this class *in any way*. This mapping is done through careful manual annotation. The other table (**T2**) contains category-wise groupings for sounding objects for example 'musical instruments', 'animal sound', 'vehicles' etc. which are the most common supercategories observed in the AVQA dataset. For MUSIC-AVQA, note that the audio files are mostly restricted to music instrument classes. Subsequently, we prepare a Table (**T3**) mapping the category information (i.e., Existential, Localization, etc.) with all the available Ground Truth answers in the MUSIC-AVQA dataset. For example, the 'Existential' category may be mapped to 'Flute', 'Piano', etc., whereas the 'Localization' category may be mapped to 'Left', 'Right', etc.

MCIT: For this task we prepare an automated script to first extract the correct response for a given question and replace that with another option from the same category. For example: if the question is 'What is the colour of the instrument at the left of the sounding object?' the correct answer 'Brown' is replaced with 'Black' which is chosen from the previously defined look-up (T2). For the AVQA dataset, we directly adapt its original options before removing the correct choice, while for MUSIC-AVQA we add the options from T3 (as defined above) depending on the question category.

ICIT: In this task, we ensure the options provided to the AVLLMs have no relevance at all to the semantics of the question. For AVQA, we again sample the options from a pre-built look-up containing category-wise object/entity

names (T1). For example, the category 'animal' contains the names of all the animals from the datasets we are dealing with. So while preparing the options for this task we ensure to choose samples from non-overlapping categories. For MUSIC-AVQA, we follow a similar strategy where we sample options based on T3 from the question categories other than the actual category under consideration.

MVIT: While preparing the samples for this task, we replace the visual content with completely unrelated visual events. We ensure that this video clip which is used to replace the original video snippet is taken from T1 containing the mapping of this category with other non-overlapping categories for AVQA. For MUSIC-AVQA, we choose options from T3 depending on the question category.

MAIT: Lastly, for AVQA we again employ T1 to find samples which are non-correlated with a sample under consideration and replace its audio content using the latter. For MUSIC-AVQA, we again select options from T3 depending on the question category.

Compositional reasoning. We leverage the AudioSet [6] dataset to prepare the samples for this task. Below we elaborate on the data preparation strategy.

COT-Stitch: We carefully choose two semantically separate audio events and concatenate them in the time dimension. The options are prepared by extracting the audio event class. For example, if a *aeroplane engine sound* is concatenated with *a person playing the guitar*, the correct option is: 'Aeroplane followed by guitar'. The remaining options are generated using LLM (e.g., GPT-4) where we ask it to swap the ordering of acoustic events, replace the preposition, or swap noun-verb associations. Consequently, the generated options serve as negatives with similar contexts but different compositions which make the task even more challenging. Such generated options in the context of the above example are:'Guitar followed by aeroplane' and 'Both events occur simultaneously'.

COT-Swap: For this task, the option preparation strategy remains the same as above while the audio components of two dissimilar videos are swapped. We pick the two samples for each case from non-overlapping sets of audio events which we prepare beforehand.

CAT: For CAT, we first create a collection of several unique audio snippets and their labels where each consists of a single audio event. Using the snippet and label corresponding to the audio events we concatenate or overlay one audio over the other. Additionally, to assure high quality we don't concatenate or overlay random events but ask an LLM to create unique audio scenes. We prepare the options in a similar fashion as described above.

Modality-specific dependency. We consider a subset of the MUSIC-AVQA dataset and only consider samples that have a dependency on both audio and visual modalities.

MVT: We systematically eliminate the video modality from each video in this task. We keep the original answer and add the remaining options by choosing entries from T3 based on the question category under consideration.

MAT: We follow the same strategy as MVT except here the audio component is eliminated.

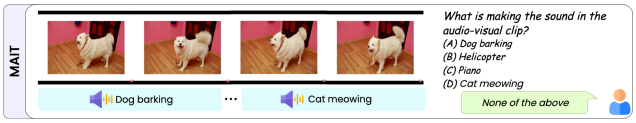


Figure 1. Impact of subtle distortion.

1.3. Diversity in the data samples.

Our dataset contains samples from a variety of datasets, e.g., AVQA, MUSIC-AVQA, and AudioSet, eventually making the data points belong to diverse distributions and categories. While our selection of AudioSet contains samples from 190 different categories, AVQA comprises 165 classes (compared to MUSIC-AVQA which comprises samples from 22 musical instruments) - which spans 355 out of a total of 377 categories making the collection of samples considerably diverse. These datasets are widely used in the majority of audio-visual tasks which lead to generalizable models due to the varied categories of events present in them. Additionally, we argue that datasets employed (e.g., CC3M, SBU, TextVQA, Kinetics, etc.) in some of the existing benchmarks do not contain meaningful audio information and hence are not suitable for our study. Finally, the size of our dataset is 40X larger than recent video benchmarks (SEED-Bench and VideoBench, etc) making it comprehensive and well round. We provide a comparison on the category-wise diversity of AVTrustBench with other existing benchmarks in the Tab. 2.

1.4. Impact of subtle distortion

We curate 100 samples targeting fine-grained scenarios with subtle audio/video distortions and partial modality degradation (Fig. 1). For MAIT, we evaluate cases like combining a video of a *dog barking* with audio partially containing a *cat meowing* and report the performance in Tab. 3

Method	MAIT ↑
Video-SALMONN	22.08
BAY-CAT	21.97
VideoLLaMA2	22.35
Gemini 1.5 Pro	**26.41**

Table 3. Comparison of methods based on MAIT.

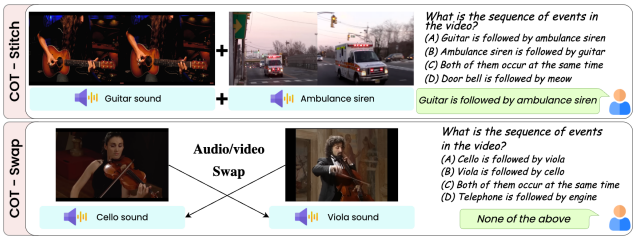


Figure 2. More examples on compositional tasks.

1.5. Examples of compositional tasks

The pseudocode (Algo.1) details the data construction steps involved in the compositional tasks (COT-Stitch and COT-Swap). We add examples (refer to Fig. 2) to demonstrate on these two tasks.

2. Additional Details on Evaluation Settings

2.1. Evaluation Settings

Unless stated otherwise, all results presented in this paper adhere to the conventional zero-shot evaluation setting. Below we provide different evaluation settings for the AVLLMs on AVTRUSTBENCH.

- **Base setting.** In this setting, neither additional instructions are provided to the model to withhold answers nor choices such as *None of the above* are provided. This setting represents the most common environment for using and the hardest scenario for evaluating AVLLMs on Adversarial attack and Modality-specific dependency suites.
- **Instruction setting.** In this setting, additional options such as "None of the above" and/or additional instruction such as "If all the options are incorrect, answer (D) None of the above." are provided to explicitly drive the model towards acknowledging the inconsistencies in the tasks present in Adversarial attack, Compositional reasoning, and Modality-specific dependency suites.

2.2. More Details on LLM-based Choice Extraction

Choice extraction strategy. We employ a two-step choice extraction strategy which we explain next. Extracting choices from free-form predictions is straightforward for human beings, but might be difficult with rule-based matching. To this end, we design a universal evaluation strategy for all AVLLMs with different instruction-following capabilities:

Step 1. *Prediction matching:* Initially, we attempt to extract choices from AVLLM predictions using heuristic matching. We aim to extract the choice label (e.g., 'A', 'B', 'C', 'D') from the AVLLM's output. If successful, we use this as the prediction. If not, we attempt to extract the choice label using GPT-4.

Step 2. *GPT-4 processing:* Previous evaluation benchmarks [10] establish the effectiveness of GPT-4 as a choice extractor. If step 1 fails, we provide GPT-4 with the question, choices, and model prediction. and instruct it to align the prediction with one of the given choices and produce the label. If there is no match found, GPT-4 returns 'No match found'.

We also employ the CircularEval strategy [10] to ensure a rigorous evaluation and effectively demonstrate the performance gap across various models.

Response matching. To apply the matching algorithm to the options we maintain the following: when an option is denoted simply by a letter such as 'A' or expressed as 'A) `<response>`', 'A. ', 'A, `<response>`', '(A) `<response>`' without the inclusion of other choices within the '`<response>`' portion, it is considered that option 'A' is being predicted.

Where does heuristic matching fail? The heuristic matching strategy typically fails in one of the following cases (i) when the AVLLM is not able to respond with any answer and asks for further clarification 'Apologies, can you please clarify ...' or its variants. (ii) when the AVLLMs respond with more than one option choice (A, B, C, etc.). In these cases we move on to Step 2 – GPT-4 based choice extraction. We provide a sample of how GPT-4 is prompted below.

Change in template for GPT-4 evaluation. Next, to identify the model prediction, we leverage GPT-4 following MMBench [10]. We query it with the template, including the question, options, and the corresponding AVLLM's prediction. As for options, we add task-specific options to recognize the model predictions.

For MCIT, we add two options: a masked correct option and the option of 'None of the above', 'Provided options are incorrect', and 'I cannot answer' and its variants.

For ICIT, we add two options: a masked correct option, and the option of 'None of the above', 'No option is correct', 'Irrelevant options', 'I cannot answer.' etc.

For MAIT and MVIT, we add an option of 'The visual/audio is incompatible with the question', or 'I cannot answer.'

For COT-Swap, we add an option of 'The visual/audio is incompatible', or 'I cannot answer.' and its variants.

Finally, for MAT and MVT we add an option of 'The audio is missing' and 'The video is missing' respectively or 'I cannot answer.' and its different variants to handle similar responses from AVLLMs.

2.3. Ensuring Robust Evaluation

Inspired by MMBench [10] we employ a CircularEval strategy to ensure robust evaluation. In AVTRUSTBENCH, the problems are presented as multiple-choice questions. Such formulation poses an evaluation challenge: random guessing

Algorithm 1 Pseudocode for COT-Stitch and COT-Swap.

```
def generate_cot_stitch(video1, audio1, video2, audio2):
    """
    Concatenates video1+video2 and audio1+audio2 into one multimodal clip.
    Returns the stitched clip and ground-truth order label.
    """
    # Concatenate video segments
    stitched_video = concatenate_videos(video1, video2)

    # Concatenate audio segments
    stitched_audio = concatenate_audios(audio1, audio2)

    # Combine video and audio into one clip
    stitched_clip = mux_video_audio(stitched_video, stitched_audio)

    return stitched_clip, label


def generate_cot_swap(video1, audio1, video2, audio2, modality_to_swap='audio'):
    """
    Swaps the temporal order of either video or audio while keeping the other unchanged.
    Returns the swapped multimodal clip and an anomaly label.
    """
    if modality_to_swap == 'audio':
        # Keep video order: video1 -> video2
        stitched_video = concatenate_videos(video1, video2)
        # Swap audio order: audio2 -> audio1
        swapped_audio = concatenate_audios(audio2, audio1)
    elif modality_to_swap == 'video':
        # Keep audio order: audio1 -> audio2
        stitched_audio = concatenate_audios(audio1, audio2)
        # Swap video order: video2 -> video1
        stitched_video = concatenate_videos(video2, video1)
    else:
        raise ValueError('Modality to swap must be audio or video')

    swapped_clip = mux_video_audio(stitched_video, stitched_audio) # Where mux_video_audio refers to the
        process of multiplexing  combining a video stream and an audio stream into a single
       synchronized media file.

    return swapped_clip, label
```

will lead to $\sim 25\%$ Top-1 accuracy for 4-choice questions. We notice the AVLLMs are prone to predict a certain choice more often introducing bias in the evaluation. Following [10] we feed each question N times to the AVLLMs where N is the number of choices by making a circular shift to the choices. We attribute the AVLLM to successfully solving a question if it correctly predicts the answer in all circular passes. Once an AVLLM fails in any of the passes there is no need to infer the remaining passes ensuring a good balance between model robustness and cost.

2.4. CircularEval vs. VanillaEval

We first compare the evaluation results under CircularEval (infer a question over multiple passes) with VanillaEval (infer a question only once) and report the average accuracy in Tab. 4 on AVTRUSTBENCH-test. We note, that for most AVLLMs switching from VanillaEval to CircularEval leads to a drop in model accuracy. In general, comparisons under CircularEval reveal a significant performance gap between different AVLLMs. The results as reported in Tab. 4 offer valuable insights, as we find the propensity in current AVLLMs to predict a certain choice when presented with a multiple-choice setup.

2.5. Human Evaluation

We manually selected 50 successful and 50 failed cases from the GPT-4o evaluation for each of the 9 tasks and conducted a manual assessment to estimate the upper bound of performance. The average accuracy we achieved was **91.27%**, suggesting that the designed tasks are synchronous to human cognition and are relatively straightforward for human subjects. This highlights the significant disparity between the current performance of the benchmark AVLLM and human capabilities.

3. Additional Results on Zero-Shot Evaluation

Considering 13 AVLLMs, we provide a leaderboard separately across all the task categories for AVTRUSTBENCH in Fig. 4. Furthermore, we provide additional results on zero-shot evaluations under *base* and *instruction* settings in Tabs. 5 - 7. We observe that for all the models the performance in the instruction setting improved considerably. However, the performance of these models is still far from satisfactory.

Choice extraction prompt for GPT-4

Can you help me match an answer with a set of options for a single correct answer type question? I will provide you with a question, a set of options, and a response from an agent. You are required to map the agent's response to the most similar option from the set. You should respond with a single uppercase character in 'A', 'B', 'C', 'D', and 'E' depending on the choice you feel is the most appropriate match. If there are no similar options you might output 'No match found'. Please refrain from being subjective while matching and do not use any external knowledge. Below are some examples:
Example 1:
Question: What color is the man's shirt who is sitting left of the object making this sound?
Options: A. Green B. Red C. Yellow D. Black
Answer: The person sitting next to the record player is wearing a black color shirt
Your output: D
Example 2:
Question: What does the audio-visual event constitute?
Options: A. A dog barking at a cat B. A dog barking on being hit by a stick C. The dog is hungry D. The dog is chasing another dog
Answer: It is a wolf
Your output: No match found

3.1. Comparison with different prompts.

In Tab. 8, we report results of zero-shot evaluation with Video-LLaMA2 on 8 additional prompts, for all the three dimensions of evaluation. We observe that the performance of the AVLLM is sensitive to the prompt used within considerable limits.

3.2. Performance differences on speech vs. non-speech audio classes.

We note that MacawLLM, which uses Whisper as their audio encoder performs relatively better under speech speech-only class due to its ability to encoder speech input better. Moroever, video-SALMONN was specifically trained for speech-related tasks and was also shown to perform better. The other models demonstrate comparable performance in both speech and non-speech classes. In Tab. 9, we report values for compositional reasoning tasks because it is the only suite containing data samples from speech (along with non-speech) classes.

4. Additional Details on Training

4.1. Under-represented categories.

We observe a non-uniformity in the distribution of categories across the AVQA and MUSIC-AVQA datasets. Such skewness leads to overemphasis of some categories on which the model's predictions are biased (as shown in Fig. 3). To mitigate such issues, we incorporate a robustness module in the proposed CAVPref (details in the main text).

s

4.2. Proof for the final objective of CAVPref.

Theorem 1. *Considering KL divergence as the discrepancy measure between Q and P, the closed-form objective becomes:*

$$\mathcal{L}_{\text{closed-form}} = -\lambda \log \left(\mathbb{E}_P \left[e^{\frac{\mathcal{L}}{\lambda}} \right] \right) \tag{1}$$

where λ is a regularization hyperparameter.

Proof. Considering the actual optimization problem:

$$\max_Q \mathbb{E}_Q[\mathcal{L}] : \mathbb{D}_{KL}(Q||P) \leq \rho \tag{2}$$

By method of Lagrangian multipliers, the problem becomes:

$$\max_Q \mathbb{E}_Q[\mathcal{L}] - \lambda(\mathbb{D}_{KL}(Q||P) - \rho) \tag{3}$$

Solving the saddle-point problem by taking partial derivative with respect to Q and equating it to 0, we obtain:

$$\frac{\partial}{\partial Q} \mathbb{E}_Q \left[\mathcal{L} - \lambda \log \frac{Q}{P} \right] = 0 \tag{4}$$
$$Q^* \propto P e^{\frac{\mathcal{L}}{\lambda}}$$

Since Q^* is a probability distribution, we obtain:

$$Q^* = \frac{P e^{\frac{\mathcal{L}}{\lambda}}}{Z} \tag{5}$$

where $Z = \mathbb{E}_P \left[e^{\frac{\mathcal{L}}{\lambda}} \right]$ is a normalizing factor or partition function.

Substituting Q^* back in the original objective, we obtain:

$$\mathbb{E}_Q[\mathcal{L}] = \sum Q^* \mathcal{L} = \sum \frac{P e^{\frac{\mathcal{L}}{\lambda}}}{Z} \mathcal{L} = \frac{1}{Z} \mathbb{E}_P \left[\mathcal{L} e^{\frac{\mathcal{L}}{\lambda}} \right] \tag{6}$$

Solving the dual problem by substituting the value of Q^*:

$$\mathbb{D}_{KL}(Q^*||P) = \lambda\rho \tag{7}$$
$$\mathbb{E}_{Q^*} \left[\log \frac{Q^*}{P} \right] = \lambda\rho$$
$$\mathbb{E}_{Q^*} [\frac{\mathcal{L}}{\lambda} - \log Z] = \lambda\rho$$
$$\frac{1}{\lambda} \mathbb{E}_{Q^*} [\mathcal{L}] - \log Z = \lambda\rho$$

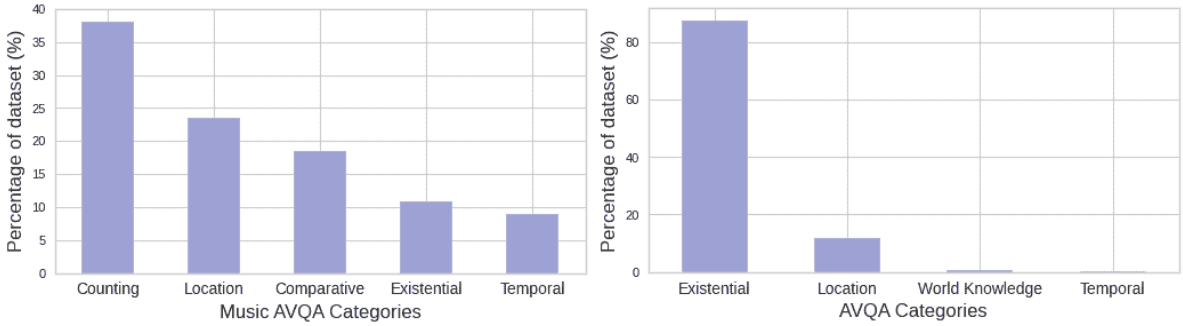


Figure 3. Distribution of different question categories across AVQA and MUSIC-AVQA datasets.

Therefore, the final closed-form objective is equivalent to minimizing:

$$\mathcal{L}_{\text{closed-form}} = -\lambda \log Z \tag{8}$$
$$\mathcal{L}_{\text{closed-form}} = -\lambda \log \left(\mathbb{E}_P \left[e^{\frac{\mathcal{L}}{\lambda}} \right] \right)$$

4.3. Simplification of the DPO objective.

DPO objective is given as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim P} \left[\log \sigma \left(\beta \log \left(\frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)} \right) - \beta \log \left(\frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)} \right) \right) \right] \tag{9}$$

Considering $f_w = \frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)}$ and $f_l = \frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)}$, putting $\sigma(x) = \frac{1}{1+\exp(-x)}$ the above equation can be rewritten and simplified as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim P} \left[\log \left(\frac{1}{1 + exp\left(-\log \left(\frac{f_w}{f_l} \right)^\beta \right)} \right) \right] \tag{10}$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim P} \left[\log \left(\frac{1}{1 + exp\left(\log \left(\frac{f_l}{f_w} \right)^\beta \right)} \right) \right]$$

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim P} \left[\log \left(\frac{1}{1 + \left(\frac{f_l}{f_w} \right)^\beta} \right) \right]$$

$$\mathcal{L}_{\text{DPO}} = \mathbb{E}_{(x,y_w,y_l)\sim P} \left[\log \left(1 + \left(\frac{f_l}{f_w} \right)^\beta \right) \right]$$

4.4. Pseudocode for CAVPref

The training pseudocode for CAVPref is shown in Algorithm 2. We employ a multimodal DPO formulation and update the objective functions as outlined below.

4.5. Results on other models

In Tab. 10 we compare the performance of 7 other open source models upon employing supervised finetuning (SFT), DPO, and CAVPref. Experimental results demonstrate a steady boost in performance upon applying CAVPref across all the models over all 9 tasks. We note that the highest performance gains are observed in the modality dependency suite - as our proposed approach guides the models to ingest modality-specific information thereby making a holistic inference.

4.6. Results on other benchmarks

We evaluate two different benchmarks, i.e., Video-Bench and MVBench before (zero-shot) and after training (following our proposed strategy - CAVPref) and report the values in Tab. 11 (using Video-LLaMA2). We observed substantial improvements with our proposed training paradigm.

5. Discussion on Bridging Networks

Bridge networks are modules used to connect the modality-specific encoders with the LLM by transforming the information from multi-modal encoders' space to LLM embedding space. For instance, VAST [2] uses text converters as the most basic and simplest bridge. Macaw-LLM uses a customized bridge network with linear layers and cross-attention-based alignment modules. VideoLLaMA(-2), Bay-CAT, video-SALMONN and X-InstructBLIP use Q-former-based bridge networks, whereas ChatBridge uses a customized perceiver network shared across all the modalities. OneLLM uses a mixture of projection experts equipped with a modality routing module, and ImageBind-LLM uses sophisticated trainable bind networks as the bridging module.

6. Performance with Different Model Variants

We experiment with the 7B and 13B variants of VideoLLaMA, PandaGPT, and X-InstructBLIP (other models employ a single variant). Experimental results confirm the

MCIT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**22.98**
2	Reka	21.18
3	Gemini 1.5 Pro	20.97
4	VideoLLaMA2	20.38
5	Bay-CAT	20.24
6	video-SALMONN	19.92
7	Unified-IO 2	19.87
8	AnyGPT	19.68
9	ImageBind-LLM	17.21
10	VideoLLaMA	13.1
11	OneLLM	12.06
12	X-InstructBLIP	11.42
13	ChatBridge	9.73
14	PandaGPT	8.24
15	Macaw-LLM	7.99
16	VAST	6.28

ICIT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**37.34**
2	Reka	36.41
3	Gemini 1.5 Pro	35.28
4	VideoLLaMA2	34.06
5	Bay-CAT	33.83
6	video-SALMONN	33.66
7	Unified-IO 2	32.93
8	AnyGPT	32.87
9	ImageBind-LLM	31.96
10	VideoLLaMA	27.41
11	OneLLM	25.01
12	X-InstructBLIP	24.85
13	ChatBridge	23.18
14	PandaGPT	22.71
15	Macaw-LLM	21.16
16	VAST	19.64

MVIT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**31.17**
2	Reka	29.76
3	Gemini 1.5 Pro	29.28
4	video-SALMONN	28.19
5	Bay-CAT	27.03
6	VideoLLaMA2	26.27
7	Unified-IO 2	25.74
8	ImageBind-LLM	25.61
9	AnyGPT	25.41
10	VideoLLaMA	20.23
11	OneLLM	18.78
12	X-InstructBLIP	17.59
13	ChatBridge	17.03
14	PandaGPT	15.87
15	Macaw-LLM	13.63
16	VAST	12.54

MAIT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**27.61**
2	Reka	27.05
3	Gemini 1.5 Pro	26.53
4	video-SALMONN	24.57
5	Bay-CAT	24.44
6	VideoLLaMA2	23.83
7	Unified-IO 2	23.31
8	AnyGPT	22.96
9	ImageBind-LLM	22.59
10	VideoLLaMA	17.81
11	OneLLM	16.28
12	X-InstructBLIP	15.6
13	ChatBridge	14.53
14	PandaGPT	13.47
15	Macaw-LLM	12.16
16	VAST	10.43

COT-Stitch

Rank	Model	Accuracy (%)
1	**GPT-4o**	**38.41**
2	Reka	37.46
3	Gemini 1.5 Pro	37.19
4	video-SALMONN	36.93
5	Bay-CAT	36.71
6	VideoLLaMA2	36.45
7	ImageBind-LLM	36.28
8	Unified-IO 2	35.95
9	AnyGPT	35.78
10	VideoLLaMA	35.24
11	OneLLM	33.55
12	X-InstructBLIP	32.57
13	ChatBridge	32.03
14	PandaGPT	31.94
15	Macaw-LLM	30.66
16	VAST	25.19

COT-Swap

Rank	Model	Accuracy (%)
1	**Gemini 1.5 Pro**	**30.69**
2	GPT-4o	30.66
3	Reka	30.63
4	VideoLLaMA2	30.52
5	Bay-CAT	30.41
6	VideoSALMONN	30.37
7	Unified-IO 2	30.17
8	ImageBind-LLM	30.09
9	AnyGPT	30.05
10	VideoLLaMA	29.81
11	OneLLM	29.45
12	Macaw-LLM	27.35
13	ChatBridge	27.32
14	PandaGPT	26.44
15	X-InstructBLIP	26.18
16	VAST	25.52

CAT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**31.52**
2	Bay-CAT	30.77
3	Reka	30.75
4	VideoLLaMA2	30.59
5	VideoSALMONN	30.48
6	ImageBind-LLM	30.45
7	Gemini 1.5 Pro	30.37
8	VideoLLaMA	30.33
9	Unified-IO 2	30.21
10	OneLLM	30.03
11	AnyGPT	29.97
12	PandaGPT	29.42
13	X-InstructBLIP	29.35
14	ChatBridge	28.92
15	Macaw-LLM	28.47
16	VAST	25.11

MVT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**52.5**
2	Reka	51.98
3	Gemini 1.5 Pro	51.59
4	VideoSALMONN	49.94
5	Bay-CAT	48.81
6	VideoLLaMA2	47.64
7	Unified-IO 2	46.55
8	AnyGPT	45.98
9	ImageBind-LLM	45.33
10	Video LLaMA	42.08
11	One LLM	40.2
12	X-InstructBLIP	39.31
13	ChatBridge	38
14	Macaw-LLM	36.46
15	PandaGPT	36.05
16	VAST	30.4

MAT

Rank	Model	Accuracy (%)
1	**GPT-4o**	**49.15**
2	Reka	48.62
3	Gemini 1.5 Pro	47.43
4	VideoSALMONN	46.55
5	Bay-CAT	45.16
6	Unified-IO 2	43.89
7	VideoLLaMA2	43.8
8	AnyGPT	43.1
9	ImageBind-LLM	41.9
10	Video LLaMA	38.76
11	One LLM	36.21
12	X-InstructBLIP	35.87
13	ChatBridge	34.78
14	Macaw-LLM	33.13
15	PandaGPT	32.85
16	VAST	26.44

Figure 4. Leaderboards for zero-shot evaluation on 9 different tasks in AVTRUSTBENCH.

Algorithm 2 PyTorch-style pseudocode for CAVPref.

```
# pi_yw_logps: winning response logprobs (policy)
# pi_yl_logps: losing response logprobs (policy)

# pi_yw_Vw_logps: winning response with correct visual logprobs (policy)
# pi_yw_Vl_logps: winning response with incorrect visual logprobs (policy)

# pi_yw_Aw_logps: winning response with correct audio logprobs (policy)
# pi_yw_Al_logps: winning response with incorrect audio logprobs (policy)

# ref_yw_logps: winning response logprobs (reference model)
# ref_yl_logps: losing response logprobs (reference model)

# ref_yw_Vw_logps: winning response with correct visual logprobs (reference model)
# ref_yw_Vl_logps: winning response with incorrect visual logprobs (reference model)

# ref_yw_Aw_logps: winning response with correct audio logprobs (reference model)
# ref_yw_Al_logps: winning response with incorrect audio logprobs (reference model)

# beta_y, beta_V, beta_A: policy regularization coefficients

# lambda_y, lambda_V, lambda_A: robustness coefficients

def CAVPref:
   # linguistic component (Eq. 1)
   pi_logratios_y = pi_yw_logps - pi_yl_logps
   ref_logratios_y = ref_yw_logps - ref_yl_logps

   loss_y = F. logsigmoid ( beta_y * ( pi_logratios - ref_logratios ))

   # visual component (Eq. 2)
   pi_logratios_V = pi_yw_Vw_logps - pi_yw_Vl_logps
   ref_logratios_V = ref_yw_Vw_logps - ref_yw_Vl_logps

   loss_V = F. logsigmoid ( beta_V * ( pi_logratios_V - ref_logratios_V ))

   # audio component (Eq. 3)
   pi_logratios_A = pi_yw_Aw_logps - pi_yw_Al_logps
   ref_logratios_A = ref_yw_Aw_logps - ref_yw_Al_logps

   loss_A = F. logsigmoid ( beta_A * ( pi_logratios_A - ref_logratios_A ))

   # Eqs. 5 and 6 combined
   CAVPref_loss = - (lambda_y * torch.log(torch.mean(torch.exp(loss_y / lambda_y))) +
       lambda_V * torch.log(torch.mean(torch.exp(loss_V / lambda_V))) + lambda_A * torch.log
       (torch.mean(torch.exp(loss_A / lambda_A))))

   return CAVPref_loss
```

performance boost with the 13B variants. A key observation is increasing the model size from 7B to 13B doesn't help in obtaining significant gain in Compositional reasoning suite of tasks. We hypothesize that LLMs are not able to capture the attribute level binding information and often work as bag-of-word models. Tab. 12 compares the two variants of the above-mentioned models.

7. More Related Works

Audio-Visual QA datasets. Deep learning for video QA relies on diverse datasets such as MSRVTT-QA [16], and ActivityNet-QA [1]. MovieQA [15] and TVQA [8] add to the diversity of available scenario-specific datasets in this space. However, these datasets often focus on specific tasks and cannot amply evaluate the comprehensive reasoning capabilities of AVLLMs. Moreover, the majority of these datasets do not contain meaningful audio and QA pairs encompassing cross-modal understanding. To this end, we leverage three public audio-visual datasets AVQA [17], MUSIC-AVQA [9] and AudioSet [6] to form the QA pairs for all our tasks. These datasets can facilitate study on spatio-temporal reasoning for dynamic and long-term audio-visual scenes, complex audio-visual reasoning, multi-modal perception and granularity (*existential*, *location*, *counting* etc.). In the face of a massive deluge of MLLMs, there is an acute shortage of benchmarks that can extensively evaluate the trustworthiness of these models. Our presented AVTRUST-

BENCH can bridge this gap by serving as a testbed to evaluate different dimensions of these models such as cross-modal comprehension, reasoning, and perception abilities.

8. Implementation Details

For open-source models, we follow their default best inference settings and hyperparameters. To evaluate GPT-4o, Gemini 1.5 Pro we utilize their official APIs. Full videos are directly passed to Gemini 1.5 Pro, as its API (using Google Cloud vertexai framework) inherently supports video inputs. For each model under evaluation, we generate responses to the questions independently and without retaining the chat history. For evaluating all open-source AVLLMs on AVTRUSTBENCH tasks, we use 1 A100 GPU. For training the open-source AVLLMs on AVTRUSTBENCH tasks, we utilize 8 A100 GPUs and follow their respective training implementation details.

9. Common Sense Reasoning

Fig. 14 shows that the current AVLLMs *lack* commonsense reasoning. There is evidence in animal study [7] that it is a natural tendency of a dog to bark at an unknown cat. In this example (refer to video 7min 50sec) most AVLLMs fail to infer this and opts for incorrect response underlying their lack of commonsense reasoning skills.

10. More qualitative Examples

We share more qualitative samples from each task in Fig. 5 - 13. As can be seen, closed-source models demonstrate an overall better performance compared to open-source counterparts with GPT-4o being the strongest performer across the majority of the tasks. We note that upon employing CAVPref, the responses of the AVLLMs improve as they tend to make fewer mistakes on the same QA pairs - which underlines the effectiveness of our proposed approach over DPO.

11. Limitations and Future Work

Although **CAVPref** incorporates AV associations, it is essentially a preference-based optimization strategy and is therefore sensitive to the quality of preference data. Moreover, it is yet to be tested whether such an approach can yield promising results for other axes of evaluation and/or fine-grained tasks. AVTRUSTBENCH currently contains coarse-grained samples e.g., QA tasks. Future work can extend this for detection/segmentation.

12. Failure Cases

Fig. 15 illustrates the failure cases of our mitigation approach CAVPref while used with video-SALMONN, Video-LLaMA2, and Bay-CAT. In the first case, the models are unable to differentiate between 'violin' in the video and 'viola' in the audio since they are semantically closely associated. Therefore, although this is a task of MVIT, the models are unable to pick the correct answer, i.e., '(E) None of the above'. In the second case, the models are unable to see the speaker (on the left) who is facing their back (i.e., their face is not visible). Therefore, they are unable to understand that the correct answer, i.e., 'left' which is not present in the set of options (MCIT task), and thus the ideal response would be '(E) None of the above'.

13. Supplementary Video Examples

In the supplementary video, we add qualitative examples for each of the tasks of AVTRUSTBENCH for each model. We find the MLLMs to produce free-form responses on many occasions. We employ our two-stage choice extraction strategy as explained in Sec. 2.2 to obtain the AVLLMs responses and process them accordingly. The use of headphones is recommended for a better audio-visual QA experience.

14. Societal Impact

In this work, we perform an extensive analysis of existing state-of-the-art AVLLMs to study their failure modes. Our study reveals that models lack sufficient audio-visual comprehension skills and most often fail to address scenarios that require common sense reasoning. We believe our work can be useful to the community and our findings can reveal the potential threats associated with deploying these models in real-time or accuracy-critical setups. The users must recognize these limitations in the new generation models and proceed with caution, especially in scenarios where the precision and neutrality of results hold significant importance. Users are encouraged to thoroughly scrutinize and validate the outputs of the model to avoid the possibility of disseminating inaccurate information. We employ the existing public datasets to curate the benchmark and we don't collect or use any personal/human subject data without their consent during our data preparation and experiments stages.

15. Human Study Details

We conducted a small study involving 20 individuals to assess the difficulty of our proposed benchmark and estimate the upper bound for the tasks proposed. The user study protocol was approved by the Institutional Review Board and we do not collect, share or store any personal information of the participants.

15.1 Data Collection and Quality Control

We form Audio-Visual QAs in the format of multiple-choice problems for each task. A problem P_i corresponds to $(Q_i, C_i, V_i, A_i, R_i)$. Q_i denotes the question, C_i represents a set with $n (2 \leq n \leq 5)$ choices $c_1, c_2, \ldots, c_n$, V_i, and A_i represents the input video and the audio respectively, and

R_i is the correct response. The number of choices varies depending on the task. For each task, we first prepare up to ~ 5 different question templates to ensure sufficient variations in the question formats. We carefully choose the questions from one of these templates. We add more details on the QA pair formation in the supplementary.

We collect the AV samples from benchmark datasets AVQA, MUSIC-AVQA, and AudioSet. While the QA pairs for AVQA and MUSIC-AVQA are adapted directly from those datasets, for AudioSet we obtain the QA pairs from a pre-designed template (Tab. 1). Finally, while forming the mismatched pairs, we follow a semi-automated (heuristics + look-up table) approach. We apriori create a dictionary of mismatched pairs by careful manual inspection to ensure that the corresponding audio-visual pairs have no association between them. To further validate, we manually investigate randomly chosen 500 samples from each of the axes of evaluation. We compute the spearman correlation coefficient between the human labels and our curated data on those samples and we obtain a mean score of 0.979 ($p < 0.05$) - indicating a significantly strong correlation.

Kindly note that samples from AudioSet were only collected for the compositional understanding tasks. For the adversarial attack and missing modality tasks, the samples are curated from the AVQA and the MUSIC-AVQA datasets. Moreover, employing AudioSet for both fine-grained and coarse-grained audio-visual tasks has been explored by the community [3, 11, 12].

AudioSet contains real-world samples under in-the-wild settings where we ensure that the constituent modalities (audio and visual) are aligned by adhering to the following strategy. We utilize the CLIP [13] and CLAP [5] scores by calculating $T_{\text{sim}} = \mathcal{S}_{\text{CLIP}}\mathcal{S}_{\text{CLAP}}^{T}$, where $\mathcal{S} \in \mathbb{R}^{N \times N}$ and denotes the pairwise cross-modal similarity scores for a batch of size N. The CLIP similarity is calculated between the chosen visual and the audio class label, similarly, the CLAP score is calculated between the audio class label and the audio snippet. The text modality acts as the bridging modality in this case. Note the range of the scores is normalized between [0,1] with 0 being the lowest. We don't consider the samples having a T_{sim} score of less than 0.70 to ensure a strong association between the two modalities. Notably, CLIP + CLAP based selection approach has been employed and accepted in the audio-visual community in recent literature [3, 4].

16. Performance on Fine-grained task

Since our adversarial setup evaluates AV consistency and robustness, it can be regardless of task granularity (we used a QA setup for simplicity). Video LLaMA-2 fine tuned with CAVPref shows promising results (Fig. 16) with zero-shot evaluation on fine-grained tasks.

17. Comparison with Other Preference Optimization Techniques

DPO, Sim-PO, GPO, β-DPO focus only on textual preferences, and *cannot handle unique challenges in AV settings*. Whereas, CAVPref **(1)** incorporates *modality-specific losses to ensure proper conditioning on AV inputs* (Eq. 2-3), addressing issues of modality absence, inconsistency and compositionality, thereby improving adversarial robustness, and optimizing preferences specific to AV tasks (Tab. 5), **(2)** integrates a *dynamic robustness mechanism* to enhance tail performance and *mitigate noisy or underrepresented preferences* (Eq. 4-5, Supp. Sec. D.3), **(3)** adaptively calibrates scaling factor for improved preference optimization using modality-specific metrics (e.g., AVSM, CLAP). Thus, direct comparisons of CAVPref with them are not meaningful. We deliberately chose DPO as the primary baseline because it serves as a foundational framework that is conceptually extendable to AV settings. Rather than a minor extension, **CAVPref fills a critical gap in AV alignment methodologies**, beyond the scope of existing uni-modal (text-based) preference optimization techniques. That said, we show an example where we further extend CAVPref within the frameworks of Sim-PO and β-DPO (note that GPO does not have a public codebase) and we outperform both of them (Tab 13).

18. Evaluation on Smaller Testset

We curated a 10k-sample subset (for labs with limited compute), preserving category-wise distribution from the original test set. We note similar performance trends on this split (Tab. 14), which we will release in the final version.

19. Bias Mitigation

While our primary focus is designing a robust framework for AV preference optimization, we recognize the importance of mitigating biases in data and ensuring generalizability. To address this, CAVPref includes a distributional robustness module to minimize worst-case risks and reduce the impact of biased or underrepresented data. In future, we plan to explore explicit bias-mitigation strategies.

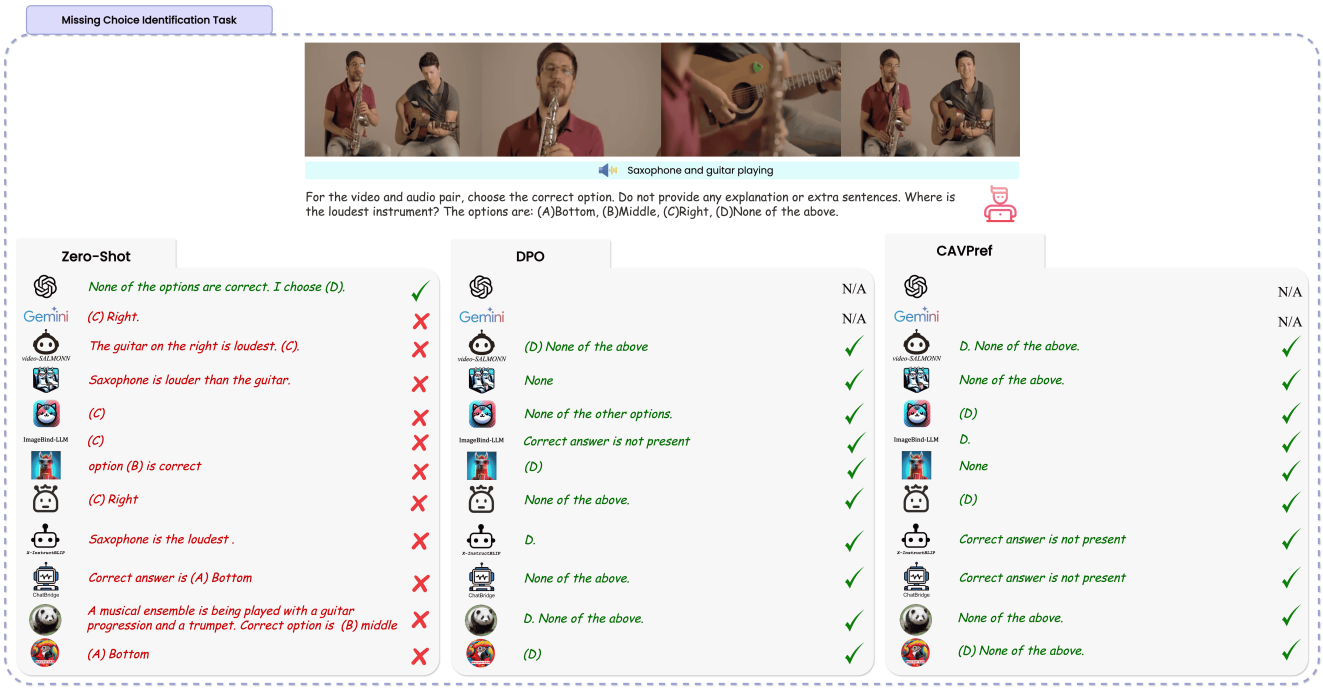


Figure 5. Performance comparison of all open source models on MCIT task under ZS, DPO and CAVPref.

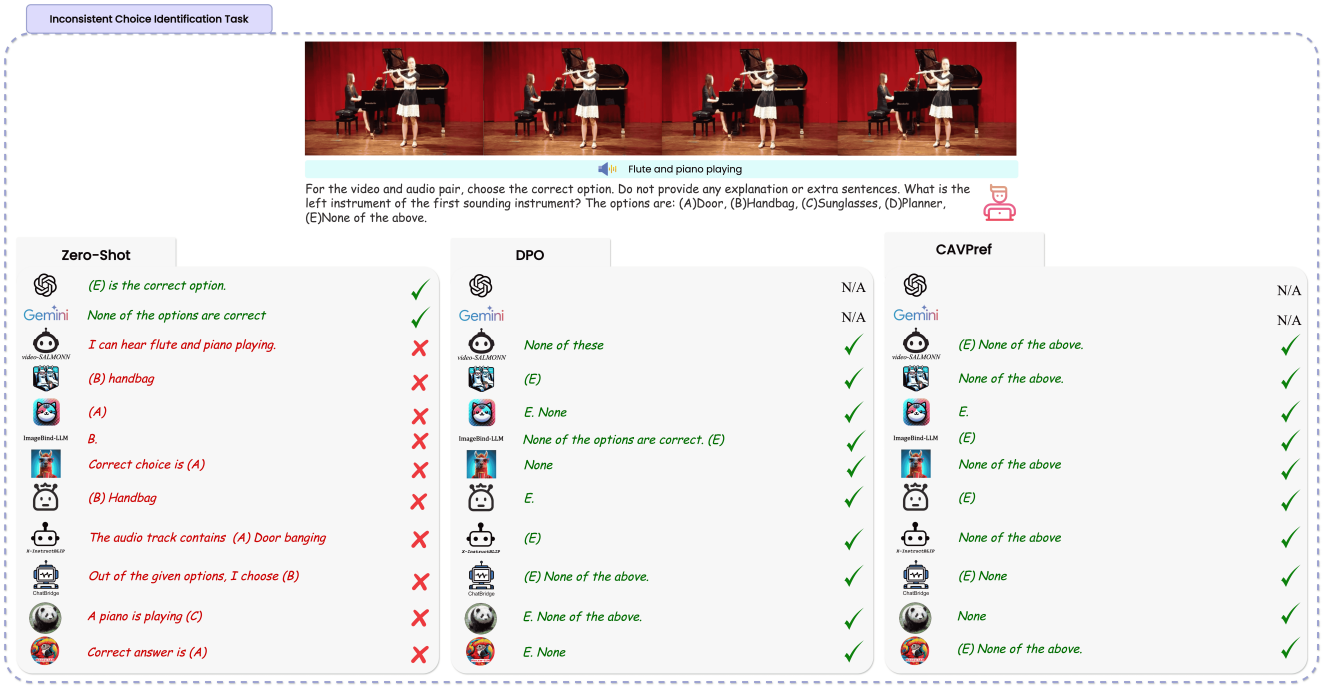


Figure 6. Performance comparison of all open source models on ICIT task under ZS, DPO, and CAVPref.

Dimension	Task	Sample Question with Options
Adversarial Attack	MCIT	Is the ukulele on the left more rhythmic than the saxophone on the right? A. Yes B. No
		Is the instrument on the right louder than the instrument on the left? A. Yes B. No
		How many sounding erhu in the video? A. Five B. Six C. More than ten D. Three E. None of the above
		Where is the lowest instrument? A. Guzheng B. Middle C. Bagpipe D. Right E. None of the above
		What are the main sources of sound in the video? A. Sound of wind B. Water flow sound C. Using a sewing machine D. None of the above
	ICIT	Is the instrument on the right louder than the instrument on the left? A. Napkin B. Container C. Calculator D. Stool E. None of the above
		Is the first sound coming from the middle instrument? A. Book B. Chair C. Wok D. Tree E. None of the above
		Is the xylophone in the video always playing? A.Blanket B. Cloud C. Computer D. Door E. None of the above
		Is the flute in the video more rhythmic than the cello? A.Calculator B. Statue C. Rag D. Kiln E. None of the above
		Is there a voiceover? A. Table B. Stapler C. Bag D. Blanket
	MAIT	Is the first sound coming from the middle instrument? A. Yes B. No
		Is the instrument on the right louder than the instrument on the left? A. Yes B. No
		Is the xylophone in the video always playing? A. Yes B. No
		Where is the performance? A. Tube B. Trumpet C. Flute D. Indoor E. None of the above
		What is the first instrument that comes in? A. Pipa B Trumpet C. Congas D. Violin
	MVIT	Is the saxophone in the video always playing? A. Yes B. No
		Is the instrument on the right louder than the instrument on the left? A. Yes B. No
		Which is the musical instrument that sounds at the same time as the pipa? A. Flute B. Guzheng C. Middle D. Acoustic guitar E. None of the above
		How many sounding flute in the video? A. Zero B. Three C. No D. One
		Is the clarinet on the right louder than the accordion on the left? A. Yes B. No
Compositional Reasoning	COT-Stitch	What is the sequence of events in the video? A. Speech is followed by Meow. B. Meow is followed by Speech. C. Both of them occur at the same time D. Toilet flush is followed by Toilet flush.
		What is the sequence of events in the video? A. Speech is followed by Meow. B. Meow is followed by Speech C. Both of them occur at the same time D. Whistle is followed by Helicopter.
	COT-Swap	What is the sequence of events in the video? A. Speech is followed by Meow. B. Meow is followed by Speech. C. Both of them occur at the same time. D. Ambulance (siren) is followed by Music. E. None of the above
		What is the sequence of events in the video? A. Speech is followed by Meow. B. Meow is followed by Speech. C. Both of them occur at the same time. D. Doorbell is followed by Moo. E. None of the above
	CAT	What is the sequence of events in the video? A. A crowd cheers and a man speaks. B. A crowd speaks and a man cheers. C. Door followed by book
		What is the sequence of events in the video? A. A man is speaking, and a crowd applauds. B. A man is applauding, and a crowd speaks. C. Boots followed by Ring.
Missing Modality	MAT	How many types of musical instruments sound in the video? A. Seven B. No C. Three D. Two E. None of the above
		Is there a voiceover? A. Yes B. No
		Which is the musical instrument that sounds at the same time as the violin? A. Suona B. Trumpet C. Middle D. Accordion E. None of the above
		Is the instrument on the right more rhythmic than the instrument in the middle? A. Yes. B. No
	MVT	How many sounding flute in the video? A. Zero B. Three C. No D. One E. None of the above
		Is the instrument on the left louder than the instrument on the right? A. Yes B. No
		Is the first sound coming from the left instrument? A. Yes B. No
		What is the first instrument that comes in? A. Acoustic guitar B. Congas C. Banjo D. Violin

Table 1. Task-wise sample templates with potential options.

MSR-VTT	LUMA	SSV2	AVTRUSTBENCH
20	50	174	**377**

Table 2. Comparison of various benchmarks with AVTRUSTBENCH on number of categories.

Task	Video LLaMA	Macaw-LLM	PandaGPT	ChatBridge	X-InstructBLIP	One LLM	VAST	ImageBind- LLM	Gemini 1.5 Pro	VideoLLaMA 2	Bay-CAT	video-SALMONN	GPT-4o
MCIT	13.1 / 15.9	7.99 / 10.94	8.24 / 10.98	9.73 / 12.63	11.42 / 12.62	12.06 / 13.8	6.28 / 8.28	17.21 / 19.76	20.38 / 22.31	20.24 / 22.19	19.92 / 21.09	20.97 / 22.06	**22.98 / 25.93**
ICIT	27.41 / 28.74	21.16 / 23.48	22.71 / 23.9	23.18 / 24.54	24.85 / 27.68	25.01 / 26.3	19.64 / 21.55	31.96 / 34.91	34.06 / 35.32	33.83 / 35.19	33.66 / 35.92	35.28 / 38.11	**37.34 / 40.33**
MVIT	20.23 / 22.12	13.63 / 16.33	15.87 / 17.61	17.03 / 19.38	17.59 / 20.08	18.78 / 21.22	12.54 / 15.16	25.61 / 26.64	26.27 / 28.68	27.03 / 28.71	28.19 / 30.25	29.28 / 30.9	**31.17 / 33.39**
MAIT	17.81 / 20.35	12.16 / 13.16	13.47 / 14.99	14.53 / 16.94	15.6 / 17.72	16.28 / 18.86	10.43 / 12.87	22.59 / 23.77	23.83 / 26.53	24.44 / 25.84	24.57 / 27.31	26.53 / 29.39	**27.61 / 30.43**
COT-Stitch	35.24 / 36.86	30.66 / 32.69	31.94 / 34.15	32.03 / 33.82	32.57 / 34.34	33.55 / 36.41	25.19 / 27.48	36.28 / 38.13	36.45 / 37.98	36.71 / 37.98	36.93 / 39.03	37.19 / 39.62	**38.41 / 40.59**
COT-Swap	29.81 / 31.47	27.35 / 30.14	26.44 / 28.17	27.32 / 29.83	26.18 / 27.24	29.45 / 31.14	25.52 / 28.25	30.69 / 32.1	30.52 / 33.36	30.41 / 32.96	30.37 / 32.15	30.69 / 32.22	**30.66 / 31.72**
CAT	30.33 / 32.05	28.47 / 31.46	29.42 / 31.73	28.94 / 31.29	29.35 / 30.4	30.35 / 32.62	25.11 / 27.63	30.45 / 31.88	30.59 / 33.43	30.77 / 32.12	30.48 / 32.87	30.37 / 32.29	**31.52 / 33.14**
MVT	42.08 / 44.16	36.46 / 39.4	36.05 / 38.77	38.2 / 41.05	39.31 / 41.66	40.2 / 42.29	30.4 / 31.86	45.33 / 47.88	47.64 / 49.98	48.81 / 50.07	49.94 / 51.27	51.59 / 54.05	**52.5 / 55.36**
MAT	38.76 / 40.93	33.13 / 34.14	32.85 / 34.03	34.78 / 36.21	35.87 / 37.23	36.21 / 38.63	26.44 / 28.7	41.9 / 43.93	43.8 / 45.79	45.16 / 47.72	46.55 / 49.03	47.43 / 48.97	**49.15 / 50.39**

Table 4. Average accuracy of each model in Circular vs Vanilla Evaluation (given as Circular / Vanilla values).

Category	Video LLaMA	Macaw-LLM	PandaGPT	ChatBridge	X-InstructBLIP	OneLLM	VAST	ImageBind-LLM	Gemini 1.5 Pro	VideoLLaMA 2	Bay-CAT	video-SALMONN	GPT-4o
Missing Choice Identification Task (MCIT)													
Existential	1.54	0.32	0.44	0.63	0.73	0.77	0.12	1.90	8.76	3.11	3.18	2.56	**10.03**
Localization	0.63	0.21	0.27	0.41	0.48	0.37	0.0	0.98	5.83	1.62	2.09	1.35	**6.12**
Temporal	0.55	0.36	0.35	0.53	0.56	0.49	0.01	1.20	3.11	1.61	1.78	2.20	**4.18**
World knowledge	0.94	0.67	0.76	0.91	0.98	0.98	0.09	1.35	6.18	2.96	2.65	1.98	**6.95**
Inconsistent Choice Identification Task (ICIT)													
Existential	3.24	2.44	2.94	2.57	4.32	3.26	1.28	4.85	11.03	5.98	5.45	5.61	**12.15**
Localization	3.17	2.19	2.86	2.99	3.51	3.24	0.88	4.78	9.14	5.96	5.11	5.67	**9.16**
Temporal	4.14	3.13	3.82	4.92	2.05	2.98	0.46	5.23	5.62	5.27	5.31	5.40	**5.91**
World knowledge	4.49	3.39	2.82	3.16	3.60	3.48	0.65	4.57	9.06	5.78	5.91	6.22	**9.42**
Mismatched Video Identification Task (MVIT)													
Existential	4.88	4.43	5.11	3.81	5.98	4.95	3.34	6.73	14.33	7.11	7.23	7.97	**14.82**
Localization	5.27	3.78	4.77	3.94	5.72	4.62	2.75	5.80	11.50	6.95	6.10	7.26	**12.11**
Temporal	5.94	4.86	5.27	6.56	3.89	3.36	2.45	6.66	6.16	6.71	6.18	6.95	**7.20**
World knowledge	6.58	3.96	3.76	4.93	5.64	4.97	2.90	5.82	12.53	5.97	6.42	6.55	**15.12**
Mismatched Audio Identification Task (MAIT)													
Existential	3.71	3.29	3.91	2.93	4.96	3.68	2.11	5.45	12.85	7.11	6.28	7.21	**13.05**
Localization	3.46	2.64	3.42	2.71	3.58	3.74	1.49	4.11	9.78	5.89	5.62	6.24	**10.12**
Temporal	4.89	3.98	4.19	3.94	3.81	2.79	1.04	4.50	5.92	4.98	4.65	5.11	**6.23**
World knowledge	5.33	2.84	2.32	3.76	4.22	3.92	1.13	5.31	9.57	5.76	5.92	5.98	**10.11**

Table 5. **Zero shot evaluation results of AVLLMs under Adversarial attack suite on AVQA dataset under *base* setting**. Models are required to demonstrate strong audio-visual comprehension capabilities to withhold answers when presented with perturbed questions/answers/input signals.

Category	Video LLaMA	Macaw-LLM	PandaGPT	ChatBridge	X-InstructBLIP	One LLM	VAST	ImageBind- LLM	Gemini 1.5 Pro	VideoLLaMA 2	Bay-CAT	video-SALMONN	GPT-4o
Missing Video Identification Task (MVT)													
Existential	7.58	5.24	6.31	6.27	6.36	6.44	3.59	9.25	12.48	10.65	11.51	10.97	**13.77**
Localization	4.22	2.30	2.20	3.51	4.43	3.27	2.42	6.50	8.74	6.91	7.01	7.13	**9.22**
Count	4.46	2.35	2.88	2.21	1.78	2.99	1.97	5.56	8.48	6.88	6.13	6.45	**10.08**
Temporal	3.37	2.23	3.36	3.46	3.15	3.67	2.76	3.44	6.19	4.98	4.87	4.91	**7.55**
Comparison	8.23	5.62	6.04	6.26	7.61	7.58	3.78	8.77	12.28	9.87	9.91	8.72	**12.96**
Missing Audio Identification Task (MAT)													
Existential	6.39	4.56	4.78	5.54	5.98	5.21	2.70	7.17	8.24	7.54	7.23	7.98	**9.06**
Localization	3.71	1.54	1.88	2.04	2.35	2.98	1.04	5.03	7.57	7.54	7.23	8.11	**8.95**
Count	3.29	1.08	1.73	1.79	2.56	2.75	0.79	4.24	7.13	6.56	5.12	7.11	**8.78**
Temporal	2.51	1.65	2.13	2.36	2.81	2.49	1.35	2.90	3.46	2.98	3.02	3.11	**3.67**
Comparison	7.71	4.84	5.34	5.72	6.26	6.91	2.47	7.46	9.84	8.52	8.76	9.03	**10.15**

Table 6. Comparison of zero-shot evaluation results on Modality-specific dependency suite for MUSIC-AVQA dataset under *base* setting.

Category	Video LLaMA	Macaw-LLM	PandaGPT	ChatBridge	X-InstructBLIP	OneLLM	VAST	ImageBind-LLM	Gemini 1.5 Pro	VideoLLaMA 2	Bay-CAT	video-SALMONN	GPT-4o
Missing Choice Identification Task (MCIT)													
Existential	1.16 / 26.72	0.31 / 14.22	0.41 / 15.34	0.62 / 16.65	0.79 / 21.59	0.73 / 23.30	0.19 / 12.36	1.45 / 27.38	8.10 / 31.88	4.12 / 29.58	5.01 / 30.01	3.62 / 30.18	**10.61 / 33.96**
Localization	0.59 / 10.26	0.27 / 7.99	0.29 / 7.96	0.40 / 8.44	0.53 / 9.80	0.39 / 9.88	0.21 / 7.22	0.97 / 13.14	5.55 / 19.39	2.16 / 16.51	3.96 / 18.11	2.67 / 18.76	**7.41 / 21.90**
Temporal	0.51/5.29	0.39 / 3.31	0.38 / 5.42	0.57 / 6.27	0.53 / 5.90	0.57 / 4.90	0.13 / 1.20	1.16 / 11.66	3.00 / 12.44	1.91 / 10.91	2.61 / 11.42	1.99 / 11.20	**5.91 / 14.93**
Count	0.82/7.10	0.65 / 4.35	0.77 / 5.45	1.04 / 7.36	0.84 / 7.87	0.95 / 7.51	0.20 / 3.78	1.27 / 13.70	6.02 / 17.10	3.61 / 15.71	4.89 / 15.98	3.90 / 14.64	**8.11 / 19.61**
Comparative	1.41 / 27.28	0.48 / 15.65	0.56 / 17.89	0.85 / 18.33	0.91 / 23.57	0.85 / 26.72	0.30 / 14.80	3.56 / 31.76	11.34 / 34.48	6.57 / 32.86	7.11 / 32.67	6.42 / 32.19	**12.91 / 36.75**
Inconsistent Choice Identification Task (ICIT)													
Existential	3.43 / 40.33	2.40 / 28.38	2.96 / 26.91	3.01 / 32.65	3.51 / 37.59	3.65 / 39.11	1.12 / 25.19	4.11 / 42.36	9.57 / 48.89	5.82 / 44.85	6.01 / 46.48	5.42 / 45.53	**10.13 / 49.65**
Localization	3.12 / 27.11	2.02 / 22.61	2.11 / 23.01	2.82 / 21.88	3.24 / 22.96	3.21 / 24.18	0.49 / 18.42	4.05 / 28.78	9.31 / 32.06	6.15 / 29.18	6.89 / 29.64	5.92 / 28.57	**10.76 / 34.66**
Temporal	2.98 / 20.27	2.38 / 13.88	2.52 / 18.87	2.91 / 19.92	2.97 / 20.05	3.28 / 14.85	0.41 / 14.16	3.92 / 27.10	6.12 / 28.14	4.95 / 27.61	4.68 / 27.67	4.15 / 27.11	**7.44 / 30.61**
Count	3.13 / 21.76	2.76 / 18.54	2.79 / 20.42	3.06 / 21.03	3.21 / 20.83	3.09 / 24.62	0.67 / 18.80	3.86 / 26.24	9.02 / 32.55	5.64 / 28.55	5.98 / 29.41	5.75 / 29.62	**11.41 / 34.56**
Comparative	4.31 / 43.54	3.16 / 29.67	3.09 / 28.26	4.15 / 34.32	3.89 / 39.44	4.41 / 40.66	1.98 / 27.22	6.78 / 44.63	11.45 / 50.90	7.23 / 46.75	8.11 / 47.11	7.86 / 49.17	**12.71 / 51.89**
Mismatched Video Identification Task (MVIT)													
Existential	4.20 / 34.80	4.03 / 22.36	5.90 / 22.14	3.64 / 26.27	5.66 / 30.37	4.48 / 30.58	3.30 / 18.27	6.47 / 37.93	13.98 / 39.77	8.42 / 38.42	8.77 / 38.91	8.18 / 39.11	**15.71 / 41.02**
Localization	5.42 / 15.33	3.31 / 11.39	5.34 / 13.48	3.38 / 14.34	5.21 / 14.91	4.56 / 16.31	2.04 / 13.56	5.98 / 20.00	11.28 / 25.25	6.11 / 21.84	6.87 / 21.91	6.45 / 20.96	**12.88 / 27.60**
Temporal	5.34 / 12.80	4.28 / 8.72	5.69 / 12.60	6.16 / 12.14	4.20 / 10.58	3.79 / 10.46	3.20 / 7.90	6.47 / 18.28	6.70 / 22.97	5.57 / 16.51	5.94 / 16.68	5.13 / 17.41	**7.19 / 23.96**
Count	6.12 / 14.28	4.62 / 12.19	4.65 / 15.14	5.75 / 14.73	5.40 / 11.03	4.24 / 17.20	2.42 / 11.25	5.49 / 21.74	12.01 / 26.20	8.32 / 22.76	8.67 / 23.13	7.18 / 23.57	**13.87 / 27.96**
Comparative	4.47 / 35.87	5.12 / 24.46	6.11 / 23.88	3.96 / 27.90	6.17 / 32.39	4.79 / 32.51	4.04 / 19.32	7.43 / 38.67	14.28 / 41.34	9.65 / 39.87	9.88 / 39.29	8.74 / 38.56	**16.41 / 42.98**
Mismatched Audio Identification Task (MAIT)													
Existential	4.68 / 31.51	3.88 / 20.67	3.47 / 21.77	2.52 / 24.24	4.62 / 28.20	3.63 / 28.35	2.35 / 15.51	5.21 / 34.34	13.61 / 38.29	6.75 / 35.78	7.42 / 36.17	6.57 / 35.57	**15.08 / 39.46**
Localization	3.15 / 13.44	2.03 / 9.37	4.21 / 12.48	2.33 / 11.03	4.36 / 14.36	3.37 / 14.00	1.03 / 11.18	4.36 / 17.76	10.38 / 23.22	6.44 / 21.76	7.12 / 22.58	6.38 / 21.69	**11.32 / 24.89**
Temporal	4.32 / 11.68	3.46 / 5.46	4.77 / 9.50	5.70 / 9.89	2.27 / 8.67	2.84 / 7.25	1.56 / 4.44	5.90 / 17.92	5.66 / 19.72	5.61 / 18.71	5.65 / 19.02	5.43 / 18.76	**7.11 / 19.89**
Count	5.88 / 13.00	2.97 / 9.39	2.53 / 12.01	4.30 / 11.27	4.72 / 10.46	3.41 / 14.76	1.83 / 8.21	4.44 / 19.95	10.28 / 24.82	6.96 / 21.67	7.24 / 22.71	6.34 / 20.98	**12.16 / 22.58**
Comparative	4.92 / 33.90	4.56 / 22.72	3.77 / 22.32	3.15 / 26.29	5.27 / 29.81	4.45 / 30.77	2.78 / 17.99	6.37 / 37.75	15.29 / 41.66	8.16 / 39.58	8.78 / 39.90	7.61 / 38.66	**16.11 / 42.71**

Table 7. **Zero shot evaluation results of AVLLMs under Adversarial attack suite on MUSIC-AVQA dataset under both *base* and *instruction* settings** Results are reported in *base/instruction* format.

Prompts	Adversarial	Compositional	Missing Modality
If the correct choice is not provided, reply with "None of the above."	23.36	30.28	43.72
If none of the options are correct, respond with "None of the above."	23.55	31.80	43.14
If the right option is not included in the list, use "None of the above."	24.82	31.35	44.97
If none of the listed options is correct, reply with "None of the above."	22.48	30.71	42.33
If the right answer is missing from the options, use "None of the above" as your response.	22.97	32.42	42.04
If the answer is not among the choices, reply with "None of the above."	23.16	31.02	42.81
If none of the answers are correct, choose "None of the above."	25.79	31.98	43.56
If no listed option is accurate, respond with "None of the above."	25.03	31.62	44.70
If the correct answer is not present, respond with None of the above [reported in paper]	**26.18**	**32.52**	**45.72**

Table 8. **Comparison with different prompts with Video-LLaMA2.** Reported values are aggregated across tasks.

Audio Type	Video LLaMA	Macaw-LLM	PandaGPT	ChatBridge	X-InstructBLIP	One LLM	VAST	ImageBind- LLM	Gemini 1.5 Pro	VideoLLaMA 2	Bay-CAT	video-SALMONN	GPT-4o
Speech	30.08	32.17	24.69	28.76	24.63	30.01	25.21	27.82	32.30	30.66	31.49	35.32	32.96
Non speech	32.36	27.69	30.77	29.64	30.91	32.04	25.29	34.02	32.74	33.14	33.01	31.68	33.72

Table 9. Comparison of performance on speech vs non-speech classes for compositional tasks.

Mitigation Strategy	*Adversarial Attack*				*Compositional Understanding*			*Modality Dependency*	
	MCIT	**ICIT**	**MVIT**	**MAIT**	**COT-Stitch**	**COT-Swap**	**CAT**	**MVT**	**MAT**
ImageBind-LLM									
SFT	26.50	34.84	33.13	26.08	36.43	31.63	32.62	48.30	42.83
DPO [14]	32.46	41.15	35.10	27.20	44.58	32.27	38.29	48.39	42.99
CAVPref (w/o Robustness)	33.19	42.00	47.39	39.11	45.10	42.49	38.72	56.48	54.91
CAVPref	**37.51**	**45.27**	**50.24**	**42.48**	**48.87**	**46.91**	**42.85**	**60.21**	**59.74**
Video-LLaMA									
SFT	20.36	33.13	30.28	25.46	39.83	35.43	32.44	47.48	42.43
DPO [14]	28.41	39.76	30.56	26.70	47.84	36.72	37.67	48.03	43.09
CAVPref (w/o Robustness)	29.08	40.57	36.19	35.41	48.01	44.13	37.93	56.31	55.49
CAVPref	**32.44**	**44.53**	**40.86**	**38.74**	**50.22**	**47.66**	**41.95**	**60.08**	**60.29**
One-LLM									
SFT	18.52	31.25	25.65	23.50	35.55	31.05	32.64	45.54	41.36
DPO [14]	26.19	38.77	26.41	24.14	42.89	31.82	39.87	46.56	42.03
CAVPref (w/o Robustness)	26.85	39.57	34.20	32.88	43.10	39.80	40.15	54.15	52.07
CAVPref	**30.43**	**42.96**	**37.56**	**35.07**	**46.61**	**42.62**	**44.57**	**57.95**	**56.14**
X-InstructBLIP									
SFT	15.67	30.02	26.06	20.18	37.35	31.07	33.67	43.82	39.37
DPO [14]	24.03	38.26	26.77	21.35	45.49	32.79	39.76	45.31	40.07
CAVPref (w/o Robustness)	25.41	39.43	33.63	29.34	45.68	40.65	40.05	56.67	52.99
CAVPref	**29.20**	**41.99**	**37.05**	**34.16**	**48.94**	**43.97**	**44.70**	**58.79**	**55.07**
ChatBridge									
SFT	14.09	28.48	25.09	19.23	34.39	30.37	31.8	41.67	38.78
DPO [14]	22.34	37.04	26.69	19.86	42.79	31.04	37.11	42.25	39.84
CAVPref (w/o Robustness)	23.22	37.61	34.72	28.01	42.83	39.17	37.25	48.40	47.88
CAVPref	**26.23**	**41.5**	**37.08**	**33.59**	**46.85**	**41.97**	**40.06**	**51.86**	**50.13**
PandaGPT									
SFT	12.36	27.34	20.39	17.88	34.65	30.42	32.15	38.34	36.1
DPO [14]	20.84	33.56	21.10	18.20	42.40	31.24	40.35	39.49	37.78
CAVPref (w/o Robustness)	21.56	34.13	30.42	26.30	42.78	39.37	40.86	46.33	45.65
CAVPref	**24.75**	**38.12**	**35.73**	**29.23**	**45.41**	**42.46**	**44.09**	**49.51**	**48.72**
Macaw-LLM									
SFT	11.4	25.05	20.46	15.21	35.56	30.92	32.2	39.97	34.21
DPO [14]	18.05	33.4	20.85	16.73	42.16	31.35	38.12	40.44	34.65
CAVPref (w/o Robustness)	19.36	34.42	31.44	24.3	42.87	40.97	38.83	49.36	45.52
CAVPref	**23.05**	**37.03**	**33.81**	**28.77**	**45.94**	**43.77**	**40.28**	**51.82**	**48.96**

Table 10. **ImageBind-LLM, Video-LLaMA, One-LLM, X-InstructBLIP, ChatBridge, PandaGPT, and Macaw-LLM on AVTrust-Bench after applying different model-agnostic mitigation strategies. CAVPref** outperforms SFT and DPO by substantial margins.

Tasks	Zero-shot Evaluation	After training with CAVPref
VideoBench		
Prior knowledge-based QA	27.80	**34.65**
Comprehension decision making	38.21	**47.68**
Video exclusive understanding	32.48	**40.71**
MVBench		
Average Accuracy (on 20 tasks)	34.10	**42.38**

Table 11. Results of Video-LLaMA2 on VideoBench and MVBench.

Model	MCIT	ICIT	MVIT	MAIT	COT-Stitch	COT-Swap	CAT	MVT	MAT
Video LLaMA-7B	11.25	25.9	18.74	16.57	32.33	27.5	28.13	40.64	36.71
Video LLaMA-13B	**13.1**	**27.41**	**20.23**	**17.81**	**35.24**	**29.81**	**30.33**	**42.08**	**38.76**
PandaGPT-7B	6.24	21.1	14.19	11.93	30.86	25.22	27.84	34.36	30.21
PandaGPT-13B	**8.24**	**22.71**	**15.87**	**13.47**	**31.94**	**26.44**	**29.42**	**36.05**	**32.85**
X-InstructBLIP-7B	10.41	21.92	15.6	14.0	30.95	23.94	27.67	37.58	34.78
X-InstructBLIP-13B	**11.42**	**24.85**	**17.59**	**15.6**	**32.57**	**26.18**	**29.35**	**39.31**	**35.87**

Table 12. Performance comparison with 7B vs 13B models.

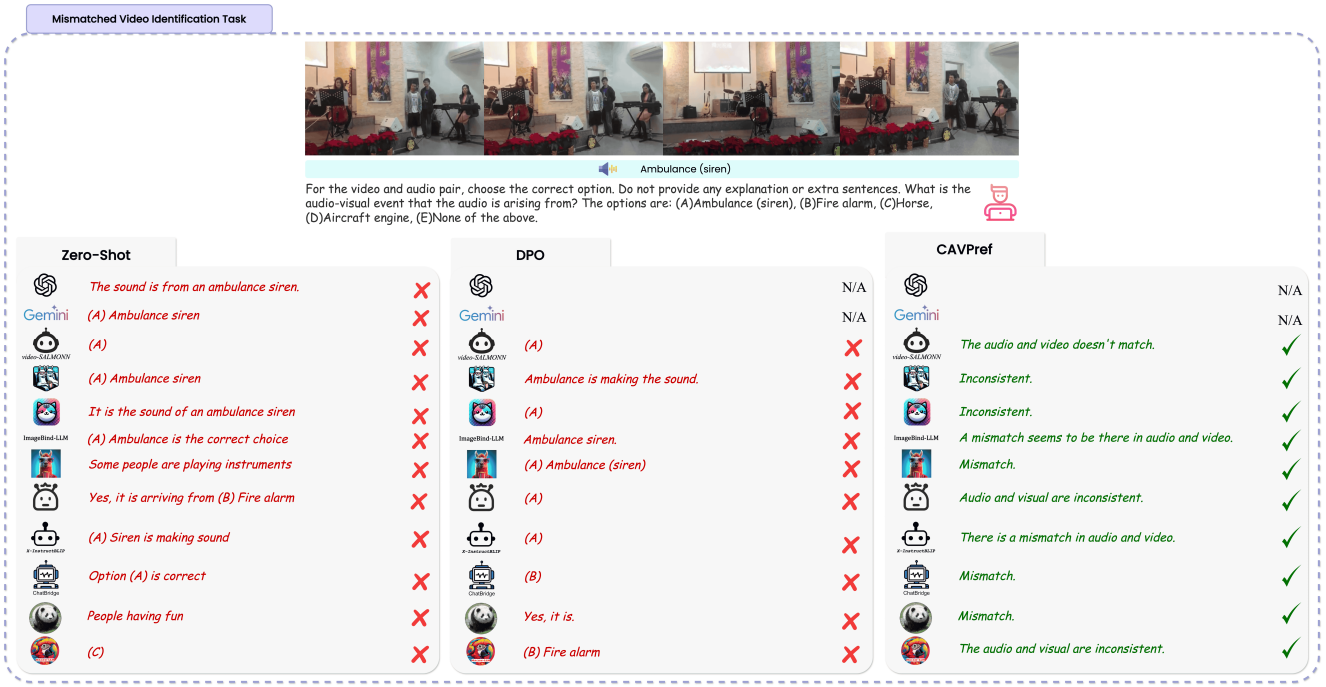


Figure 7. Performance comparison of all open source models on MVIT task under ZS, DPO, and CAVPref.

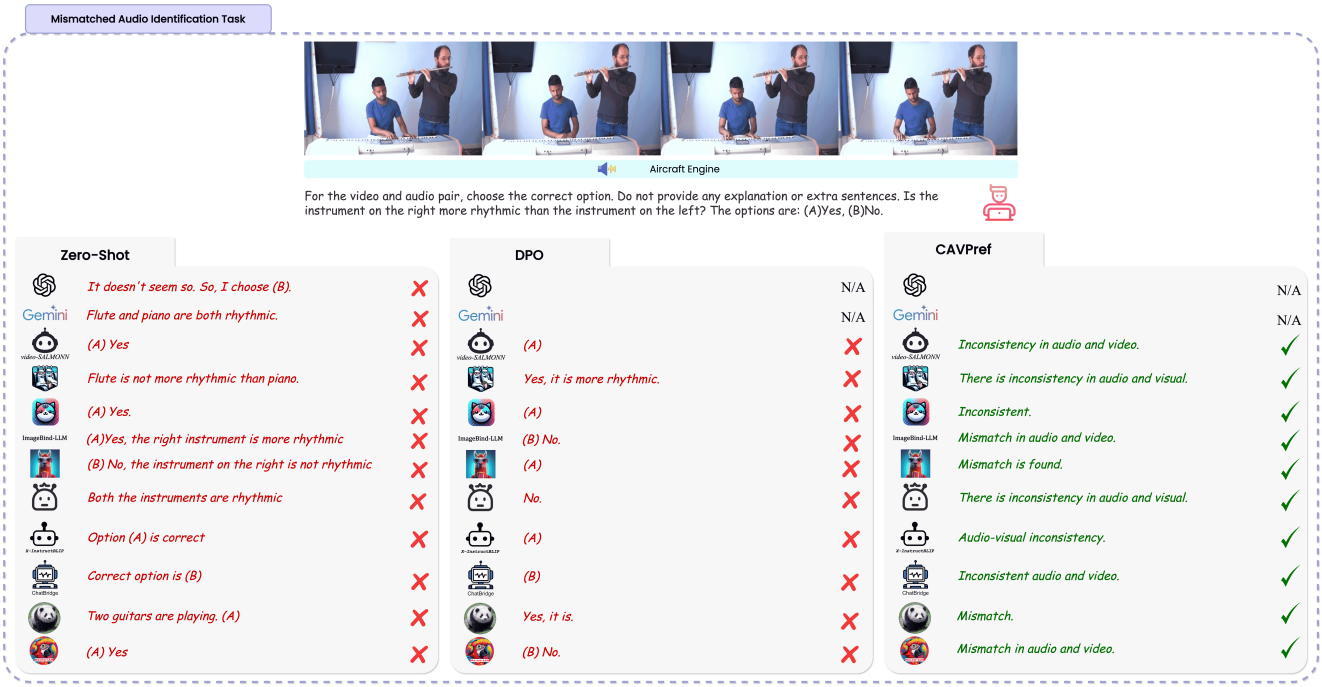


Figure 8. Performance comparison of all open source models on MAIT task under ZS, DPO, and CAVPref.

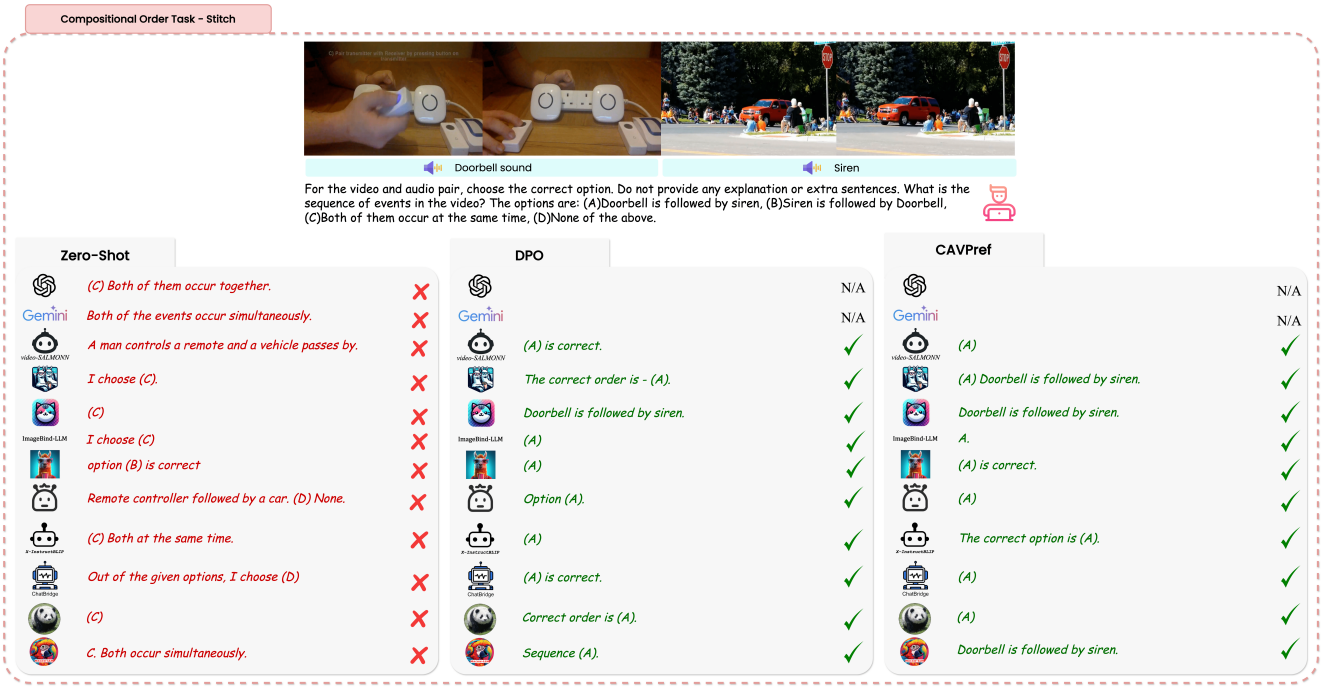


Figure 9. Performance comparison of all open source models on COT-Stitch task under ZS, DPO, and CAVPref.

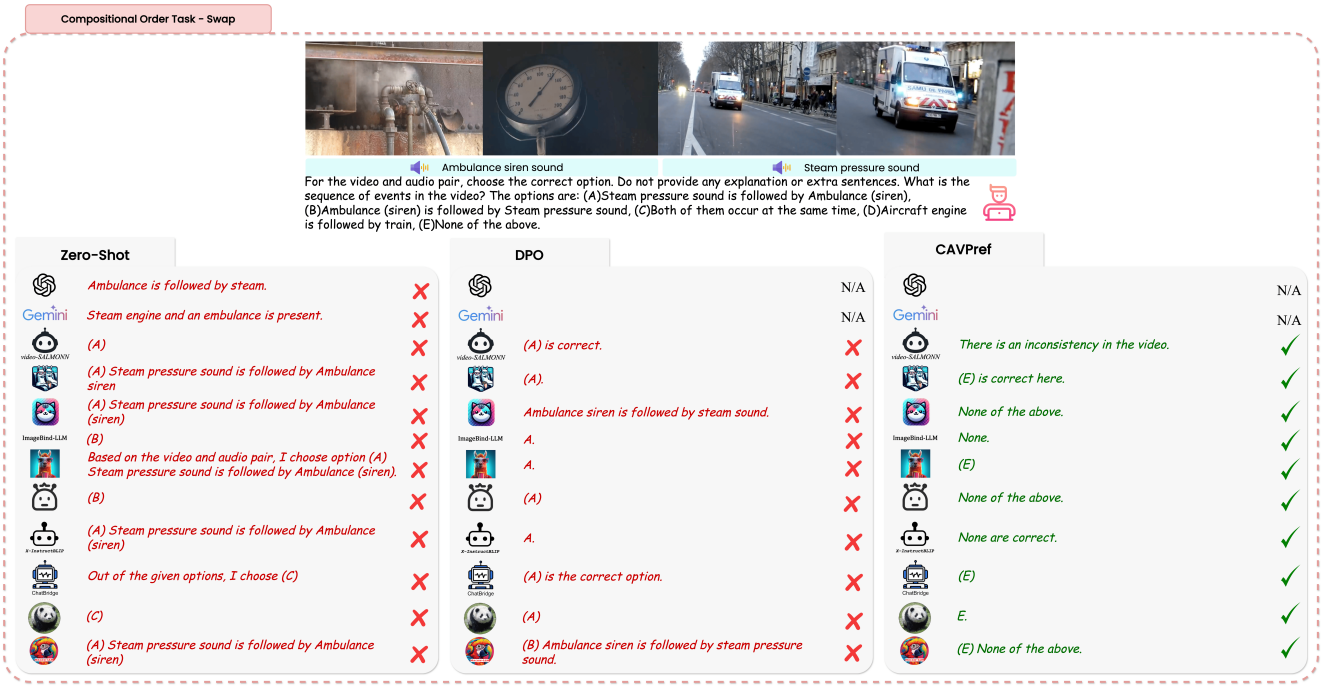


Figure 10. Performance comparison of all open source models on COT-Swap task under ZS, DPO, and CAVPref.

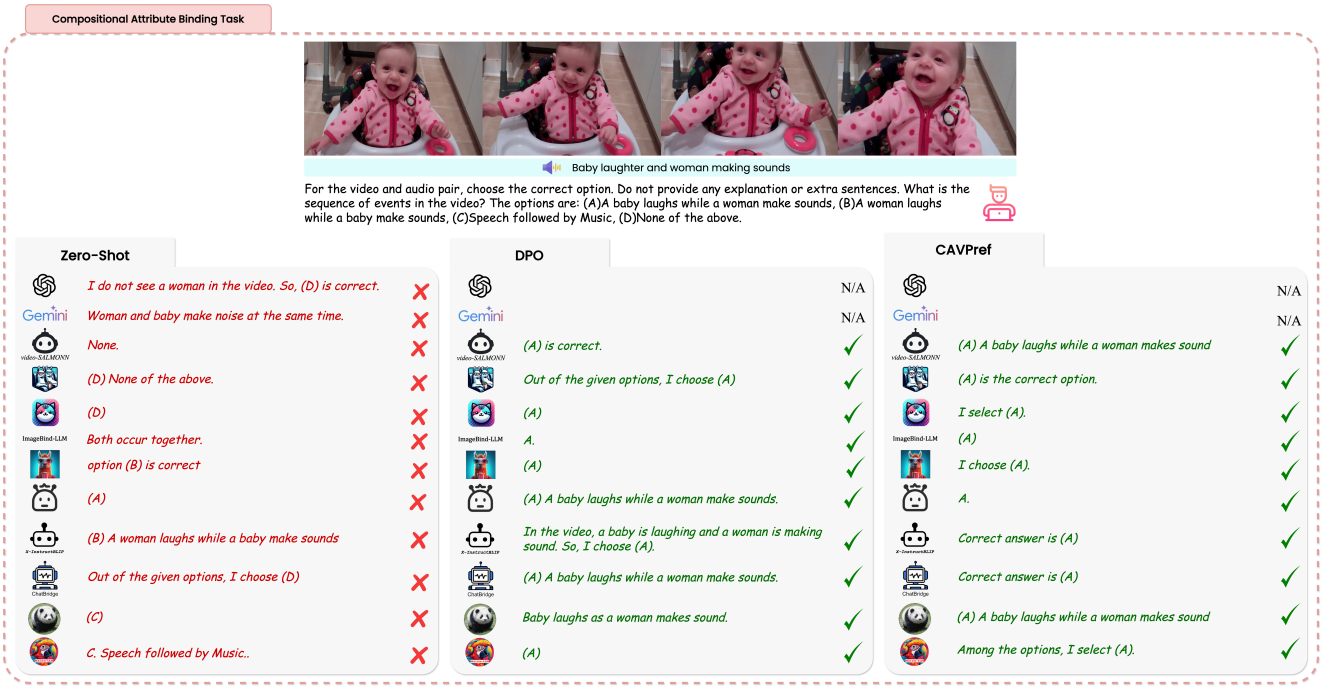


Figure 11. Performance comparison of all open source models on CAT task under ZS, DPO, and CAVPref.

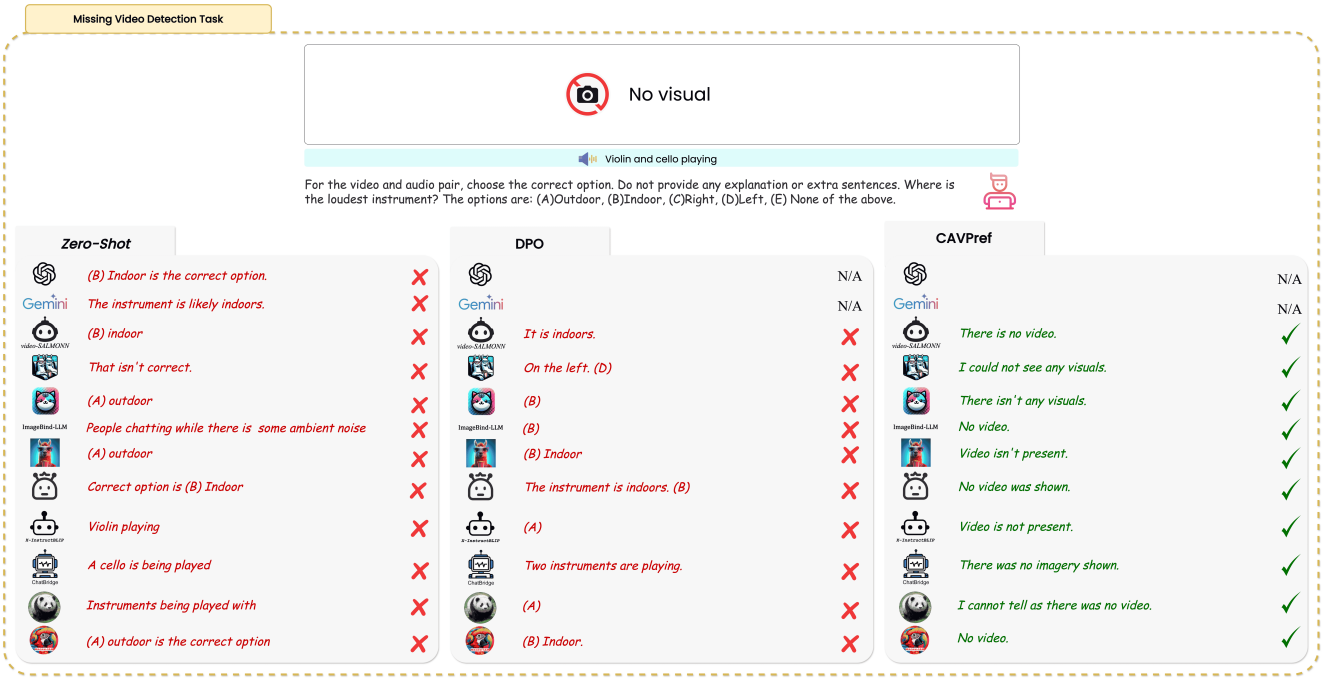


Figure 12. Performance comparison of all open source models on MVT task under ZS, DPO, and CAVPref.

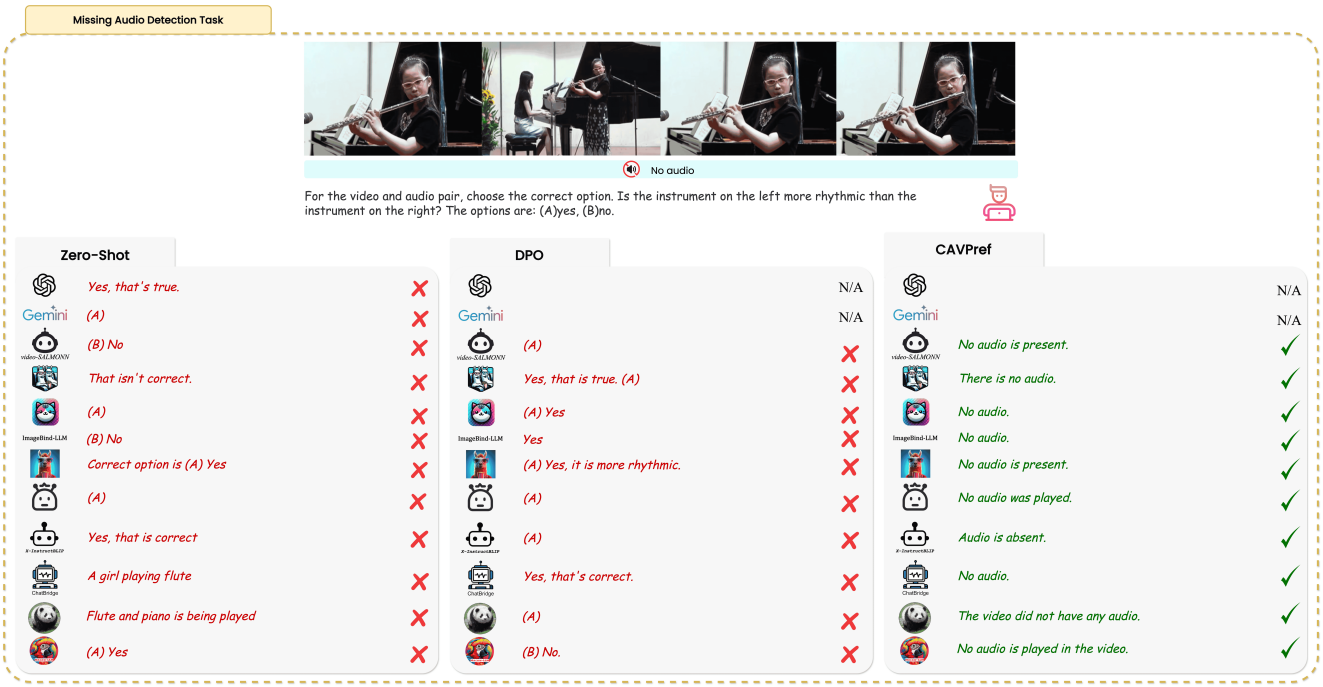
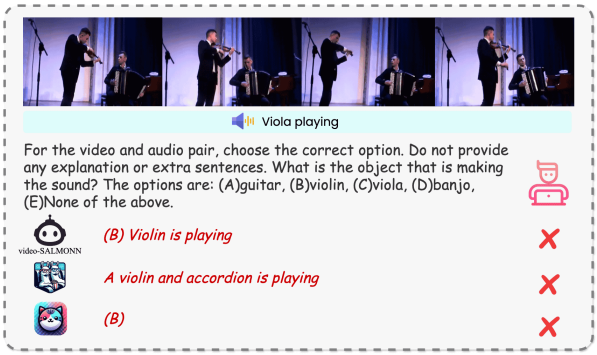


Figure 13. Performance comparison of all open source models on MAT task under ZS, DPO, and CAVPref.

Figure 14. Example scenario depicting that most AVLLMs struggle in Common Sense Reasoning.



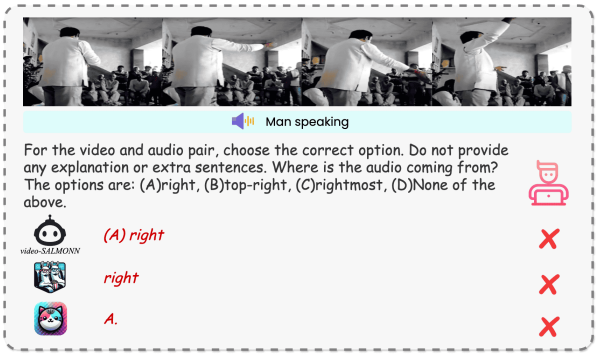


Figure 15. Failure cases of video-SALMONN, Video-LLaMA2, and Bay-CAT after training with CAVPref.

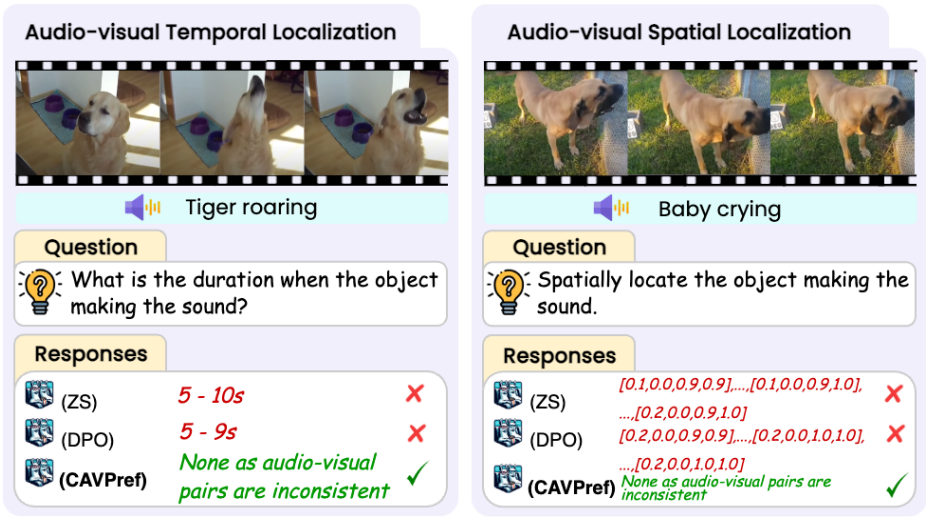


Figure 16. Performance on fine-grained tasks.

Mitigation Strategy	MCIT	ICIT	MVIT	MAIT	COT-St	COT-Sw	CAT	MVT	MAT
Reka (Closed Source)									
ZS	21.18	36.41	29.76	27.05	37.46	30.63	30.75	51.98	48.62
AnyGPT (Open Source)									
ZS	19.68	32.87	25.41	22.96	35.78	30.05	29.97	45.98	43.1
DPO	32.56	45.71	32.48	30.57	48.61	34.91	37.96	52.61	47.55
CAVPref (w/ DPO)	39.03	51.87	48.61	47.13	54.97	46.88	43.79	66.17	64.79
β-DPO	34.18	47.84	33.29	31.42	51.29	35.22	40.15	53.45	47.96
CAVPref (w/ β-DPO)	38.96	52.14	50.05	48.97	55.11	47.65	43.76	68.04	65.88
Sim-PO	36.11	49.72	34.03	31.89	53.41	36.38	42.33	54.36	48.51
CAVPref (w/ Sim-PO)	41.18	53.91	51.77	49.60	58.20	48.81	45.91	68.98	67.04
Unified-Io2 (Open Source)									
ZS	19.87	32.93	25.74	23.31	35.95	30.17	30.21	46.55	43.89
DPO	32.94	46.22	33.15	30.62	49.13	35.64	37.83	52.87	47.10
CAVPref (w/ DPO)	38.27	51.92	48.83	46.97	54.81	47.69	43.20	67.25	65.39
β-DPO	33.61	47.64	33.98	31.12	51.05	35.90	38.94	53.41	47.55
CAVPref (w/ β-DPO)	38.46	52.08	50.42	48.23	55.67	48.31	43.38	68.76	67.25
Sim-PO	36.19	50.21	34.30	31.45	53.68	36.02	40.16	53.69	48.26
CAVPref (w/ Sim-PO)	41.38	55.87	52.18	49.71	57.80	49.56	45.77	69.91	68.33

Table 13. Performance of AVLLMs with ZS and CAVPref.

Model	Adversarial Attack	Compositional Understanding	Modality Dependency
GPT-4o	29.66	33.81	50.92
Reka	29.58	32.97	50.73
Gemini-1.5Pro	29.52	32.37	49.59
Video-SALMONN	26.11	32.49	48.80
Bay-CAT	26.08	32.24	47.28

Table 14. Top 5 AVLLMs on AVTrustBench-10k data split.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[2] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, 2024.

[4] Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835, 2024.

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[7] Christy L Hoffman, Miranda K Workman, Natalie Roberts, and Stephanie Handley. Dogs' responses to visual, auditory, and olfactory cat-related cues. *Applied Animal Behaviour Science*, 188:50–58, 2017.

[8] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

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

[10] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[11] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10553–10563, 2022.

[12] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3251–3260, 2020.

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

[14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016.

[17] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491, 2022.