CAT+: Investigating and Enhancing Audio-visual Understanding in Large Language Models

Qilang Ye, Zitong Yu, Rui Shao, Yawen Cui, Xiangui Kang, Xin Liu, Philip Torr, and Xiaochun Cao

Abstract—Multimodal Large Language Models (MLLMs) have gained significant attention due to their rich internal implicit knowledge for cross-modal learning. Although advances in bringing audio-visuals into LLMs have resulted in boosts for a variety of Audio-Visual Question Answering (AVQA) tasks, they still face two crucial challenges: 1) audio-visual **ambiguity**, and 2) audio-visual **hallucination**. Existing MLLMs can respond to audio-visual content, yet sometimes fail to describe specific objects due to the ambiguity or hallucination of responses. To overcome the two aforementioned issues, we introduce the **CAT+**, which enhances MLLM to ensure more robust multimodal understanding. We first propose the Sequential Question-guided Module (SQM), which combines tiny transformer layers and cascades Q-Formers to realize a solid audio-visual grounding. After feature alignment and high-quality instruction tuning, we introduce Ambiguity Scoring Direct Preference Optimization (AS-DPO) to correct the problem of CAT+ bias toward ambiguous descriptions. To explore the hallucinatory deficits of MLLMs in dynamic audio-visual scenes, we build a new Audio-visual Hallucination Benchmark, named ***AVHbench***. This benchmark detects the extent of MLLM's hallucinations across three different protocols in the perceptual object, counting, and holistic description tasks. Extensive experiments across video-based understanding, open-ended, and close-ended AVQA demonstrate the superior performance of our method. The AVHbench is released at https://github.com/rikeilong/Bay-CAT.

Index Terms—audio-visual question answering, multimodal large language models, audio-visual hallucination.

1 Introduction

Large Language Models (LLMs) [66] show a remarkable ability to understand real-world human language and generate continuum contexts. Recent research has extended LLMs to real-world applications to assist a range of visual question-answering and video reasoning tasks. These models are characterized by their ability to understand complex visual and audio patterns and generate coherent linguistic representations. We collectively refer to such models as Multimodal Large Language Models (MLLMs). MLLMs have continuously improved the ability of large-scale models to understand other input tokens through a mixture of training strategies, high-quality multimodal corpora, and so on, leading to improved performance and generalisability. Despite these advances, correctly answering questions in dynamic audio-visual scenarios remains challenging. As shown in Fig. 1 (a), MLLMs are biased toward outputting nonsense when generating answers, where we call "Ambiguity" [75]. Such a phenomenon not only causes the model to make random guesses but also severely affects the inference of detailed answers in the Audio-Visual Question Answering (AVQA) task.

Besides this ambiguity resulting in answering the wrong question, the problem of hallucinations [6] in different modalities still affects the dependability of MLLMs and their ability applied across domains. Visual hallucinations and audio hallucinations come from the catastrophe of two different tasks. For example, visual hallucination is the phenomenon whereby MLLM generates textual content that is semantically inconsistent with its counterpart in a given visual object (e.g., image or video). Audio hallucinations are non-existent objects imagined by the MLLM's inability to understand the given audio. These two challenges stem from the fundamental problem of MLLMs, namely the over-reliance on a priori knowledge from the pre-training process and the inability of the fine-tuning strategy to correct erroneous objects.

Furthermore, we find a phenomenon of simultaneous

This study is supported by National Natural Science Foundation of China (Grant No. 62306061, No. 62306090, No. 62171309, No. 62025604, and No. 62441619), Natural Science Foundation of Guangdong Province of China (Grant No. 2024A1515010147 and No. 2514050000889), Shenzhen Science and Technology Program (CJGJZD20240729141505007), Guangdong Key Laboratory of Information Security Technology (No. 2023B1212060026), Guangdong Provincial Key Laboratory (Grant 2023B1212060076), and Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University. The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University. Qilang Ye and Zitong Yu are co-first authors and contribute equally. Corresponding authors: Zitong Yu (email: yuzitong@gbu.edu.cn) and Yawen Cui (email: yawencui@polyu.edu.hk).

Q. Ye is with Great Bay University, and also with the College of Computer Science, Nankai University, Tianjin, 300350, China. He is also with Zhongguancun Academy, Beijing, China.

Z. Yu is with Great Bay University, Dongguan 523000, China, and also with Guangdong Provincial Key Laboratory of Intelligent Information Processing & Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China. He is also with Dongguan Key Laboratory for Intelligence and Information Technology.

R. Shao is with Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China.

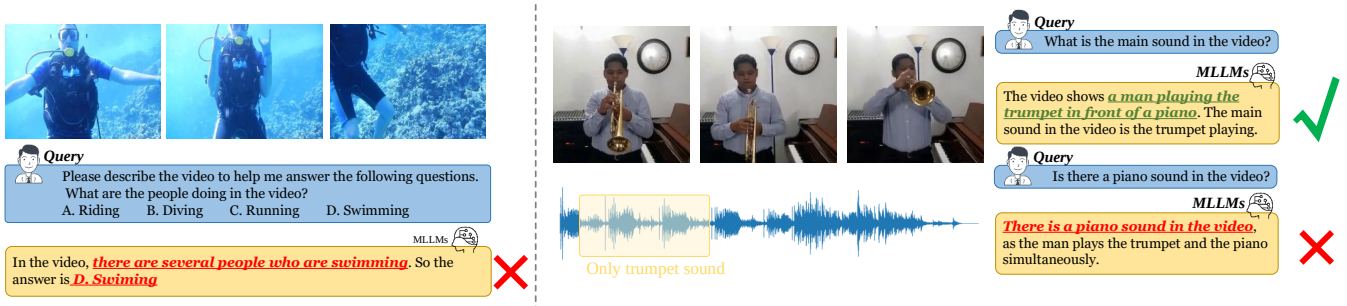
Y. Cui is with The Hong Kong Polytechnic University, Kowloon, Hong Kong.

X. Kang is with Guangdong Key Lab of Information Security, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China 51000.

X. Liu is with School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China, and also with Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Lappeenranta 53850, Finland.

P. Torr is with University of Oxford, Oxford OX12JD, UK.

X. Cao is with School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, China.



(a) Audio-visual **Ambiguity** (b) Audio-visual **Hallucination**

Fig. 1: We present examples of ambiguity and hallucinations in audio-visual scenarios. (a) When using MLLMs for audio-visual question answering, it can roughly describe the content of the video and audio, but it is difficult to choose the correct answer based on that content. (b) MLLMs can correctly describe the visual content, while hallucinating on the audio.

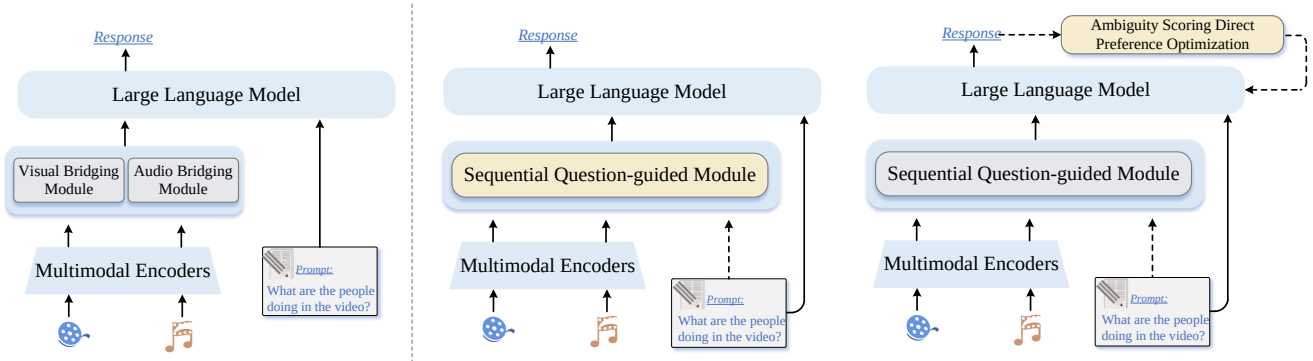
visual hallucinations and audio hallucinations, which we call Audio-Visual Hallucinations (AVH). This phenomenon exists when MLLMs cannot accommodate both visual and audio objects. As shown in Fig. 1 (b), when we ask MLLM *what the person is doing*, the model can correctly answer that the person is playing the trumpet in front of a piano. However, when we ask MLLM whether there is a piano sound in this video, the model may state that since there is a piano, there must be a piano sound. These problems not only directly affect the actual deployment of MLLM, but also make it difficult to generate accurate and reliable content for downstream tasks. We analyze earlier efforts by exploring various approaches to suppress the problem of hallucinations arising in their respective modalities. However, no paradigm collectively addresses object hallucinations that co-exist in both visual and audio scenes. Several recent studies [63] have suggested that hallucinatory target datasets are difficult to generalize and expand by performing multimodal alignment or data augmentation. Alternatively, achieving hallucination mitigation through a post-reviewer after supervised fine-tuning [66] or Reinforcement Learning from Human Feedback (RLHF) [52], [46] requires significant time costs.

Why do ambiguity and hallucinations commonly occur in audio-visual scenarios? We propose potential reasons that make the above two phenomena occur in MLLMs:

- **Audio-visual ambiguity.** As shown in Fig. 1 (a), most MLLMs can describe a video reasonably well, but can't choose the right option. We consider MLLMs to be an excellent text generator, but performance on contextual reasoning in audio-visual scenarios is still sub-optimal. This is because the foundation model that pre-trained with a text corpus often rely on a priori knowledge. This ambiguity generally manifests in ambiguous words of the corresponding audio-visual content, as well as in generating useless text that overly responds.
- **Audio-visual hallucination.** Hallucination is usually attributed to modality alignment problems, i.e., the difficulty of aligning text with audio-visuals. Under some multimodal learning scenarios, the model tends to be overly reliant on in-domain textual prior knowledge, which leads to lower confidence in the other input modalities. Specifically, given an input query that does not have any direct relationship with video and audio, the model relies on internal knowledge to output content that is coherent with that query.

In this paper, we introduce the CAT+. CAT+ mainly focuses on resolving "**Ambiguity**" and "**Hallucinations**", it consists of two stages: **1) Continuous query question-related knowledge.** As shown in Fig. 2 (a), most existing MLLMs [80], [31], [83] are mainly designed with multiple branches to handle multiple modalities individually, followed by concatenating the modality embeddings with prompts as inputs to the LLM. This paradigm mixes text and audio-visuals for training, preventing multimodalities from being associated at a low level. As shown in Fig. 2 (b), we design a Sequential Question-guided Module (SQM) to dynamically capture question-aware visual and audio hidden features to enrich the input. The SQM receives sufficient low-level question information to interact with audio-visual, it is capable of improving audio-visual grounding. **2) Retraining MLLMs to mitigate ambiguity.** Inspired by Direct Preference Optimization (DPO) [57], We further propose Ambiguity Scoring Direct Preference Optimization (AS-DPO). Although trained MLLMs are capable of multimodal understanding, they still lack the ability to correctly answer audio-visual questions. As shown in Fig. 2 (c), we recast ambiguity mitigation as a model preference optimization process. Specifically, we detect and score the ambiguity of the model by rating ambiguous responses that express the lack of clarity of a particular audio-visual object as severe and vice versa as correct. We transform the scoring system into weights and perform AS-DPO on the fine-tuned CAT+ to favor positive responses (i.e., accurate descriptions) and reject negative responses (i.e., ambiguous descriptions).

To overcome the problem of hallucinations in audio-visual scenes, we focus on collecting a dataset for the detection of audio-visual hallucinations. Current hallucination benchmarks that have been developed are those that review systems for single tasks, including Q&A [38], [16], [65], abstraction [47], image captioning [59], and visual Q&A [35]. These benchmarks measure the level of hallucination in a single modality by reference to a sample, and none of these works is able to detect the problem of hallucination in a bimodal context, i.e., AVH. Therefore, we propose ***AVHbench***, which detects whether visual and audio objects for hallucinations. Firstly, we manually annotate the visual and audio objects of each scene. Then we collect 100 samples



(a) Existing MLLMs (b) Stage-1 of CAT+ (c) Stage-2 of CAT+

Fig. 2: Comparison between existing MLLMs and CAT+. (a) Most existing MLLMs train multimodal encodings separately from the input prompt. Instead, our CAT+ includes two stages: (b) it merges all inputs in a question-guided manner. (c) We propose to retrain the entire MLLM to constrain itself to generate better content.

on the Music-AVQA dataset [29], and each of its samples includes both video and audio. To assess hallucinations more comprehensively, we propose three evaluation metrics, i.e., *Discriminative*, *Object Counting*, and *Generative*.

To sum up, the contributions of this paper are listed:

- We introduce a novel audio-visual language model, CAT+, to enhance the inference of MLLMs in audio-visual scenarios by learning question-guided audio-visual features via the proposed SQM and an ambiguity-scoring retraining strategy called AS-DPO.
- To analyze the hallucination problem in audio-visual scenarios, we propose ***AVHbench***, which is a more comprehensive assessment of the reliability of MLLMs for real-world applications compared to the current setup of unimodal hallucination detection.
- We evaluate CAT+ on a wide range of multimodal tasks. Extensive experiments demonstrate the superiority of CAT+ (e.g., outperforms the state-of-the-art on a variety of AVQA tasks, and achieves remarkable results in the evaluation of video-based text generation tasks and zero-shot video question-answering tasks). Notably, the performance of CAT+ on AVHbench shows more reliability compared to mainstream MLLMs.

This paper is an extended version of our prior publication [75] in ECCV 2024. The main differences with the conference version are as follows: 1) We explore how to learn sequential knowledge in longer videos and audio, and propose the Sequential Question-guided Module (SQM). The SQM acts as a powerful bridge to continuously query long video and audio for question-related knowledge. 2) We introduce Ambiguity Scoring Direct Preference Optimization as an efficient retraining strategy that scores ambiguous responses in a way that allows the model to select superior content based on the score. 3) We explore the problem of hallucinations in audio-visual scenarios and present ***AVHbench***, which evaluates whether MLLMs can correspond to visual and audio objects. More importantly, we outperform other similar methods on AVHbench.

2 Related Work

2.1 Large Language Models in Multimodal

Extending LLM to other multimodal tasks is a recently emerging field. This paradigm not only replaces traditional classifiers to perform logical distribution selection, but also acts as a generator to provide rich semantic text for downstream tasks. Many current works [43], [34], [70] emphasize the design of elegant bridging methods to improve the performance of downstream tasks. The simplest modules bridged for LLM use one or more linear layers for feature alignment. While this approach enables the LLM to learn more complete tokens, some overly long tokens incur significant training time. Others have designed more complex bridging networks for cross-modal queries. For example, Lyu et al. [42] proposed an alignment module for coordinating different representations before entering the LLM, and Ma et al. [43] proposed keeping the distance between all visual tokens and any linguistic tokens consistent in the LLM. This diverse pre-sampling approach to integrate cross-modal information into LLM is one of the most popular trends at present.

2.2 Alignment learning for LLMs

Alignment learning can allow multimodal large language models to correct their own intrinsic errors to accommodate more human-like description outputs. Such paradigms are also seen as integral to the fine-tuning of MLLMs. Reinforcement Learning from Human Feedback (RLHF) [52], [46] optimizes large-scale models by constructing reward models that generate content more biased towards human-written better samples. This type of alignment learning [9] enhances the LLM to generate more accurate information. Another strategy of Direct Preference Optimization (DPO) [57] enables direct optimization by bypassing the learning of reward models. This approach inspired us to solve the ambiguity problem by collecting positive and negative samples. In this paper, we further optimize the DPO strategy to make it applicable in MLLMs, and in order to emphasize the biased distribution of the positive samples, we include weight values to train the optimization process of the model.

2.3 Hallucination in Visual and Audio Scene

Indeed, the problem of hallucinations has long been defined in the field of natural language processing as generated content that deviates from the intended meaning and makes no sense [18], [38], [62], [82]. However, in visual tasks, illusions also seriously affect the accuracy of different downstream tasks, such as object detection, OCR, and image classification. LLMs fine-tuned with visual data generate plausible outputs, including mismatched or missing objects, destabilizing the original LLM. To mitigate the illusory problem of LLM in visual tasks, most approaches typically employ strategies such as fine-grained contrast learning [79], region of interest extraction [3], and data augmentation [23]. However, we believe that all such approaches struggle to unify different training paradigms and model architectures, posing a significant challenge to mitigate the visually illusory problem [21].

Generative models are applied to audio scenes mostly for audio captioning and audio object recognition. Audio captioning is the task of generating natural language descriptions from audio [22], [49], [48], as with image captioning. With the rapid development of paradigms based on LLMs, a number of audio captioning models have been proposed [48] with strong results in describing various types of audio content. Whereas audio object recognition mostly exists in audio-visual question-answering tasks, the key to such tasks is to accurately determine the objects present that correctly correspond to the audio information. However, work has been done [27] to show that generative models suffer from misdescription or non-correspondence in both audio captioning and audio object recognition. We argue that this phenomenon can be uniformly referred to as the illusion problem in audio scenes. Some researchers [68] have observed the per-word token probability and find that over-reliance on LLM internal knowledge may be one of the causes of hallucinations, and more modalities may exacerbate this phenomenon. Another work [14] has explored the issue of whether MLLMs can hear things in audiovisual scenarios via testing the model's sensitivity to sound size and frequency levels. However, they only evaluate the model's ability of audio understanding while lacking detection of simultaneous video and audio hallucination, which further motivates us to explore hallucination detection in unified audio-visual scenarios.

3 METHODS

In this section, we first present the proposed CAT + (Sec. 3.1 and 3.2) in detail. As shown in Fig. 3, CAT+ consists of three branches that extract visual knowledge, auditory knowledge, and textual knowledge as embeddings using frozen encoders, respectively. Second, we present the training strategy for CAT+ in Sec. 3.3. Lastly, we introduce the Ambiguity Scoring Direct Preference Optimization (AS-DPO) strategy, which reinforces CAT+ to favor non-ambiguous descriptions (Sec. 3.4). Notably, this section focuses on how to mitigate ambiguity to improve CAT+ performance on AVQA.

3.1 Multimodal Inputs

Audio & Visual Representation. Multimodal encoders that can extract discrete representations allow LLMs to better distinguish between each visual and audio object. In this work, we leverage the frozen ImageBind [13] as a universal encoder for all modalities. Compared to extracting each modal information separately, the ImageBind with a single joint embedding space demonstrates strong modality learning ability, and we find in our experiments that this unified encoding paradigm leads to better performance. Specifically, given an image frame $\mathbf{V}$ and a set of audio clips $\mathbf{A}$, where video frames are sampled from 2 fps and audio is derived in 3-second clips, the encoded multimodal hidden features can be obtained by:

$$h_v = \text{ImageBind}(\mathbf{V}), \quad h_a = \text{ImageBind}(\mathbf{A}), \tag{1}$$

where h_v, h_a are video and audio features, respectively.

For the input query, to highlight the question content for applying in SQM, we refactor the form of the input prompt by adding two simple tokens $< Q >$, $< /Q >$, where $< Q >$ denotes the start position of the question and $< /Q >$ denotes the end position of the question. Notably, the highlighted question content is extracted and fed into the SQM to guide the perception of audio-visual hidden features during fine-tuning.

Visual-Deepened Audio. Audio can be effective in improving video comprehension, especially in music scenes. However, there is a possible drawback that sometimes a useless piece of audio information appears in the video (e.g. no sound for a long time or irrelevant background music). Therefore, we propose a simple Visual Deepen Audio method that establishes cross-modal latent connections to attenuate audio redundancy impact. Specifically, we define Multi-Head Attention (MHA) from the Transformer as $MHA(Q, K, V)$, where Q, K, and V denote the query vector, the key vector, and the value vector, respectively. The visual-deepened audio features can be obtained by: $\hat{h_a} = \text{Expand}(h_a) + MHA(h_v, h_a, h_a)$, where $\text{Expand}(\cdot)$ is to standardize the length of h_a and h_v.

3.2 Sequential Question-guided Module

An important element of LLM for multimodal understanding is the design of efficient bridging modules. The current simple paradigm [80], [83] is to apply a linear layer to input complete visual and audio information. The advantage of this approach is that there is no need to compress the input information. However, such a way does not allow the LLM to focus on a specific video segment. This leads to the persistent generation of a large pile of irrelevant text, which is derived from a priori knowledge within the LLM. We find that the strategy of keeping audio-visual features independent of input prompts tends to a biased towards generalized text generation and an inability to specify certain content. Another cause of redundancy in LLM generation may be long video inputs. LLM itself has a strong ability to perceive long text but does not have the ability to understand long video, which leads to confusion within LLM and makes it difficult to target key content.

In this work, we design a Sequential Question-guided Module (SQM) to dynamically capture visual and audio hidden features of question perception to enrich the detailed knowledge acquired by the LLM. SQM can be divided into two steps, as illustrated in Fig. 4:

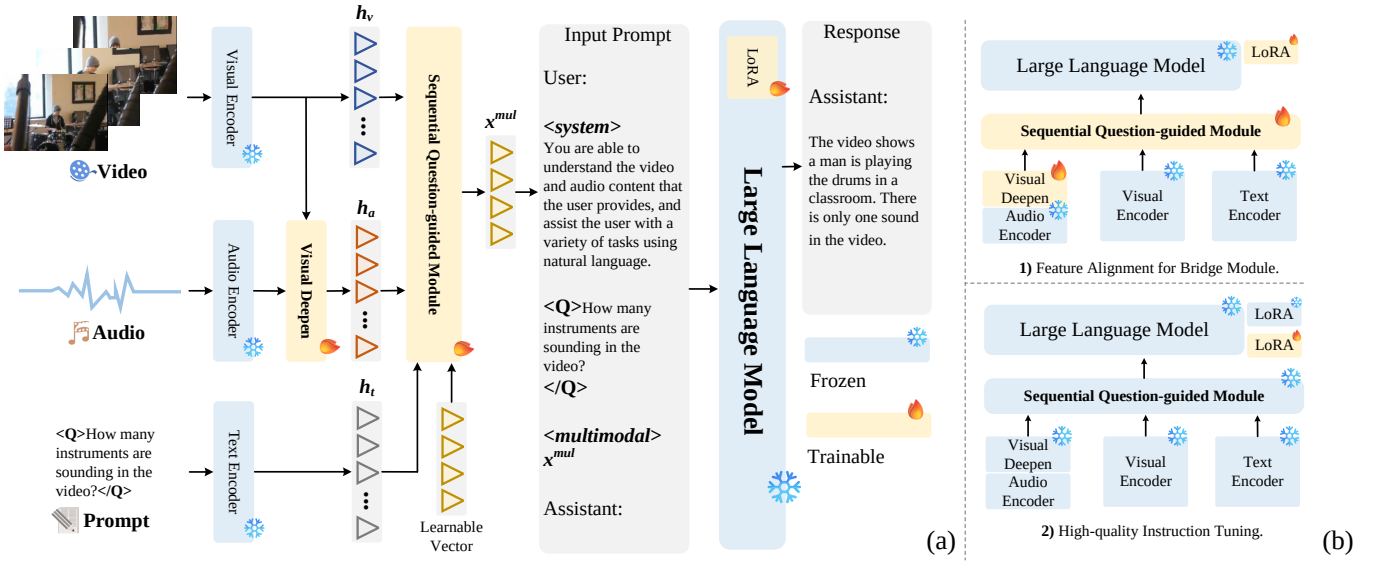


Fig. 3: Illustration of the proposed CAT+ and its training strategy. (a) Overview of CAT+. CAT+ first extracts audio, visual, and text embedding from input audio, video, and the question tagged with $< Q >$, $< /Q >$ to highlight the question content. Notably, we introduce a visual deepen operation to exacerbate the visual impact. Then, a learnable vector is fed into the Sequential Question-guided Module (SQM) to integrate all the information related to the question. Finally, we merge multimodal tokens and language and feed into the frozen large language model with LoRA [17] to output the response. (b) The training paradigm of CAT+ involves pre-alignment of the SQM and instruction tuning on the entire model.

Step1: perceiving question-related visual and audio features. Our early attempts to use an audio Q-former [30] coupled with a visual Q-former based on the idea of InstructBLIP [9] did not produce the expected results. In order to extract question-relevant details more fully from the input video and audio, we design a tiny transformer consisting of multiple layers. The tiny transformer network consists of multiple self-attention (SA), cross-attention (XA), and feed-forward networks (FFN). Further, we combine these compositions with each other in a forward and reverse order. Specifically, we define a forward order block $\mathcal{F}_1$ to perform question-aware localization based on the cross-attention map. Then, we define a reverse order block $\mathcal{F}_2$ to consolidate the question-guided audio-visual representation at the attentional level. Given a text embedding h_t, a visual embedding h_v, and a audio embedding h_a, the whole process can be represented as:

$$\mathcal{F}_1(h_t; x) = \text{XA}\left(h_t, \text{FFN}(\text{SA}(x))\right), \tag{2}$$

$$\mathcal{F}_2(x; \mathcal{F}_1(h_t; x)) = \text{SA}(\text{FFN}(\text{XA}(\text{FFN}(\text{SA}(x)), \mathcal{F}_1(h_t, x)))). \tag{3}$$

In XA$(\cdot,\cdot)$, the former represents the query vector and the latter represents the key vector and value vector. Notably, the tiny transformer for visual and the tiny transformer for audio share parameters. We consider that the correlation between the two modalities can be better established in this way. In addition, the deepened audio information covers part of the visual knowledge, and the way the parameters are shared balances the modeling process of both. Based on Eqs. 2 and 3, the question-guided visual features h'_v and audio features h'_a can be represented as:

$$h'_v = \mathcal{F}_2(h_v; \mathcal{F}_1(h_t; h_v)) \tag{4}$$

$$h'_a = \mathcal{F}_2(h_a; \mathcal{F}_1(h_t; h_a)) \tag{5}$$

Step2: sequential query key audio-visual features. Current approaches [37] present significant challenges when exploring systems based on long videos (more than one minute) due to computational complexity, training memory costs, and effective bridging modules. For example, a two-minute-plus video can be decomposed into about 400 to 500 frame-level features of length that a normal machine would not be able to process. We find that the Q-Former architecture proposed by Li et al. [30]. enables querying key information and customizing the output length.

To map the obtained audio-visual features into the embedding space of LLMs, one possible solution is to design an efficient and computationally less expensive multimodal projector that transforms the input into language tokens of suitable length. We consider the temporal dimension of audio-visuals to be constructed from a continuous series of spatial information, and pooling or arbitrary extraction would disrupt the context of the entire structure. Therefore, we introduce a novel projection method that consists of cascading Q-Formers and queries the combined audio-visual features in segments in order to avoid continuity disruption. Specifically, we first concatenate video feature h'_v and audio feature h'_a in the time dimension to obtain the combined audio-visual features $h'_{av} \in \mathbb{R}^{2T \times d_h}$. Then we set a hyperparameter k to select k time points for a combined audio-visual features segment $\{h'_{av}i | i = 1, 2, ..., t\}$, where $t = \frac{2T}{k}$. Moreover, we define the Q-Former as a function f_Q and a learnable query vector $\mathbf{q} \in \mathbb{R}^{T_q \times d_h}$ in length T_q, where k must be larger than T_q. To enable continuous querying, we utilize the previously queried audio-visual representation as the query for the next step, and the whole process can be represented as

$$\mathbf{q}_t = {f_Q}^t(\mathbf{q}_{t-1}, \mathbf{q}_t), \tag{6}$$

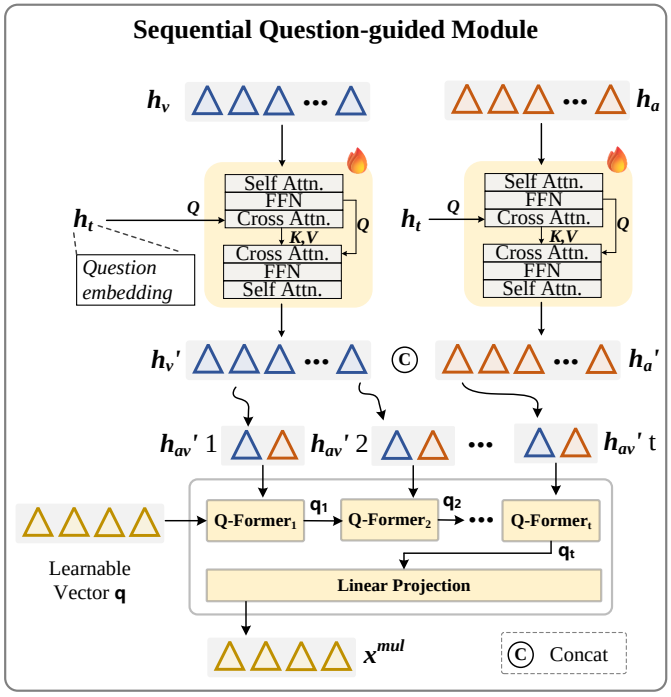


Fig. 4: Illustration of the SQM. This module includes a tiny transformer layer and cascade query modules.

where $f_Q{}^t$ denotes t-th Q-Former function, $\mathbf{q}_t \in \mathbb{R}^{T_q \times d_h}$, and all the parameters of Q-Former are shared. In practice, most videos are around one minute in length, and we set $T_q = 48$ to query question-guided audio-visual features and obtain multimodal tokens x^{mul} via a projector consisting of one linear layer $FC(\cdot)$, it denotes as $x^{mul} = FC(\mathbf{q}_t)$. With this projector, the dimension of the multimodal tokens x^{mul} is aligned with the LLM.

3.3 Multimodal Training Strategy

We follow the previous work [40] and use a two-stage training strategy to facilitate LLMs' comprehension of video and audio. As shown in Fig. 3 on the right, we first pre-training CAT+ for aligning the learning ability of the visual deepen module and the SQM in Stage 1. In Stage 2, we propose to fine-tune the entire model using a higher-quality instruction set, where the visual deepen module and SQM are frozen.

Stage-1: feature alignment for bridge module. Our goal is to update the parameters of the bridging module to put it in a state that best fits the LLM output. Therefore, we employ the dataset from 100k video instruction [45]. This dataset is carefully filtered to achieve a more balanced multimodal coverage distribution, it contains multiple video-text pairs, and the videos contain audio information. In this stage, we first freeze the LLM and the encoder to align the semantic information for vision and audio, respectively, and only the visual deepen module and the SQM are trainable. Then, we employ Low-Rank Adaptation (LoRA) to fine-tune the model to adapt the non-human language tokens with the learning objective $\mathcal{L}_{Lora}$. Specifically, we define the decoder layer from LLMs as f_{LLM}, and we input the instruction that contains multimodal tokens into f_{LLM} to train the model. $\mathcal{L}_{Lora}$ can be expressed as:

$$\mathcal{L}_{Lora} = CrossEntropy(f_{LLM}(x^{text}, x^{mul}, y), \quad (7)$$

where the entire feature alignment process is supervised by using the cross-entropy function, x^{text} denotes text tokens including system commands, questions, etc.

Stage-2: high-quality instruction tuning. While the first stage enhances CAT+'s audio-visual understanding capability, it still lacks generation accuracy. Current large-scale video and audio text datasets contain a large amount of text noise, which severely hampers the object localization ability and generation accuracy of LLMs. Second, the transformation from visual and audio information to text space is usually subject to severe bias, e.g., when video information is rich but audio is full of noise, which can make the whole model weight distribution unbalanced. Hence, we expect to balance video and audio with a higher-quality instruction set for better inference in audio-visual scenarios. In our previous publication ECCV 2024 [75], we have collected an audio-visual joint instruction dataset, named AVinstruct. This dataset emphasizes joint learning of dynamic audio-visual pairs to solve different AVQA tasks. Therefore, in our second stage, we integrate all compositions from Stage 1 and freeze them. In order to avoid conflicts between the parameters trained in Stage 1 and Stage 2, we freeze the LoRA parameters of Stage 1 and introduce a new set of LoRA parameters to learn the instruction data of AVinstruct, where the learning objectives for this stage are consistent with Stage 1.

In summary, after the first stage of training, the model starts to understand both visual and audio information. In the second stage, we enhance the model's ability to infer audio-visual scenarios with higher-quality instruction data, i.e., including visual object recognition, and audio semantic understanding, while maintaining the correspondence between the two modalities.

3.4 Ambiguity Scoring Direct Preference Optimization

The goal of making MLLMs respond accurately in a variety of dynamic audio-visual scenarios is extremely challenging. We find that RLHF-like strategies can retrain LLMs to alleviate the ambiguity that tends to arise when describing specific audio-visual objects. Furthermore, the classical DPO enables LLM to prefer positive samples directly without training a reward model. In this work, we retrain the fine-tuned CAT+ following the DPO strategy. However, three questions arise: (1) how to detect ambiguity, (2) how to get positive data, and (3) how to optimize more efficiently.

(1) how to detect ambiguity. For the data source to retrain the model, We opt for a test set in the collected AVInstruct dataset, which is rich in video and audio content and associated with a variety of open-ended questions. Such dataset is specific to a variety of scenarios that contain audio information, including music, the real world, etc., and the question-and-answer pairs include attributes of multiple objects and their spatial relationships within the visual content. As illustrated on the left in Fig. 5, to detect whether the generated content of a test model is ambiguous, we randomly select 5,000 samples from the test set, each consisting of a video, audio, question, and answer. We choose GPT 3.5 turbo as the detection model, which can

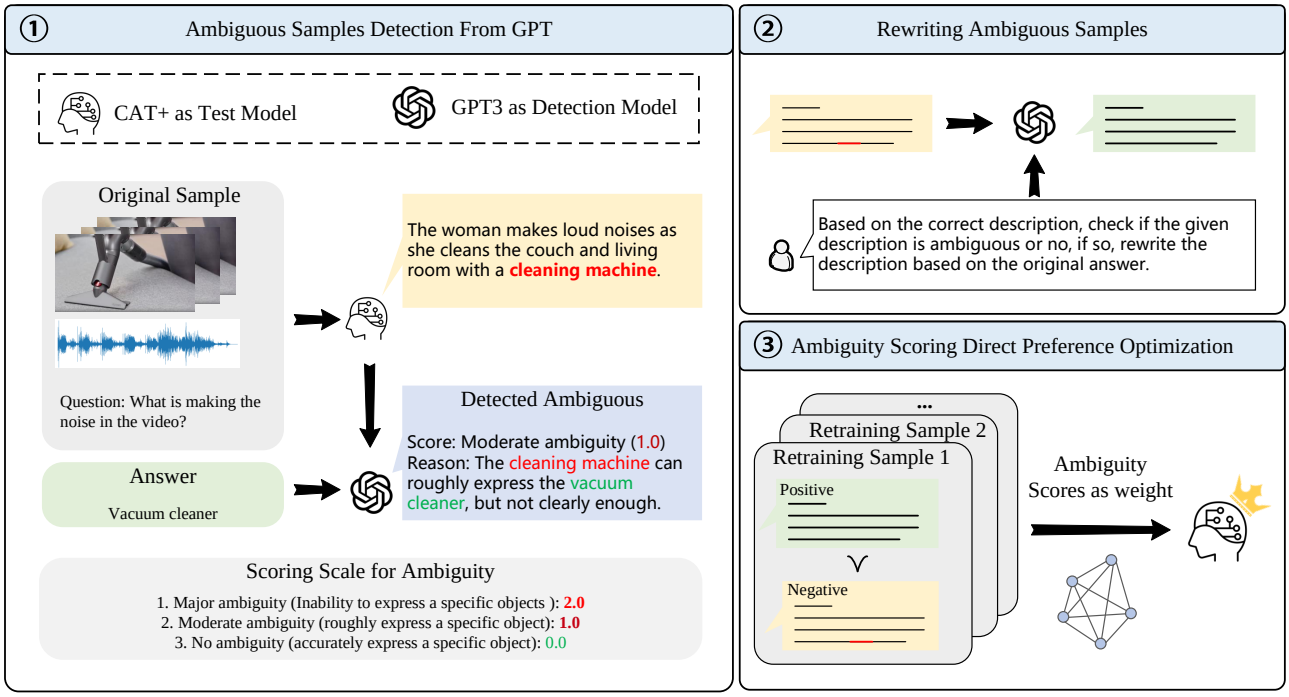


Fig. 5: Overall framework of Ambiguity Scoring Direct Preference Optimization (AS-DPO). Our work consists of three components: (1) Ambiguous Samples Detection From GPT; (2) Rewriting Ambiguous Samples; and (3) Ambiguity Scoring Direct Preference Optimization. Notably, the CAT+ in the first step denotes CAT+ after feature alignment and instruction tuning. The goal of AS-DPO is to skew CAT+ toward positive responses and reject negative responses.

efficiently play the role of a judge to select errors due to GPT's powerful logical inference capability. Specifically, given a test set D_{test}, we first input all audio-visual pairs and questions in D_{test} into the test model to obtain all descriptions to be detected. Then, we develop an ambiguity scoring table which is divided into three metrics, "Major ambiguity", "Moderate ambiguity", and "No ambiguity", where "Major ambiguity" indicates the most ambiguous with a score value of 2.0 and "No ambiguity" indicates that it can correspond to the correct option. We constrain the GPT through this table to score all the descriptions to be detected and finally generate 5,000 instances including the degree of ambiguity by combining them. **(2) how to get positive data.** After detecting all samples from the test set, we now need to collect positive samples based on the degree of ambiguity. As shown in the upper right corner of Fig. 5, we first provide GPT with a detailed prompt template for reviewing the level of ambiguity in each output in D_{test}. The rule for the review is to identify ambiguities in the generated content of "Major ambiguity" and "Moderate ambiguity" based on the given answers, i.e., responses that are significantly different from the original answer, which we refer to as negative responses y_{neg}. Then, we let GPT correct such negative responses, with the principle of generating a positive response y_{pos} that clearly describes the visual object or sound without significant modifications. Finally, repeat the above steps to get the preference dataset D_{pre} containing both positive and negative samples, where D_{pre} includes the samples of "No ambiguity". **(3) how to optimize more efficiently.** After constructing the preference dataset, we need to optimize CAT+ to mitigate the ambiguity in AVQA through the DPO strategy. However, we find a problem: assuming that all samples are weighted in the same way to force the model to learn, the model loses the ability to analyze which samples are more ambiguous and which are not. In other words, this way may cause the model to indiscriminately favor all positive samples without being able to analyze its own errors. Therefore, we propose AS-DPO, a strategy that optimizes the model with fuzzy scores as weights based on DPO, an innovative approach that allows it to learn preferences more consistently. As shown in the lower right corner of Fig. 5, given a policy model f^{pol}, and a reference model f^{ref}, they are both deep copies from CAT+. AS-DPO directly optimizes the low-rank adaptation matrix parameters [17] in the policy model f^{pol} while freezing all the parameters of the reference model f^{ref}. Formally, we define DPO loss $\mathcal{L}_{DPO}$ with ambiguity scoring as:

$$\mathcal{L}_{DPO}\left(f^{pol};f^{ref}\right) = -E_{(x^{mul},x^{txt},y_{pos},y_{neg})\sim D_{pre}} \left[log\sigma\left(\beta log\frac{f^{pol}(y_{pos}|\left[x^{mul}:x^{txt}\right])}{f^{ref}(y_{pos}|\left[x^{mul}:x^{txt}\right])} - \beta log\frac{f^{pol}(y_{neg}|\left[x^{mul}:x^{txt}\right])}{f^{ref}(y_{neg}|\left[x^{mul}:x^{txt}\right])}\right)\right]. \tag{8}$$

$[\ :\ ]$ denotes the concatenation. σ is the non-linear function, Sigmoid, and β is a hyperparameter. This objective function aims to directly optimize the model that favors the positive response y_{pos} and rejects the negative response y_{neg}. Assuming that $\beta = 1$, we simplify the reward goal $\hat{r}$ of the model based on Eq. 8 as:

$$\hat{r}(x^{mul},x^{txt},y) = log\frac{f^{pol}(y|\left[x^{mul}:x^{txt}\right])}{f^{ref}(y|\left[x^{mul}:x^{txt}\right])}. \tag{9}$$

As the variance of $\hat{r}(x^{mul},x^{txt},y_{pos}) - \hat{r}(x^{mul},x^{txt},y_{neg})$ becomes larger, the benefit of that rewarding objective is

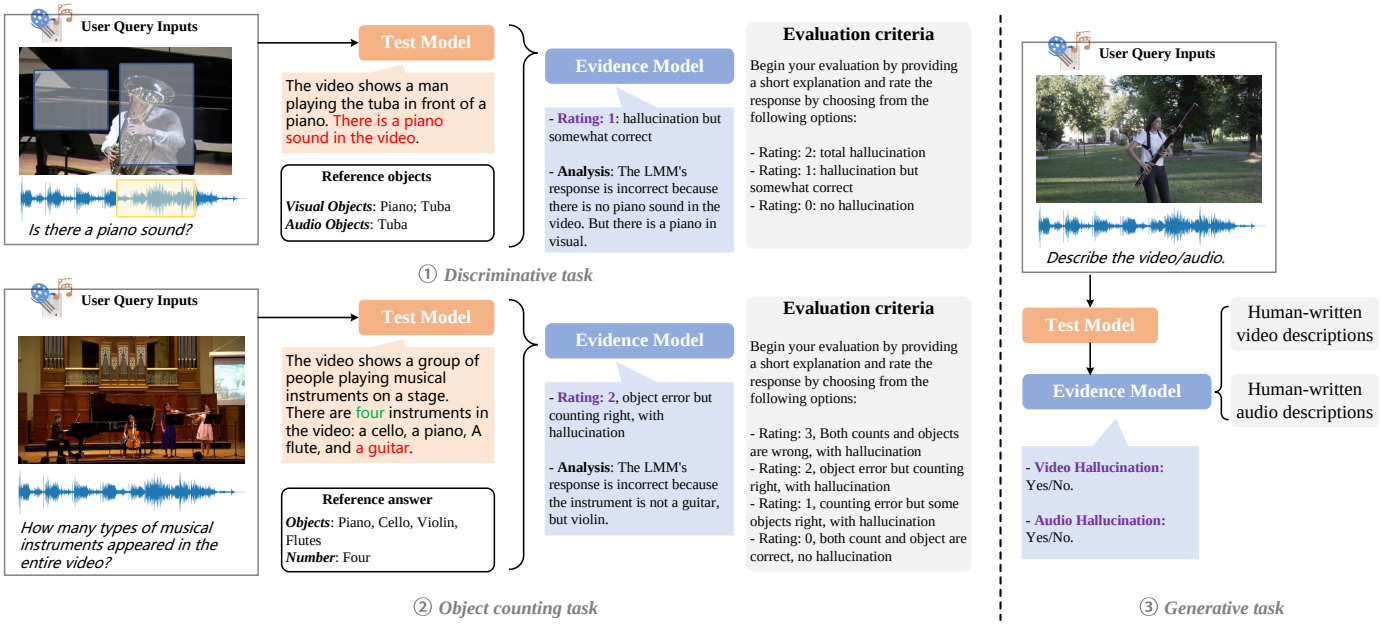


Fig. 6: Illustration of our proposed ***AVHbench*** with three evaluation protocols. ①The discriminative task is mainly used to distinguish whether all objects in a given video and audio correspond to each other. ②The object counting task focuses on assessing the number of objects and the correct object names in the given video and audio. ③The generation task primarily represents an overall description of the given video or audio, with responses scored for the degree of hallucination based on human-written answers.

maximized, when the probability that the model is guided to generate unambiguous descriptions is greater. In this supervised manner, it can achieve optimization of the model parameters to bias the output towards positive samples. However, we argue that such an optimization strategy requires a value to distinguish the severity of ambiguity, the theory being similar to weight assignment. When the ambiguity is mild, we do not expect the model to overlearn such samples where there is no difference between positive and negative responses. Conversely, when the ambiguity is more severe, the model needs to assign a larger weight value to force the model to learn.

Therefore, we introduce an additional objective function $\mathcal{L}_{AS}$, which is analogous to the logistic distribution in the supervised fine-tuning process, and to introduce the degree of ambiguity in order to supervise the model's bias in stabilizing the forward response, we define a weight $Score_{AS} = \frac{1}{\lambda} * s$, where s is the score evaluated by the detection model. It ranges from 0.0 to 2.0, and we set $\lambda = 10$ by default. Then we define $\mathcal{L}_{AS}$ as:

$$\mathcal{L}_{AS} = -\sum \log P(y_{pos} | \left[x^{mul} : x^{txt}\right]; f^{pol}), \quad \{x^{mul} : x^{txt}, y_{pos}\} \sim D_{pre}, \tag{10}$$

where $\log P$ denotes the logistic distribution, and we use this loss as an auxiliary goal in reinforcing the bias toward positive sample outputs. Finally, we combine $\mathcal{L}_{DPO}$ and $\mathcal{L}_{AS}$ through the degree of ambiguity, denoted as:

$$\mathcal{L} = \mathcal{L}_{DPO} + Score_{AS} * \mathcal{L}_{AS}. \tag{11}$$

During AS-DPO training, we introduce weight to supervise the degree of ambiguity of the model, thus responses with severe ambiguity get greater rejection to correct the model's ability to perceive the correct audio-visual objects.

4 Audio-visual Hallucination Benchmark

With the increasing popularity of large-scale fundamental models, there has been a general trend to explore the problem of defects within the field. Methods of assessment based on generated visual descriptions or audio descriptions have been widely explored in unimodal hallucination problems [36], [60], [6]. In contrast, work exploring multimodal hallucinations is still lacking. To bridge this gap, we present ***AVHbench***, a benchmark for audio-visual hallucination detection to evaluate MLLMs in audio-visual scenarios for assessing visual and audio hallucination problems. As shown in Fig. 6, we design three protocols to evaluate hallucinatory deficits under different deployment conditions, i.e., Discriminative tasks, Object Counting tasks, and Generative tasks.

Construction of *AVHbench*. All data for this benchmark comes from the multimodal datasets of audio-visual question answering: Music-AVQA [29] and AVQA [74]. Our dataset construction pipeline is illustrated in Fig. 8, it consists of only two steps in which we incorporate a series of off-the-shelf Qwen models to aid in the annotation process. Specifically, we first use Qwen2-VL [69] to detect visual objects in the video and Qwen2-Audio [8] to detect audio objects. It is worth noting that while Qwen does not make many errors in recognizing visual objects, it often mishears audio. Therefore, we include human review, a step that completely eliminates the interference of hallucinatory visual or audio objects. Furthermore, we summarize the generation of Qwen2-VL and Qwen2-Audio and obtain a list of contexts. In the final step, we use human-written examples to restrict Qwen2 [73] to generate discriminative problems and problems for determining the number of objects. For generative tasks, we remove possible misrepresentations from the context through human review and organize them

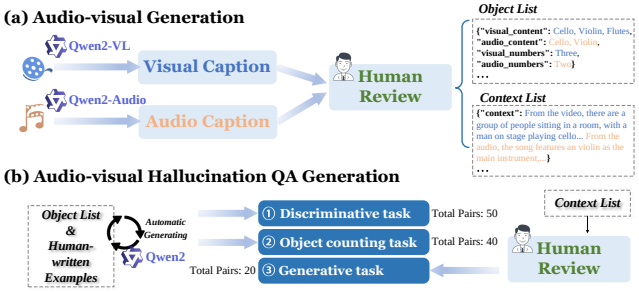


Fig. 7: Overview of the ***AVHbench*** construction pipeline. (a) employs Qwen2-VL [69] and Qwen2-Audio [8] to perform audio-visual object generation and context generation. In this process, we remove erroneous objects through human review. (b) employs Qwen2 [73] to automatically generate Q&A pairs for Tasks ①, and ②, based on human-written examples. In Task ③, we standardize the question to "describe the video" and review the context answers by human review.

into answers. In total, we collect 110 videos and generated 50, 40, and 20 questions for each of the three different tasks.

Protocol 1: Discriminative tasks. This task evaluates whether the test model has the ability to distinguish the correct object in audio-visual scenarios. As shown in Fig. 6 ①, this task is evaluated on three main guidelines with scores of 0, 1, and 2, corresponding to no hallucinations, hallucinations but partially correct, and complete hallucinations, respectively. Where a score of 1 means that we consider the response partially illusory when the object is visually present but not audio, and the test model is only illusory in a single modality. Specifically, we propose four evaluation metrics, which are: **AH**, **VH**, **MH**, and **D-Mean**, which corresponds to the meaning of "*Is there* {*audio*} *object?*" "*Is there* {*visual*} *object?*" "*Are there a* {*audio*} *object and a* {*visual*} *object?*" and the average accuracy of all the metrics. Let $R = [r_1, ..., r_i, ..., r_I]$ denotes the response of the test model $\mathcal{T}$, where r_i is the i-th sentence of the response. We re-formulate the **AH**, **VH**, **MH** below:

$$\mathcal{S}^{I}_{disc} = \frac{1}{\hat{N}} \cdot \frac{1}{\mathcal{C}} \sum \mathcal{GPT}_i(R), I = \textbf{AH}, \textbf{VH}, \textbf{MH}, \quad (12)$$

where $\hat{N}$ is the number of generated responses, $\mathcal{C}$ is the maximum value of the score, $\mathcal{GPT}$ denotes the evidence model, and the score takes a value in 0, 1, 2. Further, **D-Mean** can be expressed as the average accuracy of $\mathcal{S}^{\textbf{AH}}_{disc}$, $\mathcal{S}^{\textbf{VH}}_{disc}$, and $\mathcal{S}^{\textbf{MH}}_{disc}$. On this protocol, we have collected **20**, **20**, and **10** user input queries for **AH**, **VH**, and **MH**.

Protocol 2: Object Counting tasks. In addition to distinguishing whether a particular object is accurate or not, we introduce a setting specifically for the counting problem. This task evaluates the ability of the test model to recognize and count all objects in the audio-visual scenario. As shown in Fig. 6 ②, this task is evaluated on four main criteria with scores of 0, 1, 2, and 3, corresponding to counting and object error, incorrect number but correct object, incorrect object but correct number, and correct object and number, respectively. Specifically, we propose three evaluation metrics, which are: **AH**, **VH**, and **O-Mean**, where **AH** denotes the correctness of the test model in responding to audio objects and quantities, i.e. "*How many* {*sounding*} *in the video?*", **VH** denotes the correctness in responding to visual objects and quantities, i.e. "*How many* {*instruments*} *in the video?*" and **O-Mean** denotes the average accuracy of the above two metrics. We re-formulate the **AH**, **VH** below:

$$\mathcal{S}^{I}_{count} = \frac{1}{\hat{N}} \cdot \frac{1}{\mathcal{C}} \sum \mathcal{GPT}_i(R), I = \textbf{AH}, \textbf{VH}. \quad (13)$$

O-Mean can be expressed as the average accuracy of $\mathcal{S}^{\textbf{AH}}_{count}$ and $\mathcal{S}^{\textbf{VH}}_{count}$. On this protocol, we have collected **20**, and **20** user input queries for **AH**, and **VH**.

Protocol 3: Generative tasks. As shown in Fig. 6 ③, this task serves to assess the prevalence of hallucinations in responses to MLLMs conditioned on video, audio, and text prompts. Hallucinatory universality assessment tests the generality as well as the portability of the large-scale model, and its goal is to find out if there are errors in generating descriptions such as imaginative objects, factually inaccurate objects, etc. based on the correct descriptions. To ensure the accuracy of the evaluation, we formulate holistic descriptions of each video and audio for the collected user input queries through human annotation, and each description must contain all objects as well as verbs in the given audio-visual content. This task contains only one accuracy metric **G-Mean**. We let the evidence model score the test model based on human-written visual descriptions and audio descriptions. The ranking scores take values in the range 0, 1, indicating no hallucinations and hallucinations, respectively. On this protocol, we have collected **20** user input queries for **G-Mean**.

5 Experiments

5.1 Datasets

Video-based text generation tasks. We choose to perform a holistic evaluation of videos, a task whose goal is to assess the model's ability to understand the video. We evaluate our CAT+ on the video understanding benchmarks [45]. Specifically, this benchmark proposes five metrics: *Correctness of Information*, *Consistency*, *Detail Orientation*, *Contextual Understanding*, and *Temporal Understanding*.

Zero-shot on video question answering tasks. To evaluate whether CAT+ is capable of reasoning in unseen scenarios, we choose MSRVTT-QA [71] and ActivityNet-QA [76] for the zero-shot test. MSRVTT-QA and ActivityNet-QA consist of 10k and 5.8k videos respectively containing audio information, and all the questions are derived from daily life and are relatively challenging.

Closed-ended AVQA tasks. The goal of the closed AVQA task is to test MLLMs' ability to recognize specific objects as well as contextual reasoning. We choose Music-AVQA [29] and AVQA [74] as the evaluation datasets, each of which consists of up to 42 candidate answers, with questions that require the model to select the most appropriate answer based on the given visual and auditory content.

Open-ended AVQA tasks. Open-ended AVQA requires models to answer given questions in a free-form format; these types of tasks need the model to interpret, correlate, and reason about cross-modal information using accurate language. We evaluate CAT+'s zero-shot reasoning ability on these benchmarks by opting for audio-visual dialogue (AVSD [1]), and audio-visual captioning (VALOR [4]) tasks.

TABLE 1: GPT-based evaluation [45] for video-based text generation and zero-shot video question answering. For clarity, five scores are reported ("Cr.": Correctness of Information, "Cs.": Consistency, "De.": Detail Orientation, "Ct": Contextual Understanding, "Te.": Temporal Understanding). Score indicates the confidence.

Method	LLM Size	Video-based text generation					Zero-shot on MSRVTT-QA		Zero-shot on ActivityNet-QA	
		Cr.	De.	Ct.	Te.	Cs.	Acc.	Score	Acc.	Score
LLaMA-VID [34]	13B	61.4	61.0	72.0	51.6	52.6	58.9	3.3	47.5	3.3
PLLaVA [72]	34B	72.0	64.0	78.0	53.4	65.0	68.7	3.6	60.9	3.7
CAT+ 13B (Ours)	13B	70.2	**64.2**	75.2	**58.0**	60.0	64.8	3.6	53.4	3.6
Video-LLaMA [80]	7B	39.2	43.6	43.2	36.4	35.8	29.6	1.8	12.4	1.1
LLaMA-Adapter [81]	7B	40.6	46.4	46.0	39.6	43.0	43.8	2.7	34.2	2.7
VideoChat [31]	7B	44.6	50.0	50.6	38.8	44.8	45.0	2.5	26.5	2.2
Video-ChatGPT [45]	7B	48.0	50.4	52.4	39.6	47.4	49.3	2.8	35.2	2.7
VISTA-LLaMA [43]	7B	48.8	52.8	63.6	45.2	46.2	60.5	3.3	48.3	3.3
VideoChat2 [32]	7B	60.4	57.8	70.2	53.2	56.2	54.1	3.3	49.1	3.3
Chat-UniVi [20]	7B	57.8	58.2	69.2	**57.8**	56.2	54.6	3.1	45.8	3.2
LLaMA-VID [34]	7B	59.2	60.0	70.6	49.2	50.2	57.7	3.2	47.1	3.3
CAT [75]	7B	61.6	59.0	69.8	56.2	57.8	62.1	3.5	50.2	3.5
VideoGPT+ [44]	7B	65.4	63.6	**74.8**	56.6	57.8	60.6	3.6	50.6	3.6
CAT+ 7B (Ours)	7B	**66.8**	**61.2**	73.4	57.2	**59.2**	**62.6**	**3.6**	**51.4**	**3.6**

TABLE 2: Comparison with non-LLMs-based methods on Music-AVQA dataset with fully fine-tuned.

Method	Language Model	Audio avg.	Visual avg.	Audio-Visual avg.	Overall avg.
FCNLSTM [12]	CLIP [56]	68.9	56.2	60.4	60.8
GRU [2]	CLIP	68.3	67.0	63.0	65.0
HCAttn [41]	CLIP	64.9	65.3	60.3	62.5
MCAN [77]	CLIP	70.6	71.8	61.5	65.8
PSAC [33]	CLIP	72.0	69.4	63.6	66.6
HME [11]	CLIP	69.9	68.8	64.8	66.7
HCRN [26]	CLIP	63.7	65.2	49.8	56.3
AVSD [61]	CLIP	68.8	70.3	65.4	67.3
Panp-AVQA [78]	CLIP	72.1	73.2	67.0	69.5
AVST [29]	CLIP	73.9	74.4	69.5	71.6
PSTP-Net [28]	BERT [10]	70.9	77.3	72.6	73.5
LAVISH [39]	CLIP	77.1	77.3	77.0	-
CAD [50]	GLoVE [54]	78.1	79.7	76.9	78.2
VALOR [4]	BERT	-	-	-	78.9
VAST [5]	BERT	-	-	-	80.7
CAT 7B [75]	LLaMA2	84.9	86.1	83.2	84.3
CAT+ 7B (Ours)	LLaMA2	**86.2**	**86.8**	**84.6**	**86.4**

5.2 Experimental Setup

Evaluation metrics. For video-based text generation and zero-shot on video question-answering tasks, we follow the evaluation pipeline proposed by Video-ChatGPT [45] that uses GPT-3.5 turbo to evaluate predictive descriptions against correct descriptions, with a score from 0 to 5 indicating accuracy. In this paper, we standardize 0 to 5 as 0 to 100 to align common accuracy rubrics. For closed-ended AVQA tasks, we report the accuracy of the correct sample. For open-ended AVQA tasks, we report the CIDEr [67] that specializes in evaluating visually descriptive tasks.

Architecture. We use frozen ImageBind [13] and LLaMA2 7B [66] as audio-visual encoders and LLM, respectively. The size of modality embeddings for visual is $\mathbb{R}^{T\times 1024}$, where the output of the audio embedding that has been visually deepened is also $\mathbb{R}^{T\times 1024}$. The outputs x^{mul} is $\mathbb{R}^{48\times 4096}$.

Training details. We conduct **Instruction-tuning** and **AS-DPO traing** for CAT+ on 2 NVIDIA A100 GPUs. During instruction tuning, we use the AdamW optimizer with a cosine learning rate decay and a warm-up period. We set r = 128 and alpha = 256 for the LoRA parameters, and the total batch size is set to 128 for training 1 epoch with a learning rate of $2e^{-5}$. During AS-DPO, we set the computation type of the model to $bfloat16$ for faster runtime, and the total batch size is set to 128 for training 1 epoch. Both the policy model and the reference model need to be run during training to compute the difference in the data of the same batch. The reference model does not need to add LoRA parameters as it does not need to be updated.

5.3 Comparison to State-of-the-Art Methods

5.3.1 *Results on video-based text generation tasks*

We follow the benchmark proposed by Video-ChatGPT [45] to evaluate our CAT+ on video-based text generation tasks. As shown in Table 1, in comparison to the current state-of-the-art method VideoGPT+, our CAT+ consistently outperforms the state-of-the-art results in terms of the accuracy of the generated content (Cr.), the ability to describe details (De.), and the coherence of the described video content (Cs.). Notably, we find that CAT+ has the ability to rival work employing the LLaMA-13B scale, and we analyze that thanks to our proposed AS-DPO strategy, which allows CAT+ to gain a huge boost in comprehension. However, our CAT+'s ability to parse video context and focus on temporal information is still suboptimal. ChatUniVi [20] models temporally fine-grained information by combining embeddings of single-frame and long-sequence images, which allows it to better understand long videos. Similarly, VideoGPT+ [44] samples both local and global feature inputs and compresses the resulting tokens, which we believe is a paradigm that improves the model's understanding of the whole.

We include LLaMA2-13B as the large language model decoder to compare other large-scale models, e.g., LLaMA-VID 13B [34] and PLLaVA 34B [72]. The results show that

TABLE 3: Results of different MLLMs on audio-visual hallucination benchmark ***AVHbench***. We expect a lower evaluation result to indicate a lower level of hallucination.

Methods	Discriminative tasks (%)				Object Counting tasks (%)			Generative tasks (%)
	AH	**VH**	**MH**	**D-Mean**	**AH**	**VH**	**O-Mean**	**G-Mean**
Close-source models								
Gemini 1.5 Flash [58]	47.5	37.5	45.0	43.0	45.0	35.0	40.0	45.0
Gemini 1.5 Pro	45.0	35.0	45.0	41.0	**38.3**	30.0	**34.2**	40.0
Gemini 2.0 Flash	**40.0**	**30.0**	**40.0**	**36.0**	43.3	**28.3**	35.8	**30.0**
Open-source models								
Video-LLaMA 7B [80]	75.0	57.5	85.0	70.0	53.3	33.3	43.3	60.0
Video-LLaMA2 7B [7]	45.0	**37.5**	40.0	41.0	**43.3**	36.7	40.0	**35.0**
Video-SALMONN 7B [64]	62.5	65.0	45.0	60.0	45.0	38.3	41.7	45.0
ChatBridge 13B [83]	67.5	50.0	55.0	58.0	48.3	40.0	44.2	65.0
CAT 7B [75]	55.0	42.5	45.0	48.0	51.7	35.0	43.3	50.0
CAT+ 7B (Ours)	**42.5**	40.0	**35.0**	**40.0**	46.7	**31.7**	**39.2**	45.0

TABLE 4: Comparison with LLMs-based methods on zero-shot Music-AVQA dataset.

Method	zero-shot	Acc.
OneLLM 7B [15]	✓	43.0
ChatBridge 13B [83]	✓	47.6
CAT 7B [75]	✓	48.6
CAT+ 7B (Ours)	✓	**50.1**

TABLE 5: Evaluation on zero-shot open-ended AVQA datasets.

Method	Acc.
HGA + HAVF [19]	87.7
HCRN + HAVF [74]	89.0
PSTP-Net [28]	90.2
CAT 7B [75]	92.0
CAT+ 7B (Ours)	**92.6**

TABLE 6: Evaluation on zero-shot open-ended AVQA datasets.

Method	zero-shot	AVSD CIDEr	VALOR CIDEr
VALOR [4]	×	-	61.5
VAST [5]	×	-	62.2
FA+HRED [51]	×	84.3	-
MTN [25]	×	98.5	-
COST [55]	×	108.5	-
OneLLM 7B [15]	✓	74.5	29.2
ChatBridge 13B [83]	✓	75.4	24.7
CAT 7B [75]	✓	79.0	32.4
CAT+ 7B (Ours)	✓	**80.2**	**32.7**

CAT+ 13B outperforms PLLaVA on the video comprehension task for detail modeling, And thanks to SQM's ability to model long video sequentially, it further outperforms PLLaVA in temporal comprehension by 4.6%.

5.3.2 Results on zero-shot question answering tasks

We show the zero-shot question-answering performance of CAT+ on several public video quiz datasets on the right side of Table 1. According to the observation, recently designed MLLMs with bridging modules such as LLaMA-VID [34], VISTA-LLaMA [43], etc., which interact with embedding-level information through improved cross-attention mechanisms, have achieved respectable results. Further, our proposed SQM queries the key information on the basis of the attention level in a continuously modeling manner, which makes CAT+ far ahead of them in recognizing accurate answers. Our approach achieves 62.6% and 51.4% accuracy in the MSRVTT-QA [71] and ActivityNet-QA [76] benchmarks, respectively.

5.3.3 Results on on closed-ended AVQA tasks

We choose to evaluate CAT+'s ability to perceive specific audio-visual objects to answer questions on a close-ended AVQA task (i.e., Music-AVQA ([29])) by means of a fully fine-tuned approach. This task requires the model to produce a description of the audio-visual content in addition to selecting the correct answer among multiple options. According to the results in Table 2, we find that our CAT+ can accurately retrieve specific objects in dynamic audio-visual scenes after fully learning such a recognition task. Compared to other non-LLMs-based approaches, CAT+ completely outperforms all state-of-the-art results. In addition, compared to our previous version of CAT [75], the SQM, which is more adept at reasoning about long sequences, and the improved retraining strategy bring a 2.1% improvement to CAT+ in overall average accuracy. Thanks to LoRA ([17]) which reduces the trainable parameters of the fine-tuning process, we can improve the quality of the evaluation with the help of the world knowledge in LLaMA2 ([66]). By comparison, our CAT+ has only 5.2 million trainable parameters, a significant reduction compared to other paradigms.

Furthermore, we examine CAT+'s ability to perform zero-shot inference on Music-AVQA [29] in Table 4, which shows the results compared with other LLMs-based methods. Notably, to avoid data leakage, we have removed the LoRA parameters fine-tuned by CAT+ on AVinstruct. Compared to other more large-scale LLMs such as OneLLM and ChatBridge, CAT+ still performs slightly better. In addition, we test CAT+'s performance in audio-visual tasks in realistic scenarios in Table 5. The AVQA dataset contains a wide variety of four-option problems that require the model to perform contextual reasoning and select the correct choice based on what it understands from the video and audio. The results show that CAT+ continues to perform well in multi-choice scenarios.

5.3.4 Results on open-ended AVQA tasks

To evaluate the portability and groundedness of CAT+, we test zero-shot reasoning on the open AVQA task in Table 5. We select experiments on the zero-shot complex text reasoning tasks AVSD [1] and VALOR [4], which require responses to everyday events based on audio-visual content. The tests are divided into "0-shot" and "fully supervised" and the results show that, unlike other multimodal-based

TABLE 7: Ablations about SQM and visual deepen on the video understanding dataset [45]. $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the corresponding tiny transformer layers in the SQM, respectively.

Method	Component	Cr.	De.	Te.
w/o visual deepen	-	64.8	59.6	56.8
w/ visual deepen	-	**66.8**	**61.2**	**57.2**
w/o SQM	-	58.2	56.4	52.2
w/ SQM	only $\mathcal{F}_1$	63.8	60.2	56.2
	$\mathcal{F}_1$ + $\mathcal{F}_2$ w/o h_v	28.4	30.6	27.6
	$\mathcal{F}_1$ + $\mathcal{F}_2$ w/o h_a	58.8	57.6	54.0
	$\mathcal{F}_1$ + $\mathcal{F}_2$	**66.8**	**61.2**	**57.2**

TABLE 8: Comparison experiment between Clue Aggregator (CA) [75] and SQM on video understanding [45], where **A** and **V** are average accuracy for audio and video, respectively. Notably, we keep CAT+ in the same setting as CAT, i.e., visual-deepen is removed.

Multimodal-tokens modeling	Video Task		Music-AVQA	
	Cr.	Te.	**A**	**V**
Temporal cross-attention [34]	41.4	32.6	72.9	74.3
Feed-forward network	58.2	52.2	75.1	73.7
one Q-Former [30]	59.8	52.4	80.4	84.6
CA [75]	60.6	54.0	84.9	86.1
SQM	**64.8**	**56.8**	**85.7**	**86.4**

TABLE 9: Trainable parameters in different connectors and FLOPs for Video-LLaMA [80], Video-LLaMA2 [7], CAT [75] and our CAT+.

Model	Decoder	Params. (M)	FLOPs (T)
Video-LLaMA [80]	LLaMA2-7B	137.31	1.73
Video-LLaMA2 [7]	Qwen2-7B [73]	182.88	2.95
CAT [75]	LLaMA2-7B	148.13	1.94
CAT+ (Ours)	LLaMA2-7B	142.04 (↓ 6.09)	**1.21** (↓ 0.73)

LLMs such as OneLLM [15] and ChatBridge [83], we have a superior performance on 0-shot, which shows a realistic application of CAT+. In addition, our method is also close to the results of FA [51] under fully supervised training, further demonstrating CAT's powerful multimodal understanding. Compared to our previous version of CAT, although our improvement is not significant, our proposed AS-DPO is more effective in suppressing the ambiguity of information interactions between modalities, which demonstrates our potential for open-ended AVQA tasks.

5.4 Experiments on Audio-visual Hallucination

5.4.1 Experimental Setting

The main audio-visual-to-text hallucination experiments are conducted using a subset of data from Music-AVQA and AVQA. To avoid data conflicts with AVinstruct, we rewrite the questions based on the original audio-visual objects and assign different query inputs based on three different test tasks (Discriminative tasks, Object Counting tasks, and Generative tasks). We choose 4 open-source audio-visual large language models and our CAT+ as test subjects, 4 of which are based on LLaMA-2 7B [66] (e.g., Video-LLaMA [80], ChatBridge [83], and CAT [75]), and 1 model commercialized open-source large model [58]. In addition, we choose GPT 3.5 turbo as the evidence model to verify whether each generated sentence supports the other or if there is a hallucination.

In order to assess the degree of hallucination of the ranking methodology, the reference samples for the three protocol's metrics are derived from manual annotations and obtained by filtering, where the overall ranking of the table is obtained by averaging the rankings for each metric.

As mentioned in Sec. 4, we report for each test model the **AH** (degree of hallucination of audio objects), **VH** (degree of hallucination of visual objects), **MH** (degree of hallucination of multiple objects), and **D-Mean** (average accuracy of discriminative) on *Discriminative tasks*, the **AH** (recognition and counting of audio objects), **VH** (recognition and counting of visual objects), **O-Mean** (average accuracy of counting tasks) on *Object Counting tasks*, and **G-Mean** (accuracy of overall description) on *Generative tasks*.

5.4.2 Hallucination Ranking Results

We apply the three proposed protocols of ***AVHbench*** to investigate hallucination ranking in audio-visual LLMs, where all the test models allow the processing of video inputs alone or video inputs including audio. Four models, Video-LLaMA [80], Video-LLaMA2 [7], Video-SALMONN [64] and ChatBridge [83], are used as open-source extensible models with different sources of fine-tuning data. Gemini 1.5 Flash [58], Gemini 1.5 Pro, and Gemini 2.0 Flash are used as commercial LLMs, where we characterize the user inputs by invoking the official API.

First, the results of discriminative hallucinations are shown in Table 3, as measured by the proposed **AH**, **VH**, **MH**, and **D-Mean**. Based on the observation that earlier work [83], [80] are poor for the perception of audio-visual object correspondences, e.g., Video-LLaMA scores 75.0% and 85.0% on audio object hallucination and multiple object discrimination, respectively. In contrast, our CAT+ and Video-LLaMA2 improve considerably over other methods. For close-source models, Gemini 2.0 Flash outperforms the other two models overall.

Regarding the object counting task, the overall comparison between Gemini 1.5 Pro and our CAT+ have yielded the best results, with average accuracies of 34.2% and 39.2%, respectively. This task requires strong knowledge-based reasoning within the model, and we observe that Gemini 1.5 Pro outperforms CAT+ overall. We analyze this to be due to the fact that Gemini is trained by a much larger corpus, allowing it to have richer contextual reasoning capability. For the generative description experiments. Video-LLaMA2 achieves state-of-the-art results even beyond the Gemini 1.5 series. We believe that Video-LLaMA2 uses larger-scale training data, making it more capable of localizing in audio-visual scenarios. Due to our designed high diversity of instruction sets, our CAT+ get sub-optimal performance.

5.5 Ablations and Analyses

In this subsection, we explore the impact of the various components of CAT+. We first test the impact of the input multimodal token x^{mul} length on the video-based text generation task and zero-shot question-answering capability. Then, we test the impact of each component within SQM on the overall model. Finally, we test whether AS-DPO brings a boost to the generation task and its generalization.

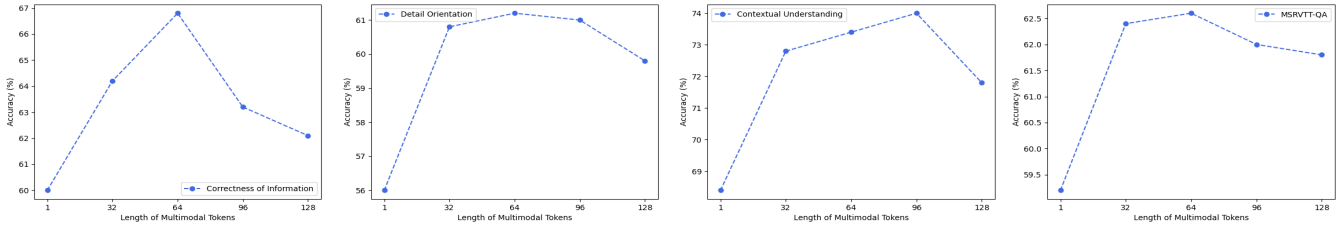


Fig. 8: Different number of multimodal tokens on video-based text generation tasks and zero-shot question-answering tasks.

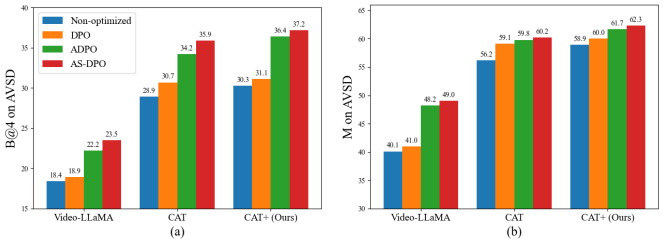


Fig. 9: Ablation experiment about DPO [57], ADPO [75] and AS-DPO with different MLLMs on AVSD. It is worth noting that DPO refers to not adding the $\mathcal{L}_{AS}$ (Section 3.4) and the training strategy remains consistent.

TABLE 10: Performance of CAT+ of varying sizes on the video understanding task [45].

Model	LLM Sizes	Video Understanding		
		Cr.	De.	Te.
CAT+	LLaMA3.2-3B	60.4	48.6	42.0
	LLaMA2-7B	66.8	61.2	57.2
	LLaMA2-13B	**70.2**	**64.2**	**58.0**

5.5.1 *Length of input multimodal token*

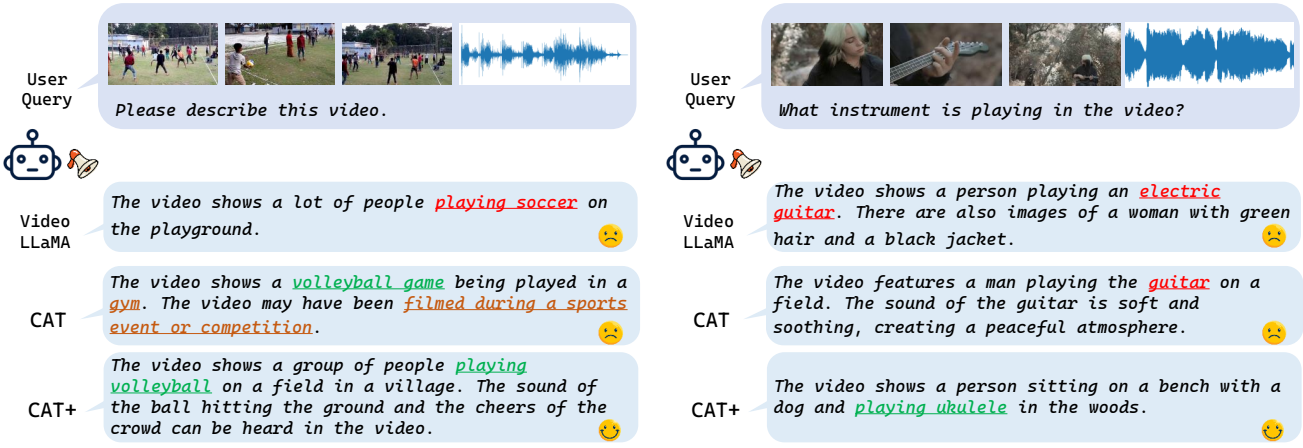
We evaluate the effect of the number of input multimodal tokens on different video-based understanding tasks in Fig. 8. In this study, we test 5 different input lengths: 128, 96, 64, 32, and 1 (time compression) to find the most suitable input length. Observations show that different lengths of inputs bring almost the same effect on CAT+, both of them presenting a weakening of the generative capacity of the model when they are too short or too long. From the stage of 1 to 64, CAT+ shows signs of increasing by growing the input length. This further suggests that the LLM is more sensitive to inputs of appropriate length. However, when we try to expand the input length continuously, CAT+ starts to show signs of decreasing, which suggests that LLM may not be good at modeling non-language tokens with longer temporal information. In addition, an interesting finding is that CAT+ gives the worst results when we compress the length of the input multimodal tokens to 1. We believe that this operation of compression results in a complete loss of input detail, and that many video or audio messages are too similar to be compressed, making these data more homogeneous with each other, which results in the inability of the LLM to distinguish between specific objects. Therefore, after thorough comparison and analysis, we define the length of the input multimodal tokens x^{mul} as 64.

5.5.2 *Ablation on the connector*

Different multimodal tokens modeling. Earlier works [37], [31] connect visual and audio branches in a relatively simple way to large language models. The research [72] has pointed out that modeling long temporal sequences of video is not an optimal strategy, yet we argue that frames contain continuous information, and averaging operations lose this correlation. Traditional methods such as "linear layers", "context attention", and "Q-former" can serve as connectors between multimodal information and LLMs. To verify the effectiveness of SQM compared with them, we have conducted additional experiments to compare the token arrangements in CAT+ while keeping all other settings consistent, i.e., we removed the visual-deepen module from CAT+. As shown in Table 8, the proposed SQM has a temporally longer comprehension of the representation of multimodal tokens. The results show that SQM improves 4.2% in terms of correctness of comprehension and 2.8% in terms of time comprehension over CA [75]. It further demonstrates that the dynamically aligned multi-modal tokens in CAT+ lead to better cross-modal integration and more effective utilization of the language model. Furthermore, our SQM provides an overall improvement over other multimodal tokens modeling strategies in audiovisual comprehension tasks.

An interesting finding is that audio and vision are least capable of characterizing the test set when they are directly employed cross-attention. In contrast, when allocating them to different linear layers, this simple approach is even comparable to the Q-former architecture. This suggests that connecting different video and audio features to the LLM is of intrinsic importance and that different adaptation methods allow the LLM to gain high-level knowledge of different discrete types.

Internal ablation of SQM. We present the ablation results in Table 7 on the impact of SQM and its internal components. To validate the effects of the internal components of SQM, we split $\mathcal{F}_1$ and $\mathcal{F}_2$ to explore the effects of different constituents on dynamic audio-visual pairs of reasoning. We find that its video comprehension is worse when visual features are removed, suggesting that the model relies much more on visual contextual knowledge than auditory. We argue that vision has more mineable information compared to audio, and thus has less impact when audio features are discarded. Next, we explore the impact of visual-deepen on audio expansion. Based on the results, it is shown that utilizing visual enhancement to attenuate some of the noise can bring 2.0%, 1.6%, and 0.4% improvement in video understanding, respectively.



(a) Video Understanding (b) Audio-visual Understanding

Fig. 10: Error cases of Video-LLaMA [80] and CAT [75], and our CAT+ results in audio-visual scenario question-answering. Orange font indicates the hallucinatory descriptions that appeared in CAT.

Trainable parameters. We present the trainable parameters and FLOPs for different MLLMs in Table 9. The result shows that while the simple architecture of Video-LLaMA [80] (containing only a Q-Former and a fully connected layer) achieves faster inference, it underperforms compared to other MLLMs. Our proposed CAT+ rivals Video-LLaMA in terms of the parameters of connectors and provides a much higher performance. Compared to CAT [75], while the trainable parameters of the model have not been reduced much, we discard the separate visual tokens and separate audio tokens, making it run 0.73T faster. More importantly, we further compress the length of multimodal tokens through SQM while ensuring sufficient audiovisual information.

5.5.3 Impact of AS-DPO

Unlike the previous version of the ADPO strategy [75], we introduce a weight value to adaptively correct for samples with severe ambiguity. To show the advantages of AS-DPO over the traditional DPO, we evaluate the generalization of AS-DPO on Video-LLaMA [80], CAT [75] and the effect on AVSD [1] in Fig 9. Specifically, the evaluation indicators B@4 and M denote BLEU-4 [53], and METEOR [24], respectively. We find that supervised fine-tuning alone still underperforms in audio-visual scenarios, and introducing DPO did bring an improvement in values. However, we argue that it is difficult to achieve large gains just by excluding negative responses in the DPO strategy. Therefore, we include a loss toward positive response learning in both ADPO and AS-DPO to supervise the reinforcement process. It can be seen from the results that ADPO already delivers satisfactory performance, and the proposed AS-DPO can further improve the model's descriptive ability without much extra learning cost. This might be because the weight bias allows the model to modify only the negative outputs and leave the parameters unchanged for the positive outputs.

5.5.4 Impact of Different scalability of LLMs

We include LLMs of different sizes as decoders for CAT+ performance on a video comprehension task in Table 10. Compared to a small-scale model, i.e., 3B, the LLaMA2-7B that has been fine-tuned on specific tasks, overall outperforms the LLaMA3.2-3B. We argue that LLaMA3.2-3B struggles with tasks that demand deeper contextual understanding due to the constraints of fewer resources and training parameters, making it difficult to generalize other scenarios. Moreover, CAT+ 13B does not gain much performance in consistency understanding and zero-shot reasoning over CAT+ 7B. We argue that a large number of parameter scales may not directly improve performance, it is necessary to ensure robust connectors to transfer non-language tokens to the LLM.

5.6 Qualitative Analysis

Our CAT+ emphasizes the ability to both provide a simple description of the given audio-visual content and to give answers based on questions. In Fig. 10, we analyze the different qualitative results of Video-LLaMA [80], CAT [75], and CAT+ in different scenarios. In the video understanding example, Video-LLaMA gives a similar but wrong activity answer. In contrast, both CAT and CAT+ can answer volleyball. However, we find that CAT's over-reliance on LLM internal coherence knowledge leads to responses that are hallucinatory, e.g., the activity scene is outdoors rather than in a gym. CAT+ has a keen ability to perceive complex outdoor scenes and our descriptions are able to accurately present information about people, places, and behaviors in the video. In the example of audio-visual understanding, Video-LLaMA and CAT fail to answer the right object in audio. CAT+ benefits from correct audio-visual grounding to accurately localize the ukulele's sound period based on the audio content and respond with an answer after describing the visual content. This further demonstrates the robustness of CAT+'s ability to capture specific objects in audio-visual scenarios.

6 DISCUSSION

Limitation. While CAT+ performs well in the current task, its ability to mitigate audio-visual hallucinations has not yet reached human-level performance, such as the ability of the human ear to clearly distinguish which instruments are playing and which are not. Our future work will focus

on more precise sound source localization, as well as finer-grained sound recognition.

Social impact. Our approach may have both positive and potentially negative social impacts. Regarding potential negative impacts, CAT+ still performs poorly on AVHbench, i.e., there is a possibility that it may respond incorrectly in some scenarios. When MLLMs are applied to automate annotation, this significantly reduces their dependability while users need to perform a lot of data cleansing during their use. In addition, CAT+ is still not a substitute for expert networks for downstream tasks, which poses a further challenge to MLLMs. As for the positive impacts, we identify the shortcomings of existing MLLMSs in real-world scenarios and propose corresponding solutions, which are conducive to the promotion of AI-assisted human society.

7 Conclusion

In this work, we introduce CAT+ to enhance LLMs' generalizability for video-based understanding, as well as open-ended and closed-ended AVQA. To model non-language tokens more robustly, we propose SQM to implement question-guided long sequence queries. Besides instruction fine-tuning, we introduce an optimization strategy guided by the ambiguity scores as weights, called AS-DPO. To investigate hallucinations in audio-visual scenarios, We establish a benchmark called ***AVHbench***. Extensive experiments on diverse audio-visual datasets demonstrate CAT+'s effectiveness. Moreover, our results in AVHbench show CAT+ is more plausible than other MLLMs in audio-visual scenarios.

References

[1] AlAmri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., Lee, S., Parikh, D.: Audio visual scene-aware dialog. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 7558–7567 (2019)

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: IEEE International Conference on Computer Vision, ICCV. pp. 2425–2433. IEEE Computer Society (2015)

[3] Biten, A.F., Gómez, L., Karatzas, D.: Let there be a clock on the beach: Reducing object hallucination in image captioning. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV. pp. 2473–2482. IEEE (2022), https://doi.org/10.1109/WACV51458.2022.00253

[4] Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: VALOR: vision-audio-language omni-perception pretraining model and dataset. CoRR **abs/2304.08345** (2023)

[5] Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. CoRR **abs/2305.18500** (2023)

[6] Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Liang, L., Gu, J., Chen, H.: Unified hallucination detection for multimodal large language models. In: Ku, L., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024. pp. 3235–3252. Association for Computational Linguistics (2024)

[7] Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. CoRR **abs/2406.07476** (2024)

[8] Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., Zhou, J.: Qwen2-audio technical report. CoRR **abs/2407.10759** (2024)

[9] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. CoRR **abs/2305.06500** (2023)

[10] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 4171–4186. Association for Computational Linguistics (2019)

[11] Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 1999–2007. Computer Vision Foundation / IEEE (2019)

[12] Fayek, H.M., Johnson, J.: Temporal reasoning via audio question answering. IEEE ACM Trans. Audio Speech Lang. Process. **28**, 2283–2294 (2020)

[13] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind one embedding space to bind them all. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 15180–15190 (2023)

[14] Gong, K., Feng, K., Li, B., Wang, Y., Cheng, M., Yang, S., Han, J., Wang, B., Bai, Y., Yang, Z., Yue, X.: Av-odyssey bench: Can your multimodal llms really understand audio-visual information? CoRR **abs/2412.02611** (2024)

[15] Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. CoRR **abs/2312.03700** (2023)

[16] Han, M., Kang, M., Jung, H., Hwang, S.J.: Episodic memory reader: Learning what to remember for question answering from streaming data. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics. pp. 4407–4417. Association for Computational Linguistics (2019)

[17] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR (2022)

[18] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 248:1–248:38 (2023), https://doi.org/10.1145/3571730

[19] Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: AAAI. pp. 11109–11116. AAAI Press (2020)

[20] Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. CoRR **abs/2311.08046** (2023)

[21] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. CoRR **abs/2001.08361** (2020)

[22] Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 119–132. Association for Computational Linguistics (2019)

[23] Kim, J., Koepke, A.S., Schmid, C., Akata, Z.: Exposing and mitigating spurious correlations for cross-modal retrieval. In: CVPRW. pp. 2585–2595. IEEE (2023)

[24] Lavie, A., Agarwal, A.: METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Callison-Burch, C., Koehn, P., Fordyce, C.S., Monz, C. (eds.) Proceedings of the Second Workshop on Statistical Machine Translation. pp. 228–231. Association for Computational Linguistics (2007)

[25] Le, H., Sahoo, D., Chen, N.F., Hoi, S.C.H.: Multimodal transformer networks for end-to-end video-grounded dialogue systems. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL. pp. 5612–5623. Association for Computational Linguistics (2019)

[26] Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: IEEE/CVF

Conference on Computer Vision and Pattern Recognition, CVPR. pp. 9969–9978. Computer Vision Foundation / IEEE (2020)
[27] Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR. pp. 13872–13882. IEEE (2024)
[28] Li, G., Hou, W., Hu, D.: Progressive spatio-temporal perception for audio-visual question answering. In: El-Saddik, A., Mei, T., Cucchiara, R., Bertini, M., Vallejo, D.P.T., Atrey, P.K., Hossain, M.S. (eds.) ACM MM. pp. 7808–7816. ACM (2023)
[29] Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: CVPR. p. 19086–19096 (2022)
[30] Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) International Conference on Machine Learning, ICML. pp. 19730–19742. PMLR (2023)
[31] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. CoRR **abs/2305.06355** (2023)
[32] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. CoRR **abs/2311.17005** (2023)
[33] Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: AAAI. pp. 8658–8665. AAAI Press (2019)
[34] Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. CoRR **abs/2311.17043** (2023)
[35] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics (2023)
[36] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.: Evaluating object hallucination in large vision-language models. pp. 292–305. Association for Computational Linguistics (2023)
[37] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. CoRR **abs/2311.10122** (2023)
[38] Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. pp. 3214–3252. Association for Computational Linguistics (2022)
[39] Lin, Y., Sung, Y., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 2299–2309. IEEE (2023)
[40] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. CoRR **abs/2304.08485** (2023)
[41] Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems NIPs. pp. 289–297 (2016)
[42] Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., Tu, Z.: Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. CoRR **abs/2306.09093** (2023)
[43] Ma, F., Jin, X., Wang, H., Xian, Y., Feng, J., Yang, Y.: Vista-llama: Reliable video narrator via equal distance to visual tokens. CoRR **abs/2312.08870** (2023)
[44] Maaz, M., Rasheed, H.A., Khan, S., Khan, F.S.: Videogpt+: Integrating image and video encoders for enhanced video understanding. CoRR **abs/2406.09418** (2024)
[45] Maaz, M., Rasheed, H.A., Khan, S.H., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. CoRR **abs/2306.05424** (2023)
[46] MacGlashan, J., Ho, M.K., Loftin, R.T., Peng, B., Wang, G., Roberts, D.L., Taylor, M.E., Littman, M.L.: Interactive learning from policy-dependent human feedback. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML. pp. 2285–2294 (2017)
[47] Maynez, J., Narayan, S., Bohnet, B., McDonald, R.T.: On faithfulness and factuality in abstractive summarization. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL. pp. 1906–1919. Association for Computational Linguistics (2020)
[48] Mei, X., Liu, X., Huang, Q., Plumbley, M.D., Wang, W.: Audio captioning transformer. In: Font, F., Mesaros, A., Ellis, D.P.W., Fonseca, E., Fuentes, M., Elizalde, B. (eds.) Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events 2021 (DCASE 2021), Online, November 15-19, 2021. pp. 211–215 (2021)
[49] Mei, X., Liu, X., Sun, J., Plumbley, M.D., Wang, W.: Diverse audio captioning via adversarial training. In: ICASSP. pp. 8882–8886. IEEE (2022)
[50] Nadeem, A., Hilton, A., Dawes, R., Thomas, G., Mustafa, A.: CAD - contextual multi-modal alignment for dynamic AVQA. CoRR **abs/2310.16754** (2023)
[51] Nguyen, D.T., Sharma, S., Schulz, H., Asri, L.E.: From film to video: Multi-turn question answering with multi-modal context. CoRR **abs/1812.07023** (2018)
[52] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, NeurIPS (2022)
[53] Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. ACL (2002)
[54] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 1532–1543. ACL (2014)
[55] Pham, H., Le, T.M., Le, V., Phuong, T.M., Tran, T.: Video dialog as conversation about objects living in space-time. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. vol. 13699, pp. 710–726. Springer (2022)
[56] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the International Conference on Machine Learning, ICML. vol. 139, pp. 8748–8763. PMLR (2021)
[57] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. CoRR **abs/2305.18290** (2023)
[58] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T.P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A.M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. CoRR **abs/2403.05530** (2024)
[59] Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045. Association for Computational Linguistics (2018)
[60] Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. pp. 4035–4045. Association for Computational Linguistics (2018)
[61] Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 12548–12558. Computer Vision Foundation / IEEE (2019)
[62] Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.: REPLUG: retrieval-augmented black-box language models. In: Duh, K., Gómez-Adorno, H., Bethard, S. (eds.) NAACL. pp. 8371–8384. Association for Computational Linguistics (2024)
[63] Sun, G., Manakul, P., Liusie, A., Pipatanakul, K., Zhang, C., Woodland, P.C., Gales, M.J.F.: Crosscheckgpt: Universal hallucination

ranking for multimodal foundation models. CoRR **abs/2405.13684** (2024)
[64] Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., Zhang, C.: video-salmonn: Speech-enhanced audio-visual large language models. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024 (2024)
[65] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 809–819. Association for Computational Linguistics (2018)
[66] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. CoRR **abs/2307.09288** (2023)
[67] Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 4566–4575 (2015)
[68] Wang, C., Chen, X., Zhang, N., Tian, B., Xu, H., Deng, S., Chen, H.: MLLM can see? dynamic correction decoding for hallucination mitigation. CoRR **abs/2410.11779** (2024)
[69] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. CoRR **abs/2409.12191** (2024)
[70] Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z.: Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. Visual Intelligence **2**(1), 37 (2024)
[71] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM MM. pp. 1645–1653 (2017)
[72] Xu, L., Zhao, Y., Zhou, D., Lin, Z., Ng, S., Feng, J.: Pllava : Parameter-free llava extension from images to videos for video dense captioning. CoRR **abs/2404.16994** (2024)
[73] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., Fan, Z.: Qwen2 technical report. CoRR **abs/2407.10671** (2024)
[74] Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: AVQA: A dataset for audio-visual question answering on videos. In: ACM MM. pp. 3480–3491. ACM (2022)
[75] Ye, Q., Yu, Z., Shao, R., Xie, X., Torr, P., Cao, X.: CAT: enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. CoRR **abs/2403.04640** (2024)
[76] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. pp. 9127–9134 (2019)
[77] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 6281–6290. Computer Vision Foundation / IEEE (2019)
[78] Yun, H., Yu, Y., Yang, W., Lee, K.I., Kim, G.H.: Pano-avqa: Grounded audio-visual question answering on 360° videos. In: CVPR. p. 2031–2041 (2021)
[79] Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, vol. 162, pp. 25994–26009. PMLR (2022)
[80] Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the Empirical Methods in Natural Language Processing, EMNLP. pp. 543–553 (2023)
[81] Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. CoRR **abs/2303.16199** (2023)
[82] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A.T., Bi, W., Shi, F., Shi, S.: Siren's song in the AI ocean: A survey on hallucination in large language models. CoRR **abs/2309.01219** (2023)
[83] Zhao, Z., Guo, L., Yue, T., Chen, S., Shao, S., Zhu, X., Yuan, Z., Liu, J.: Chatbridge: Bridging modalities with large language model as a language catalyst. CoRR **abs/2305.16103** (2023)



Qilang Ye is a Ph.D. candidate at the College of Computer Science, Nankai University, China, since 2025. He is also a student under the supervision of Zhongguancun Academy, Beijing, China. Previously, he was a Research Assistant at Great Bay University, China. His research interests include multimodal large language models, embodied AI, and human-centric computer vision.



Zitong Yu (Senior Member, IEEE) received the Ph.D. degree in Computer Science and Engineering from the University of Oulu, Finland, in 2022. Currently, he is an Assistant Professor at Great Bay University, China. He was a Postdoctoral researcher at ROSE Lab, Nanyang Technological University. He was a visiting scholar at TVG, University of Oxford. His research interests include human-centric computer vision and biometric security. He was area chairs of ICME 2023, BMVC 2024, and IJCB 2024. He was a recipient of IAPR Best Student Paper Award, IEEE Finland Section Best Student Conference Paper Award, Best Paper Candidate of ICME 2024, second prize of the IEEE Finland Jt. Chapter SP/CAS Best Paper Award, and World's Top 2% Scientists 2023-2024 by Stanford University.



Rui Shao received the BEng degree from the University of Electronic Science and Technology of China, in 2015, and the PhD degree from Hong Kong Baptist University, in 2021, respectively. He is currently a professor with the Harbin Institute of Technology (Shenzhen). His research interests lie primarily in multimodal learning and computer vision. He has published more than 20 papers in top journals and conferences, such as IEEE Transactions on Pattern Analysis and Machine Intelligence,International Journal of Computer Vision, CVPR, and ECCV.



Yawen Cui received B.E. degree in computer science and technology from Jiangnan University, Wuxi, China, and the M.S. degree in software engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2016 and 2019, respectively, and the Ph.D. degree from the University of Oulu, Finland, in 2023. She is currently a postdoctoral researcher at the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. Her research interests include few-shot learning and continual learning.



Xiangui Kang (Senior Member, IEEE) received the B.S. degree from Peking University, Beijing, China, and the Ph.D. degree from Sun Yat-sen University, Guangzhou, China. He was a Visiting Scholar with The University of British Columbia in 2014 and the University of Maryland, College Park, from 2011 to 2012. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University. His current research interests include information forensics/hiding and multimedia communications and security. He was a member of the IEEE Information Forensics and Security Technical Committee. He was the Chair of the Multimedia Security and Forensics Technical Committee of the Asian–Pacific Signal and Information Processing Association. He is a Senior Area Editor of IEEE SIGNAL PROCESSING LETTERS.



Xin Liu (Senior Member, IEEE) received the Ph.D. degree in Computer Science and Engineering from the University of Oulu, Finland, in 2019. He worked at the University of Sydney as an Endeavour Research Fellow from 2017 to 2018. He was a visiting scholar at Imperial College London in 2019. He held a postdoc fellow position at the Academy of Finland. Notable accolades include the ICME 2017 Best Paper Award and Runner-Up of ECCV 2020 VIPriors Challenges. His research interests include human behavior analysis, affective computing, social signal processing, emotion AI, and biometrics.



Philip Torr (Senior Member, IEEE) received the PhD degree from Oxford University. After working for another three years at Oxford, he worked for six years for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He is now a professor with Oxford University. He has won awards from top vision conferences, including ICCV, CVPR, ECCV, NIPS, and BMVC. He is a Royal Society Wolfson Research Merit Award holder.



Xiaochun Cao (Senior Member, IEEE) is a Professor of School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University. He received the B.E. and M.E. degrees both in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. Before joining SYSU, he was a professor at Institute of Information Engineering, Chinese Academy of Sciences. He has authored and coauthored over 200 journal and conference papers. In 2004 and 2010, he was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition. He is on the editorial boards of IEEE Trans. on Image Processing and IEEE Trans. on Multimedia, and was on the editorial board of IEEE Trans. on Circuits and Systems for Video Technology.