ALIGNED BETTER, LISTEN BETTER FOR AUDIO-VISUAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

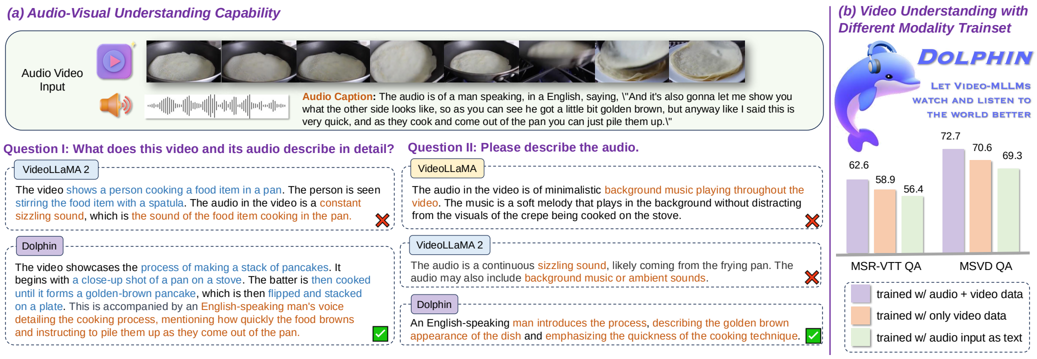
Audio is essential for multimodal video understanding. On the one hand, video inherently contains audio and audio supplies complementary information to the visual modality. Besides, video large language models (Video-LLMs) can encounter many audio-centric settings. However, existing Video-LLMs and Audio-Visual Large Language Models (AV-LLMs) exhibit deficiencies in exploiting audio information, leading to weak understanding and hallucination. To solve the issues, we delve into the model architecture and data aspects. (1) From the architectural perspective, we propose a fine-grained AV-LLM, namely Dolphin. The concurrent alignment of audio and visual modalities in both temporal and spatial dimensions ensures a comprehensive and accurate understanding of videos. Specifically, we devise an audio-visual multi-scale adapter for multi-scale information aggregation, which achieves spatial alignment. For temporal alignment, we propose audio-visual interleaved merging. (2) From the data perspective, we curate an audio-visual caption & instruction-tuning dataset, called AVU. It comprises 5.2 million diverse, open-ended data tuples (video, audio, question, answer) and introduces a novel data partitioning strategy. Extensive experiments show our model not only achieves remarkable performance in audio-visual understanding, but also mitigates hallucinations. Our codes and dataset will be made publicly available.

1 INTRODUCTION

Humans perceive the dynamic world through their eyes and ears, with visual and auditory information complementing each other, both are indispensable. Similarly, the audio modality proves crucial for the comprehensive understanding capabilities of Multimodal Large Language Models (MLLMs) (Liu et al., 2023). On the one hand, audio modality can provide complementary information to the visual modality, aiding MLLMs in more accurate comprehension. On the other hand, there are many audio-centric tasks (Tian et al., 2018; Yang et al., 2022; Mo & Morgado, 2022) in audio-visual (AV) understanding, *e.g.*, AV question answering, AV event localization and AV segmentation. However, most existing Video-LLMs (Lin et al., 2023; Li et al., 2023b; Zhang et al., 2023a) directly neglect audio, with only a small portion incorporating both visual and audio modality. A natural question arises: *How proficient are these models in their audio-visual comprehension capabilities?*

To answer the question, we design two progressive experiments, as in Figure 1. First, we evaluate the AV understanding abilities of current AV-LLMs, *e.g.*, Video-LLaMA (Zhang et al., 2023a) and Video-LLaMA 2 (Cheng et al., 2024). We found they consistently neglect audio and the descriptions solely come from the visual content. Furthermore, when directly querying the audio content, the responses are always the speculations and associations derived from the visual input, rather than the audio itself. For example, when Video-LLaMA is presented with a cooking video, the response is "background music playing throughout the video." But actually, the audio features a male commentator's narration. Besides, when replacing the background sound with white noise, the model's responses remained unchanged, indicating that the model did not extract information from the audio.

Based on the results, it naturally begs the question: *Why do AV-LLMs tend to neglect audio modality?* We deduce the following reasons: **(1) For alignment**, most models lack fine-grained alignment and interactions between modalities, and simply concat the visual and audio tokens. **(2) For datasets size**, large-scale audio-visual instruction-following datasets are scarce, and most works align vision-language and audio-language separately, resulting in less coordinated audio-video representations,



Audio Caption: The audio is of a man speaking, in a English, saying, \"And it's also gonna let me show you what the other side looks like, so as you can see he got a little bit golden brown, but anyway like I said this is very quick, and as they cook and come out of the pan you can just pile them up.\"

Question I: What does this video and its audio describe in detail?

VideoLLaMA 2: The video shows a person cooking a food item in a pan. The person is seen stirring the food item with a spatula. The audio in the video is a constant sizzling sound, which is the sound of the food item cooking in the pan. ✗

Dolphin: The video showcases the process of making a stack of pancakes. It begins with a close-up shot of a pan on a stove. The batter is then cooked until it forms a golden-brown pancake, which is then flipped and stacked on a plate. This is accompanied by an English-speaking man's voice detailing the cooking process, mentioning how quickly the food browns and instructing to pile them up as they come out of the pan. ✓

Question II: Please describe the audio.

VideoLLaMA: The audio in the video is of minimalistic background music playing throughout the video. The music is a soft melody that plays in the background without distracting from the visuals of the crepe being cooked on the stove. ✗

VideoLLaMA 2: The audio is a continuous sizzling sound, likely coming from the frying pan. The audio may also include background music or ambient sounds. ✗

Dolphin: An English-speaking man introduces the process, describing the golden brown appearance of the dish and emphasizing the quickness of the cooking technique. ✓

Figure 1: (a) Audio visual capability of previous AV-LLMs and our Dolphin. We pose questions separately for audio-video and audio, discovering that VideoLLaMA and VideoLLaMA 2 exhibit significant hallucinations for audio understanding, while Dolphin produces accurate responses. (b) Audio could provide complementary information compared to video. Incorporating audio into training greatly enhances video understanding.

(3) In current datasets, visual modality has relatively higher information density, where audio does not provide necessary content for video understanding, AV-LLMs tend to disregard audio.

To solve these problems, we explore from two perspectives, *i.e.*, model architecture and training dataset. From the model perspective, we propose a novel fine-grained AV-LLM, namely Dolphin, which aligns audio and visual modalities both spatially and temporally and effectively harnesses both complementary modalities. For spatial alignment, we propose an audio-visual multi-scale adapter, which extracts multi-scale features and implements audio-visual interaction and merging at various scales. For temporal alignment, we propose audio-visual interleaved merging, where both audio and visual serve for context for each other through interleaved tokens. Finally, the fine-grained tokens aligned both spatially and temporally are projected into the input space of LLM to achieve remarkable audio-visual joint understanding.

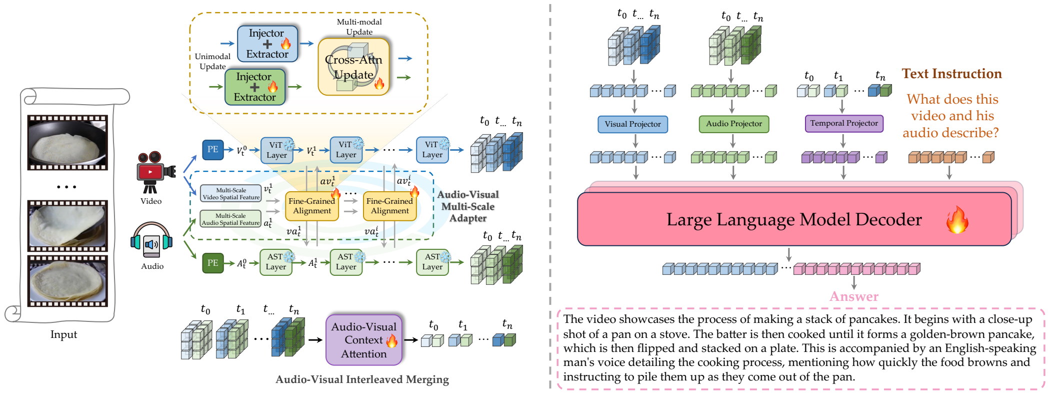
From the dataset perspective, we propose a large-scale audio-visual understanding caption and instruction-following dataset, called AVU, which consists of 5.2M AV caption and question-and-answer (Q&A) tuples. We extract video and audio meta-information and generate high-quality caption and Q&A pairs. The dataset is divided into several splits according to AV consistency for different training objectives. Specifically, we incorporate datasets from several AV tasks, *e.g.*, AVE (Tian et al., 2018), AVL (Mo & Morgado, 2022), AVS (Zhou et al., 2022) and AVVP (Tian et al., 2020), and convert them to fine-grained instruction-following data. Besides, some negative samples for rejection are devised to avoid potential hallucinations. To comprehensively evaluate audio-visual understanding, we further propose a benchmark, called AVU-Bench, for AV-LLMs. We highlight the importance of interaction from both modalities, as in Fig. 1 (b).

In summary, we contribute in the following aspects:

- We propose Dolphin, a fine-grained AV-LLM for audio-video multimodal and unimodal understanding. Dolphin could remarkably exploit audio information for understanding.
- The core innovation lies in the architecture of audio-visual multi-scale spatial alignment and contextual and temporal alignment, which ensures fine-grained extraction of two complementary modalities and interaction between them.
- We curate the first large-scale audio-visual caption and instruction-following dataset. It contains 5.2M samples with several splits, which is suitable for training AV-LLMs.
- Extensive experiments show that Dolphin could not only achieve outstanding audio-visual understanding performance, but also be competent in unimodal tasks, which validates that Dolphin effectively exploits the information of two complementary modalities.

2 RELATED WORKS

Multi-Modal Large Language Models for Video Understanding. A series of studies first decompose videos into different representation dimensions and then integrate the inputs to enrich the



What does this video and his audio describe?

The video showcases the process of making a stack of pancakes. It begins with a close-up shot of a pan on a stove. The batter is then cooked until it forms a golden-brown pancake, which is then flipped and stacked on a plate. This is accompanied by an English-speaking man's voice detailing the cooking process, mentioning how quickly the food browns and instructing to pile them up as they come out of the pan.

Figure 2: Overview of our Dolphin, which aligns on both spatial and temporal dimensions to fully exploit the natural consistency of videos and enhance the complementary roles of vision and hearing. Specifically, for spatial alignment, we introduced an audio-visual multi-scale adapter using a dual-feature pathway design, which extracts multi-scale features from both visual and auditory inputs and achieves fine-grained alignment with the respective modality.

prompts for MLLMs. For example, Video-ChatGPT (Maaz et al., 2023) splits videos into spatial and temporal branches for pooling. VideoChat (Li et al., 2023b) decomposes videos into descriptions and video embeddings. LLaMA-VID (Li et al., 2023c) represents each frame with context tokens and content tokens. Video-LaVIT (Jin et al., 2024) performs video tokenization using keyframes and motion vectors. Other works incorporate images into unified video training to enrich the training data. Chat-UniVi (Jin et al., 2023) and Video-LLaVA (Lin et al., 2023) employ strategies such as object-based adaptive cluster-based tokens and aligning before projection to unify image and video inputs, thereby achieving more powerful visual understanding.

Audio-Visual Large Language Models. An early work VideoChat (Li et al., 2023b) simply encodes audio inputs using Whisper (Radford et al., 2023) and directly overlays them to the textual input. Later works seek to align the output of audio and visual encoders before feeding them into LLMs. MACAW-LLM (Lyu et al., 2023) aligns the encoder outputs to the textual space through a learnable alignment module. Audio-Visual LLM (Shu et al., 2023) activates the embeddings of different modalities with different tags. Moreover, some works (Sun et al., 2023a; Zhang et al., 2023a) explore temporal alignment between video and audio using a Q-Former (Li et al., 2023a) structure, but most of them (Tang et al., 2024) neglect fine-grained spatial modeling. Meerkat (Chowdhury et al., 2024) explores fine-grained understanding but only focuses on images. To sum up, most existing AV-LLMs neither struggle to capture fine-grained local information nor handle temporal alignment, which motivates us to delve into the design and training of AV-LLMs.

3 MODEL ARCHITECTURE

Overview. Dolphin primarily focuses on effectively strengthening the fine-grained alignment and interaction between visual and auditory modalities. It effectively exploits the complementary information of two modalities and prevents overlooking any of them. Specifically, Dolphin primarily comprises three components: (1) Audio-visual (AV) multi-scale adapter (Section 3.1), which aims to align audio and visual features across various spatial scales in a fine-grained manner. (2) Audio-visual (AV) interleaved merging (Section 3.2) for temporal alignment and extracting complementary information of two modalities. (3) Large Language Model (LLM) to handle the interacted audio-visual tokens and output responses according to the instructions.

Notations. We split each video into $T = 8$ visual frames and audio clips. Let H_v, W_v denote the height and width of each frame, while each audio clip is transformed into a spectrogram of $H_a \times W_a$. N denotes the number of ViT blocks. In the AV multi-scale adapter, the visual modality contains dual-pathway input features, *i.e.*, global feature $\boldsymbol{\mathcal{V}}_t^i$ and multi-scale local feature $\boldsymbol{v}_t^i$, where i denotes the i-th block and t denotes the t-th frame. Besides, $\widehat{\boldsymbol{\mathcal{V}}}_t^i$ and $\hat{\boldsymbol{v}}_t^i$ are features after modality interaction. Similar expressions of $\boldsymbol{\mathcal{A}}_t^i, \boldsymbol{a}_t^i, \widehat{\boldsymbol{\mathcal{A}}}_t^i, \hat{\boldsymbol{a}}_t^i$ are applicable to audio. The joint feature after modality interaction could be represented as $\boldsymbol{av}_t^i$. In AV interleaved merging, visual and audio tokens after contextual interaction could be written as $\boldsymbol{\mathcal{V}}_t^{temp}$ and $\boldsymbol{\mathcal{A}}_t^{temp}$, respectively.

3.1 SPATIAL: AUDIO-VISUAL MULTI-SCALE ADAPTER

The AV multi-scale adapter is designed to enhance the fine-grained alignment of spatial features. Taking the spirit of ViT-Adapter (Chen et al., 2023), we first independently extract each modality's global and pyramid multi-scale features. Then fine-grained alignment is performed across different scales. In this section, we introduce the data flow with the visual modality as an example, and the same principle applies to audio as well.

Visual global and initial multi-scale features. For ViT-L (Dosovitskiy et al., 2021), we divide it into $N = 4$ blocks, each with 6 layers. To acquire multi-scale spatial features, we feed the images into the spatial module, *i.e.*, pyramid convolutional network (Lin et al., 2017), to extract various multi-resolution features, *i.e.*, 1/8, 1/16 and 1/32 of $H_v \times W_v$, forming the initial D-dimensional multi-scale features $\boldsymbol{v}_t^1$ as an input to the adapter, as shown below:

$$\boldsymbol{v}_t^1 \in \mathbb{R}^{\left(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2}\right)\times D}, \quad \boldsymbol{a}_t^1 \in \mathbb{R}^{\left(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2}\right)\times D}. \tag{1}$$

Inter-modality feature interaction. To achieve fine-grained alignment across scales, audio features are injected into the multi-scale visual features $\boldsymbol{v}_t^i$, where i denotes the i-th block, we incorporate audio global feature $\boldsymbol{\mathcal{A}}_t^i$ into $\boldsymbol{v}_t^i$ through cross-attention, and obtain audio-guided multi-scale visual feature $\boldsymbol{av}_t^i$ as follows:

$$\boldsymbol{av}_t^i = \mathrm{CrossAttn}(\boldsymbol{v}_t^i, \boldsymbol{\mathcal{A}}_t^i, \boldsymbol{\mathcal{A}}_t^i). \tag{2}$$

Intra-modality feature fusion. As shown in Figure 2, in the fusion process, the audio-guided multi-scale visual features $\boldsymbol{av}_t^i$, which contains visual spatial priors and audio information, are injected into original global features $\boldsymbol{\mathcal{V}}_t^i$ through cross-attention and added to $\boldsymbol{\mathcal{V}}_t^i$, as follows:

$$\widehat{\boldsymbol{\mathcal{V}}}_t^i = \boldsymbol{\mathcal{V}}_t^i + \beta^i\,\mathrm{CrossAttn}(\boldsymbol{\mathcal{V}}_t^i, \boldsymbol{av}_t^i, \boldsymbol{av}_t^i). \tag{3}$$

Here, β is a learnable vector initialized as 0, ensuring the initial outputs are the same as ViT's global features $\boldsymbol{\mathcal{V}}_t^i$. Subsequently, the global features $\widehat{\boldsymbol{\mathcal{V}}}_t^i$ are transmitted to the next block through the i-th standard ViT block and the fine-grained features $\hat{\boldsymbol{v}}_t^i$ are passed to the next block as $\boldsymbol{v}_t^{i+1}$ through cross-attention and FFN layers:

$$\boldsymbol{\mathcal{V}}_t^{i+1} = \text{ViT-Block}^i(\widehat{\boldsymbol{\mathcal{V}}}_t^i) \tag{4}$$

$$\boldsymbol{v}_t^{i+1} = \hat{\boldsymbol{v}}_t^i + \mathrm{FFN}\left(\hat{\boldsymbol{v}}_t^i\right), \quad \hat{\boldsymbol{v}}_t^i = \boldsymbol{v}_t^i + \mathrm{CrossAttn}\left(\boldsymbol{v}_t^i, \boldsymbol{\mathcal{V}}_t^{i+1}, \boldsymbol{\mathcal{V}}_t^{i+1}\right). \tag{5}$$

Final outputs. The AV multi-scale adapter obtains $\boldsymbol{\mathcal{V}}_t^{N+1} \in \mathbb{R}^{B\times T\times L_v\times D}$ and $\boldsymbol{\mathcal{A}}_t^{N+1} \in \mathbb{R}^{B\times T\times L_a\times D}$, where B, T, D denote batch size, number of frames and feature dimension, and L_v, L_a are the number of tokens of two modalities. We perform average pooling across the temporal dimension T and obtain the final spatial tokens. In this way, we gradually incorporate audio information into the visual features and enhance the interaction between audio and visual modalities. The guidance provided by audio to the multi-scale visual features facilitates fine-grained alignment.

3.2 TEMPORAL: AUDIO-VISUAL INTERLEAVED MERGING

In this stage, we merge the final audio and visual features to implement alignment and interaction at the temporal axis. Specifically, by concatenating visual and audio tokens in the same frame, we could obtain T pairs of audio-visual interleaved tokens, each referred to as an AV group, as shown in Figure 2. Within each group, we perform bi-directional contextual attention on visual and audio tokens, which produce visual-contextualized audio tokens and audio-contextualized visual tokens.

$$\boldsymbol{\mathcal{V}}_t^{temp} = \mathrm{CrossAttn}(\boldsymbol{\mathcal{A}}_t, \boldsymbol{\mathcal{V}}_t, \boldsymbol{\mathcal{V}}_t), \quad \boldsymbol{\mathcal{A}}_t^{temp} = \mathrm{CrossAttn}(\boldsymbol{\mathcal{V}}_t, \boldsymbol{\mathcal{A}}_t, \boldsymbol{\mathcal{A}}_t). \tag{6}$$

Then each AV token group is mapped into the input space of LLM through a joint audio-visual projector. Here, we condense each frame of the video into two tokens. In this way, we achieve integration and merging of audio and visual information, which enhances the audio-visual information exploitation of AV-LLMs.

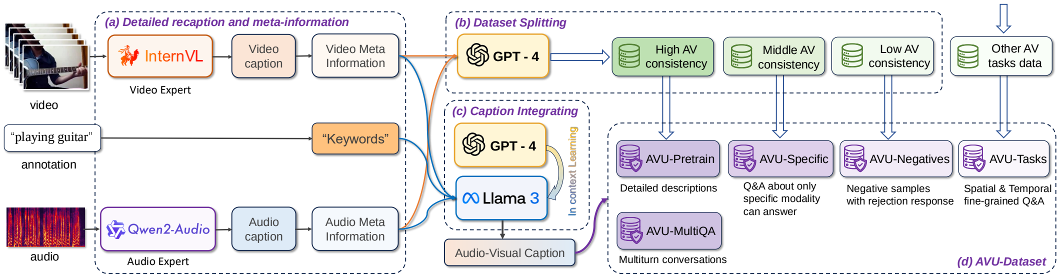


Figure 3: The integration pipeline of the audio-visual understanding dataset (AVU-dataset).

Table 1: The detailed statistics for our AVU (Audio-Visual Understanding Dataset).

Subsets	AVU-Pretrain	AVU-MultiQA	AVU-Specific		AVU-Negetive	AVU-Tasks	
Instruction Type	Detailed Description	Multiturn Conversation and Reasoning	Visual-Specific Info	Audio-Specific Info	Negative Samples	Temporal and Spatial	Total
Audio-Video Statistics	1.11M	500k (subset as AVU-Pretrain)	554k		186k	283k	2.13M
Instruction Statistics	1.11M	1M	1.09M	1.11M	370k	559k	5.24M

3.3 TRAINING STRATEGY

We employ Vicuna-v1.5 (Chiang et al., 2023) as the LLM. The video encoder is ViT-L/14 from CLIP (Radford et al., 2021), and the audio encoder adopts ImageBind (Girdhar et al., 2023). (1) In the initial pre-training phase, we freeze the visual encoder, audio encoder, and LLM and only update all the projectors as well as the AV multi-scale adapter to achieve alignment across visual, auditory, and LLM modal spaces. We use the AVU-Pretrain dataset. (2) For the instruction tuning phase, only the visual and audio encoders are frozen, while other modules are updated. We employ AVU-Multi Q&A, AVU-Specific, AVU-Negatives and AVU-Tasks subsets. See Section 4.3 for more details.

4 AVU: AUDIO-VISUAL UNDERSTANDING DATASET

Motivation. Currently, there is a shortage of large-scale audio-visual instruction-following and fine-grained captions, which hinders the model from focusing on modality-specific information and potentially leads to audio hallucinations. To solve the issues, we propose an audio-visual understanding (AVU)-dataset, a large-scale AV understanding and instruction-following dataset.

Overview. In this section, we introduce the construction of the AVU-dataset (Figure 3). Specifically, we first re-caption (Section 4.1) audio and video data and generate meta-information (Section 4.2). Then the dataset is split (Section 4.3) into several subsets according to the similarity between audio and visual meta-information, and we feed different prompt templates to generate the corresponding instruction-following dataset and training stages.

Dataset statistics. As shown in Table 1, AVU-dataset contains 2.13M audio-video pairs, each has several Q&A pairs, resulting in 5.24M Q&A pairs in total. AVU-dataset has four subsets, *i.e.*, AVU-Pretrain (1.11M samples and Q&A), AVU-Multi Q&A (500k samples and 1M Q&A pairs), AVU-Specific (554K samples and 1.09M Q&A pairs for video and 1.11M for audio), AVU-Negative (186Ksamples and 370K Q&A pairs) and AVU-Tasks (283K samples and 559K Q&A pairs).

Datasets quality and verification. We design three types of filtering mechanisms, including CLIP-Score filtering, Self-consistency filtering, and Annotation filtering, to filter noisy samples and guarantee the , and then human verification is implemented. Details are shown in the Appendix B.

4.1 DETAILED RE-CAPTION GENERATION

Source datasets. We collect widely used audio-visual datasets, including AudioSet-2M (Gemmeke et al., 2017), VGG-Sound (Chen et al., 2020a) and task-specific datasets, *i.e.*, MUSIC (Li

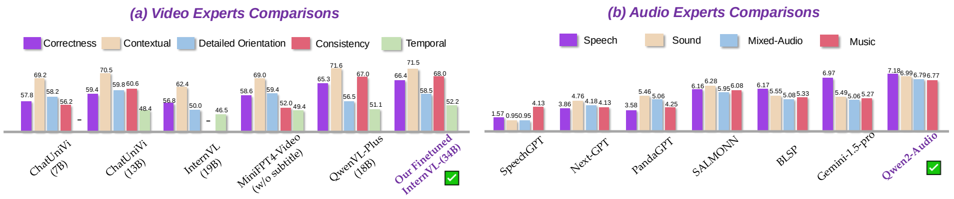


Figure 4: Performance comparison of different task-specific experts.

[Audio-video information]:
This is the **detailed caption of the video**: {...}; and **the meta-information**: {...}. This is the **detailed caption of the audio:** {...}; and **the meta-information:** {...}. These above contain important information that needs to be read carefully.*

AVU-Specific

INPUT:

[Audio-video information]

You have to finish the following two tasks. First, carefully compare the information typically included for audio versus video. Second, Identify information that is included for audio but not for video, and generate a QA pair for each. Similarly, identify information that is included for video but not for audio, and generate a QA pair for each.

Each pair should simulate a dialogue that focuses on various aspects of the content without introducing information not explicitly mentioned. The purpose of this exercise is to fine-tune a language model so that it can generate accurate and relevant responses.

The QA pair should start with **"Q:"** and contain an instruction related to the audio and video, while the second line should start with **"A:"** and provide a response to the instruction.

Please ensure that your instructions are diverse and of high quality. Avoid making assumptions or adding details that are not explicitly given.

Figure 5: Examples of prompt templates for generating AVU-dataset, others are in the appendix.

AVU-Pretrain

Please provide a detailed description of the video.

In the video, a person is seen interacting with a small frog on a wooden surface. The person's finger is seen touching the frog's back, and the frog responds by moving around. The video captures the interaction between the person and the frog, showcasing the frog's movements and the person's gentle touch.

AVU- Specific

What kitchen tools or utensils are prominently featured in the **video** during pancake preparation?

The video showcases basic kitchen tools such as a mixing bowl, whisk, spatula, and frying pan, demonstrating their use in making pancakes.

Discuss the overall tone and style of the English **audio** commentary. How does it enhance the tutorial?

The video showcases basic kitchen tools such as a mixing bowl, whisk, spatula, and frying pan, demonstrating their use in making pancakes.

AVU-Negatives

Can you describe the color of the frog in the audio?

The color of the frog is not mentioned in the audio. The background audio is funny music.

What is the child doing in the video?

The video does not show any children. Please confirm if this is the correct statement of the issue.

AVU-Tasks

When does the dog bark?

In the video, the dog barks from the fifth to the eighth second.

What happened in seconds five to eight?

A dog is barking from the fifth to the eighth second.

Figure 6: Some examples of AVU-dataset.

et al., 2022) for AVQA (Yang et al., 2022), Flickr-SoundNet (Arandjelovic & Zisserman, 2017), VGG-SoundSource for AV source localization (Mo & Morgado, 2022), AVE dataset for AV event localization (Tian et al., 2018), AVS-dataset for AV segmentation (Zhou et al., 2022), and LLP for AV video parsing (Tian et al., 2020). Then we utilize expert models to re-caption audio and video to obtain audio, video and audio-video captions.

Expert models. For MLLM, we employ a fine-tuned version of InternVL-34B (Chen et al., 2024). For the audio expert captioners, we choose Qwen2-Audio-7B-Instruct (Chu et al., 2023). The performance of these models is illustrated in Figure 4.

Prompt templates. We prompt expert models with hand-crafted templates to generate detailed captions, and set some regularizations to mitigate hallucinations. Figure 5 displays a subset: AVU-Specific prompt templates. See Figure 8 and Figure 9 in the appendix for more details.

4.2 META-INFORMATION GENERATION AND INTEGRATION

Meta-information generation. The function of meta-information is to maintain audio and video details when integrating unimodal audio and video captions into multimodal audio-video captions.

Meta-information could be utilized to guide subsequent AV caption integration and dataset splitting. Specifically, we design options of ‘event’, ‘object’, ‘scene’, ‘place’, ‘action’ and ‘emotion’, *etc*., and transform original annotations to ‘keywords’. Then, we employ GPT-4 (Achiam et al., 2023) to judge consistency based on keywords and meta-information and filter noisy samples, which might be caused by experts’ bias and hallucinations.

Meta-information integration. We choose LLaMA3-70B-Instruct (Dubey et al., 2024) to integrate meta-information. First, we employ GPT-4 to generate multiple examples of integrating keywords and meta-information. After human adjustments, we feed them to LLaMA3. The in-context learning abilities enable the generation of audio-visual captions.

4.3 DATASET SPLITS

Splitting process. The dataset is split according to the consistency between audio and video meta-information. **Notably, we do not require strict audio-visual consistency, which is the main novelty of this work**. For different types of data, we create various types of instruction-tuning datasets for corresponding training stages. The whole dataset is divided into three subsets based on the audio-visual consistency. Specifically, we feed audio and video meta-information to GPT-4 (Achiam et al., 2023) to obtain the consistency score. Figure 6 shows examples of subsets of AVU-dataset.

AVU-Pretrain comprises samples with high AV consistency. The audio and video information are nearly the same. These samples are suitable for the pre-training stage to align AV modalities. We design fixed question templates (Figure 8 (a)) and randomly select one each time, and use the previously integrated AV captions as the answers. In this way, we obtain AVU-Pretrain subsets.

AVU-Multi Q&A also consists of high consistency samples. Different from AVU-Pretrain, we design templates (Figure 8 (b)) to transfer AV caption to multi-turn Q&A and reasoning.

AVU-Specific comprises samples with medium AV consistency. Both audio and video carry relatively additional information compared to each other. Questions are posed regarding this additional information to generate Q&A pairs. These Q&A pairs construct AVU-Specific subsets and could only be answered by focusing on a specific modality (Figure 9 (c)).

AVU-Negatives consist of low-consistency samples, *e.g.*, the sounding object is not present in the frame. Taking the spirit of contrastive learning (Chen et al., 2020b; He et al., 2020), we create the negative sample dataset, *i.e.*, AVU-Negatives (Figure 9 (d)), whose answers are primarily used as rejection. This subset could teach LLMs the rejection option, mitigating potential hallucinations.

AVU-Tasks are curated directly from downstream AV tasks. Specifically, we transform the original annotations to the format of Q&A, including accurate details like time, spatial, and event. Notably, the AVU-Tasks are derived from the fine-grained annotations, and significantly contribute to the model’s fine-grained alignment capabilities.

For pre-training, we mainly employ high-consistency samples to enhance modality alignment, while for instruction-tuning, we make use of samples that audio and visual do not exactly overlap and mix AVU Multi Q&A, AVU-Specific, AVU-Negatives and AVU-Tasks for training. In this way, the issue of neglect of audio and hallucination is mitigated.

5 EXPERIMENTS

We introduce the experimental setup and comparisons among models. In the ablation studies, we explored and validated that fine-grained alignment significantly aids LLMs in multimodal understanding. Temporal contextual alignment is an effective way to leverage the inherent consistency of videos and to exploit complementary audio-visual information. Additionally, we also conducted ablations on the proposed dataset, showcasing its assistance in audio-visual joint perception and its high quality.

5.1 EXPERIMENTAL SETUP

Implementation details. For each video, we extract 8 frames of 224×224 resolution, and audio is sampled into 8 frames, each turned into a 128×204 spectrogram. Both pre-training and fine-tuning are conducted for one epoch, with batch sizes of 256 for pretraining and 128 for finetuning. Projectors

Table 2: Comparison with existing Video-LLMs. We conducted a performance comparison with the existing Video-LLM and reported the scoring results of GPT on four zero-shot video-QA datasets.

Method	LLM	MSVD-QA		MSRVTT-QA		ActivityNet-QA		TGIF		POPE
		Acc	Score	Acc	Score	Acc	Score	Acc	Score	
LLaMA-Adapter (Zhang et al., 2023b)	LLaMA-7B	54.9	3.1	43.8	2.7	34.2	2.7	-	-	-
Video-Chat (Li et al., 2023b)	Vicuna-7B	56.3	2.8	45.0	2.5	26.5	2.2	-	-	-
Video-ChatGPT (Maaz et al., 2023)	Vicuna-7B	64.9	3.3	49.3	2.8	35.2	2.7	-	-	-
BT-Adapter (Liu et al., 2024)	Vicuna-7B	67.5	3.7	47.0	3.2	45.7	3.2	-	-	-
LLaMA-VID (Li et al., 2023c)	Vicuna-7B	69.7	3.7	57.7	3.2	47.4	3.3	-	-	-
LLaMA-VID (Li et al., 2023c)	Vicuna-13B	70.0	3.7	58.9	3.3	47.5	3.3	-	-	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	70.7	**3.9**	59.2	**3.5**	45.3	3.3	70.0	**4.0**	84.4
PandaGPT (Sun et al., 2024)	Vicuna-7B	46.7	-	23.7	-	-	-	-	-	78.5
VideoLLaMA (Zhang et al., 2023a)	Vicuna-7B	51.6	2.5	29.6	1.8	12.4	1.1	-	-	82.9
AV-LLM (Shu et al., 2023)	Vicuna-7B	67.3	-	53.7	-	47.2	-	-	-	-
AVicuna (Tang et al., 2024)	Vicuna-7B	70.2	-	59.7	-	**53.0**	-	-	-	84.1
OneLLM (Han et al., 2024)	LLaMA2	56.5	-	53.8	-	-	-	-	-	-
FAVOR (Sun et al., 2023b)	Vicuna-7B	67.8	-	59.3	-	-	-	-	-	-
VideoLLaMA 2 (Cheng et al., 2024)	Mistral-Instruct	71.7	-	57.4	-	49.9	-	-	-	85.4
video-SALMONN (Sun et al., 2024)	Vicuna-7B	67.9	3.7	59.5	3.4	-	-	-	-	-
Dolphin-7B-LoRA (Ours)	Vicuna-7B	71.6	**3.9**	61.3	3.4	47.9	**3.4**	70.9	3.9	85.1
Dolphin-7B (Ours)	Vicuna-7B	72.7	**3.9**	62.6	3.5	49.1	**3.4**	71.2	3.9	85.9
Dolphin-13B (Ours)	Vicuna-13B	**74.8**	**3.9**	**63.5**	**3.5**	49.6	**3.4**	**71.3**	**4.0**	**86.2**

Table 3: Comparison with Audio-LLMs. We conducted closed-ended and open-ended auditory tasks with LTU and LTU-AS, where ZS denotes zero-shot evaluation.

Method	Audio Classif.	Audio Caption	Speech Recognition	Emotion Recognition	Gender Classif.	Age Pred.	Music Genre Classif.	Audio Question	Speech Question	Audio Hallucination
	ESC-50 (ACC↑)	AudioCaps (SPICE↑)	Librispeech (WER↓)	IEMOCAP (ACC↑)	Voxceleb2 (maro-F1↑)	Voxceleb2 (MAE↓)	GTZAN (ACC↑)	(ACC↑)	(ACC↑)	Random (ACC↑)
Best specialized models trained supervisedly on each dataset. Not generalizable to unseen label sets and tasks.										
TASK-SOTA	97.0	17.7	1.4	70.6	98.3	9.4	-	-	-	
CLIP-like audio-text model.										
AudioClip (Guzhov et al., 2022)	69.4	-	-	-	-	-	-	-	-	-
CLAP (Huang et al., 2013)	82.6	-	-	-	-	-	25.2	-	-	-
LTU-Audio (Gong et al., 2023)	82.8	17.0	104.2	38.2	77.0	-	29.8	96%	69%	-
LTU-Speech (Gong et al., 2023)	10.9	0.5	12.9	69.8	90.1	7.9	23.5	65%	93%	-
LTU-AS (Gong et al., 2023)	80.8^{zs}	15.0	4.9	65.2	90.8	7.3	50.3^{zs}	96%	94%	50.1
VideoLLaMA (Zhang et al., 2023a)	62.6^{zs}	6.2	128.4^{zs}	23.4^{zs}	43.5^{zs}	8.8	22.2^{zs}	56%	27%	43.2
VideoLLaMA 2 (Cheng et al., 2024)	74.8^{zs}	15.8	9.8^{zs}	63.9^{zs}	89.7^{zs}	7.3	34.9^{zs}	92%	91%	56.8
video-SALMONN (Sun et al., 2024)	77.6^{zs}	16.6	3.9^{zs}	65.5^{zs}	90.6^{zs}	7.4	36.7^{zs}	95%	94%	58.2
Dolphin-LoRA (Ours)	81.6^{zs}	17.2	12.8^{zs}	67.4^{zs}	91.2^{zs}	7.2^{zs}	33.6	96%	93%	58.6
Dolphin (Ours)	83.1^{zs}	17.8	8.3^{zs}	69.2^{zs}	92.5^{zs}	7.0^{zs}	37.8	96%	94%	63.2

for audio, video, and audio-video use two-layer MLPs with a GELU (Hendrycks & Gimpel, 2016) activation. Training is performed on NVIDIA A100 GPUs. More details are in the appendix.

Dataset. During training, we enhanced our model by mixing inputs from multiple modalities. Apart from using AVU, we employ LLaVA (Liu et al., 2023), 10% of Valley, and audio clips for pre-training, LLaVA_instruct, Video-ChatGPT (Maaz et al., 2023), and ClothoV2 (Drossos et al., 2020) for instruction-based finetuning. We assessed our model's zero-shot capabilities on single-modality video and audio Q&A benchmarks and compared it against existing audio-visual LLMs using a tailored downstream task benchmark.

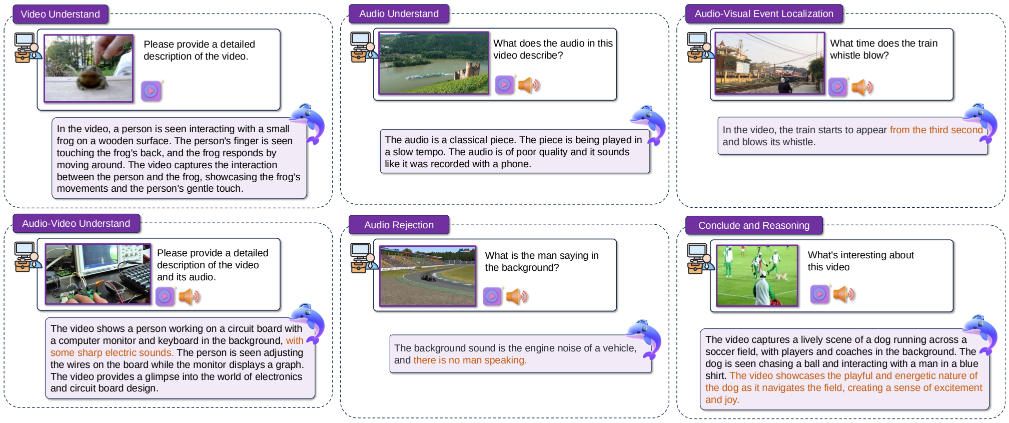
5.2 COMPARISON WITH STATE-OF-THE-ARTS

Zero-Shot Video Understanding. To validate our model's efficacy, we compared its performance in video comprehension with existing video-LLMs. Specifically, we showcased Dolphin's comparison with various methods on MSRVTT-QA, MSVD-QA, and ActivityNet-QA benchmarks. As indicated in Table 2, our model demonstrated superior video understanding, proving that auditory information complemented visual comprehension in the training stage. Moreover, POPE Results show our method could mitigate hallucinations.

Closed and Open-Ended Audio Tasks. To test our model's audio understanding, we followed LTU and LTU-AS, evaluating closed-ended and open-ended audio tasks like classification and captioning in table Table 3. Despite zero-shot settings on many datasets, our model showed effective understanding, achieving or exceeding SOTA performance. We use the pre-trained ImageBind-aud as our audio

Table 4: Results on the proposed audio-visual understanding bench and AVSD.

Task	AVU		AVSL		AVE		AVVP		AVSD(Alamri et al., 2019)
Dataset	MUSIC		AVL, AVS		AVE		LLP		AVSD(Alamri et al., 2019)
Metric	acc	score	acc	score	acc	score	acc	score	acc
Video LLaMA (Zhang et al., 2023a)	65.4	3.3	35.6	2.3	40.8	2.8	21.5	2.1	36.7
Video LLaMA 2 (Cheng et al., 2024)	65.4	3.3	35.6	2.3	40.8	2.8	21.5	2.1	57.2
video-SALMONN (Sun et al., 2024)	65.4	3.3	35.6	2.3	40.8	2.8	21.5	2.1	51.6
Dolphin (Zero-shot)	76.8	3.7	42.2	2.4	49.6	2.8	28.1	2.6	59.1
Dolphin	**78.2**	**3.9**	**51.8**	**3.0**	**52.1**	**3.2**	**31.4**	**2.8**	**59.1**



Please provide a detailed description of the video.

In the video, a person is seen interacting with a small frog on a wooden surface. The person's finger is seen touching the frog's back, and the frog responds by moving around. The video captures the interaction between the person and the frog, showcasing the frog's movements and the person's gentle touch.

What does the audio in this video describe?

The audio is a classical piece. The piece is being played in a slow tempo. The audio is of poor quality and it sounds like it was recorded with a phone.

What time does the train whistle blow?

In the video, the train starts to appear from the third second and blows its whistle.

Please provide a detailed description of the video and its audio.

The video shows a person working on a circuit board with a computer monitor and keyboard in the background, with some sharp electric sounds. The person is seen adjusting the wires on the board while the monitor displays a graph. The video provides a glimpse into the world of electronics and circuit board design.

What is the man saying in the background?

The background sound is the engine noise of a vehicle, and there is no man speaking.

What's interesting about this video

The video captures a lively scene of a dog running across a soccer field, with players and coaches in the background. The dog is seen chasing a ball and interacting with a man in a blue shirt. The video showcases the playful and energetic nature of the dog as it navigates the field, creating a sense of excitement and joy.

Figure 7: Qualitative cases of Dolphin.

encoder and freeze it during training, whose structure is AST (also used in LTU). It is less effective in speech recognition compared with Whisper (used by LTU-AS). However, our model showed improved speech recognition, benefiting from our dataset mixing audio and speech.

Audio-Visual Understanding Bench. To assess our model's competency in audio-visual understanding, we developed a minibench tailored for Large Language Models (LLMs) that aligns with existing audio-visual tasks. We meticulously gathered test datasets from labeled audio-visual tasks, comprising MUSIC (AVQA), LLP (AVVP), AVE (AVE), AVS-Bench (AVS), Flickr-SoundNet, and VGG-SoundSource (AVL), to ensure a fair and precise evaluation. Inspired by zero-shot question-answer evaluations and Video-ChatGPT, we transformed these tasks' ground truths into open-ended question-answer formats. We also evaluate our model on AVSD(Alamri et al., 2019) bench.

This method evaluates the precision of our model's predictive output, awarding scores on a 1-5 scale. Comparing the performance of existing methods, like Video-LLaMA (Zhang et al., 2023a), Video-LLaMA 2 (Cheng et al., 2024), and video-SALMONN (Sun et al., 2024), our results, presented in Table 4, demonstrate a notable performance gap favoring our model on audio-centric vision tasks, which highlights our model's enhanced comprehension of audio and visual modalities.

Qualitative evaluation. The qualitative cases of the proposed Dolphin are illustrated in Figure 7. Results show that Dolphin could remarkably comprehend both audio and visual modalities, together with enhanced video understanding.

5.3 ABLATION STUDIES AND ANALYSIS

In this section, through detailed ablation studies, we explored how to enhance an audio-visual LLM's integration of video and audio for better video understanding. We also validated the effectiveness of our proposed methods and datasets through model and dataset ablations.

Fine-grained spatial alignment effectively aids AV-LLMs in understanding multimodal semantic information. In our prior analysis, we found that without fine-grained interaction between visual and auditory information, LLMs struggle to learn relevant information between videos and audios, as the lack of prior knowledge regarding the two modalities leads the model to focus more on the information-rich video content while neglecting audio. To investigate this issue, we conducted ablation studies on our fine-grained alignment module in Table 5a. The results reveal that inter-modality interaction indeed enhances the LLM's attention to both modalities. However, in comparison, when the two modalities undergo fine-grained alignment before projection, the model is not only able to simultaneously focus on both modalities but also better extract the complementary information from visual and audio. Table 5a illustrates that fine-grained audio-visual alignment aids LLMs in cross-modal understanding. Inter-modal interaction boosts attention to both modalities, but fine-grained alignment before projection further allows simultaneous focus and better extraction of visual and auditory complementary information. Compared to removing temporal alignment, understanding of videos significantly improves with the help of temporal alignment and interaction.

Contextual alignment effectively leverages the inherent consistency of videos. We analyzed our temporal context attention module in Table 5a and found that models without it or with only the chronological arrangement of tokens underperform. In contrast, employing temporal context attention, which aligns video and audio features over time, significantly boosts performance by tapping into their inherent temporal consistency, thus enhancing LLM's understanding of both modalities together.

Table 5: Ablations on model architecture designs (a) and various dataset subsets (b).

Module	Variants	AVU		AVE	
		acc	score	acc	score
Original	Dolphin (All)	**78.2**	**3.9**	**52.1**	**3.2**
Spatial	w/o inter-modal	77.6	3.8	51.8	3.2
	w/o AV adapter	75.4	3.6	51.6	3.2
Temporal	w/o bi-dir context attn	76.1	3.6	50.3	3.1
	w/o AV inter-merging	32.3	2.5	22.6	2.2

(a) Model architecture designs.

Variants	MSR-VTT Q&A		AV Understand		POPE
	acc	score	acc	score	
w/o AVU-Pretrain	58.8	3.3	69.8	3.6	81.7
w/o AVU-Specific	59.3	3.4	72.6	3.6	82.1
w/o AVU-Negatives	61.0	3.5	77.8	3.8	84.8
Full AVU	**62.6**	**3.5**	**78.2**	**3.9**	**85.9**

(b) AVU-dataset subsets.

Table 6: Detailed ablation on datasets and models. Here 'V-L' denotes Video-LLaMA.

Datasets	MSRVTT Q&A		Audiocaps	AV Understand	
Metrics	acc	score	SPICE	acc	score
V-L + V-L dataset	29.6	1.8	11.8	65.4	3.3
V-L + AVU (Ours)	55.3 (+25.7)	3.1 (+1.3)	15.9 (+4.1)	73.2 (+7.8)	3.6 (+0.3)
Dolphin + V-L dataset	42.8 (+13.2)	2.7 (+0.9)	14.4 (+2.6)	69.9 (+4.5)	3.4 (+0.1)
Dolphin + AVU (Ours)	62.6 (+19.8)	3.4 (+1.6)	17.8 (+3.4)	78.2 (+6.3)	3.9 (+0.5)

Effectiveness of model structure and dataset quality. We conducted an ablation study on our model and dataset in Table 5b. First, training without the audio dataset diminished the model's understanding of pure video content, indicating that our model effectively utilizes complementary information from audio. Moreover, comparing the performance without our dataset and training Video-LLaMA with our dataset (Table 6) showed a significant performance decline without the AVU dataset, whereas Video-LLaMA achieved better performance on our dataset. These result, along with the human validation in Appendix.B, jointly demonstrates the quality and effectiveness of our dataset.

6 CONCLUSION

In this paper, we primarily explored and studied how existing AV-LLMs can effectively overcome insufficient attention to audio information. We innovatively proposed Dolphin, an Audio-Visual Large Language Model for video understanding, which features fine-grained interaction and alignment on both spatial and temporal levels. We designed a multi-scale audio-visual adapter and a temporal context module to fully leverage the inherent consistency of videos and realize the complementary function of visual and auditory information. Extensive experiments indicate the effectiveness of the model. Additionally, we collected and labeled an audio-visual caption and instruction fine-tuning dataset for video understanding, containing 2.13M pairs of AV samples and 5.24M Q&A pairs and providing diverse training modalities for audio-visual LLMs. This instruction fine-tuning dataset suitable for large model learning has been proven to effectively enhance the audio-visual understanding capabilities of existing models (such as Video-llama), thereby proving the quality of our dataset. Finally, in the experimental section, we explored and concluded that fine-grained temporal and spatial alignment can effectively help audio-visual LLMs better acquire visual and auditory abilities.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7558–7567, 2019.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020a.

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

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. *arXiv preprint arXiv:2407.01851*, 2024.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

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

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

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

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

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jeff Huang, Charles Zhang, and Julian Dolby. Clap: Recording local executions to reproduce concurrency failures. *Acm Sigplan Notices*, 48(6):141–152, 2013.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023c.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.

Ruyang Liu, Chen Li, Yixiao Ge, Thomas H Li, Ying Shan, and Ge Li. Bt-adapter: Video conversation is feasible without video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13658–13667, 2024.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

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

Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv preprint arXiv:2310.05863*, 2023a.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv preprint arXiv:2310.05863*, 2023b.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6420–6429, 2023c.

Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 247–263, 2018.

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 436–454. Springer, 2020.

Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *ACM International Conference on Multimedia*, 2020.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023b.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2022.

A DIFFERENCES FROM EXISTING METHODS

Comparison between Dolphin and LTU (Gong et al., 2023) and LTU-AS. We summarize the following four differences: (1) Different model types. LTU and LTU-AS are audio/speech-specific models, trained specifically with audio/speech-language models for audio or speech tasks. In contrast, our Dolphin model is an AV-LLM that comprehends both audio and video, encompassing a broader range of modalities. (2) Different audio backbones. LTU-AS employs Whisper (Radford et al., 2023), which is pre-trained on a large-scale speech-language dataset, resulting in stronger speech recognition performance. We use the ImageBind-aud encoder (used by LTU), which has not been pre-trained on a speech dataset. Moreover, the primary tasks in Table 3, such as speech recognition, emotion recognition, and gender classification, are closely related to speech. Therefore, our performance reflects zero-shot results. (3) Compared to LTU-AS, our audio encoder's zero-shot recognition performance still achieves state-of-the-art results in the majority of tasks. This demonstrates our model's ability to pay attention to and comprehend audio effectively. (4) Compared to LTU, our performance surpasses theirs by a considerable margin, which can be attributed to our generated dataset that includes both audio and speech samples. This highlights the effectiveness of our dataset.

B DETAILS OF DATASET CURATION AND VERIFICATION

For the dataset filtering, we design three types of filtering mechanisms: CLIP-Score filtering, self-consistency filtering and annotation filtering. Subsequently, human verification is implemented to quantitatively verify the quality of the generated caption.

- **CLIP-score filtering**. In this stage, the visual and audio experts first generate captions based on the input video and audio, then we employ CLIP and CLAP to assess the similarity score for each caption, respectively. Captions with lower scores might have noise and hallucinations.
- **Self-consistency filtering**. We further prompt the visual and audio experts to summarize the meta-information of the generated captions, and utilize GPT4 to assign the matching score given the initial caption the its meta-information. Captions with lower scores might have noise and hallucinations.
- **Annotation filtering**. The original annotations of the datasets are transformed to ‘keywords’. Then, we employ GPT-4 to judge consistency based on keywords and meta-information and filter noisy samples, which might be caused by experts' bias and hallucinations.
- We synthesized the aforementioned three factors to calculate a weighted confidence score for each sample and subsequently ranked them. The bottom 25% of samples, based on this ranking, will be eliminated. The weights for the screening criteria are as follows: CLIP-Score filtering: 2, Self-consistency filtering: 1, Annotation Filtering: 5.

Human verification. After the former three filtering steps, 100 human annotators are employed to give scores (1 to 5) to each of the video-caption and audio-caption pair, considering the completeness and accuracy (related to hallucinations). We randomly sample 100 video-caption and 100 audio-caption pairs in the generated dataset. Besides, we also verify the caption after the integration from two modalities (Integration Effect). The results are shown in Table 7.

Table 7: The mean scoring of human verification on completeness and accuracy of audio (A) and visual (V) modalities.

Captions	V: Completeness	V: Accuracy	A: Completeness	A: Accuracy	Integration Effect
Mean scoring	4.27	4.45	4.23	4.40	4.31

The verification results show that the generated dataset is precise and accurate.

C FURTHER IMPLEMENTATION DETAILS

Training details. Stage-1: Multi-modal text alignment pre-training. The model needs to read audio, video and corresponding textual instructions, which is related to the understanding of audio, video and audio-video contents. The ground truth annotations are the captions of AVU-dataset. Stage-2: Audio-visual dynamic instruction-tuning. The model is required to respond accordingly to various types of instructions. The instructions comprise complex visual, audio and audio-visual understanding tasks. Both projector and LLM are updated. Dataset prompts, audio or/and video and questions are fed to the model to generate answers. The generated answers are then supervised by the ground truth captions of the datasets to update both the projectors and LLMs. For both two stages, the learning objective is autoregressive next-word-prediction loss.

The proposed spatial and temporal modules are mutually-promoted. The two modules are utilized to align fine-grained spatial and temporal information for audio-visual data. The AV multi-scale adapter separately extracts multi-scale features from visual and auditory modalities and interacts with the other modality for fine-grained alignment. This promotes fine-grained alignment between audio and video, avoiding the neglect of auditory information and effectively enhancing the model's performance in audio-visual dense prediction tasks. The primary innovation of the temporal interleaved merging lies in simultaneously calculating bidirectional attention for both audio and video features, enabling the alignment of video and audio features in the temporal dimension. This effectively leverages the consistency of videos, improving the LLM's joint understanding of video and audio.

D FURTHER EXPERIMENTAL RESULTS

Comparison with task-specific expert models. Dolphin is a general-purpose AV-LLM, which could handle various audio-visual downstream tasks. Here, we compare Dolphin with task-specific expert models. Specifically, we compare with FNAC (Sun et al., 2023c) on audio-visual localization (AVL) on the Flickr dataset, the metric is CIoU. For audio-visual event localization (AVE), we compare with CMRAN (Xu et al., 2020) and report accuracy metric. For audio-visual question answering (AVQA), we compare with AVQA (Li et al., 2022) on MUSIC dataset using accuracy metric. Results are shown in Table 8.

Table 8: Comparison with task-specific models on AVL, AVE and AVQA.

AVL (CIoU)	AVE (acc)	AVQA (acc)
FNAC: 85.14	CMRAN: 78.3	AVQA: 69.51
Dolphin: 86.22	**Dolphin: 79.5**	**Dolphin: 76.56**

Comparison with various audio and visual encoder variants. We compared various variants of audio and visual encoders and reported the performance for video/audio/audio-video understanding, as shown in Table 9.

Then the following conclusions are drawn: (1) If both audio and visual aspects are aligned with language (as in CLIP (Radford et al., 2021), CLAP (Elizalde et al., 2023)), the performance on A/V understanding tasks tends to be good, but the performance on AV tasks is not high. We attribute this to the fact that the AV encoder, due to lack of alignment, is unable to effectively utilize the AV adapter. (2) Building upon this, we added stage 0.5 before pretraining and pre-trained for one epoch on AudioSet-2M (Gemmeke et al., 2017) using the audio-visual contrastive loss. We observed performance improvement (especially in audio), which validates that fine-grained audio-visual alignment effectively helps the model pay attention to the audio modality. (3) If both audio and video use ImageBind (Girdhar et al., 2023), which has undergone audio-visual alignment, the A/V/AV understanding capabilities decline. We believe that in Machine Learning Language Models, since the output format is language, using a backbone that has not been aligned with language will affect the final understanding results. (4) If the audio modality is aligned with language (CLAP) and the visual modality is aligned audio-visually (ImageBind-vid), the results are inferior compared to the other mode (ImageBind-aud+CLIP). We attribute this to the fact that the semantic density of video is much higher than audio, and video-language alignment can better enhance the model's understanding and instruction-following capabilities.

In summary, we believe that selecting a visual encoder aligned with language and an audio encoder aligned with visual can effectively balance A/V/AV understanding capabilities. Therefore, we chose CLIP+ImageBind as our backbone encoder.

Table 9: Comparison with various audio and visual encoder variants.

id	Encoders	MSRVTT Q&A	Audiocaps	AV Understand
(a)	CLIP+CLAP	62.5	17.3	74.2
(b)	ImageBind-vid+ImageBind-aud	58.2	15.8	72.3
(c)	CLIP+ImageBind-vid	59.8	17.6	76.5
(d)	ImageBind-aud+CLIP (Ours)	61.3	17.2	78.2

E DATASET DETAILS

The motivation of the dataset. We list the purpose of the proposed AVU-dataset as follows: (1) There is a lack of relevant datasets. Currently, there are no large-scale audio visual captioning and instruction tuning datasets available. Existing methods have only been trained on vision-language and audio-language datasets, resulting in a deficiency of audio-visual alignment, which consequently leads to suboptimal performance in audio visual tasks. (2) We analyze that one of the primary reasons existing models overlook the audio modality is that, in most cases, audio does not provide more information than video. Therefore, one of our significant innovations lies in our dataset, which specifically selects samples where audio conveys more information than video. We have transformed this audio-specific information into question-and-answer pairs, effectively addressing the issue of AV-LLM's neglect of audio. (3) Existing audio-visual datasets exhibit diverse annotation formats (bounding boxes, timestamps, masks). Consequently, we have standardized the input and output formats for audiovisual tasks such as AVE, AVL, AVQA, and AVVP, facilitating the training of AV-LLM. Additionally, we have provided a dataset with fine-grained temporal and spatial granularity (AVU-tasks).

The contribution of the dataset. The contribution of our AVU-dataset is summarized as follows: (1) The AVU dataset addresses the current lack of large-scale, high-quality audiovisual instruction tuning datasets within the community. (2) The AVU dataset features a rich variety of subsets, effectively addressing the issue of AV-LLM neglecting the audio modality and significantly reducing the occurrence of hallucinations. (3) The introduction of meta-information in the AVU dataset, which is categorized based on AV consistency, allows for the extraction of modality-specific information (AVU-specific) and the generation of negative samples (AVU-negative). This approach can be widely applied in the process of creating other datasets.

Prompt templates. The prompt templates are shown in Figure 8 and Figure 9.

[Audio-video information]:
This is the **detailed caption of the video**: {...}; and **the meta-information**: {...}. This is the **detailed caption of the audio:** {...}; and **the meta-information:** {...}. These above contain important information that needs to be read carefully.*

I . AVU-Pretrain

INPUT:

[Audio-video information]

Your task is to generate a detailed caption for an audio-video segment based on provided meta-information and detailed descriptions. Ensure the caption is comprehensive and accurate, using only the provided information without adding any fictional content.

EXAMPLE:

- The video starts with a woman standing in the kitchen, holding a knife and a potato. She begins slicing the potato thinly, showcasing her skill and precision. The camera captures her movements as she continues slicing, demonstrating the food preparation process. Soft ambient sounds fill the kitchen, including faint simmering from a nearby stove and occasional utensil clinks.

II. AVU-MultiQA

INPUT:

[Audio-video information]

Your task is to generate ten pairs of instructions and responses based on the provided meta-information and detailed captions of the audio and video content. Each pair should simulate a dialogue that focuses on various aspects of the content without introducing information not explicitly mentioned. The purpose of this exercise is to fine-tune a language model so that it can generate accurate and relevant responses.

The QA pair should start with **"Q:"** and contain an instruction related to the audio and video, while the second line should start with **"A:"** and provide a response to the instruction.

Please ensure that your instructions are diverse and of high quality. Avoid making assumptions or adding details that are not explicitly given.

EXAMPLE:

- **Q:** Describe how the video integrates visual steps and audio tips to guide viewers in making pancakes.
  A: The video combines close-up visuals of pancake preparation with descriptive audio commentary, emphasizing quick cooking and stacking instructions.

- **Q:** Provide a summary of the video's approach to combining visual demonstrations and audio explanations.
  A: The video seamlessly integrates detailed visual demonstrations of pancake preparation with insightful audio explanations, offering viewers a comprehensive guide to achieving perfect pancakes.

Figure 8: Prompt templates for generation of AVU-dataset subsets.

[Audio-video information]:

This is the **detailed caption of the video**: {...}; and **the meta-information**: {...}. This is the **detailed caption of the audio:** {...}; and **the meta-information:** {...}. These above contain important information that needs to be read carefully.*

II. AVU-Specific

INPUT:

[Audio-video information]

You have to finish the following two tasks. First, carefully compare the information typically included for audio versus video. Second, Identify information that is included for audio but not for video, and generate a QA pair for each. Similarly, identify information that is included for video but not for audio, and generate a QA pair for each.

Each pair should simulate a dialogue that focuses on various aspects of the content without introducing information not explicitly mentioned. The purpose of this exercise is to fine-tune a language model so that it can generate accurate and relevant responses.

The QA pair should start with **"Q:"** and contain an instruction related to the audio and video, while the second line should start with **"A:"** and provide a response to the instruction.

Please ensure that your instructions are diverse and of high quality. Avoid making assumptions or adding details that are not explicitly given.

EXAMPLE:

- **Q:** What kitchen tools or utensils are prominently featured in the **video** during pancake preparation?
  A: The video showcases basic kitchen tools such as a mixing bowl, whisk, spatula, and frying pan, demonstrating their use in making pancakes.

- **Q:** Discuss the overall tone and style of the English **audio** commentary. How does it enhance the tutorial?
  A: The commentary's friendly and instructive tone makes pancake-making feel approachable and enjoyable, encouraging viewers to try the recipe themselves.

IV. AVU-Negatives

INPUT:

[Audio-video information]

The information provided for audio and video does not fully match. Find out what the audio and video information does not match. Create negative question-answer pairs where the question asks for information not included in the provided data. Provide negative responses or rejections indicating the information is unavailable. Don't design questions that can be answered with yes or no.

The QA pair should start with **"Q:"** and contain an instruction related to the audio and video, while the second line should start with **"A:"** and provide a response to the instruction.

EXAMPLE:

- **Q:** Can you describe the sound of the frog in the audio?
  A: The color of the frog is not mentioned in the audio. The background audio is funny music.

Figure 9: Prompt templates for generation of AVU-dataset subsets.