

SilVar: Speech-Driven Multimodal Model for Reasoning Visual Question Answering and Object Localization

Tan-Hanh Pham^{1,2,*}, Le Hoang Nam^{3,†}, Nguyen Phu Vinh^{4,†}, Chris Ngo⁵, Truong-Son Hy⁶

¹Harvard Medical School, Harvard University, ²Massachusetts General Hospital,

³FPT University, ⁴Uppsala University, ⁵Knovel Engineering Lab, ⁶Alabama University

*Correspondence: tpham33@mgh.harvard.edu; †Equal contribution

Abstract

Visual Language Models have demonstrated remarkable capabilities across various tasks, including visual question answering and image captioning. However, most models rely on text-based instructions, limiting their effectiveness in natural human-machine interactions. Moreover, the quality of language models primarily depends on reasoning and prompting techniques, such as chain-of-thought, which remain underexplored when using speech instructions. To address these challenges, we propose **SilVar**, an end-to-end multimodal model that leverages speech instructions for reasoning-based visual question answering. Additionally, we investigate reasoning techniques at different levels, including conversational, simple, and complex speech instructions. SilVar is built upon CLIP, Whisper, and LLaMA 3.1-8B, enabling more intuitive interactions by allowing users to provide verbal or text-based instructions. To this end, we introduce a new dataset designed to challenge models with speech-based reasoning tasks for object localization. This dataset enhances the model’s ability to process and explain visual scenes from spoken input, moving beyond simple object recognition to reasoning-based interactions. To our knowledge, SilVar is the first open-source, speech-driven VLM. We believe SilVar will inspire the next generation of multimodal reasoning models, advancing toward expert artificial general intelligence. Our code and dataset are publicly available [here](#).

1 Introduction

Visual Language Models (VLMs) have gained significant attention due to their capacity to bridge the gap between visual and textual modalities, facilitating more intuitive interactions between humans and machines. These models are valuable in tasks like visual question answering (VQA), which may involve yes/no answers, multiple-choice questions, or even generating image descriptions. With advances in deep learning, VLMs can now effectively

handle visual scenes and provide meaningful textual outputs that explain or describe those scenes in natural language.

Recent advancements in VLMs, such as CLIP (Radford et al., 2021), have enabled Large Language Models (LLMs) to process images and text simultaneously (Ranasinghe and Ryoo, 2023; Alayrac et al., 2022; Awadalla et al., 2023). Models like Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), LLaVA (Liu et al., 2023), LocVLM (Ranasinghe et al., 2024), and LISA (Lai et al., 2024) demonstrate strong multimodal capabilities, excelling in tasks such as image captioning and VQA through improved reasoning and prompting techniques. Despite these advances, most VLMs still rely on text-based interactions, limiting usability in scenarios where text is impractical. While prompting and reasoning for LLMs have been studied extensively in text, their adaptation to speech remains underexplored. Recent models like GPT-4o (OpenAI, 2024b) support speech-based interaction, greatly enhancing user experience, but open-source efforts in this area are still limited.

To enable speech interaction with LLMs, speech instruction models such as Qwen2-Audio (Chu et al., 2024), SALMONN (Tang et al., 2023), and Llama-Omni (Fang et al., 2024) have been developed to process speech instead of text-based instructions. Although these models enable speech instruction, they are not capable of understanding both images and text simultaneously. Inspired by VLMs and ARS models, we propose SilVar, a multimodal model that can understand both images and audio or images and text, at the same time. In addition, we investigate reasoning techniques for speech instructions on image description and object localization. To this end, we further provide a dataset for speech instruction. The key contributions of our paper are summarized as follows:

- We propose a multimodal model with speech

instruction for text generation and object localization.

- Speech reasoning – we investigate the effect of reasoning for speech instruction, which includes conversation level, simple reasoning, and complex reasoning.
- We propose a training pipeline and publicly release the reasoning speech instruction dataset.

2 Related Work

With the advent of large language models (LLMs) like GPT-3 (Brown, 2020), GPT-4 (Achiam et al., 2023), and open-source models such as the Llama family (Touvron et al., 2023a,b; Dubey et al., 2024) and Vicuna (Zheng et al., 2023b), the foundation was laid for the development of vision-language models (VLMs). VLMs extend LLMs by enabling learning from both text and visual inputs (Li et al., 2023b), accelerating progress in multimodal modeling, particularly for integrating vision or speech with language. Early efforts like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) introduced frameworks for combining vision and language. Building on these, models such as Flamingo (Alayrac et al., 2022), BLIP (Liu et al., 2024), MiniGPT-v2 (Chen et al., 2023), MiniGPT-4 (Zhu et al., 2023), and LLava (Liu et al., 2023) advanced tasks like visual question answering and image captioning. Applications have since expanded to object detection, segmentation, and reasoning-based localization (Liu et al., 2023; Lai et al., 2024; Zhu et al., 2023; Wang et al., 2024; Ranasinghe et al., 2024). While early models emphasized scale, recent efforts focus on smaller, high-performance models and effective prompting techniques (Wei et al., 2021; Dubey et al., 2024; Li et al., 2022).

In parallel, speech recognition has emerged as a crucial area of research, particularly with the development of automatic speech recognition (ASR) systems such as Whisper (Radford et al., 2023) and Wav2Vec (Baevski et al., 2020). In addition, there are lot of work that has been done on speech-related tasks, including speech-to-text translation, speech emotion recognition (SER), and vocal sound classification (VSC) (Tang et al., 2023; Wang et al., 2023a; Ao et al., 2021). Beyond speech generation, recent models have been investigated to enhance emotion and voice interactions, such as AudioPaLM (Rubenstein et al., 2023) and LauraGPT (Du et al., 2023), fostering more natural communication. Innovations such as VALL-E (Wang et al.,

2023b) and MusicGen (Copet et al., 2024) further illustrate how audio generation can enrich text-based interactions. Furthermore, the challenge of low-resource conversational telephony speech corpora has been investigated using unsupervised learning and fine-tuning techniques of large pre-trained models (Vieting et al., 2023).

The integration of ASR with language models has led to the development of multimodal models (Chu et al., 2024; Xie and Wu, 2024). For instance, SpeechGPT (Zhang et al., 2023a) allows users to engage with large language models using speech. Additionally, HuggingGPT (Shen et al., 2024) enhances this interaction by discretizing speech into tokens and expanding the LLM’s vocabulary to accommodate speech inputs. Furthermore, the study in (Adedeji et al., 2024) demonstrated that LLMs have the potential to improve the accuracy of ASR systems, particularly in medical transcription.

While VLMs have significantly advanced vision-text integration, incorporating audio and ASR has enabled more dynamic multimodal interactions. The release of GPT-4o (OpenAI, 2024b) showcases real-time speech-based interaction with LLMs, improving user experience over text-only models. However, its closed-source nature limits further development. Currently, few open-source models support speech-based interaction with VLMs. To address this, we introduce **SilVar**, a novel multimodal model that integrates speech instructions directly into the reasoning process, enhancing tasks like image interpretation and object localization. We also provide a complete pipeline leveraging open-source foundation models, aiming to support and inspire research in multimodal reasoning.

Prompting techniques play a vital role in enhancing LLMs’ reasoning abilities for tasks like question answering (Liu et al., 2023; Lai et al., 2024), with methods such as chain-of-thought (CoT) and zero-shot prompting significantly influencing performance (Wei et al., 2022; Yao et al., 2024). In ASR contexts, prompting has also been applied to complex tasks like diarization and error correction (Adedeji et al., 2024). Building on this, we explore reasoning techniques, particularly zero-shot and CoT prompting, for speech-based instruction. As LLMs, VLMs, and instruction tuning evolve, benchmarks such as ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), and LLaVA (Liu et al., 2023) have been developed to evaluate multimodal reasoning, yet they primarily focus on text-image input. To address the lack of benchmarks

for speech-based instruction, we introduce **SilVar**, a new dataset that incorporates text, images, and speech instructions for text generation. We also augment existing reasoning datasets like MMMU, LISA, and ScienceQA by converting their text inputs into speech.

3 Data Generation

With the development of multimodal models, there has been a surge in datasets that support model training, such as Flickr30K (Young et al., 2014), Visual Genome (Krishna et al., 2017), and MovieQA (Tapaswi et al., 2016). However, these datasets are limited to tasks like automatic image description, image or video captioning, and simple visual question-answering. To explore the understanding and explainability of multimodal models, more intricate datasets such as LAION (Schuhmann et al., 2022), SEED (Li et al., 2023a), and LLaVA (Liu et al., 2023) have been created, enabling LLMs to generate detailed responses. Despite this progress, the available data are insufficient for guiding LLMs in querying and responding to users' input, particularly for tasks requiring complex instructions. As a result, techniques like hard prompting (Wen et al., 2024) or prompt engineering (Wei et al., 2022; Yao et al., 2024) have been proposed. For dataset, LLaVA (Liu et al., 2023) is one of the recent datasets that use strong prompting baselines. However, this type of data is a text-based instruction and not well-suited for speech instruction, especially in the context of human-machine interaction. For this, we propose a unique speech instruction dataset that emphasizes natural conversation.

Inspired by the success of recent GPT models in text-based tasks and GPT-assisted datasets (Liu et al., 2023), we developed our dataset with GPT-4 (OpenAI, 2024a) assistance, as shown in Figure 1. Notably, we only used text as input for GPT-4 to generate different types of questions. To ensure the model can effectively perform reasoning-based object localization and generate coherent responses from speech instructions, we designed the dataset according to the following criteria:

- **Human-machine conversation:** The dataset is designed to reflect natural human-machine conversations, enabling the agent to interpret and respond to verbal instructions in a conversational context.
- **Reasoning instructions and responses:** Unlike traditional datasets focused on simple ob-

ject recognition, our dataset contains reasoning instructions and explanations behind the responses. This is crucial for enhancing the model's ability to explain not just what the object is, but also why it is located in a particular place.

- **Detailed descriptions:** The dataset includes both simple questions and complex reasoning scenarios, requiring the model to provide detailed descriptions of visual scenes.

The SilVar dataset contains 4,161 speech-driven samples based on 2,518 randomly selected images from the COCO 2014 dataset (Lin et al., 2014). It includes two task types: spatial object localization (2,361 samples), which helps models learn spatial reasoning, and object detection with bounding boxes (1,800 samples), the primary focus of this work. All samples are generated using GPT-4 with a structured prompting strategy. For spatial tasks, object positions are described using directional terms (e.g., left, right, front, behind), while detection tasks include bounding box outputs. Rather than naming objects directly, GPT-4 generates scenario-based questions that highlight an object's role, as illustrated in Figure 1, with answers providing detailed explanations.

- **Object's characteristics:** A description of the object's features, shape, and functionality.
- **Background context:** An explanation of how the object is relevant to the broader environment depicted in the image.
- **Interaction with surroundings:** An analysis of how the object interacts with other elements in the scene, such as people, activities, or other objects.

Following the generation of questions and answers, four people corrected the data to ensure quality and coherence. Human reviewers verified that each set of questions consistently referred to the same object and that the answers offered detailed reasoning, including the object's characteristics, background, and interactions. In addition, bounding boxes were manually labeled using Roboflow (Dwyer et al., 2024) to accurately pinpoint the specific objects being referred to in each image. After ensuring data quality through human verification, we convert text to speech using Google Cloud APIs,

Question type 1: Conversation

What object known for its comfort and joy, often in colorful outfits, is frequently found in cozy gatherings? Let's explain it in detail.

Question type 2: Simple reasoning

What object known for comfort and joy, often in colorful outfits, is frequently found in cozy gatherings? Let's think step by step.

Question type 3: Complex reasoning

As friends come together for a laid-back evening of sharing stories and enjoying refreshments, what soft and cuddly figure symbolizes friendship and nostalgia, often bringing a smile to those in its presence?

Response:

The soft and cuddly figure referred to is the teddy bear, located at [BBOX]. This plush toy evokes feelings of nostalgia and friendship, characterized by its rounded form, soft fur, and friendly expression. Its placement among friends and the wine bottles signifies its role in creating a warm and inviting atmosphere, where stories can be shared. The bear's presence encourages a sense of camaraderie and emotional connection among guests, reinforcing how such objects often serve to bridge interactions and enhance the joy of gathering.

BBOX: [[304,36,628,364], [28,230,423,384], [76,96,258,292]]

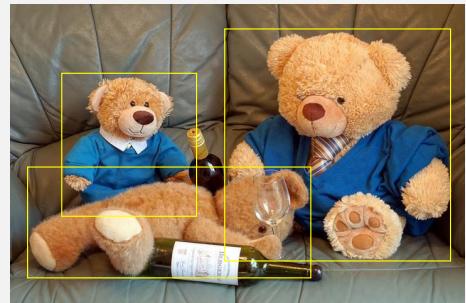


Figure 1: An example from our SilVar-bench dataset, focusing on reasoning speech instructions with different types: conversation, simple reasoning, and complex reasoning. The detected objects are highlighted in yellow bounding boxes. The dataset not only focuses on reasoning instructions but also generates visual explanations, enhancing spatial understanding and interpretability.

which support over 50 voices. SilVar-Bench provides a robust evaluation framework, challenging models in object recognition and context-specific reasoning. It is designed to advance multimodal models requiring spatial understanding and detailed reasoning in speech-driven interactions. It is worth noting that our dataset is larger than smaller-scale datasets such as LISA, which contains only 1,218 samples. This highlights that our dataset ensures not only quality but also sufficient quantity.

Dataset	Train	Validation	Test
ScienceQA	6,218	2,097	2,017
MMMU	150	900	10,500
LISA	239	200	779
SilVar	3,461	-	700

Table 1: The datasets used in this study. It is important to note that the SilVar training set contains 1,100 bounding box localization samples and 2,361 spatial localization samples, while the test set includes 700 bounding box localization samples.

In addition to our dataset, we utilize existing text-based reasoning datasets—MMMU (Yue et al., 2024), LISA (Lai et al., 2024), and ScienceQA (Lu et al., 2022)—for pretraining, as they focus on reasoning, localization, and description. MMMU includes 11,500 samples across 30 subjects and 183 subfields, while LISA provides 239 training samples (we use only the training split). ScienceQA

spans 26 topics, 127 categories, and 379 skills; we filter it to include only samples with image-text pairs. To adapt these datasets for spoken input, we apply preprocessing steps to ensure clarity and vocalizability for speech generation.

- **Handling special characters:** We converted complex symbols like LaTeX and non-standard characters into formats suitable for speech instruction, enabling correct processing and vocalization.
- **Punctuation and text normalization:** We standardized punctuation in the text to ensure that it would generate smooth, natural speech outputs, enhancing the clarity and coherence of the spoken content.

4 Speech-Driven Visual Language Model

4.1 Architecture

SilVar is a multimodal model designed for image interpretation and object localization based on various input modalities, including speech and text. This approach enables SilVar to understand and respond comprehensively to various inputs and prompts. The model architecture is illustrated in Figure 2, consisting of several key components: an audio encoder, a visual encoder, a projector, and a large language model.

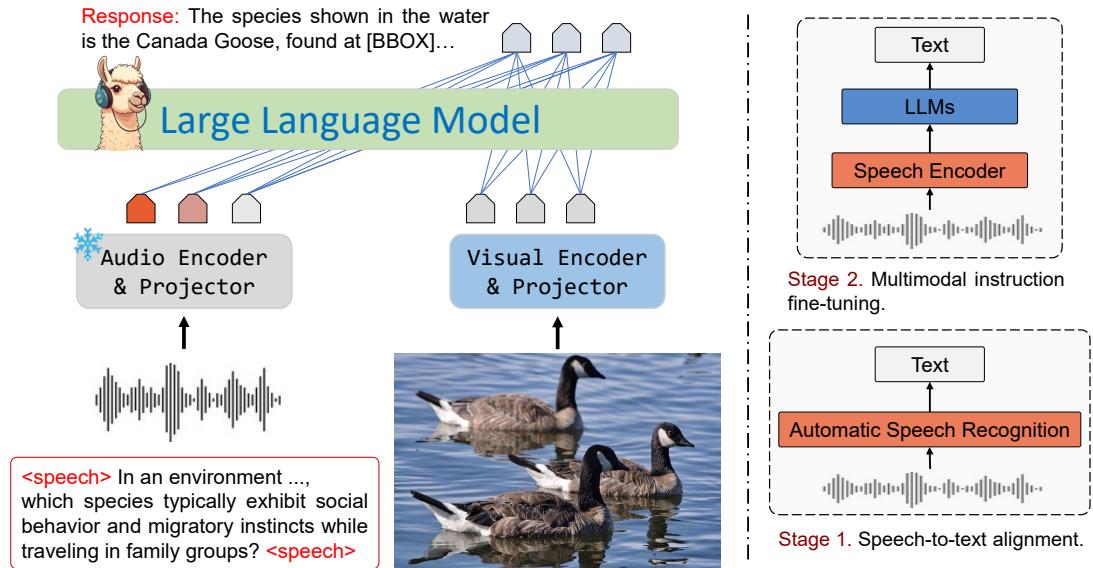


Figure 2: Illustration of the SilVar’s architecture, integrating images and audio instruction for reasoning description and object localization.

Audio Encoder: This module extracts features from speech instructions to provide context beyond text-based inputs. We use the lightweight Whisper-tiny model (Radford et al., 2023; Moor et al., 2023), with 39M parameters, known for its strong performance in speech recognition. Given an input audio $\mathbf{A} \in \mathbb{R}^{1500}$, Whisper encodes it into a 768-dimensional feature vector. This is passed through a Linear layer (audio projector) to match the LLM input size (\mathbb{R}^{4096}) and is later concatenated with image embeddings. We also explore different audio adapters, including MLP and Transformer layers.

Visual Encoder: This module extracts meaningful features from input images using the pre-trained CLIP model (ViT-B/32) (Radford et al., 2021), which was trained on over 400 million image-text pairs. For each image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where $C = 3$, the input is resized to 224×224 and passed through CLIP to produce a sequence of 768-dimensional visual tokens. To align with the language model, these tokens are projected to \mathbb{R}^{4096} using a visual adapter. Following MiniGPT-v2 (Zhu et al., 2023), the adapter consists of two Linear layers with GELU activation. The resulting visual features are concatenated with audio embeddings to form a unified multimodal representation.

Large Language Model: At the core of SilVar is a language model, responsible for generating text and bounding boxes by processing information from the audio and visual encoders. To this end,

we decided to use LLama 3.1-8B (Dubey et al., 2024), which serves as the foundational model, allowing us to effectively transfer its pre-trained knowledge into domain-specific tasks such as localization where understanding both verbal and visual inputs. By combining visual and audio tokens, we provide a diverse set of token embeddings for the language model, requiring it to process and generate a final representation of textual tokens. The integration of these components enables SilVar to generate human-quality text responses by leveraging the complementary strengths of each modality, making it a robust system for multimodal instruction and interaction.

4.2 Training Pipeline

We propose a two-step training process for speech instruction: (1) speech-to-text alignment and (2) LLM training response, as shown in Figure 2. Since speech plays an important role as an instructional modality in our model, we start training a speech-to-text system to align audio with text in the domain of reasoning text generation. We use the Whisper models for speech recognition due to their efficiency in real-time speech processing. As mentioned in Section 4.1, to develop a foundation model that can understand and handle numerous tasks, we train the model on the ScienceQA and MMMU datasets using speech instructions, as these datasets are reasoning benchmarks for human-level understanding and explanation (Lu et al., 2022; Yue et al., 2024). Furthermore, we

aim to specialize the model for the task of object localization; therefore, we further train the Whisper on the LISA dataset and our benchmark. The data type of each dataset and the stages used in the training process are shown in Table 2.

Dataset	Data type	Stage 1	Stage 2
ScienceQA	Text, image	x	x
MMMU	Text, image	x	x
LISA	Text, image	x	x
SilVar	Text, image, speech	x	x

Table 2: Summary of datasets and corresponding stages used in our training process.

In stage 2, we use the pre-trained weights from stage 1 to train our model for the visual question-answering task, using direct audio input from the audio encoder for reasoning. Similar to stage 1, we use the ScienceQA, MMMU, LISA, and SilVar datasets for training text generation. In our experiments, we use AdamW optimizer to update the parameters (Loshchilov and Hutter, 2019), accompanied by a learning rate scheduler with a linear warmup followed by a cosine decay. The model is trained for up to 20 epochs with a batch size of 4 and 2 workers. The model is trained on a computing system equipped with four A100 GPUs, around twenty-two hours.

5 Experiment and Result

5.1 Speech To Text Alignment and Quality

We trained Whisper for 10 epochs on converted-to-speech datasets, and the results are shown in Table 3 using WER and CER metrics. While Whisper Tiny performs well overall, its accuracy varies across datasets. For example, ScienceQA exhibits higher error rates (WER: 7.77 train, 7.98 test), with Whisper Small significantly outperforming it. Despite minor misalignments affecting embeddings and final predictions, we chose Whisper Tiny due to its compact size and suitability for our end-to-end training process later.

Dataset	Models	WER		CER	
		train	test	train	test
ScienceQA	Whis. Small	4.30	4.17	5.25	4.32
ScienceQA	Whis. Tiny	7.77	7.98	9.43	10.21
MMMU	Whis. Tiny	6.45	6.72	5.18	5.76
SilVar	Whis. Tiny	2.16	3.58	2.59	3.54

Table 3: Evaluation of speech-to-text alignment of audio encoders. Whis. denotes Whisper.

5.2 Speech Instruction & Text Instruction

To evaluate the impact of instructional modality, we compare SilVar’s performance using speech- and text-based instructions across various reasoning and conversational tasks. This includes assessing its ability to handle both simple and complex reasoning, as well as conversational prompts, in both formats. SilVar is trained only on complex reasoning, while other techniques are used for evaluation. We use CIDEr, BLEU, METEOR, and ROUGE for text generation, and bounding box accuracy at an IoU threshold of 0.5 for object localization, providing a comprehensive performance overview.

As shown in Table 4, text-based instructions consistently outperform speech-based ones across most metrics, with complex reasoning achieving the highest scores: ROUGE-1 (37.23), BLEU-1 (38.12), METEOR (28.66), and CIDEr (0.08). In object detection, the IoU0.5 accuracy reaches 27.56% for text-based complex reasoning, indicating better localization. Speech-based inputs also perform best with complex reasoning but score slightly lower. These results highlight SilVar’s effectiveness and the challenges of aligning synthetic speech with images, as discussed in Section 5.1.

5.3 MMMU-bench

MMMU is a benchmark with tasks that demand college-level knowledge and deliberate reasoning, which enables the test of model performance in terms of expert-level perception and reasoning. The dataset includes 30 subjects across various disciplines, such as art, science, and engineering. In particular, we compare the SilVar to models that have similar number of parameters.

As shown in Table 5, our text-based model achieves a validation score of 31.8, outperforming several similarly sized baselines, including Adept Fuyu-8B (27.9), OpenFlamingo2-9B (28.7), MiniGPT4-Vicuna-13B (26.8), and LLaMA-Adapter2-7B (29.8), demonstrating strong performance in complex reasoning. While it trails behind SOTA models like LLaVA-1.5-13B (36.4), it offers a solid balance of accuracy and efficiency. With speech-based instructions, SilVar also performs well, scoring 30.2, which is better than using pure transcribed text from Whisper (30.0), and showcasing its multimodal versatility. End-to-end training further improves performance to 30.4. Replacing Llama3.1 with DeepSeek-R1-Distill-Llama-8B boosts the score to 32.8. Al-

Instruction Type	ROUGE-1	BLEU-1	METEOR	CIDEr	Accuracy (IoU = 0.5)
Complex reasoning (text)	37.23	38.12	28.66	0.08	27.56
Simple reasoning (text)	34.08	36.11	27.58	0.07	26.54
Conversation (text)	35.15	36.57	26.95	0.07	26.56
Complex reasoning (speech)	34.44	37.62	26.99	0.06	24.44
Simple reasoning (speech)	33.42	34.83	24.74	0.05	23.02
Conversation (speech)	34.16	34.25	23.86	0.04	22.71
<i>Ablation study when using DeepSeek-R1-Distill-Llama-8B as the language model</i>					
Complex reasoning (speech)	38.62	39.05	28.17	0.09	28.41

Table 4: Performance of SilVar model on various instructional types (conversational, simple, and complex) using speech- and text-based modalities. The highlighted values in purple represent the highest scores achieved for each metric in text-based complex reasoning, while values in orange highlight the highest scores for speech-based complex reasoning.

Model	Instruction	Val	Test
LLaVA-1.5-13B (Liu et al., 2024)	Text	36.4	33.6
Qwen-VL-7B-Chat (Bai et al., 2023)	Text	35.9	32.9
LLaMA-Adapter2-7B(Zhang et al., 2023b)	Text	29.8	27.7
OpenFlamingo2-9B (Awadalla et al., 2023)	Text	28.7	26.3
Adept Fuyu-8B (Bavishi et al., 2023)	Text	27.9	27.4
MiniGPT4-Vicuna-13B (Zhu et al., 2023)	Text	26.8	27.6
<i>Our experiments</i>			
SilVar	Text	31.8	-
SilVar-transcribed	Text	30.0	-
SilVar	Speech	30.2	-
SilVar-e2e	Speech	30.4	-
GPT-4o mini	Speech	39.3	-
Gemini Flash 1.5	Speech	36.9	-
<i>Ablation study when using DeepSeek as the language model</i>			
SilVar-e2e (DeepSeek)	Speech	32.8	-

Table 5: Performance of SilVar and other models on the MMMU benchmark. e2e denotes end-to-end training, and transcribed denotes the usage of Whisper-transcribed text.

though GPT-4o mini (39.3) and Gemini Flash 1.5 (36.9) achieve higher scores, SilVar shows strong potential as an open-source, speech-driven VLM suitable for fine-tuning on downstream tasks.

5.4 ScienceQA Benchmark

In this benchmark, we compare SilVar to SoTA models with a similar parameter count, using both speech- and text-based instructions. As shown in Table 6, text-based SilVar achieves an average accuracy of 86.32, outperforming models like LLaMA-Adapter (85.19) and approaching Chat-UniVi (88.78). Speech-based SilVar scores 63.21%, surpassing several text-instruction models, including MiniGPT-4 (47.71), LLaVA-7B (41.10), and OpenFlamingo (39.27). End-to-end training yields a slight improvement, and replacing the language model with DeepSeek-R1-Distill-Llama-8B raises the score to 65.82. We also compare SilVar to speech-driven GPT-4o mini (80.15) and Gemini

Flash 1.5 (78.31). While a performance gap remains between speech- and text-based models, SilVar proves competitive and promising in scenarios where spoken instructions are necessary.

5.5 Chatbot models

After training SilVar, we evaluated its performance by comparing its predictions with those of commercialized chatbots such as GPT-4o mini and Gemini Flash 1.5. As shown in Table 8, all three models successfully inferred the underlying activity in the image. SilVar demonstrated an enriched reasoning process by incorporating both speech and spatial awareness. This highlights SilVar’s ability and potential in VQA tasks and enables us to fine-tune it for downstream applications. Furthermore, this capability suggests that SilVar can effectively integrate visual and auditory information in an end-to-end manner, making it particularly suited for vision-language tasks that require speech commu-

Model	Instruction	Average Score
LLaVA-13B (Liu et al., 2023)	Text	90.92
LaVIN-13B (Luo et al., 2024)	Text	90.83
Chat-UniVi (7B) (Jin et al., 2024)	Text	88.78
LLaMA-Adapter (Zhang et al., 2023b)	Text	85.19
MiniGPT-4 (Zheng et al., 2023a)	Text	47.71
LLaVA-7B (Yang et al., 2023)	Text	41.10
OpenFlamingo (Yang et al., 2023)	Text	39.27
<i>Our experiments</i>		
SilVar	Text	86.32
SilVar	Speech	63.21
SilVar-e2e	Speech	63.45
GPT-4o mini	Speech	80.15
Gemini Flash 1.5	Speech	78.31
<i>Ablation study when using DeepSeek as the language model</i>		
SilVar-e2e (DeepSeek)	Speech	65.82

Table 6: Performance comparison of our model and other models on the ScienceQA Benchmark. e2e denotes end-to-end training.

nication like medical surgery and self-driving cars.

6 Ablation study

While visual and audio encoders are key to processing image-speech pair instructions, the visual and audio adapters are equally vital for transferring encoded information to LLMs. In Llava (Liu et al., 2024), linear and MLP-based adapters effectively bridge encoders and LLMs. Model performance can also benefit from different neural network types (Chen et al., 2020b,a), and adapter choice is critical in multimodal models like the Q-Former in BLIP-2 (Li et al., 2023b) and the Perceiver Resampler in Flamingo (Alayrac et al., 2022). Building on this, we design both MLP- and Transformer-based adapters for the audio encoder.

Table 7 shows SilVar’s performance using different audio adapters in end-to-end speech instruction training. The Transformer-based adapter requires 1.5 \times more training time than the MLP-based one. However, performance differences are minimal, suggesting that using Whisper’s final encoder layer eliminates the need for complex adapter designs. We also explore different MLP hidden layer sizes (2816 and 5632), finding performance variation within ± 0.1 on the MMMU dataset.

Adapter	SilVar	MMMU (val)	ScienceQA
Linear layer	24.44	30.04	63.45
MLP	25.17	31.16	63.41
Transformer	25.66	31.05	63.78

Table 7: Performance of SilVar with different audio adapters on VLM benchmarks.

Additionally, we replace Llama 3.1 with DeepSeek-R1-Distill-Llama-8B, resulting in modest gains, about 8% on MMMU and 4% on ScienceQA, as shown in Table 4, Table 5, and Table 6.

7 Conclusion

In this study, we present **SilVar**, the first end-to-end speech-driven VLM designed to enhance human-machine interaction by enabling effective reasoning from both images and speech instructions. Unlike most existing VLMs and LLMs that depend on text inputs, SilVar is an open-source model supporting verbal communication, allowing for more intuitive and natural interactions. However, building speech-based VLMs is challenging due to input quality and the alignment of audio and visual embeddings, leading to a performance gap compared to text-based models. However, SilVar performs comparably to text-based VLMs on the MMMU and ScienceQA benchmarks, showing its potential in scenarios where text input is unavailable.

Limitations

One limitation of this work is the small size of the training dataset. Despite it being enough to fine-tune our model for the new task, the usage of this dataset outside of our research scope could be limited.

Potential risks

The core of Silvar is a pre-trained LLM, which can suffer from hallucination and sometimes produce incorrect information. The usage of this model should be carefully revised in Silvar’s answer.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, and 1 others. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

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

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifang Guo, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, and 1 others. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

B. Dwyer, J. Nelson, T. Hansen, and 1 others. 2024. Roboflow (version 1.0). Available from <https://roboflow.com>. Software.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

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

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

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittweiser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, and 1 others. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR 2019)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2024. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

OpenAI. 2024a. Gpt-4. Available at <https://openai.com/gpt-4>. Model used for dataset generation.

OpenAI. 2024b. Hello, gpt-4o. Accessed: 2024-10-21.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Kanchana Ranasinghe and Michael S Ryoo. 2023. Language-based action concept spaces improve video self-supervised learning. *Advances in Neural Information Processing Systems*, 36:74980–74994.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. 2024. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quirky, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueling Zhuang. 2024. Hugging-gpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwala Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peter Vieting, Christoph Lüscher, Julian Dierkes, Ralf Schlüter, and Hermann Ney. 2023. Efficient utilization of large pre-trained models for low resource asr. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP-W)*, pages 1–5. IEEE.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023b. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and 1 others. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Dal-ing Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2023. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. *arXiv preprint arXiv:2310.09036*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

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

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023a. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Chatbot Comparison

Question: What activity might be taking place that encourages interaction and bonding between the two individuals in the image?

Silvar: The two individuals share a playful moment and joy, they are located [142 459 284 359]. In the environment like a zoo, they might be engaging in an interactive activity like visiting the zoo or feeding animals. This activity may enhance their experience through shared laughter and companionship.

GPT-4o mini: The two young children in the image appear to be at a zoo, standing in front of a giraffe enclosure. They are holding hands, smiling, and leaning close to each other, suggesting a shared moment of enjoyment. The activity that might be encouraging interaction and bonding between them is a visit to the zoo, where they are likely exploring and experiencing animals together. This shared experience fosters a sense of excitement, curiosity, and connection.

Gemini 1.5 Flash: The two individuals in the image are encouraged to visit a zoo or animal park together. Observing and reacting to the animals fosters interaction and bonding through shared experience, communication, and physical closeness.

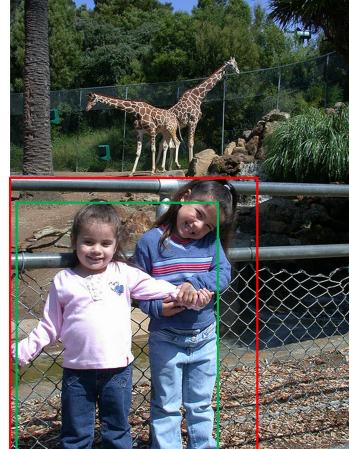


Table 8: Comparison of predictions between our models and the other speech-driven model on the reasoning abnormal detection. Unlike GPT-4o mini and Gemini 1.5 Flash, our Silvar is an open-sourced end-to-end speech-driven VLM.