# Audio-visual training for improved grounding in video-text LLMs

**Shivprasad Sagare, Hemachandran S., Kinshuk Sarabhai,**
**Prashant Ullegaddi, Rajeshkumar SA**
PhroneticAI
**Correspondence:** shivprasad.sagare@phronetic.ai, rajesh.kumar@phronetic.ai

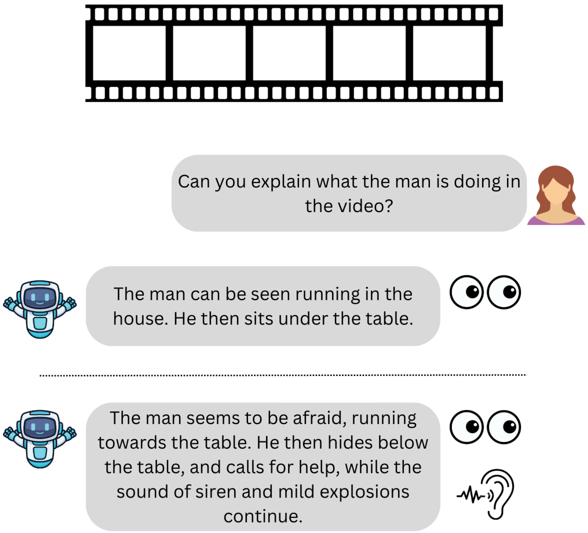
## Abstract

Recent advances in multimodal LLMs, have led to several video-text models being proposed for critical video-related tasks. However, most of the previous works support visual input only, essentially muting the audio signal in the video. Few models that support both audio and visual input, are not explicitly trained on audio data. Hence, the effect of audio towards video understanding is largely unexplored. To this end, we propose a model architecture that handles audio-visual inputs explicitly. We train our model with both audio and visual data from a video instruction-tuning dataset. Comparison with vision-only baselines, and other audio-visual models showcase that training on audio data indeed leads to improved grounding of responses. For better evaluation of audio-visual models, we also release a human-annotated benchmark dataset, with audio-aware question-answer pairs.

## 1 Introduction

Conversational agents fueled by LLMs have made it possible for us to interact in a new way with data from multiple modalities (Yin et al., 2024)(Wadekar et al., 2024). Image-text multimodal LLMs(MLLMs) like LLaVA (Liu et al., 2023) have demonstrated the effectiveness of visual instruction-tuning(IT) data. Several works like VideoChatGPT (Maaz et al., 2023), VideoChat (Li et al., 2024), PLLaVa (Xu et al., 2024) have extended the image-text model architecture for video related tasks.

However, most of the above works rely only on the visual input, and do not consider audio signal for video understanding. In real world, listening to audio while playing the video, adds immensely to our perception of the video. We propose a video-text MLLM, with Phi-2 (Gunasekar et al., 2023) as the LLM backbone. It supports both audio and visual inputs, using Whisper (Radford et al., 2022) and sigLIP (Zhai et al., 2023) encoders respectively. Unlike previous works, we train the model using audio data explicitly, in addition to the visual data. We aim to explore the role of audio in video understanding and if audio input can be utilized for better grounding of video-text LLMs. We also explore the creation of better benchmarks that encompass variety of question-answer pairs. Evaluation on several benchmarks demonstrates the effectiveness of audio as an additional signal in better understanding of the video content.



Can you explain what the man is doing in the video?

The man can be seen running in the house. He then sits under the table.

The man seems to be afraid, running towards the table. He then hides below the table, and calls for help, while the sound of siren and mild explosions continue.

Figure 1: An example of improved grounding in the video-text LLM outputs, due to the additional audio signal as input.

Overall we make the following key contributions:
**1**.We propose an efficient video-text MLLM architecture consisting of separate encoders to process the audio and visual inputs.
**2**.We train our video-text model using both audio and visual signals simultaneously, aiming to explore the effect of audio input on model outputs.
**3**.We release a human-annotated benchmark dataset containing video instruction-tuning samples, which are audio-aware.

*Proceedings of the 17th International Natural Language Generation Conference*, pages 440–445
September 23–27, 2024. ©2024 Association for Computational Linguistics

| Models | Visual | Audio | Audio-visual |
|---|---|---|---|
| VideoChatGPT | ✓ | – | – |
| LLaSM | – | ✓ | – |
| Video-LLaMA | ✓ | × | × |
| NExT-GPT | ✓ | ✓ | × |
| our | ✓ | ✓ | ✓ |

Table 1: Comparing MLLMs based on the input modalities supported, and the training data. – indicates that the input modality isn't supported. × indicates that the input modality is supported, but the model isn't trained using such data. ✓indicates that the model architecture supports the input modality, and has also been explicitly trained on such data.

## 2 Related work

**Vision-text MLLMs**: LLaVA (Liu et al., 2023), MiniGPT4 (Zhu et al., 2023) have showcased the efficacy of visual instruction-tuning datasets for image-text tasks. Bunny (He et al., 2024) explores a similar idea but using lightweight LLM backbones. Several works like PLLaVA (Xu et al., 2024) build on the top of image-text MLLMs to support video input. VideoChatGPT (Maaz et al., 2023) extends the CLIP image encoder (Radford et al., 2021) to videos by averaging the representations across spatial and temporal dimensions.

**Audio-text MLLMs**: Similar to vision-text, there has been recent work in fusing audio input features with text LLM for several audio-text tasks (Zhang et al., 2023a). LLaSM (Shu et al., 2023) demonstrates the effectiveness of pretraining the projector layers using speech-to-text data. Some previous works like AudioGPT (Huang et al., 2023) build on LLM-based planning and tool-use to solve several audio tasks at once.

**Audio-vision-text MLLMs** Similar to our work, Video-LLaMA (Zhang et al., 2023b), and NExT-GPT (Wu et al., 2023) support audio and visual input simultaneously, both relying on unified modality encoder ImageBind (Girdhar et al., 2023). However, Video-LLaMA is trained only on visual IT datasets, assuming the audio branch learns implicitly. NExT-GPT is trained using cross-modal IT dataset, but doesn't utilize audio-visual simultaneous input from videos. Unlike previous works, we explore training using audio-visual input from videos simultaneously, and explore the grounding effect it has on model outputs.

## 3 Model architecture

Following the idea of fusing the modality inputs into LLM (Liu et al., 2023)(Zhang et al., 2023b), we build a video-text MLLM architecture consisting of two separate branches for audio and visual inputs. Each branch consists of modality encoder, projector layers to transform the encoder representations into LLM embedding space, followed by the backbone LLM.

We use Whisper (Radford et al., 2022) as an audio encoder, and use its last hidden state as audio representations (Shu et al., 2023). To encode the video, we use sigLIP image encoder (Zhai et al., 2023). Following (Maaz et al., 2023), we treat video as a sequence of images, and compute frame representations using sigLIP. We then compute spatial and temporal average of representations across 100 uniformly sampled frames, and use it as a video representation. Inspired from Bunny (He et al., 2024), we rely on low-cost, efficient, lightweight LLM backbone with 2.7 Billion parameters, phi-2 (Gunasekar et al., 2023). Projector layer for both vision and audio branch is mlp2x-gelu (He et al., 2024).

The exact flow of input data through both the audio and visual branches is shown in the form of tensor dimensions, in figure 2. Audio and visual input is converted into 64 and 829 token embeddings respectively. Audio, visual, and text token embeddings are then concatenated before passing to the backbone LLM.

## 4 Training setup and datasets

Training different components of our model with appropriate data is a key focus of our research. Typically, these MLLMs go through a pretraining stage, followed by the finetuning stage.

**Pretraining**: Pretraining aims to align different modalities to text LLM space, by training on some generic modality-to-text task. Only projector layer weights are trained during this phase, while encoders, and LLM weights are frozen. We pretrain our audio projector layers using a combination of Speech-to-Text(STT) dataset(CommonVoice (Ardila et al., 2020)) and audio captioning dataset(AudioCaps (Kim et al., 2019)) with 50K samples each. We convert these datasets into our instruction-tuning prompt template by creating 10 instructions each for transcription and captioning. Since our visual branch relies on image encoder, we employ already trained

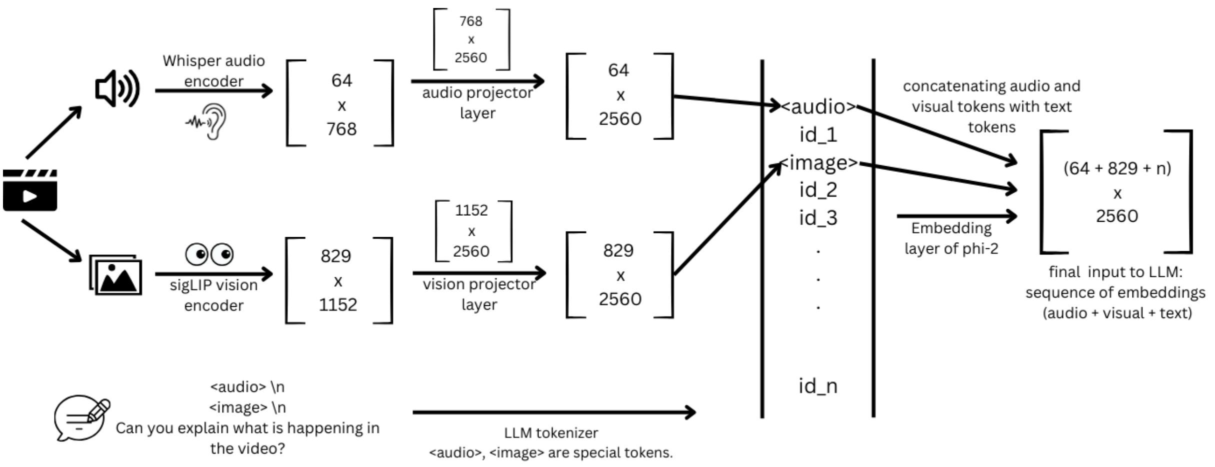


Figure 2: Tensor dimensions in the figure denote the flow of data through the encoder and projector layers. Audio encoder(Whisper) and video encoder(using sigLIP) produce 64 and 829 token embeddings respectively, which are then concatenated with the text token embeddings as the final input to the LLM. Unlike previous works, we train both the audio and vision branch simultaneously using a video instruction tuning dataset.

checkpoint by Bunny (He et al., 2024) to initialize vision projector layers. It has been trained on 2M subset of an image-text dataset LAION (Schuhmann et al., 2022). We freeze the vision branch while pretraining audio projector layers, and vice versa.

**Finetuning**: Finetuning or instruction tuning is aimed to train the LLM model to follow the exact requests or questions in the user prompt (Ouyang et al., 2022). Unlike previous works, we explicitly train both the audio and visual branches of the model simultaneously, using video instruction-tuning dataset containing both the audio and visual data. We rely on VideoInstruct100K (Maaz et al., 2023) dataset with 100K samples containing video and question answer pair. Although the dataset authors had used the dataset only for visual instruction tuning, we extract the audios(wav format) from the videos(mp4 format) for our use-case.

We aim to explore if including audio features during training helps the model to better understand the video. To measure this effect, we also train a baseline vision-only model, without the audio branch. We train the vision branch of the model, using the visual data from same dataset.

**Experiment details** We implement the audio and video functionality by extending the codebases of Bunny and LLaSM. We use Whisper-small, siglip-so400m-patch14-384, and phi-2 models from HuggingFace. Pretraining for audio projector layer was done using A100, with global batch size of 128. Finetuning was implemented using LoRA for training LLM weights, on A40 machine.

## 5 Benchmark dataset

Several evaluation criteria and datasets have been introduced to benchmark the vision-text MLLMs (Chen and Dolan, 2011)(Maaz et al., 2023)(Heilbron et al., 2015). VideoChatGPT has released a human verified benchmark dataset consisting of 500 videos and corresponding question-answer pairs for video-text tasks. However, these benchmarks do not consider audio information while creating the question-answer pairs based on videos. Thus, it is challenging to evaluate the capability of model to attend to both the audio and visual signals while generating the output.

Therefore, we annotate such an audio-visual instruction-tuning dataset that contains question-answer pairs based both on audio and visual information in the video. We include both generic questions, like ’What is happening in the video?’, as well as more specific questions related to the video. Answer of each question is around 2 sentences, with most of the videos available on YouTube. We release a set of 120 such samples, as we intend to scale the size and quality of the data in future. Example samples from our benchmark dataset are shown below.

**Sample 1**
**Question**: What is the man doing in the video?
**Answer**: In the video, the man fires his gun upwards, producing the sharp sound of a bullet being shot. The echo reverberates through the air, adding tension and intensity to the scene.

| Metrics | visual-only model (our) | video-llama | audio-visual model (our) |
|---|---|---|---|
| Correctness of Information | 2.34 | 1.96 | **2.69** |
| Detail Orientation | 2.35 | 2.18 | **2.49** |
| Contextual Understanding | 2.74 | 2.16 | **3.04** |
| Temporal Understanding | 1.97 | 1.82 | **2.22** |
| Consistency | 2.45 | 1.79 | **2.71** |
| Average | 2.37 | 1.98 | **2.63** |

Table 2: Results on VideoChatGPT evaluation framework. Our audio-visual training setup shows impressive results when compared with other audio-vision model(Video-LLaMA), as well our vision-only baseline.

| Metrics | visual-only model (our) | video-llama | audio-visual model (our) |
|---|---|---|---|
| Correctness of Information | 2.34 | 1.49 | **2.77** |
| Detail Orientation | 2.36 | 1.7 | **2.44** |
| Contextual Understanding | 2.75 | 1.92 | **3.04** |
| Temporal Understanding | 2.17 | 1.4 | **2.4** |
| Average | 2.40 | 1.62 | **2.66** |

Table 3: Results on our benchmark dataset. Results illustrate similar trend as above, where training on audio signals helps the model to generate more accurate responses. We haven't yet incorporated evaluation for consistency metric in our benchmark dataset.

**Sample 2**
**Question**: What is the man on the stage mentoring about in the video?
**Answer**: The workshop leader, mentors a student on speaking louder for clarity. He asks the student to raise the volume from level 3 to level 7. Finally, the student earns an applause from the audience in the communication workshop.

## 6 Evaluation

We extensively evaluate our model using VideoChatGPT evaluation framework across 5 key metrics. It relies on LLM-based evaluation(using GPT-3.5) which rates the output on the scale of 1-5. We compare our audio-visual model with the visual-only baseline that we have trained, as well as other audio-visual model, Video-LLaMA. The evaluation results are summarized in the table 2. Similarly, we evaluate on our benchmark dataset, and observe similar trends, as summarized in 3.

The audio-visual model clearly performs better than the vision-only baseline by a margin. Interestingly, Video-LLaMA which is also an audio-visual model performs poorly on both the benchmarks. Video-LLaMA does not utilize the audio inputs explicitly, and instead rely on visual signals only during training. We could not compare against another audio-visual model, NExT-GPT, as it relies on LLaMA-v0 weights which couldn't be available to us due to licensing.

Qualitative analysis of audio-visual model outputs demonstrate better overall quality compared to vision-only model. We also analyze the model outputs at intermediate stages, i.e. after pre-training. Our model could very well generate the captions of audio data, which showed the efficacy of pre-training step. There is scope for better encoding strategies and training regimes for utilizing audio information even more.

## 7 Conclusion and future work

We performed several experiments and evaluations to specifically study how audio signal can be utilized for better video understanding. Training the MLLM simultaneously on audio-visual signals of the video indeed results in a better performance, as seen in quantitative evaluation using several metrics. We also contributed a benchmark dataset curated to evaluate the video-understanding capability using both visual and audio information.

Based on these results, we are motivated to experiment with sophisticated ways of incorporating audio and visual signals together for video related tasks. Future work also consists of the extensive analysis of the type of question-answer pairs in video IT datasets, and work on creating better evaluation benchmarks catering to wide range of video-related use-cases.

## References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. *Preprint*, arXiv:2305.05665.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar, Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *Preprint*, arXiv:2402.11530.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *Preprint*, arXiv:2304.12995.

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videochat: Chat-centric video understanding. *Preprint*, arXiv:2305.06355.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. 2023. Llasm: Large language and speech model. *Preprint*, arXiv:2308.15930.

Shakti N. Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. 2024. The evolution of multimodal model architectures. *Preprint*, arXiv:2405.17927.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *Preprint*, arXiv:2309.05519.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *Preprint*, arXiv:2404.16994.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

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. *Preprint*, arXiv:2305.11000.

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

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