MoonWhisper: Using Audio Inputs to Augment Vision-Language Models

Babak Abtahi [*1] **Guneev Arora** [*1] **Hakim Temacini** [*1] **Umar Rajguru** [*1]

Abstract

This paper presents the integration of additional modalities into pre-existing large language models (LLMs) to create more robust multimodal architectures capable of processing novel input types beyond their original design. Specifically, we incorporate the Whisper encoder to enable audio input processing within the Moondream LLM framework. We explore the performance on tasks such as speech transcription and audio recognition. Furthermore, we demonstrate the model's ability to handle tasks that require seamless integration of multiple modalities, even in the presence of noisy inputs.

1. Introduction

Recently, the emergence of Large Language Models (LLMs) have brought significant advances to both language understanding and task solving, demonstrating robust ability in solving Natural Language Processing (NLP) tasks (Zhao et al., 2024). Various LLMs based on the Transformer architecture (Vaswani et al., 2023) like GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019), and LLaMA (Touvron et al., 2023) have played a pivotal role in these advancements, achieving strong performance in text generation, sentiment analysis, text classification, logical reasoning, and contextual comprehension (Chang et al., 2023).

Likewise, Vision Language Models (VLMs) (Zhang et al., 2024) that align image text features have shown considerable results in tasks such as zero-shot image classification, image captioning, and visual question answering (Bordes et al., 2024).

Numerous papers have explored extending existing LLM architectures to incorporate new modality sources (Yin et al., 2024). For example, LLaVA (Liu et al., 2023) integrates visual inputs by utilizing an image encoder and projector before passing the processed representations into a pre-existing LLM. Similarly, Qwen2 (Yang et al., 2024) incorporates audio inputs via the Whisper encoder, which generates embeddings that are subsequently fed into an LLM. Microsoft has proposed an end-to-end speech recognition pipeline built around an existing LLM architecture (Ling et al., 2023).

However, these approaches typically rely on significant preprocessing or architectural modifications to the original LLM. Many methods involve altering the training procedure or employing techniques such as LoRA (Hu et al., 2021) to fine-tune components of the LLM at scale. These modifications often introduce substantial complexity to the overall system.

In contrast, our work adopts a simpler approach by introducing new modality sources as modular components and examining the outcomes of integrating pre-existing models without major architectural changes. We demonstrate that this method achieves acceptable performance while showcasing the potential for seamless integration of multimodal inputs without requiring extensive adjustments to the underlying LLM architecture. Furthermore, we show that this model is robust to different introductions of noise into a modality source, further showing the synergy of the entire model.

2. Methodology

Our proposed architecture consists of two main components: the Whisper speech encoder (Radford et al., 2022) and the pre-trained Moondream VLM (Moondream). As shown in Figure 1, audio representations produced by the Whisper speech encoder are directly processed by the VLM, enabling seamless integration of audio signals into the multimodal processing framework.

The Whisper encoder and processor handle audio inputs by extracting feature embeddings, which are then projected into the same representational space as text embeddings. These projected embeddings, along with text and image tokens, are tokenized and input to the VLM for joint processing across all modalities.

To maintain the integrity of the Moondream VLM architecture, minimal modifications were made. Textual inputs are tokenized using a pre-trained tokenizer, and tokens are pro-

[1]McMaster University. Correspondence to: Babak Abtahi <abtahib@mcmaster.ca>, Guneev Arora <arorag12@mcmaster.ca>, Hakim Temacini <temacinih@mcmaster.ca>, Umar Rajguru <rajguruu@mcmaster.ca>.

,

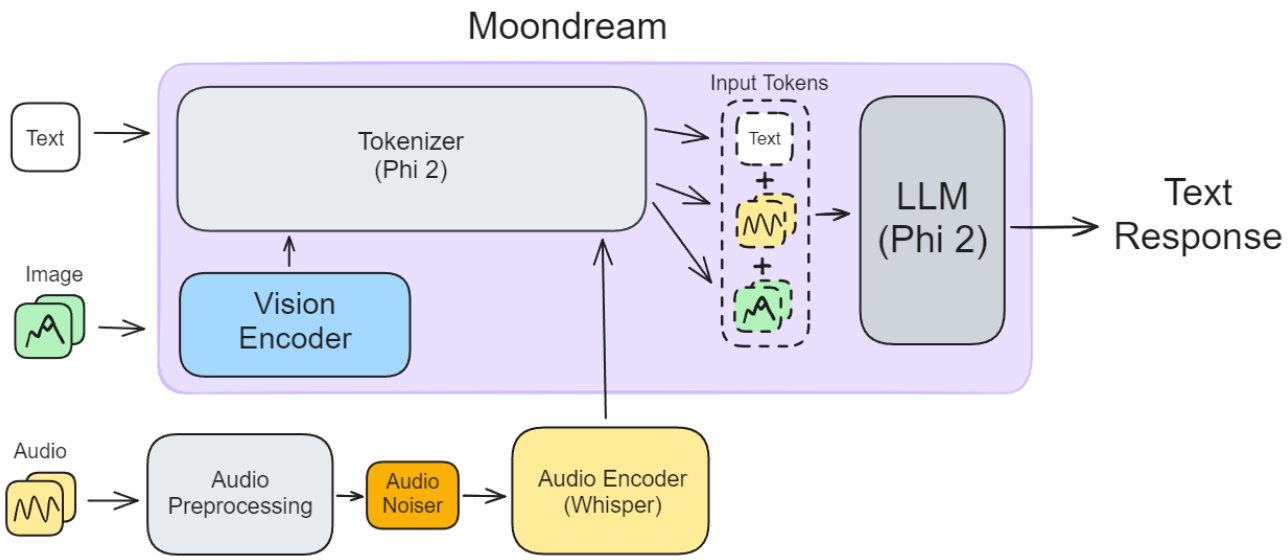


Figure 1. Overview of our approach: text, image, and audio are passed to the Phi-2 tokenizer to generate input tokens, which are then passed to the Phi-2 LLM, resulting in a text response. The image data is passed to a vision encoder before being received by the tokenizer, while the audio data undergoes preprocessing, is fed to the Audio Noiser which adds garbage data, where the result is then passed to the Whisper audio encoder before being received by the tokenizer.

cessed through the internal pre-trained Phi language model (Li et al., 2023), which generates final hidden states. This strategy extends the model's functionality to accommodate new modalities without altering its core design.

For training, we employed the same cross-entropy loss function used in the original training of the Moondream VLM. Gradients were accumulated over two batches during training, followed by back-propagation to update model parameters. This approach ensures reproducibility and consistency while adapting the model to handle multimodal inputs.

We utilized the OmniInstruct dataset (Li et al., 2024), which includes a diverse collection of images, audio samples, and textual descriptions, each paired with a single correct answer. Some samples include empty options, requiring the VLM to generate a suitable completion autonomously. The dataset is partitioned into training and testing splits.

To evaluate the model's resilience to noisy interference, we trained variants of the model with varying probabilities of audio noise. Noise was introduced into the training data via uniform random sampling according to a specified probability p as detailed in Figure 2.

For benchmarking, we compared the noise-trained models to a baseline model fine-tuned on a speech transcription dataset without noise. This allowed us to evaluate whether fine-tuning improves the model's generalization capability in the presence of noisy or multimodal inputs.

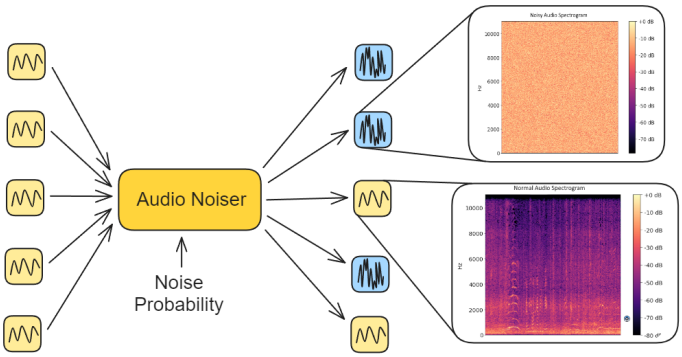


Figure 2. Audio samples are substituted with randomly selected audio inputs drawn from a uniform distribution based on a probability p

3. Results and Discussions

We evaluate the performance of these models on a validation split consisting of image, text, and audio inputs. In each sample, there are a range of 0 to 4 different options to choose from, and only a perfect match is rendered as a correct answer.

We summarize the results in Table 1. We see that all of the models demonstrate similar performance on the testing set. In fact, even as the noise probability increases the accuracy drops but not significantly enough. This suggests, that perhaps the introduction of audio to this dataset is not useful or helpful to the overall performance of the model.

However, when we tested the models with all 3 modalities,

Audio Noise Probability	Image and Text	Image, Audio and Text
0%	47.36%	48.16%
20%	47.10%	47.60%
40%	47.49%	47.89%
60%	47.29%	47.27%

Table 1: Accuracy of 4 models trained on the OmniInstruct test set. The first column shows the varying noise probabilities of each model. The second column indicates the results when tested without the audio modality, and the third column shows the results when tested on all 3 modalities.

we observe higher overall performance, with an accuracy reaching its peak at 48.16%, though differences in accuracy were small (less than 1%).

Additionally, we notice that the model with 60% noise probability performed worse when it had access to the audio modality as opposed to only image and text.

4. Further Work

This paper highlights the potential of adapting vision-language models (VLMs), originally designed for textual and image inputs, to accommodate new and unintended modality sources. Our primary objective was to examine the strengths and limitations of combining existing architectures. Moving forward, we hope to see further exploration into generalizable methods for synergizing disparate models to achieve more unified tasks. The human brain processes not only auditory and visual stimuli but also a wide array of sensory inputs. Similarly, enabling AI systems to handle a diverse range of inputs is an essential step toward achieving more versatile and adaptive multimodal systems.

Future research could also focus on scaling the efficient integration of multiple modalities during training. Recent advancements, such as GMC(Poklukar et al., 2022), offer techniques for geometrically aligning different modality inputs, improving robustness when some modalities are missing. However, these methods face challenges in non-classification tasks where text prompts may not semantically align with image or audio prompts. While alternative fusion techniques exist, they often lack the same efficiency and scalability. Developing approaches that balance alignment, robustness, and scalability across a broad spectrum of tasks and input types remains an open and critical area for exploration.

5. Conclusions

Our findings demonstrate that existing LLM and VLM architectures can be effectively integrated with dedicated modules to support new modality sources. These integrated models achieve acceptable performance on tasks requiring the simultaneous processing of diverse input streams, enabling a range of applications previously beyond the scope of the original models' design.

Notably, the introduction of noise to the audio modality had minimal impact on overall model performance, underscoring the robustness of the approach to variations in the new modality. This resilience further highlights the potential for adapting existing architectures to handle complex, multimodal environments.

The simplicity and scalability of building such enhanced models pave the way for the development of smarter, more versatile systems capable of interpreting and responding to a wide array of environmental inputs. This work establishes a foundation for future research into creating unified, multimodal AI systems that are robust and adaptable to real-world challenges.

References

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., Padthe, K., Sharma, V., Xu, H., Tan, X. E., Richards, M., Lavoie, S., Astolfi, P., Hemmat, R. A., Chen, J., Tirumala, K., Assouel, R., Moayeri, M., Talattof, A., Chaudhuri, K., Liu, Z., Chen, X., Garrido, Q., Ullrich, K., Agrawal, A., Saenko, K., Celikyilmaz, A., and Chandra, V. An introduction to vision-language modeling, 2024. URL https://arxiv.org/abs/2405.17247.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang,

Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. A survey on evaluation of large language models, 2023. URL https://arxiv.org/abs/2307.03109.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., and Lee, Y. T. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Li, Y., Zhang, G., Ma, Y., Yuan, R., Zhu, K., Guo, H., Liang, Y., Liu, J., Yang, J., Wu, S., Qu, X., Shi, J., Zhang, X., Yang, Z., Wang, X., Zhang, Z., Liu, Z., Benetos, E., Huang, W., and Lin, C. Omnibench: Towards the future of universal omni-language models, 2024. URL https://arxiv.org/abs/2409.15272.

Ling, S., Hu, Y., Qian, S., Ye, G., Qian, Y., Gong, Y., Lin, E., and Zeng, M. Adapting large language model with speech for fully formatted end-to-end speech recognition, 2023. URL https://arxiv.org/abs/2307.08234.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Moondream. Moondream. https://moondream.ai. HHek.

Poklukar, P., Vasco, M., Yin, H., Melo, F. S., Paiva, A., and Kragic, D. Geometric multimodal contrastive representation learning, 2022. URL https://arxiv.org/abs/2202.03390.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey, 2024. URL https://arxiv.org/abs/2304.00685.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.