Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, Bryan Catanzaro
{zkong, wping, rafaelvalle}@nvidia.com

Summary

- We propose **Audio Flamingo**: a Flamingo-based audio language model for audio understanding.
- Audio Flamingo achieves SOTA results on several close-ended and open-ended audio understanding tasks.
- We design a series of methodologies for efficient use of in-context learning and retrieval, which lead to SOTA few-shot learning results.
- Audio Flamingo has strong multi-turn dialogue ability.

- Code and checkpoints at https://github.com/NVIDIA/audio-flamingo
- Sound demos at https://audioflamingo.github.io/

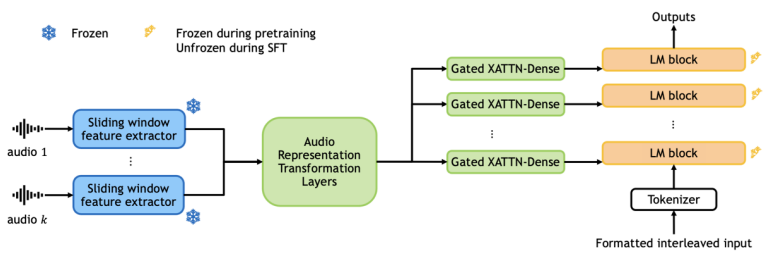
Motivation

We aim to build an audio language model that can understand sound beyond speech transcriptions.

Tasks

- ✓ Audio Captioning
- ✓ Audio Question Answering
- ✓ Audio Classification
- ✓ Retrieval-augmented few-shot learning
- ✓ Multi-turn dialogues

Architecture



Training

Our training objective combines non-interleaved and interleaved samples.

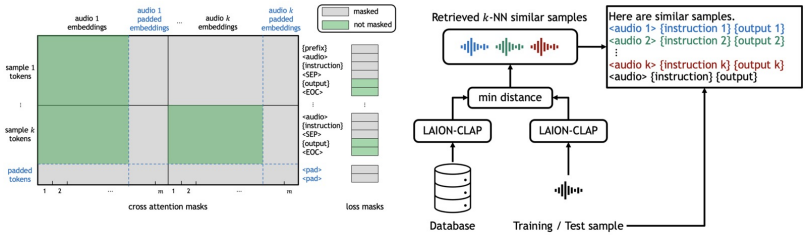
$$\mathcal{L}(z) = \sum_{t=1}^{|y_{\text{out}}|} \log p_\theta\left((y_{\text{out}})_t | x, y_{\text{ins}}, (y_{\text{out}})_{<t}\right)$$

$$\mathcal{L}_{\text{int}}(z_{\text{int}} = \{z^1, \cdots, z^J\}) = \sum_{j=1}^{J} \sum_{t=1}^{|y^j_{\text{out}}|} \log p_\theta\left((y^j_{\text{out}})_t | z^{<j}, x^j, y^j_{\text{ins}}, (y^j_{\text{out}})_{<t}\right)$$

$$L = -\sum_{i \in \mathcal{I}} \lambda_i \mathbb{E}_{z \sim \mathcal{D}^i} \mathcal{L}(z) - \sum_{i' \in \mathcal{I}_{\text{int}}} \lambda_{i'} \mathbb{E}_{z_{\text{int}} \sim \mathcal{D}^{i'}_{\text{int}}} \mathcal{L}_{\text{int}}(z_{\text{int}})$$

Retrieval Augmented Generation and In-Context Learning

We use a block-triangular cross-attention mask for interleaved data (**left**), and a retrieval method to construct interleaved training samples (**right**).



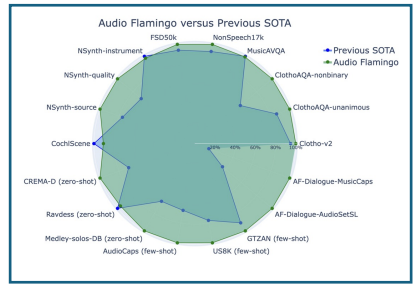
Templates

```
<s>{task description}<audio>{instruction}
Options:\n- option1\n··· - optionm
<SEP>{output}<EOC></s>
```

```
<s>The task is dialogue.<audio>
user: {instruction1}
assistant: <SEP>{output1}<EOC>
...
user: {instructions}
assistant: <SEP>{outputs}<EOC></s>
```

```
<s>{task description}Here are similar samples.
<audio>{instruction1}<SEP>{output1}<EOC>
...
<audio>{instructionk}<SEP>{outputk}<EOC>
<audio>{instruction}
Options:\n- option1\n··· - optionm
<SEP>{output}<EOC></s>
```

Results



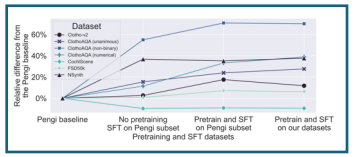
Dataset	Task	Metric	Previous SOTA ↑	Ours ↑
Clotho-v2	CAP	CIDEr	0.441 (Chu et al., 2023)	**0.465**
$\text{ClothoAQA}_{\text{unanimous}}$	AQA	ACC	74.9% (Chu et al., 2023)	**86.9%**
$\text{ClothoAQA}_{\text{non-binary}}$	AQA	ACC	29.1% (Deshmukh et al., 2023)	**49.5%**
$\text{ClothoAQA}_{\text{numerical}}$	AQA	ACC	26.2% (Deshmukh et al., 2023)	**36.4%**
$\text{MusicAVQA}_{\text{audio-only}}$	AQA	ACC	**72.1%** (Chu et al., 2023)	71.6%
CochlScene	CLS	ACC	**91.6%** (Deshmukh et al., 2023)	83.0%
NonSpeech7k	CLS	ACC	79.0% (Rashid et al., 2023)	**85.1%**
FSD50k	CLS	$\text{F1}_{\text{approx}}$	65.6% (Deshmukh et al., 2023)	**69.7%**
$\text{NS}_{\text{instrument}}$	CLS	ACC	**78.8%** (Chu et al., 2023)	77.1%
$\text{NS}_{\text{quality}}$	CLS	F1	46.3% (Deshmukh et al., 2023)	**66.7%**
$\text{NS}_{\text{source}}$	CLS	ACC	60.1% (Deshmukh et al., 2023)	**78.7%**

Dataset	Task	Metric	Previous SOTA (0-shot) ↑	Ours (0-shot) ↑
AudioCaps (Kim et al., 2019)	CAP	CIDEr	0.281 (Salewski et al., 2023)	**0.502**
CREMA-D (Cao et al., 2014)	CLS	ACC	18.5% (Deshmukh et al., 2023)	**26.8%**
Ravdess (Livingstone & Russo, 2018)	CLS	ACC	**21.7%** (Elizalde et al., 2023b)	20.9%
US8K (Salamon et al., 2014)	CLS	ACC	71.9% (Deshmukh et al., 2023)	**75.0%**
GTZAN (Sturm, 2013)	CLS	ACC	**71.0%** (Han et al., 2023)	67.9%
Medley-solos-DB (Lostanlen et al., 2019)	CLS	ACC	61.3% (Deshmukh et al., 2023)	**92.7%**

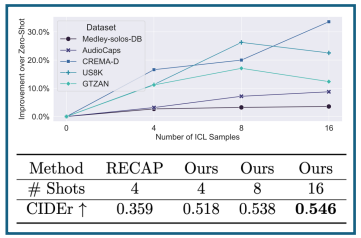
Audio Flamingo achieves SOTA generation quality on several audio understanding tasks. **Left**: overview of results, where 100% means the best of all baseline and our models. **Upper right**: in-domain tasks (evaluated on test splits). **Lower right**: 0-shot evaluation results.

Testset	Method	CIDEr ↑	Bleu4 ↑	R-L ↑
A	Qwen-Audio	0.507	0.060	0.292
A	LTU†	0.823	0.153	0.403
A	Ours†	**1.622**	**0.237**	**0.473**
M	MU-LLaMA	0.585	0.083	0.348
M	LTU†	0.419	0.108	0.336
M	Ours†	**1.143**	**0.142**	**0.417**

Dialogue evaluation results on our `Audio-Dialogues` dataset. † indicates the model is finetuned on `Audio-Dialogues`. `A` is the audio subset and `M` is the music subset.



Effects of data scaling.



Method	RECAP	Ours	Ours	Ours
# Shots	4	4	8	16
CIDEr ↑	0.359	0.518	0.538	**0.546**

Retrieval augmented in-context learning improves Audio Flamingo's generation quality. **Upper**: relative improvements with respect to #few-shot samples. **Lower**: retrieval-augmented audio captioning results.

References

[1] Alayrac, Jean-Baptiste, et al. *Flamingo: a visual language model for few-shot learning.* NeurIPS 2022.
[2] Iyer, Srinivasan, et al. *OPT-IML: Scaling language model instruction meta learning through the lens of generalization.* arXiv 2022.
[3] Elizalde, Benjamin, Soham Deshmukh, and Huaming Wang. *Natural language supervision for general-purpose audio representations.* ICASSP 2024.
[4] Wu, Yusong, et al. *Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation.* ICASSP 2023.
[5] Deshmukh, Soham, et al. *Pengi: An audio language model for audio tasks.* NeurIPS 2023.
[6] Chu, Yunfei, et al. *Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models.* arXiv 2023.
[7] Tang, Changli, et al. *SALMONN: Towards Generic Hearing Abilities for Large Language Models.* ICLR 2024.
[8] Gong, Yuan, et al. *Listen, Think, and Understand.* ICLR 2024.
[9] Liu, Shansong, et al. *Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning.* ICASSP 2024.
[10] Ghosh, Sreyan, et al. *RECAP: retrieval-augmented audio captioning.* ICASSP 2024.