Audio Dialogues:
Dialogues dataset for audio and music understanding

Arushi Goel[†], *Zhifeng Kong*, *Rafael Valle*[‡], *Bryan Catanzaro*

NVIDIA

[†]`arushig@nvidia.com`, [‡]`rafaelvalle@nvidia.com`

Abstract

Existing datasets for audio understanding primarily focus on single-turn interactions (*i.e.* audio captioning, audio question answering) for describing audio in natural language, thus limiting understanding audio via interactive dialogue. To address this gap, we introduce `Audio Dialogues`: a multi-turn dialogue dataset containing 163.8k samples for general audio sounds and music. In addition to dialogues, `Audio Dialogues` also has question-answer pairs to understand and compare multiple input audios together. `Audio Dialogues` leverages a prompting-based approach and caption annotations from existing datasets to generate multi-turn dialogues using a Large Language Model (LLM). We evaluate existing audio-augmented large language models on our proposed dataset to demonstrate the complexity and applicability of `Audio Dialogues`. Our code for generating the dataset will be made publicly available. Detailed prompts and generated dialogues can be found on the demo website[1].
Index Terms: Multi-turn dialogues, Instruction-tuning, AudioLLMs, Audio and Music understanding

1. Introduction

Audio, a fundamental component of human communication and interaction, carries vast amounts of information, ranging from speech and music to general and ambient sounds. The development of models for audio understanding plays a vital role in various tasks such as audio and sound monitoring [1], speech recognition [2], music recommendation systems [3], and even aiding individuals with hearing impairments. To aid progress in audio understanding, the research community has developed large-scale datasets [4, 3, 5, 6] (*e.g.* AudioSet [6], WavCaps [5], AudioCaps [7] *etc.*). Hence, models [8, 9, 10, 11] built on these datasets have shown great potential in learning audio representations for tasks such as audio retrieval [12], audio and music captioning [13], sound event classification [14] and so on.

Recent developments have enhanced audio understanding models by integrating them with Large Language Models (LLMs) [15]. This integration has demonstrated the potential to harness the capabilities of LLMs for robust knowledge retention, reasoning, and task execution in audio-related domains [16, 14, 7]. A notable advancement in LLMs is their adeptness to engage in dialogues with humans [17, 15]. It is essential to extend this capability for multi-turn dialogues to audio-augmented LLMs [10, 18], as it is pivotal for constructing models capable of listening and interacting effectively. Such advancement in audio models requires conversation-based dialogue datasets tailored for audio applications.

Although there are extensive datasets describing audio or music in natural language [19], they are primarily designed for audio captioning [7] or single-turn audio question answering [20]. For instance, these datasets typically include questions such as "What is the emotion in this audio?" or "What does the audio convey?". However, fine-tuning models on such datasets limits the potential of audio-augmented Large Language Models (LLMs) to engage in more complex interactions regarding the audio content. Addressing this limitation, Chu *et al.* [10] introduced a multi-turn dialogue dataset consisting of 20,000 samples to train their proposed model. Nonetheless, the dataset's size remains relatively small, and there is a lack of information regarding its generation process.

To address these limitations, in this paper, we propose `Audio Dialogues`: an audio-based dialogue dataset with *multi-turn dialogues* and *comparison questions* for general sounds and music. Similar in spirit to how instruction-tuning datasets are generated to train vision and language assistants [21], we use prompting-based approach to generate a multi-turn dialogue dataset for audios using a pre-trained LLM [15]. Specifically, we utilize caption annotations sourced from the AudioSet strongly labeled dataset [6] and the MusicCaps dataset [13] to guide the dialogue generation process leveraging GPT-4 [15]. Additionally, we implement a data filtration strategy to filter out noisy synthetic dialogues, promoting the retention of the most reliable ones. In total, our proposed dataset comprises of 163.8k samples, each containing between one to four dialogues.

Our main contributions are as follows: 1) a multi-turn dialogue dataset, `Audio Dialogues` for general sounds and music understanding with training and evaluation splits, 2) a detailed data generation pipeline to foster the generation of dialogue datasets, and 3) evaluation of existing audio-augmented large language models [10, 18, 22] on our proposed dataset.

2. Related work

Instruction tuning datasets. Instruction following Large Language models (LLMs) [15] have shown remarkable capabilities in zero-shot and few-shot tasks in the language domain such as machine translation [23], summarization [24] and so on. This idea of developing models that can follow instructions has then been ex-

[1]`https://audiodialogues.github.io/`

tended to other domains such as vision [21] and audio [18, 17]. LLaVA [25] made the first attempt at generating instruction-following data involving visual content using GPT-4. Specifically, they use image captions and bounding box localization as meta information for the image to be used as query for the language model. Overall, they collect 158k samples for language-image instruction-following data. Since then, there has been growing interest in developing instruction following datasets such as VALLEY [26], Macaw-LLM [27] and Video-ChatGPT [28].

In the audio domain, LTU [11] generated, using GPT, an open-ended question-answering dataset that tries to capture general knowledge and reasoning ability about general sounds. LTU's audio based question answering dataset is limited, given that it only has single-turn conversations, lacks in complex context between conversations and does not have strong correlations between rounds (*e.g.* use of pronouns). Qwen-Audio [10] curates a 20k audio-based instruction-following dataset, but there is little to no discussion about the curation process and dataset. Our `Audio Dialogues` dataset addresses all the above mentioned limitations by generating multi-turn conversations for an audio sample, covering both general sounds and music domains. Compared to existing datasets, `Audio Dialogues` has multi-turn dialogues with strong correlations between rounds through the presence of pronouns (*e.g.* he, she, it), follow-up questions based on the previous answer, and complex context.

Audio augmented LLMs. Recent research has focused on advancing audio foundation models [9, 18] capable of comprehending audio content by harnessing Large Language Models (LLMs) [15]. Typically, these models employ an audio encoder [19, 2] to convert audio into tokens, which are then integrated with textual instructions within an LLM to generate the final response. These models are pretrained on various tasks such as audio captioning [7], emotion recognition [29], sound event classification [14], speech recognition [2], music understanding [13], among others, and have shown significant gains in zero-shot and few-shot performances using a unified model. While these models exhibit robust audio comprehension, recent works like Audio Flamingo [18] have introduced techniques such as in-context learning [21] and retrieval-augmented generation [30] to enhance the model's instruction-following capabilities through fine-tuning with interleaved audio-text pairs. To measure the importance of our proposed `Audio Dialogues` dataset, we evaluate the performance of audio foundation models such as LTU [11], Qwen-Audio [10] and Audio Flamingo [18] on these multi-turn dialogues.

3. Data generation pipeline

3.1. Pipeline

In this section, we discuss our data generation pipeline illustrated in Figure 1. We construct `Audio Dialogues` using the strongly labeled AudioSet-SL [31] and MusicCaps [13]. The Audioset-SL dataset [31] has time-stamped annotations for the 10-second audio clips which we preprocess to describe sound events for the audio samples.

Following [11], we augment timestamped sound events with acoustic features for each sound event. Specifically, given a sound class name, *e.g.* howl, we prompt GPT-4 with "describe the acoustic characteristic of a *howl* sound in less than 10 words." This gives us an acoustic feature description of that sound class. An augmented sound event with acoustic features example is shown below:

> **Sound events:** Sound of Howl (Loud, prolonged, mournful, echoing sound.): [0.406s-9.237s], [9.575s-10.000s]; Sound of Wind noise (microphone) (Low frequency, random, broadband sound.): [2.128s-2.584s], [9.288s-9.850s]; Sound of Animal (Loud, diverse, and often rhythmic.): [8.174s-9.221s], [9.778s-10.000s].

Similar to AudioSet-SL, the MusicCaps dataset [13] has detailed descriptions or captions for the music samples which we directly use as input for dialogue generation.

We utilize these audio sound events and music description information along with prompt templates to guide GPT-4 [15] in generating multi-turn dialogues. Next, we discuss the prompt design and propose a data filtration strategy to retain only high quality dialogues.

3.2. Prompts

We design specific prompt templates to generate 1) multi-turn dialogues and, 2) audio comparison question-answer pairs. To generate multi-turn dialogues for the Audioset-SL [31] and MusicCaps [13] dataset, the prompt template consists of a *system prompt* and *examples* of hand-crafted dialogues. Due to space constraints, we only show an example of system prompt to generate the subset AudioSet Dialogues and Music Dialogues. Detailed prompts are on our demo website.

> **System prompt**
> Based on the sound events, create a dialogue between you (the assistant) and a person (the user) about the events in the audio. Each dialogue should consist of:
> 1. A user examines the audio and sends a reasonable and creative message to the assistant.
> 2. Once the audio is provided, the assistant thoroughly perceives and comprehends them, responding with helpful answers that provide comprehensive reasoning. Do not include timestamps in the answer provided by the assistant.
> 3. Considering the past dialogue i.e. the question and the answer in the previous timestep, the user chooses another question or further inquiry.
> 4. The assistant provides a helpful answer providing comprehensive reasoning regarding the audio content. All the questions asked by the user should be diverse and complex.

Apart from the multi-turn dialogues, the `Audio Dialogues` dataset also consists of question-answer pairs for multiple (two or more) audios, called as AudioSet comparison subset. Below we show the prompt template used to generate these question-answer pairs given the meta information from the audio clips.

In order to cluster multiple audios to be used as meta information for generating question-answer pairs, we find similar and dissimilar neighbours (2 or more) for the query audio sample using CLAP [8] embedding features of the audio clips. Specifically, we randomly sample k between

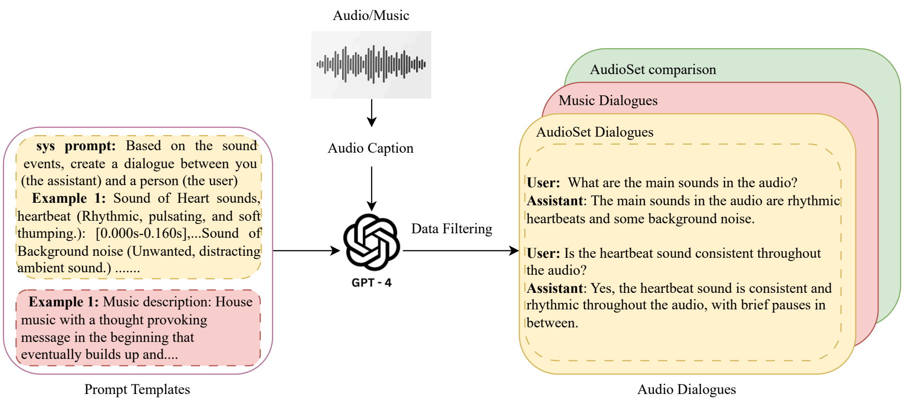

Figure 1: *Illustration of our data generation pipeline.* `Audio Dialogues` *is generated using GPT-4 which takes text-only inputs to generate subsets AudioSet dialogues, Music dialogues and AudioSet comparison subsets of our proposed dataset.*

> **System prompt**
> Based on the description of audios, create a dialogue between you (the assistant) and a person (the user) about the events in the audio.
> **Example 1:**
> Audio 1: Sound of Car (Engine hum and tire noise.)
> Audio 2: Sound of Car (Engine hum and tire noise.)
> User: What's the common type of sound in these two audios?
> Assistant: Both of them have sounds of car and engine humming.

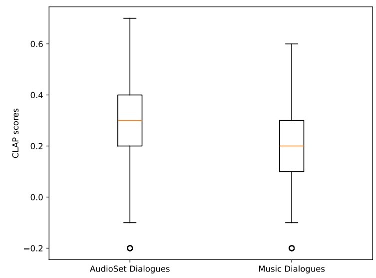

Figure 2: *LAION-CLAP similarities before filtration for AudioSet Dialogues (left) and Music Dialogues (right).*

2 and 4 to select top-k or bottom-k audios given CLAP cosine similarity scores.

3.3. Data filtration

Our data generation pipeline in Section 3.2 generates answers that have phrases such as "difficult to infer", "not specified", "no specific", "no information", and so on. To ensure desirable outputs from the model, we filter these QA pairs following [32]. This is done by manually designing a list of phrases denoting low confidence or uncertainty in the generated answer.

In order to further improve data quality especially the accuracy of the generated answer to the audio, we additionally pick samples in which the answer is highly relevant to the query audio sample. This is done by computing the cosine similarity between the CLAP text-embeddings and audio-embeddings [33] for each QA pair in each dialogue. In Figure 2, we plot the distributions of these similarities before filtration. The samples that have a similarity of less than 0.3 are removed from the dataset.

4. Audio Dialogues dataset

Statistics. In Table 1, we show dataset statistics of Qwen Audio [10] and its instruction fine-tuning dataset for dialogues with 20,000 samples. Apart from this, there is no other information provided as the dataset is not open-source. We also compare the different subsets of our `Audio Dialogues` dataset. For the AudioSet Dialogues subset, there are 76,642 dialogues in the train split and 1,442 dialogues in test split. The Music Dialogues subset has 3,358 dialogues in the train split and 1,641 dialogues in the test split. Each sample has one to four rounds (QA pairs) per dialogue and 1 audio sample as input. Our AudioSet Comparison subset has 64,085 dialogues in the train split and 16,249 dialogues in the test split. Compared to all the other subsets, the AudioSet Comparison dataset has 2 to 4 audio samples as input per dialogue, with an average of 3 audios in the entire subset.

Dataset	Avg. turn or audio per dialogue	#Samples Train	Test
Qwen-Audio [10]	-/1	∼20,000	N/A
Audio Dialogues (Ours)			
AudioSet Dialogues	2.21/1	76,642	1,442
Music Dialogues	1.67/1	3,358	1,641
AudioSet Comparison	1/3.00	64,085	16,249

Table 1: *Average number of turns or average number of audios per dialogue, and total number of train and test samples for all the subsets of* `Audio Dialogues`*.*

Generated dialogues. We demonstrate an example of generated dialogue on the Audioset-SL [31] dataset in Table 2. The generated dialogue captures various aspects of the audio by capturing the frequency of events in the audio, their variation and impact, and the ordering of the events. More examples of all the other subsets are on the demo website.

5. Experiments

In this section, we evaluate three recent audio understanding LLMs on `Audio Dialogues`: LTU [22], Qwen-Audio [10] and Audio Flamingo [18]. We used the pretrained

LTU and Qwen-Audio checkpoints and re-implemented Audio Flamingo. The results are in Table 3.[2] We first do zero-shot evaluation.

```
[Audio annotations:  Sound of Wind
(Whistling, rustling, airy sound.):
[0.000s-10.000s]; Sound of Change ringing
(campanology) (Melodic, rhythmic bell
sequences.):  [0.116s-10.000s]; Sound
of Hubbub, speech noise, speech babble
(Loud, overlapping, indistinct chatter.):
[0.116s-10.000s]; Sound of Tap (Sharp,
rhythmic, resonant sound.):  [0.465s-0.693s],
[2.394s-2.543s]; Sound of Clapping (Sharp,
brief, percussive sound.):  [7.551s-7.693s]]
```

User: What are the main sounds in the audio?
Assistant: The main sounds in the audio are wind, change ringing or campanology, hubbub or speech noise, tapping, and clapping.

User: What is the nature of the speech noise?
Assistant: The speech noise is loud, overlapping, and indistinct, which is often referred to as hubbub or speech babble.

User: Is there any specific event that the clapping sound might be associated with?
Assistant: Given the presence of change ringing and hubbub, the clapping sound could be associated with a celebration or event, possibly in a public place like a town square or church.

Table 2: *Example dialogue from `Audio Dialogues`.*

We then fine-tuned LTU and Audio Flamingo on the training set of AudioSet Dialogues and Music Dialogues. For the AudioSet comparison subset, we only fine-tune Audio Flamingo as LTU does not support multiple audios as input. Compared to the zero-shot results, the fine-tuned model achieves better performance across all metrics (marked as Δ). As Audio Flamingo is trained with retrieval and in-context learning, it shows better performance and can use context better than LTU. This shows fine-tuning on `Audio Dialogues` enables an audio understanding LLM to have much stronger dialogue capabilities.

In Table 4, we also present qualitative results on the test set of the `Audio Dialogues` dataset. We find Audio Flamingo generates more accurate answers than LTU in the multi-turn dialogue setting.

6. Discussion

In this paper, we introduce the Audio Dialogues dataset, designed to enhance audio understanding within the context of multi-turn dialogues, covering a broad spectrum of general sounds and music. By leveraging a prompting-based approach and utilizing caption annotations from existing datasets, we generate a substantial volume of high-quality dialogues suitable for training and evaluating audio-based models.

[2]We do not report Qwen-Audio results on Music Dialogues as it was not trained on music understanding datasets.

Subset	Method	CIDEr ↑	Bleu4 ↑	R-L ↑
AudioSet Dialogues	Qwen-Audio	0.507	0.060	0.292
	LTU	0.580	0.122	0.324
	LTU$^\Delta$	0.823	0.153	0.403
	Audio Flamingo	0.615	0.072	0.341
	Audio Flamingo$^\Delta$	**1.672**	**0.244**	**0.477**
Music Dialogues	LTU	0.168	0.065	0.217
	LTU$^\Delta$	0.419	0.108	0.336
	Audio Flamingo	0.395	0.036	0.250
	Audio Flamingo$^\Delta$	**1.191**	**0.155**	**0.426**
AudioSet Comparison	Audio Flamingo$^\Delta$	1.885	0.444	0.627

Table 3: *Evaluation of LTU [11], Qwen-Audio [10] and Audio Flamingo [18] on the subsets of `Audio Dialogues` test sets. We report the following metrics: CIDEr, Bleu4, and Rouge-L (R-L). Scores improve for all models fine-tuned on `Audio Dialogues`, marked Δ.*

While `Audio Dialogues` enables models to engage in interactions with the user, leading to better comprehension and more nuanced responses, there are certain limitations. One limitation is the absence of timestamps in the generated questions and answers. In the future, an interesting direction would be to temporally ground an answer about an audio. Another limitation is the need for strong annotations to generate dialogues from LLMs. A potential direction to scale generated data is to generate detailed meta descriptions from an existing audio understanding model and designing human feedback-based data filtration strategies.

```
[Audio annotations:  Sound of Rattle
(Sharp, rapid, repetitive, percussive
sound.):  [0.378s-1.346s]; Sound of Spray
(High-pitched, hissing, and continuous sound.):
[1.402s-1.921s], [2.024s-4.346s]]
```

User: *Is the rattle sound followed immediately by the spray sound?*
GT: *Yes, the rattle sound is immediately followed by the spray sound with a very short gap in between.*
LTU: *Yes, the rattle sound is followed immediately by the spray sound. They occur in quick succession.*
Audio Flamingo: *Yes, the rattle sound is followed immediately by the spray sound.*

User: *Does the spray sound occur only once?*
GT: *The spray sound occurs twice in the audio.*
LTU: *Yes, the spray sound occurs only once in the audio, between the 1.414 second and 2.036 second marks. It's a high-pitched, rapid, and hissing sound.*
Audio Flamingo: *No, the spray sound occurs twice in the audio.*

User: *Is the second spray sound longer than the first?*
GT: *Yes, the second spray sound lasts longer than the first one.*
LTU: *Yes, the second spray sound is longer than the first one.*
Audio Flamingo: *Yes, the second spray sound is longer than the first one.*

Table 4: *Qualitative results on `Audio Dialogues` dataset.*

7. References

[1] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2015.

[2] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," *arXiv preprint arXiv:2307.03183*, 2023.

[3] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," *arXiv preprint arXiv:2308.11276*, 2023.

[4] S. Doh, K. Choi, J. Lee, and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," *arXiv preprint arXiv:2307.16372*, 2023.

[5] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[7] C. D. Kim, B. Kim, H. Lee, and G. Kim, "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)*, 2019, pp. 119–132.

[8] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *arXiv preprint arXiv:2305.11834*, 2023.

[10] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[11] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," *arXiv preprint arXiv:2305.10790*, 2023.

[12] S. Ghosh, S. Kumar, C. K. R. Evuru, R. Duraiswami, and D. Manocha, "Recap: Retrieval-augmented audio captioning," *arXiv preprint arXiv:2309.09836*, 2023.

[13] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[14] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[17] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," *arXiv preprint arXiv:2304.03277*, 2023.

[18] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," *arXiv preprint arXiv:2402.01831*, 2024.

[19] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," 2023. [Online]. Available: https://arxiv.org/abs/2309.05767

[20] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1140–1144.

[21] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[22] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," *arXiv preprint arXiv:2305.10790*, 2023.

[23] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, 2022.

[24] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert systems with applications*, vol. 165, p. 113679, 2021.

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[26] R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[27] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv preprint arXiv:2306.09093*, 2023.

[28] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[30] Z. Yang, W. Ping, Z. Liu, V. Korthikanti, W. Nie, D.-A. Huang, L. Fan, Z. Yu, S. Lan, B. Li *et al.*, "Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning," in *EMNLP*, 2023.

[31] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[32] J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "Llark: A multimodal foundation model for music," *arXiv preprint arXiv:2310.07160*, 2023.

[33] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.