Discrete Multimodal Transformers with a Pretrained Large Language Model for Mixed-Supervision Speech Processing

Anonymous submission to Interspeech 2024

Abstract

Recent work on discrete speech tokenization has paved the way for models that can seamlessly perform multiple tasks across modalities, e.g., speech recognition, text to speech, speech to speech translation. Moreover, large language models (LLMs) pretrained from vast text corpora contain rich linguistic information that can improve accuracy in a variety of tasks. In this paper, we present a decoder-only Discrete Multimodal Language Model (DMLM), which can be flexibly applied to multiple tasks (ASR, T2S, S2TT, etc.) and modalities (text, speech, vision). We explore several critical aspects of discrete multimodal models, including the loss function, weight initialization, mixed training supervision, and codebook. Our results show that DMLM benefits significantly, across multiple tasks and datasets, from a combination of supervised and unsupervised training. Moreover, for ASR, it benefits from initializing DMLM from a pretrained LLM, and from a codebook derived from Whisper activations.

Index Terms: discrete tokens, multimodal language models

1. Introduction

The recent introduction of discrete speech tokens [1, 2] has facilitated the integration of large language model (LLM) capabilities into speech models. Consequently, these speech models not only possess knowledge from the labeled speech-text pairs on which they were trained, but also inherit the extensive linguistic knowledge from the LLM. Furthermore, discrete tokenization opens the possibility for the seamless integration of speech, text, and vision within a single framework, enabling the model to perform multiple translation tasks from one modality to another (see Related Work section below). However, within this new research direction, crucial questions arise regarding how to manage training when dealing with multi-modal inputs encompassing speech, text, and/or vision; whether mixed supervision (supervised and/or unsupervised) on multiple tasks can benefit prediction accuracy; and how the discrete codebook should be created.

In this paper, we contribute to the growing field of discrete speech models and (1) propose a Discrete Multimodal Language Model (DMLM) (see Figure 1), which is a unified model that can flexibly translate data from one modality to another to perform multiple tasks for speech processing and more. DMLM is one of the first discrete token-based decoder-only models that can translate to and from text, speech, and images. As such, it can benefit from pretrained large language models, such as OPT [3], that were optimized on vast text corpora. We illustrate DMLM on the tasks of automatic speech recognition (ASR), text-to-speech (T2S), speech-to-text translation (S2TT) from 21 different languages, and image captioning (I2T). In addition, we (2) propose a length-normalized tri-modal loss, a modification of the standard cross-entropy loss for large language models. This adaptation enhances the efficiency of training multi-modal systems, particularly when diverse modalities exhibit significant differences in sequence length. We (3) conduct extensive experiments to explore how mixed-supervision training improves performance on ASR, T2S, S2TT and I2T. Finally, (4) we introduce a discrete codebook based on Whisper [4] and compare with current start-of-the-art codebook [2] for the ASR task.

2. Related Work

Discrete Speech Tokens: Motivated by advances in text generation via large language models as well as the goal of crafting more robust and space-efficient speech representations, researchers have recently explored ways of discretizing input speech [1, 5, 2]. One encoder-decoder method [1] employs an encoder to compress speech into embeddings, and then a decoder toreconstruct the input speech signal. The embeddings undergo quantization and are expressed as a combination of basis vectors in a codebook. While this method can yield high-quality speech with a specific number of quantizers (e.g., 8 or 12 quantizers), it requires a substantial number of tokens (codebook IDs) to capture just one second of speech – typically in the range of 8×65 or 8×75 discrete tokens per second. An alternative discretization strategy [2] involves clustering on a particular layer of a pre-trained speech model, using the clustering ID as the discrete token index. In particular, [2] cluster the activations of a speech model trained through self-supervised learning with contrastive and diversity loss [6, 7].

Speech Processing Based on Pretrained LLMs: Advancements in large language models (LLMs) have prompted efforts to leverage pretrained LLMs for speech-related tasks [8, 2]. For instance, [8] utilize PaLM-2 [9], a non-open-source LLM, by fine-tuning it on speech recognition and speech-to-speech translation (S2ST) datasets. However, they employ loss masking on the Automatic Speech Recognition (ASR) task. Loss masking is a special case of the modal-norm loss function that we use (see Section 3); in particular, it corresponds to setting one or more of the λ values to 0. In our work, instead of hard-coding these λ to a particular value, we optimize them using hyperparameter search to reach potentially more accurate models. VoxtLM [10] use OPT, which is an open-source LLM [3], training it on unsupervised speech data with balanced sampling, ensuring that the number of speech tokens is comparable to the number of text tokens to address length differences. Our approach, on the other hand, ensures that the loss is independent of token length for each modality (speech, text), which may enhance training stability across batches.

Multimodal processing: Efforts in training multimodal

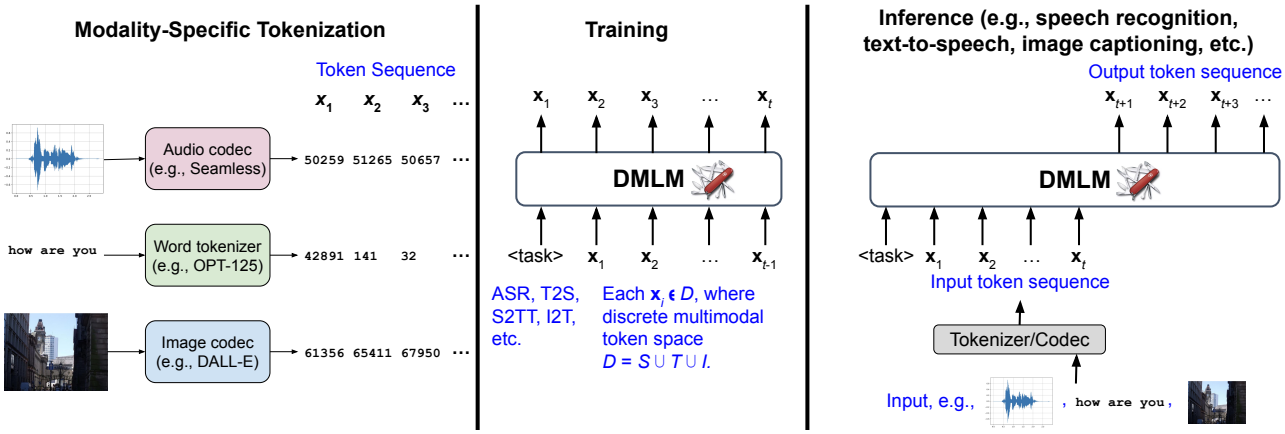

Figure 1: *Discrete Multimodal Language Model (DMLM): The input (text, speech, image) is tokenized by a modality-specific tokenizer/codec. The discrete token sequence* $\mathbf{x}_1, \mathbf{x}_2, \ldots$ *is concatenated to a task description and then passed to DMLM for processing. Training is conducted based on next-token prediction. Inference can be performed for a variety of tasks (ASR, S2TT, T2S, I2T, etc.).*

models have predominantly focused on the integration of text and vision. An effective approach to combining text and vision is to leverage pretrained single modals for each modality separately. Subsequently, one can either perform cross-attention between these domains within the embedding space [11] or concatenate the embeddings of the two modalities [12]. This methodology has been extended to incorporate audio in [13].

While various approaches exist for combining latent spaces across different modalities, some efforts have explored the integration of discrete tokens [14, 15, 16]. Specifically, discrete tokens have been employed in tasks related to vision and text [15, 17]. What sets our work apart is our emphasis on a tri-modal approach, involving speech, vision, and text. In parallel with our research, [16] has also introduced a tri-modal framework. However, our approach distinguishes itself by harnessing the capabilities of LLM and addressing multilingual aspects, particularly in tasks such as speech-to-speech translation. Another advantageous aspect of our model is its ability to facilitate input as a combination of three modalities: speech, vision, and text. In contrast, in [16], the input is limited to either speech, vision, or text individually, lacking the capability to accommodate a combination of these modalities.

3. Discrete Multimodal Language Model

Our discrete multimodal language model (DMLM) is a Transformer decoder in the manner of open-source models such as OPT [3]; as such, it can harness knowledge from LLM pretraining on large text corpora. The model both inputs and outputs discrete tokens across multiple modalities. It can also translate to and from multiple languages. The model can be used for both speech- and non-speech processing tasks.

Let $\mathcal{T}$, $\mathcal{S}$, and $\mathcal{I}$ represent the (mutually disjoint) sets of discrete tokens for the text, speech, and image modalities, respectively, and let $\mathcal{D} = \mathcal{T} \cup \mathcal{S} \cup \mathcal{I}$. These tokens are computed from raw inputs by modality-specific codecs/tokenizers, e.g., Seamless [2] for audio, OPT-125 for text, and DALL-E [15] for images. Associated with each discrete input token $\mathbf{x}_i$ is a d-dimensional embedding vector, which we can implement within the first layer of the decoder by extending the existing embedding matrix (of shape $d \times |\mathcal{T}|$) of a pretrained LLM by adding more columns, with a final shape of $d \times |\mathcal{D}|$. Moreover, the space of possible output tokens is also $\mathcal{D}$. To convert output tokens to speech, we can use a vocoder such as [18].

We frame each task (ASR, T2S, S2TT, etc.) as a next-token prediction problem. The data is organized as a sequence of tokens $\mathbf{x}_1, \mathbf{x}_2, \ldots$, and depending on the specific task at hand, the number and order of modalities may vary. During training the model to translate from a modality A to modality B, the general format is: *Task InputTokens EndA OutputTokens EndB*
where EndA and EndB represents the end delimiter tokens for the source and target modality, respectively. Note that more advanced uses of the model are also possible (e.g., input both audio and images to perform context-dependent speech recognition), but we leave these to future work.

Loss function: Naively applying the cross-entropy loss summed over the multimodal sequence of tokens $\mathbf{x}_1, \mathbf{x}_2, \ldots$ leads to poor results due to large differences in the number of tokens between domains. For ASR, for example, the number of audio tokens is typically much larger than the corresponding number of text tokens. To account for these differences, we devised a new loss function that normalizes the lengths of tokens within each modality:

$$\mathcal{L} = -\frac{\lambda_{\mathcal{S}}}{n_{\mathcal{S}}} \sum_{j \in \mathcal{S}} p_j \log \hat{p}_j - \frac{\lambda_{\mathcal{T}}}{n_{\mathcal{T}}} \sum_{j \in \mathcal{T}} p_j \log \hat{p}_j - \frac{\lambda_{\mathcal{I}}}{n_{\mathcal{I}}} \sum_{j \in \mathcal{I}} p_j \log \hat{p}_j$$

where $\hat{p}$ is the vector of softmax probabilities over the entire output token space ($\mathcal{D}$), and $\lambda_{\mathcal{T}}$, $\lambda_{\mathcal{S}}$, and λ_I are the weights of text, speech, and image modalities, respectively. Meanwhile, $n_{\mathcal{T}}$, $n_{\mathcal{S}}$, $n_{\mathcal{I}}$ represent the number of text, speech and image tokens. This loss function resembles that in DALL-E, albeit with a distinction: In DALL-E [15], the loss is weighted between text and vision tokens, and the weights λ are determined heuristically. In our approach, we utilize hyperparameter search to determine the contribution of each modality to the loss.

Example inputs & outputs: Figure 2 shows example inputs & outputs of a DMLM model, trained on the datasets mentioned in Section 4, when applied to different tasks. Note that for these examples, different model weights were optimized for each task; however, in principle a single set of DMLM model weights could be trained to perform all tasks simultaneously.

Also, for the specific task of automatic speech recognition (ASR), we performed supervised training LibriSpeech (960 hours), TED-LIUM (777 hours), SPGI (5000 hours), along with unsupervised on a 3% subset of Gigaspeech, and thereby

Figure 2: *Example inputs/outputs of a trained DMLM.*

produced an ASR model that attains 10.6% word error rate (WER) on LibriSpeech-other. This is slightly better than the [10] VOXTLM model, which was trained on 4x as much supervised and 160x as much unsupervised data and attains a WER of 11.9%.

4. Experiments

Before presenting experiments on (1) the loss function weights λ, (2) model pretraining, (3) unsupervised training, and (4) tokenization codebooks, we first describe the experimental setup.

Datasets: For **ASR**, we used the following datasets: (1) Librispeech [19], encompassing 960 hours (460 hours clean, 500 hours noisy) of speech from public-domain audio books; (2) AMI Meeting [20], featuring 100 hours of speech from meetings; (3) TED-LIUM ASR corpus [21] containing a total of 777 hours of recordings of TED talks; (4) SPGISpeech [22], encompassing 5000 hours of corporate earnings calls along with their associated transcriptions; and (5) Gigaspeech [23]. For **T2S** tasks, we used LibriTTS-R [24], containing 960 hours of speech. For the **S2TT** tasks, we used the CoVoST2 dataset [25]. CoVoST2 consists of 861 hours of transcribed speech from 21 languages and the corresponding English translation. Finally, for **I2T** we used COCO [26] and SBU [27].

Network, Hyperparameter Settings, & Codecs: We employ the OPT model [3] as our pretrained large language model (LLM) with 125 million parameters. We leverage Hugging Face [28] as our training framework. Our experiments run with a batch size of 4, a learning rate set at 5×10^{-5}, weight decay of 0.0001, and a patience of 1. Model training is conducted on a single Nvidia 40 GB or 80 GB A100 GPU.

To convert from raw waveform to discrete speech tokens, we use the pretrained Seamless codec [2] except in Experiment 4, where we compare to Whisper [4]. To convert from images to discrete tokens and vice-versa, we use DALL-E [15].

Evaluation and Metrics: For ASR, we utilize the standard WER metric. For T2S, we transcribe the synthesized speech and use pretrained Whisper large-v2 to transcribe and evaluate the character error rate (CER). For image captioning, we use CIDEr [29], BLEU-4 [30] and METEOR [31]. For speech-to-text translation, we use BLEU-4 and sacreBLEU [32].

4.1. Experiment 1: Tri-Modal Loss Function

Early during development of DMLM we found that the loss function played a critical role in ensuring the stability of training. In particular, it was necessary to select the λ to account for the varying lengths of tokens depending on modality. Hence, in our first experiment, we assess how $\lambda_{\mathcal{S}}, \lambda_{\mathcal{T}}$ affect the test accuracy of DMLM when applied to speech recognition. In particular, we train DMLM for ASR on 100 hours of LibriSpeech, and we compare (1) heuristically setting $\lambda_{\mathcal{S}} = \lambda_{\mathcal{T}}$ (similar to Votxtlm [10]); (2) setting $\lambda_{\mathcal{S}} = 0, \lambda_{\mathcal{T}} = 1.00$ (as with Google AudioPaLM [8]); and (3) automatic hyperparameter selection using Ax (`https://ax.dev`). With Ax, we conducted 25 trials using early stopping on the LibriSpeech-clean development set with WER as the metric to choose the best model.

Table 1: *Effect of $\lambda_{\mathcal{S}}, \lambda_{\mathcal{T}}$ on ASR WER(%) of LibriSpeech-clean test set, after training on LibriSpeech (100 hours).*

Speech weight $\lambda_{\mathcal{S}}$	Text weight $\lambda_{\mathcal{T}}$	LS-c
0.25	0.93	**12.9**
1.00	1.00	21.9
0.00	1.00	15.7

Table 2: *Effect of model pretraining on ASR WER(%), comparing random initial model weights to OPT-pretrained weights, followed by training/fine-tuning on LibriSpeech (960 hours).*

Initialization	LS-clean	LS-other	AMI
Random	7.1	14.3	658.7
With LLM	**5.6**	**12.5**	**94.8**

Results: Ax selected hyperparameters $\lambda_{\mathcal{S}} = 0.25$ and $\lambda_{\mathcal{T}} = 0.93$. We thus used these values in this as well as all other experiments in this paper. Furthermore, as a heuristic choice, we also set $\lambda_{\mathcal{I}} = 0.25$ for image captioning in Experiment 3.

Table 1 shows the WER on the LibriSpeech-clean test set for different values of $\lambda_{\mathcal{S}}, \lambda_{\mathcal{T}}$. It is clear that applying a loss to the predictions of both speech and text tokens yields the best WER (12.9%), outperforming both the setting (used in AudioPalm [8]) where only text is modeled (15.7%) and the setting (used in VoxtLM [10]) with balanced weights between speech and text (21.9%).

4.2. Experiment 2: Pretraining the LLM

A potential benefit of a decoder-only model such as DMLM is that it can harness the knowledge from pre-trained LLMs such as OPT. Here, we compare a DMLM with random initialization to a model initialized from OPT. In particular, we train DMLM for speech recognition (ASR) on all 960 hours of LibriSpeech, and we evaluated it on the LibriSpeech-clean, LibriSpeech-other, and AMI test sets.

Results: The results in Table 2 illustrate a consistent accuracy improvement due to LLM pre-training. Notably, the LLM pre-training makes a large difference for the out-of-domain AMI dataset (WER of 94.8% vs. 658.7%).

4.3. Experiment 3: Mixed-Supervision Training

In contrast to state-of-the-art speech recognition systems such as Whisper, DMLM can be trained in an unsupervised manner for next-token prediction, without needing labeled pairs that link across modalities. In this experiment we explore the utility of mixed-supervision training for speech recognition (ASR), speech synthesis (T2S), and image captioning (I2T) tasks.

For **ASR**, we compare training only supervised on LibriSpeech (either 100 hours or 960 hours), to training both supervised on LibriSpeech as well as unsupervised with either speech-only data from the AMI training set, LibriSpeech-500 speech-only data, or Gigaspeech (we used just 3% of the whole dataset due to computational constraints) text-only data. For

Table 3: *Effect of unsupervised training (plus supervised training on LibriSpeech-100) on ASR WER(%) on the LibriSpeech-other test set.*

Training set		Test set
Supervised	**Unsupervised**	**LS-o**
LS-100	None	28.6
LS-100	20% of LS-500 (Speech only)	**25.1**

Table 4: *Effect of unsupervised training (plus supervised training on LibriSpeech-960) on ASR WER(%) on the AMI test set.*

Training set		Test set
Supervised	**Unsupervised**	**AMI**
LS-960	None	94.8
LS-960	AMI (Speech only)	92.6
LS-960	3% Gigaspeech (Text only)	**59.3**

T2S, we train either just supervised on LibriTTS-R, or both supervised and unsupervised using Gigaspeech (4% of the whole dataset). For **I2T**, we compare training just supervised on COCO, to supervised on COCO combined with LibriSpeech 960 (text-only) and SBU (image-only) data.

Results: Table 3 shows that unsupervised training with speech-only data from LibriSpeech-500, in conjunction with supervised training on LibriSpeech-100, improves WER on LibriSpeech-other for **ASR**. Similarly, Table 4 shows that leveraging speech-only data from AMI, in conjunction with supervised training on LibriSpeech-960, improves WER on the AMI test set. Notably, unsupervised training on a text-only dataset (3% of Gigaspeech) decreases the WER even further to 59.3%.

T2S: Table 5 shows that, by adding additional 4% of Gigaspeech for speech-only unsupervised training, the CER is reduced from 22.9% to 20.2%.

S2TT: Table 6 shows a small improvement of speech-to-text translation with unsupervised speech-only data from LibriSpeech-960 (line #2), as well as a larger improvement (line #3) by mixing two forms of supervised training (CoVoST2 for S2TT and LibriSpeech-960 for ASR).

I2C: Table 7 shows that, by adding the text-only data from LS-960, we improve the CIDEr score from 58.1 to 61.5. Furthermore, the inclusion of image-only data from SBU increases the CIDEr score to 62.8.

4.4. Experiment 4: Codebook

Finally, we investigate the use of an alternative codebook of speech tokens derived from the state-of-the-art Whisper [4] model. We use the same approach as [2], namely K-means clustering on the activations from a single layer of the encoder – we used the final layer of Whisper large-v2. We use the mini-batch k-means algorithm in scikit-learn [33] with k=10000 and 64 utterances per minibatch. Tokens are then generated by passing audio to the encoder and assigning a token based on the nearest cluster for each timestep. For comparison, we train equivalent k-means models on XLS-R as used in [2]. We use Librispeech-960h to train the k-means for both Whisper and XLS-R.

Results: We find that using Whisper features as the basis for discrete tokens yields a lower WER on Librispeech test than when the units are trained on XLS-R (Table 8). This improvement may be attributed to the fact that Whisper was trained with ASR as one of its tasks, thereby learning a more effective representation for this task.

Table 5: *Effect of unsupervised training (plus supervised training on LibriTTS-R) on T2S CER(%) on the LibriTTS-R test set*

Training set		Test set
Supervised	**Unsupervised**	**LibriTTS-R**
LibriTTS-R	None	22.9
LibriTTS-R	4% Gigaspeech (Speech only)	**20.2**

Table 6: *Effect of mixed-supervision training on S2TT B@4 (BLEU-4) and sacreBLEU. BLEU and sacreBLEU scores are calculated based on the average performance across 21 languages within the CoVoST2 X→en test set.*

Training set		Test set	
Supervised	**Unsupervised**	**B@4(↑)**	**sacreBLEU (↑)**
CoVoST2	None	5.71	5.75
CoVoST2	LS-960 (speech)	5.74	5.76
CoVoST2 (S2TT) + LS-960 (ASR)	None	**6.18**	**6.22**

Table 7: *Effect of unsupervised training (plus supervised training on COCO) on I2T metrics on the COCO test set. B@4: BLEU-4, M: METEOR.*

Training set		Test set		
Supervised	**Unsupervised**	**CIDEr (↑)**	**B@4 (↑)**	**M (↑)**
COCO	None	58.1	19.5	19.2
COCO	LS-960 (text)	61.5	20.6	19.9
COCO	SBU (image)	**62.8**	**21.1**	**20.1**

5. Discussion and Conclusions

We introduced DMLM, a decoder-only framework that can harness pre-trained LLMs and can be used for multiple tasks across multiple modalities. DMLM uses discrete speech, vision and text tokens as the inputs and treats all the tasks as a next-token prediction task, thus enabling the framework to train with mixed supervision in a unified manner. We demonstrate the effectiveness of using unsupervised data on various of tasks: speech recognition, speech synthesis, speech to text translation and image captioning. We further investigate the choice of embedding space leading to discrete speech tokens, in particular between Whisper and XLS-R for speech recognition task.

Our work leads to interesting future research directions, e.g., expand the model to handle different audio tasks beyond speech such as music generation, as well as audiovisual speech recognition through multimodal input streams.

Table 8: *WER(%) using Whisper-derived tokens*

Model	K-means data	Training set	LS-c	LS-o
Seamless	LS-960	LS-960	31.8	61.2
Whisper	LS-960	LS-960	**14.9**	**23.1**

6. References

[1] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[2] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[3] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023.

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[6] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. M. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2021.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[8] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[9] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[10] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, "Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks," *arXiv preprint arXiv:2309.07937*, 2023.

[11] 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.

[12] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[13] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.

[14] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[16] M. Kim, J.-w. Jung, H. Rha, S. Maiti, S. Arora, X. Chang, S. Watanabe, and Y. M. Ro, "Tmt: Tri-modal translation between speech, image, and text by processing different modalities as different languages," *arXiv preprint arXiv:2402.16021*, 2024.

[17] Y. Ge, S. Zhao, Z. Zeng, Y. Ge, C. Li, X. Wang, and Y. Shan, "Making llama see and draw with seed tokenizer," *arXiv preprint arXiv:2310.01218*, 2023.

[18] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[20] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[21] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*. Springer, 2018, pp. 198–208.

[22] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman *et al.*, "Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv preprint arXiv:2104.02014*, 2021.

[23] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[24] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. A. U. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: Restoration of a large-scale multi-speaker tts corpus," 2023.

[25] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv preprint arXiv:2007.10310*, 2020.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[27] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[29] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[31] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[32] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319

[33] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1177–1178. [Online]. Available: https://doi.org/10.1145/1772690.1772862