

BLSP: Bootstrapping Language-Speech Pre-training via Behavior Alignment

Anonymous ACL submission

Abstract

The emergence of large language models (LLMs) has sparked significant interest in extending their remarkable language capabilities to speech. However, modality alignment between speech and text remains an open problem. Current solutions can be categorized into cascaded approaches, which limit the interaction between speech and LLMs, and end-to-end approaches that rely on scarce speech instruction data. In this paper, we propose the **BLSP** approach that **Bootstraps Language-Speech Pre-training via behavior alignment**, leveraging existing ASR training data. We achieve this by developing a lightweight modality adapter between a frozen speech encoder and an LLM, optimized to ensure that the LLM exhibits the same generation behavior irrespective of the modality of input: a speech segment or its transcript. We primarily focus on the continuation writing behavior as it closely resembles next-token prediction in a broad sense but also found that introducing other behaviors could lead to improved performance. We demonstrate that this simple process can extend the capabilities of LLMs to speech and achieve competitive performance compared to cascaded systems, enabling speech recognition, speech translation, spoken language understanding, and speech conversation, even in zero-shot cross-lingual scenarios.¹

1 Introduction

Large Language Models (LLMs), trained on massive amounts of textual data, have achieved significant success on various natural language processing tasks (Chowdhery et al., 2022; OpenAI, 2023; Gao et al., 2023). Recent research efforts have attempted to expand LLMs’ capabilities to comprehend diverse modalities (Yin et al., 2023; Latif et al., 2023). Speech, as an important modality, offers a plethora of benefits that complement

text-based communication. Speech not only serves as the primary mode of human interaction but also conveys rich emotions, tones, and intentions that cannot be fully captured in text. Thus, enabling LLMs to understand speech could greatly enhance their utility in real-world scenarios.

However, effectively integrating and aligning speech with LLMs remains a significant challenge. Current approaches can be classified into two categories. One adopts a cascade paradigm, where the LLM is equipped with an automatic speech recognition (ASR) model to convert speech into text (Huang et al., 2023; Shen et al., 2023), or the LLM is fed output states from a separately trained recognition system (Chen et al., 2023). In this setup, the transfer of knowledge from the LLM to the speech modality is hindered due to the separation between ASR and LLM training. Recent efforts explore training end-to-end speech-language LLMs for direct speech interaction (Zhang et al., 2023; Shu et al., 2023). Yet, this approach heavily relies on scarce speech instruction data, which is challenging to collect in large quantities, and struggles to generalize robustly across languages and speakers. In this work, we address the question of whether it is possible to align speech and text in a generalized manner using existing cross-modal datasets like ASR data, which is available in large volumes.

Our preliminary investigation has revealed that a model trained to predict the ground-truth transcript with speech input loses the ability to follow instructions. To achieve effective cross-modal alignment, we introduce the BLSP approach, which bootstraps language-speech pre-training via behavior alignment. The key idea is to develop a lightweight modality adapter between a frozen speech encoder and an LLM, optimized to ensure that the LLM exhibits the same generation behavior irrespective of the modality of input: a speech segment or its transcript. Specifically, we first prompt an LLM to generate text responses from speech transcripts.

¹Video demos are available at <https://anonymous4blsp.github.io/arr/>.

Then, we use these responses as supervised signals to optimize the parameters of the modality adapter. Our primary focus is on the continuation writing behavior as it prompts the LLM to generate text that resembles the broad data it is trained on, without biasing toward a specific task. However, we have observed that incorporating other behaviors, specifically repetition that mirrors the speech recognition task, could yield advantages in fine-grained lexical modeling. Our experiments reveal that the BLSP approach can effectively achieve cross-modal alignment and achieve competitive performance compared to cascaded systems, enabling LLMs to understand speech while retaining their language capabilities.

The contributions of our work are as follows:

- We introduce a novel approach to bootstrap language-speech pre-training through behavior alignment, providing a new direction for cross-modal alignment in LLMs.
- We develop a simple process that requires training only a lightweight modality adapter, leveraging a pretrained speech encoder and LLM, and utilizing existing speech recognition data, thus eliminating the need to acquire speech instruction data.
- We conduct quantitative evaluations and provide video demonstrations to showcase that our BLSP approach effectively extends LLMs to speech inputs and achieves competitive performance compared to cascaded systems, enabling speech recognition, speech translation, spoken language understanding, and speech conversation, even in zero-shot cross-lingual scenarios.

2 Background

Due to the scarcity of speech instruction data, a natural approach to align speech and text for leveraging LLMs is to connect a pre-trained speech encoder to an LLM through a modality adapter trained on large volumes of speech-transcript pairs collected for speech recognition, as explored in (Shu et al., 2023; Xue et al., 2024). Similar methods have achieved considerable success in vision-language models. Notably, BLIP-2 (Li et al., 2023) and MiniGPT-4 (Zhu et al., 2023) have demonstrated that training a learnable interface using aligned image caption data can effectively bridge

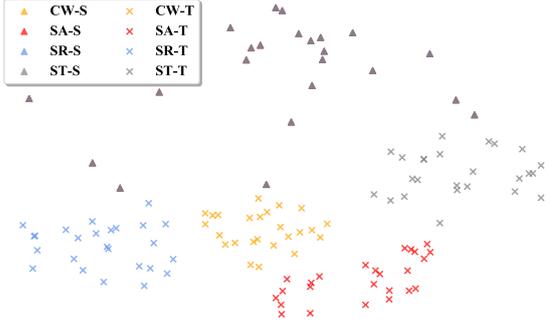


Figure 1: T-SNE visualization of feature representations learned from ASR task. Colors denote task instructions: orange for continuation writing (CW), red for sentiment analysis (SA), blue for speech recognition (SR), and gray for speech translation (ST). Shapes distinguish input modality: triangles for speech, crosses for text. Note that speech inputs result in overlapping features.

the modality gap between vision and language, enabling an LLM to comprehend images while retaining its capacity to follow text prompts.

However, this approach proves to be more intricate when it comes to effectively achieving speech and text alignment, crucial for extending the language capabilities of LLMs to speech inputs. Our preliminary investigation, detailed in Appendix A, has found that training a modality adapter to predict the ground-truth transcript from speech input can inadvertently restrict the LLM to performing solely speech recognition tasks. This issue arises despite the variety of transcription instructions used during training, as the LLM tends to overlook any textual instructions provided before the speech segment. We hypothesize that the reliance on homogeneous ASR training data results in a strong bias in the learned speech representations, confining the LLM’s functionality to the transcription task only.

To substantiate our hypothesis, we conducted an analysis of the representations learned from ASR task on speech and text pairs from the LibriSpeech dataset (Panayotov et al., 2015). We consider four distinct tasks: continuation writing (CW), sentiment analysis (SA), speech recognition (SR), and speech translation (ST), prompt by instructions. For each task, the same task-specific instruction is employed to prompt the LLM to process either speech or its corresponding transcript. The cross-modal prompt is formatted as follows:

```
###[Human]:<instruction><speech/transcript>\n\n\n###[Assistant]:
```

The learned representations of these inputs are obtained by extracting the hidden state of the final

###[Assistant]:<response>

In this work, we primarily focus on the continuation behavior, as it resembles universal generation in next-token prediction and can produce a diverse range of texts reflecting the extensive dataset used to train LLMs. This characteristic is crucial as it avoids over-fitting to more specific behaviors that exhibit a strong bias in the response, akin to issues encountered with ASR pretraining, as discussed in Section 2.

While not inherently beneficial on their own, incorporating data from certain specific behaviors alongside continuation data in modest proportions can enhance the performance of the model. For instance, integrating the repetition behavior, which resembles ASR pretraining, into the continuation data can assist the adapter in capturing fine-grained lexical information, as explored in this study. We leave the systematic investigation of other behaviors for future research.

See Table 1 for two instructions used to prompt behaviors. It is worth noting that since the response for repetition behavior is simply the original transcript with minor changes based on how closely the LLM follows the repetition instruction, we can skip the first step and directly use the speech transcript as the response.

Continuation: *Continue the following text in a coherent and engaging style with less than 40 words.*

Repetition: *Please repeat the following words.*

Table 1: Instructions used to prompt LLM behaviors.

4 Experiment Setup

4.1 Training Details

We utilize the encoder part of Whisper-small (Radford et al., 2022) as the speech encoder and employ Llama-2-7B (Touvron et al., 2023) as the large language model (LLM). To induce instruction-following capability², we employ the publicly accessible dataset Alpaca-52K (Taori et al., 2023) to fine-tune the LLM. The Alpaca-52k dataset consists of 52K (text instruction, text input, response) triplets, which we convert into (text instruction, response) pairs by combining the instructions and

²We could have also used the official chat version of Llama-2, but we opted to perform instruction finetuning using publicly available data, as it offers flexibility for future research involving multi-modal instruction data.

inputs. During this stage, we fine-tune all parameters of the LLM for 3 epochs with a batch size of 128.

The modality adapter is composed of three 1-dimensional convolution layers followed by a bottleneck layer (Houlsby et al., 2019) with a hidden dimension of 512. The convolution layers are designed to reduce the length of the speech features by a factor of 8, with each layer having a stride size of 2, a kernel size of 5, and a padding of 2. To train the modality adapter, we utilize publicly available speech recognition datasets, including LibriSpeech (Panayotov et al., 2015), GigaSpeech (Chen et al., 2021), and Common Voice 2.0 (Ardila et al., 2020).

We train two BLSP models. The primary BLSP model is trained solely on continuation behavior, using 8.8 million (speech, text continuation) pairs obtained by performing the continuation writing task on the ASR datasets with the fine-tuned Llama-2 model. The secondary BLSP+RP model, included for comparison, is trained on a 1:9 mixing ratio of repetition data in the form of (speech, transcript) pairs and the aforementioned continuation data. During this stage, we fine-tune the modality adapter for one epoch with a batch size of 768.

4.2 Baselines

We compare our method with the following baselines.

Text+LLM The input to the LLM is the ground-truth speech transcripts.

Whisper+LLM The input to the LLM is the speech recognition output from whisper-small, which is comprised of both an encoder (utilized as the speech encoder in BLSP) and a decoder (not employed in BLSP). When comparing BLSP models to this baseline, it is important to note that BLSP’s speech training data is much smaller than that for Whisper models.

CTC+LLM The input to the LLM is the speech recognition output from an in-house CTC-based ASR model. This ASR model consists of a speech encoder and adapter identical to those in BLSP, in addition to a CTC projector. We freeze the speech encoder and fine-tune the adapter and projector on the same ASR datasets used for BLSP training. We consider CTC+LLM as the most realistic cascaded baseline for demonstrating the modeling power of the BLSP approach.

Method	ASR (WER↓)				ST (BLEU↑)		SLU (ACC↑)		
	LibriSpeech test-clean		TED-LIUM 3		MUST-C	CoVoST 2.0	SNIPS light-close	FSC	SLUE VoxCeleb
Text+LLM	0.0	5.6	0.0	14.5	19.7	21.9	86.3	72.4	75.0
Whisper+LLM	3.4	5.9	4.3	20.4	16.6	16.9	83.2	56.3	74.1
CTC+LLM	6.2	10.8	8.4	20.7	13.3	13.2	79.0	60.4	74.7
ASR pretraining	—	3.7	—	4.5	0.0	0.0	0.0	0.0	0.0
BLSP	—	10.4	—	23.1	12.3	12.7	75.8	60.9	76.0
+RP	—	6.4	—	8.1	14.9	13.8	78.8	77.5	75.5

Table 2: Overview of BLSP results on zero-shot speech-to-text tasks. For each ASR test set, we report two WER scores: on the left for the standalone ASR component of a cascaded system, and on the right for instructing the LLM to repeat the words recognized by the ASR component. The BLEU scores for the ST test sets are averaged across multiple translation directions.

ASR pretraining The model utilizes the same architecture as BLSP, except the modality adapter is trained by predicting the ground-truth transcript, as discussed in Section 2 and detailed in Appendix A.

5 Results

We have found through experiments that the proposed BLSP approach is capable of empowering the LLM with speech understanding capabilities while maintaining fidelity to text instructions, achieving competitive performance compared to the cascaded baseline CTC+LLM. We conduct evaluations on multiple speech-related downstream tasks, including speech recognition (ASR), speech translation (ST), and spoken language understanding (SLU). It is important to note that the primary BLSP model is trained solely on the continuation writing task; therefore, all evaluations are conducted in a **zero-shot** setting, where we utilize text instructions to perform various speech-to-text generation tasks. For the BLSP+RP model, all evaluations except the ASR task are conducted in a zero-shot setting. We also demonstrate the open-ended generation capability of BLSP by conducting general-purpose QA. Additionally, we demonstrate that our model supports cross-modal conversations and develops multilingual capabilities, even though the alignment training is carried out only in English.

5.1 Quantitative Evaluations

Instructed Zero-Shot Speech-to-Text Tasks

We perform speech-to-text tasks by prompting the BLSP model with task-specific instructions, detailed in Table 11 in the Appendix, followed by the speech features as input to the LLM. The same

instructions are also used in the baseline systems.

For the ASR task, we conduct quantitative evaluations on both in-domain (LibriSpeech, Panayotov et al., 2015) and out-of-domain (TED-LIUM 3, Hernandez et al., 2018) test sets, utilizing Word Error Rate (WER) as the evaluation metric. In a cascaded system, the ASR task can be performed in two ways: either directly using the standalone ASR component or by instructing the LLM to repeat the words recognized by the ASR component. We compare these methods to assess the LLM’s ability to follow ASR instructions. For the ST task, we use SacreBLEU (Post, 2018) as the evaluation metric, and report in-domain results on CoVoST-2 (Wang et al., 2020) and out-of-domain results on MUST-C (Di Gangi et al., 2019), averaged across five and eight translation directions, respectively, as detailed in Appendix B. For the SLU task, we evaluate on intent classification (IC) datasets SNIPS (Saade et al., 2019) and FSC (Lugosch et al., 2019), and sentiment analysis (SA) dataset SLUE-VoxCeleb (Shon et al., 2022), using accuracy as the evaluation metric.

Results are presented in Table 2. Despite a significant performance disparity in ASR and ST tasks when compared to the cascaded system of Whisper+LLM, our primary BLSP model demonstrates respectable outcomes across all evaluated tasks. It’s important to note that the Whisper model, incorporating a decoder absent in our BLSP model, benefits from training on a substantially larger corpus of speech data than the BLSP model. However, when contrasted with the most directly comparable cascaded baseline, CTC+LLM, which has a similar architecture and is trained on an equivalent volume of speech data, the performance gap narrows considerably. Remarkably, the BLSP model surpasses CTC+LLM in the FSC and SLUE-VoxCeleb test

sets for the SLU tasks. Conversely, the ASR pre-training method, frequently employed in prior research to facilitate cross-modal alignment in LLMs (Zhang et al., 2023; Shu et al., 2023), proves ineffective in maintaining any capability for instruction-following in non-ASR tasks.

Incorporating Additional Behaviors We observe that the ASR component of a cascaded system has a significantly lower WER score than the cascaded system itself. This suggests that the LLM’s insufficient ability to closely follow the ASR instruction is one of the reasons the BLSP method performs less effectively than traditional ASR models in ASR tasks. As shown in Table 2, the BLSP+RP model, which utilizes repetition training data at a 1:9 mixing ratio with the continuation training data, achieves WER scores comparable to the ASR component of the CTC+LLM model, and significantly better scores than those achieved by the CTC+LLM method through prompting. Moreover, the inclusion of repetition data also leads to improved performance on other tasks, achieving better scores on both ST test sets and two of the three SLU test sets compared to the CTC+LLM baseline.

General-Purpose QA We also evaluated the performance of our BLSP models on a general-purpose question-answering (QA) task. This task focuses on grasping the semantics conveyed through speech and encompasses a broader range of textual instructions. For this evaluation, detailed in Appendix C, we selected 1460 samples from the GigaSpeech test set and employed ChatGPT to create a question for each sample based on its transcript. We then utilized ChatGPT again to determine the acceptability of the responses generated by different methods. The evaluation findings are summarized in Table 3. Both BLSP models exhibited competence in this task, achieving scores comparable to the cascaded baseline CTC+LLM (88.5%/88.3% vs 88.6%). This performance underscores our approach’s ability to endow the LLM with a general comprehension of speech, thereby equipping it to adeptly handle diverse cross-modal instructions and produce satisfactory responses.

5.2 Analysis

Effectiveness as a Pre-Training Strategy We evaluate the BLSP method’s effectiveness as a pre-training strategy for downstream tasks, with a focus on speech translation. To do this, we follow the

Method	Accept Rate (%)
Text+LLM	94.5
Whisper+LLM	91.3
CTC+LLM	88.6
ASR pretraining	0.0
BLSP	88.5
+RP	88.3

Table 3: ChatGPT evaluation using acceptable rate.

same translation instruction as used in zero-shot translation tasks and fine-tune the primary BLSP model to predict target language translations directly from speech inputs. This fine-tuning process utilizes training data across eight language pairs from the MUST-C dataset. We apply LoRA (Low-Rank Adaptation) (Hu et al., 2021) to modify the key, query, value, and output layers of the LLM’s self-attention mechanism, setting LoRA hyperparameters to $R = 16$ and $\alpha = 16$. We also update the speech encoder and the modality adapter’s parameters to enhance model performance. For context, we compare these results with a commonly used pre-training approach, specifically ASR pre-training, as detailed in Section 4.2.

As illustrated in Table 4, our primary BLSP model exhibits an advantage in pre-training the modality adapter for the downstream speech translation task, achieving substantial improvements over random initialization. While pre-training the modality adapter using the ASR task proves beneficial, it may introduce a bias that hinders its ability to generalize across different downstream tasks. This limitation is highlighted by the superior performance of our BLSP approach over ASR pre-training. Our method achieves notably higher COMET scores across all translation directions and higher BLEU scores in six out of the eight directions evaluated.

Effectiveness in Speech-Text Alignment We evaluate the effectiveness of the BLSP method in aligning speech and text inputs, using the procedures outlined in Section 2. As illustrated in Figure 3, the distribution of learned representations from speech inputs by the primary BLSP model no longer significantly differs from that of text inputs. This is a departure from the results observed with the ASR task, as depicted in Figure 1. The representations of speech inputs now share the same distribution as those of text inputs, with the

Method	en-de	en-es	en-fr	en-it	en-nl	en-pt	en-ro	en-ru
w/o pretraining	21.1 / 74.4	25.4 / 76.1	29.9 / 75.6	20.6 / 76.1	23.6 / 76.8	25.3 / 76.7	16.4 / 74.7	13.7 / 73.5
ASR pretraining	22.7 / 76.6	27.9 / 78.7	32.1 / 77.7	22.3 / 78.2	25.4 / 78.7	27.3 / 79.6	18.6 / 77.4	14.9 / 76.2
BLSP	23.3 / 77.7	27.4 / 79.5	31.9 / 78.5	23.2 / 79.0	26.4 / 80.0	28.5 / 80.4	19.2 / 78.6	15.6 / 77.3

Table 4: ST results (BLEU / COMET) of fine-tuned models on MUST-C.

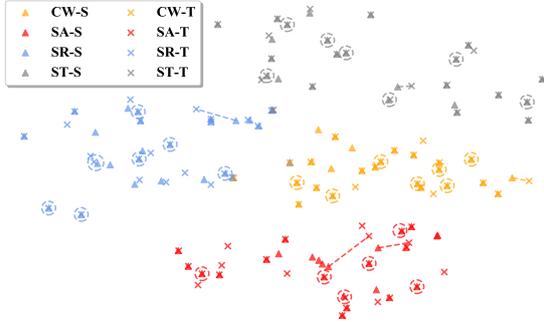


Figure 3: T-SNE visualization of feature representations learned from BLSP. Selected paired speech and text inputs are highlighted using dashed lines and circles.

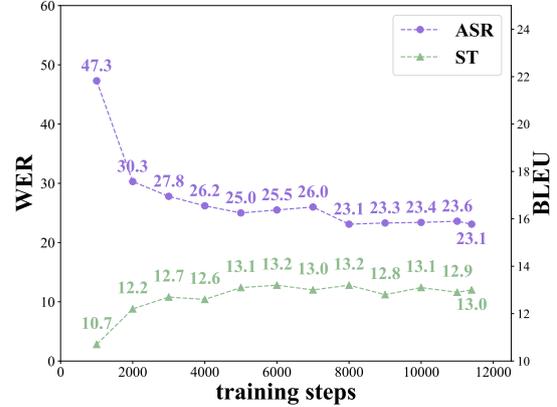


Figure 4: ASR and ST results at different training steps in one epoch.

representations of paired speech and text inputs being closely aligned, often overlapping. In Appendix D, we provide quantitative evidence that our BLSP model can generate distinct representations for the same speech input under different instructions, and that the representations for paired speech and text inputs closely match when given the same instructions. These results indicate that the BLSP approach effectively aligns speech and text inputs within the same space, thereby extending the instruction-following capabilities of LLMs to speech inputs.

Impact of Data Size We evaluate the impact of data size on model performance within the BLSP approach, utilizing measurements on out-of-domain datasets, specifically TED-LIUM 3 for zero-shot ASR performance and MUST-C en-de direction for zero-shot ST performance. In our experimental setup, we limit model training to a single epoch since the training loss converges well before the completion of one epoch. Consequently, we employ its performance at various training steps (approximately 0.8 million training samples for every 1,000 updates) as an estimate of its performance at different data scales. As shown in Figure 4, we observe rapid improvement in model performance during the early stages of training, followed by convergence after approximately 8,000 updates (equivalent to around 6 million training samples).

5.3 Cross-Modal Conversation

We have observed that the BLSP approach can enable multi-turn conversation capabilities with LLMs using speech, thereby extending their remarkable conversational capabilities learned from text-only data to spoken languages. Figure 5 illustrates an example of engaging in a spoken conversation in English with the model. More examples are presented in Appendix E. Longer video demonstrations are available online³.

5.4 Emergence of Multilingual Capabilities

Despite being trained solely on English ASR data for behavior alignment in continuation writing, we have observed that the BLSP model demonstrates an understanding of non-English speech inputs. This can be attributed to the multilingual capabilities of both the speech encoder (Whisper, Radford et al. (2022)) and the LLM (Llama-2, Touvron et al. (2023)), as well as the specific design of the BLSP training process. Note that both the speech encoder and the LLM remain frozen during BLSP training, suggesting that despite training solely on English data, the modality adapter can learn to project the multilingual space in Whisper encoder’s output to the multilingual space for the LLM.

To quantitatively measure the multilingual ca-

³Video demos are available at <https://anonymous4blsp.github.io/arr/>

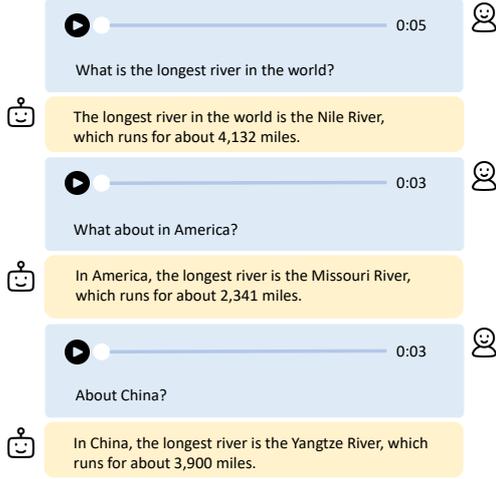


Figure 5: Speech conversation in English.

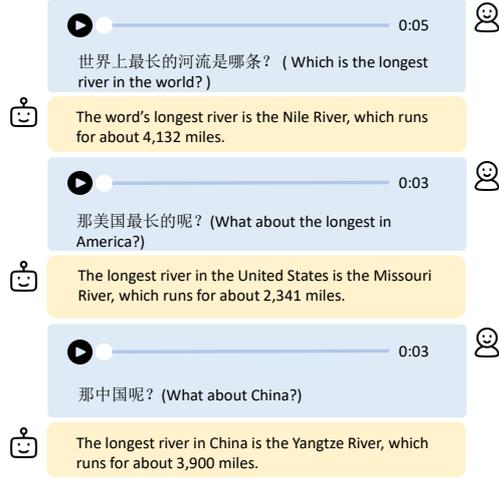


Figure 6: Speech conversation in Mandarin.

Method	BSTC		MSLT	
	zh-en	de-en	fr-en	
Text+LLM	16.1 / 58.7	32.8 / 84.2	29.6 / 76.3	
Whisper+LLM	11.1 / 54.3	25.3 / 79.4	24.1 / 71.5	
CTC+LLM	1.1 / 41.3	5.3 / 60.4	4.0 / 53.4	
BLSP	5.0 / 49.8	13.1 / 70.9	13.4 / 64.8	

Table 5: ST results (BLEU / COMET) in X-to-English directions.

528 pabilities, we evaluate the speech translation performance of our BLSP model in the Chinese (zh) to English (en) direction on BSTC (Zhang et al., 2021) and in the German (de) and French (fr) to English (en) directions on MSLT (Federmann and Lewis, 2016). As shown in Table 5, the BLSP model demonstrates reasonable multilingual translation competency for source languages that were not observed during behavior alignment training. We note that there is a significant gap in translation quality, as measured by both BLEU and COMET, when compared to Whisper+LLM and Text+LLM. This highlights the potential for further advancements in multilingual training. On the other hand, the cascaded model CTC+LLM, which was trained on English data, does not have cross-lingual capability.

545 As illustrated in Figure 6, our model is capable of engaging in multi-turn conversations with non-speech (Mandarin) speech input. It is worth mentioning that the model’s responses are always in English. This is a direct result of the English-only training procedure in BLSP, where the continuations are consistently in English. This observation also suggests that there is benefit in incorporating

553 multilingual training in behavior alignment for future research. 554

6 Related Works 555

556 Due to the lack of space, please see Appendix F for a discussion on related works. 557

7 Conclusion 558

559 In this paper, we introduce the BLSP approach, which bootstraps language-speech pre-training through behavior alignment. Our training procedure is straightforward, requiring only learning of a lightweight modality adapter through a novel utilization of speech recognition training data. As evidenced by quantitative evaluations in speech recognition, speech translation, spoken language understanding, and illustrated through multi-turn conversation demonstrations, BLSP effectively extends the remarkable language capabilities of LLMs to speech, enabling direct interaction with LLMs using speech input. BLSP represents a fresh and valuable perspective for achieving cross-modal alignment in LLMs, and there are numerous directions for expansion and improvement in future research. 574

Limitations 575

576 Although our BLSP approach can extend the remarkable language capabilities of LLMs to speech, as evidenced by quantitative evaluations and illustrative demonstrations, there are several limitations in our current study. 578 579 580

581 **Alignment Quality.** As indicated by the quantitative evaluations, there exists a substantial perfor- 582

mance gap when using speech input as opposed to the cascaded approach. Our approach to behavior alignment of continuation writing, in its current form, tends to align speech and text at a semantic level that restricts its capacity to capture detailed phonetic information. Exploring more fine-grained loss designs or approaches for constructing more fine-grained training data, including in combination with speech recognition, speech translation, or general speech instruction data, is worthy of further investigation.

Paralinguistic Information. In this study, we mainly focus on aligning speech and text in the semantic space, without addressing the paralinguistic aspects of spoken language that cannot be simply described by words, such as emotions, tones, and intentions. It is possible to capture and incorporate paralinguistic information with LLMs by leveraging data from more diverse speech-related tasks, such as speaker identification, keyword spotting, and speech emotion recognition.

Safety and Ethics. The use of continuous speech representations in our BLSP model could make it more susceptible to adversarial attacks and can potentially compromise the LLM’s established adherence to the HHN criteria (Harmless, Helpful, Honest). This is an area that is worthy of future research, both in identifying weaknesses and searching for solutions.

Broader Applicability. While our study has focused on the behavior alignment of continuation writing for speech-text alignment, the fundamental principles underlying this approach could have broader applicability. This involves expanding existing paired data in creative ways with the assistance of LLMs, ultimately benefiting LLMs. We leave it to future studies to extend this approach to diverse scenarios, including vision-language and multilingual alignments.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022.

[Flamingo: a visual language model for few-shot learning.](#) 632
633

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215. 634
635
636
637
638
639

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*. 640
641
642
643
644

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Giga-speech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*. 645
646
647
648
649
650

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.* 651
652
653
654
655
656

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*. 657
658
659
660
661
662

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. 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 2012–2017. Association for Computational Linguistics. 663
664
665
666
667
668
669
670

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 671
672
673
674
675
676
677

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*. 678
679
680
681
682

Christian Federmann and William Lewis. 2016. Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *Proceedings of the 13th International Conference on Spoken Language Translation*. 683
684
685
686
687

688	Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. <i>arXiv preprint arXiv:2304.15010</i> .	744
689		745
690		746
691		747
692		748
693	Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. <i>arXiv preprint arXiv:2305.10790</i> .	749
694		750
695		751
696	François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In <i>Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20</i> , pages 198–208. Springer.	752
697		753
698		754
699		755
700		756
701		757
702		758
703	Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In <i>International Conference on Machine Learning</i> , pages 2790–2799. PMLR.	759
704		760
705		761
706		762
707		763
708		764
709	Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. <i>arXiv preprint arXiv:2106.09685</i> .	765
710		766
711		767
712		768
713		769
714	Rongjie Huang, Mingze Li, Dongchao Yang, Jia-tong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. <i>arXiv preprint arXiv:2304.12995</i> .	770
715		771
716		772
717		773
718		774
719		775
720	Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. <i>arXiv preprint arXiv:2308.12792</i> .	776
721		777
722		778
723		779
724		780
725	Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. <i>arXiv preprint arXiv:2301.12597</i> .	781
726		782
727		783
728		784
729	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. <i>arXiv preprint arXiv:2304.08485</i> .	785
730		786
731		787
732	Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. <i>arXiv preprint arXiv:1904.03670</i> .	788
733		789
734		790
735		791
736		792
737	Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. <i>arXiv preprint arXiv:2305.15023</i> .	793
738		794
739		795
740		796
741		797
742	R OpenAI. 2023. Gpt-4 technical report. <i>arXiv</i> , pages 2303–08774.	798
743		799
		800
	Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In <i>2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)</i> , pages 5206–5210. IEEE.	
	Matt Post. 2018. A call for clarity in reporting BLEU scores. In <i>Proceedings of the Third Conference on Machine Translation: Research Papers</i> , pages 186–191, Belgium, Brussels. Association for Computational Linguistics.	
	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In <i>International conference on machine learning</i> , pages 8748–8763. PMLR.	
	Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. <i>arxiv. arXiv preprint arXiv:2212.04356</i> .	
	Alaa Saade, Joseph Dureau, David Leroy, Francesco Caltagirone, Alice Coucke, Adrien Ball, Clément Doumouro, Thibaut Lavril, Alexandre Caulier, Théodore Bluche, et al. 2019. Spoken language understanding on the edge. In <i>2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMCC2-NIPS)</i> , pages 57–61. IEEE.	
	Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugging-gpt: Solving ai tasks with chatgpt and its friends in huggingface. <i>arXiv preprint arXiv:2303.17580</i> .	
	Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In <i>ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)</i> , pages 7927–7931. IEEE.	
	Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. 2023. Llaslm: Large language and speech model.	
	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca .	
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	
	Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. <i>arXiv preprint arXiv:2007.10310</i> .	

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. *E-chat: Emotion-sensitive spoken dialogue system with large language models*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset. *arXiv preprint arXiv:2104.03575*.

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

A Implementation Details of ASR Pretraining

In the pre-experiments, the model architecture and training dataset used are the same as in BLSP. The only difference is that in BLSP, the modality adapter is trained using the continuation task, while in the pre-experiments, the modality adapter is trained to predict the ground-truth transcript. Additionally, similar to the approach in SpeechGPT (Zhang et al., 2023), we utilize GPT-4 to generate 100 distinct text instructions for prompting ASR tasks. These instructions are concatenated before the speech input. For visualization purposes, we construct cross-modal prompts to extract features using the instructions shown in Table 6. We then apply t-SNE dimensionality reduction mapping to all samples from the LibriSpeech test set.

CW: Please continue the following sentence.

SA: Please classify the emotional tone of the following text.

SR: Please transcribe the following audio into English text.

ST: Please translate the following English text into German text.

Table 6: Instructions used for extracting cross-modal representations.

To further demonstrate the overfitting problem in the ASR task, as illustrated in Figure 1, we present

the average cosine similarity between the learned representations of the same input across different task instructions in Table 7. Notably, the representations for speech input are remarkably similar regardless of the task instruction used, indicating a deficiency in following instructions. Additionally, Table 8 highlights consistently low similarity scores between paired speech and text input representations under the same task instructions, suggesting a lack of alignment between the representations of speech and text inputs.

	CW-S	SA-S	SR-S	ST-S
CW-S	1.000	0.997	0.997	0.991
SA-S	0.997	1.000	0.997	0.992
SR-S	0.997	0.997	1.000	0.993
ST-S	0.991	0.992	0.993	1.000

Table 7: Average similarity between representations of the same speech inputs under different task instructions learned from ASR task.

	CW	SA	SR	ST
	0.270	0.106	0.328	0.176

Table 8: Average similarity between representations of paired speech/text inputs under the same task instructions learned from ASR task.

B Detailed Results for Zero-Shot Speech-to-Text Tasks

We present the performance of speech translation in each direction, as illustrated in Tables 9 and Table 10. For CoVoST-2, we evaluate our method on five translation directions: English (en) to Catalan (ca), German (de), Indonesian (id), Slovenian (sl), and Swedish (sv). Additionally, we conduct experiments on MUST-C for all eight translation directions: English (en) to Catalan (ca), German (de), Indonesian (id), Slovenian (sl), and Swedish (sv). The instructions used for each speech-to-text generation task are presented in Table 11.

C The Data Construction and Evaluation Process of General-Purpose QA

For our evaluation on the general-purpose question answering task, we selected 1460 speech-text pairs from the GigaSpeech test set. The selected texts contain 40-60 words to ensure that the samples encompass relatively complete semantics. To

Method	en-ca	en-de	en-id	en-sl	en-sv
Text+LLM	24.8	21.9	22.0	12.9	27.8
Whisper+LLM	19.2	17.3	15.8	10.0	22.2
CTC+LLM	14.5	14.0	12.3	7.9	17.5
BLSP	13.9	14.0	12.1	7.0	16.6
+ RP	14.8	14.9	12.3	7.9	19.3

Table 9: ST results on in-domain dataset CoVoST-2.

Method	en-de	en-es	en-fr	en-it	en-nl	en-pt	en-ro	en-ru
Text+LLM	20.2	21.7	24.2	20.5	24.4	20.3	13.3	13.3
ASR+LLM	16.9	19.2	20.5	16.4	20.7	17.1	10.9	11.4
CTC+LLM	13.6	15.3	15.7	12.6	17.0	13.9	8.2	9.9
BLSP	13.0	14.7	14.4	11.6	17.0	12.6	6.2	9.2
+RP	14.2	16.4	17.5	13.2	17.5	16.1	9.2	10.2

Table 10: ST results on out-of-domain dataset MUST-C.

ASR: Please repeat the following words.

ST: Please translate the following English text into <target> text.

SNIPS: Please classify the intent of the text, choose from [DecreaseBrightness, IncreaseBrightness, SetLightBrightness, SetLightColor, SwitchLightOff, SwitchLightOn].

FSC: Please classify the intent of the text, choose from [bring newspaper, deactivate lamp, change language English, deactivate music, increase heat, change language Korean, change language none, bring shoes, change language German, activate lights, bring socks, change language Chinese, decrease heat, decrease volume, increase volume, activate music, activate lamp, bring juice].

SLUE-VoxCeleb: Please classify the emotional tone of the text as either positive, negative, or neutral.

Table 11: Instructions used for speech-to-text generation tasks.

formulate questions based on these transcripts, we utilized ChatGPT. As shown in Listing 1, we provided ChatGPT with the transcript as the input, and the task for ChatGPT was to generate a suitable question based on the given text input.

In the next step, we used different models to generate responses to the questions posed by ChatGPT in the previous step. We then employed ChatGPT again to evaluate the acceptability of these generated answers. As shown in Listing 2, we provided ChatGPT with the question, the ground-truth transcript, and the answer generated by different models. The task for ChatGPT was to evaluate whether the generated answer is acceptable.

D Quantitative Analysis of Representations from BLSP

As depicted in Table 12, the representations of speech inputs learned from BLSP are distinct under various task instructions, unlike in Table 7 for ASR task. Table 13 illustrates the average cosine similarity between representations of paired speech and text inputs learned from BLSP, revealing a high level of similarity between the two, as opposed to the low similarity depicted in Table 8 for ASR task. We want to point out that there remains a notable gap between the representations for speech and text inputs that is worthy of future research.

	CW-S	SA-S	SR-S	ST-S
CW-S	1.000	0.494	0.745	0.381
SA-S	0.494	1.000	0.501	0.278
SR-S	0.745	0.501	1.000	0.477
ST-S	0.381	0.278	0.477	1.000

Table 12: Average similarity between representations of the same speech inputs under different task instructions learned from BSLP.

CW	SA	SR	ST
0.785	0.866	0.808	0.900

Table 13: Average similarity between representations of paired speech/text inputs under the same task instructions learned from BLSP.

E Selected Examples of Cross-Modal Conversation

As demonstrated in Figure 7, BLSP provides expanded mechanisms to interact with LLMs. Users can freely switch between text and speech inputs, and directly employ speech instructions to carry out speech-to-text tasks.

F Related Works

Multi-Modal Large Language Models Current multi-modal large language models have been prominently focusing more on visual modality (OpenAI, 2023; Yin et al., 2023). These models utilize a pre-trained visual encoder to extract key visual features from images, which are then combined with text inputs to generate relevant outputs. PaLM-E (Driess et al., 2023) combines the huge 540B PaLM (Chowdhery et al., 2022) and the 22B Vision Transformer (ViT) (Dosovitskiy et al., 2020) to create the largest vision-language model currently reported. Since it would be costly to train a large multi-modal model in an end-to-end manner, many works introduce a learnable interface between the pre-trained visual encoder and LLM to connect different modalities while freezing the parameters of the pre-trained models. Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023) and X-LLM (Chen et al., 2023) leverage a group of learnable query tokens to extract information in a query-based manner. LLaVa (Liu et al., 2023) connects the pre-trained CLIP (Radford et al., 2021) encoder and Vicuna (Chiang et al., 2023) with a simple projection layer. LLaMA-Adapter (Gao et al., 2023) and LaVIN (Luo et al., 2023) explore a parameter-efficient tuning manner, introducing a lightweight adapter module during training. Recent research has extended the above-mentioned approach to “audio” (Gong et al., 2023), which refers to natural sound, such as thunder and chirp. However, there is still a lack of exploration when it comes to human speech.

Interact with LLMs through Speech After the introduction of ChatGPT, several studies have focused on combining specialized speech models with LLMs, allowing for speech interaction with these language models. Initial endeavors in this field (e.g., HuggingGPT (Shen et al., 2023), AudioGPT (Huang et al., 2023)) employed a cascading model structure, linking LLMs with additional ASR and TTS models to enable speech input and

output. These models showcase heightened intricacy, require substantial resources, and are susceptible to the inevitable issue of error accumulation. Recent works have started to explore end-to-end model architectures. SpeechGPT (Zhang et al., 2023) takes the discretized output of a speech model in self-supervised training and treats it as a specialized linguistic unit, training it alongside a large language model. However, due to the high sampling frequency of the discrete unit, it is difficult for this method to achieve multiple rounds of dialogue. LLaSM (Shu et al., 2023) has constructed an extensive speech instruction dataset intended for training the modal adapter to attain modality alignment. Their methodology is predominantly data-driven, with a lesser emphasis on the explicit design of modality alignment.

Listing 1: The prompt used to generate general-purpose QA data.

```
Please ask a question about the input and then answer the question based on the
input. The output format should be in json and contains question and the response.
Example:
input: ah yeah good day and welcome to this instructional video on how to ah wash
your car um with a baby . basically , you just ask them to do it . you know they
love this kind of stuff this bubbles and a brush and .
output: {"question": "What is the video about?", "answer": "the video is about how to
wash your car um with a baby."}
input: it is the gibraltar strait where you lost control and then you dived down ...
one of those cases where you let the wings go in the clouds but you lose
orientation completely
output: {"question": "Where did the incident occur?", "answer": "Gibraltar Strait."}
BEGIN:
input: ${transcript}
output:
```

Listing 2: The prompt used to evaluate whether a response is acceptable.

```
Given a question, related input, and answer, please help determine whether the
answer is acceptable.
The output choose from acceptable or unacceptable.
Question: ${question}
Input: ${transcript}
Answer: ${answer}
Your output:
```

Hello!

Hello! How can I help you?

0:03

I am going to Beijing, can you give me some suggestions?

Sure, here are some suggestions for your trip to Beijing:

- Visit the Forbidden City and explore the ancient palace complex.
- Take a stroll along the Wangfujing Street for some shopping and snacks.
- Visit the Tianmen Square and take a picture of the Monument of the People's Heroes.
- Take a rickshaw ride in the hutongs and experience the local life.
- Enjoy a Peking Duck dinner at one of the local restaurants.
- Take a day trip to the Great Wall of China and hike along the wall.
- Visit the Summer Palace and admire the beautiful gardens.

0:01

Tell me about alpacas.

Alpaca are South American camelids that are known for their soft, luxurious wool. They are a domesticated species and have been bred for their fiber since pre-Incan times. Alpacas are very social animals and live in herds of up to 15 individuals. They are an important source of income for many Peruvian and Bolivian farmers.

0:05

Translate the sentence 'I have no mouth but I must scream' into Spanish.

No tengo ninguna boca, pero debo gritar.

0:03

List all Canadian provinces in alphabetical order.

Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, Saskatchewan.

Which province has the largest area?

The province with the largest area is Saskatchewan.

Figure 7: Selected examples of cross-modal conversation using BLSP model.