APPENDIX

6 RESULTS FOR MULTIMODAL CONNECTIONS

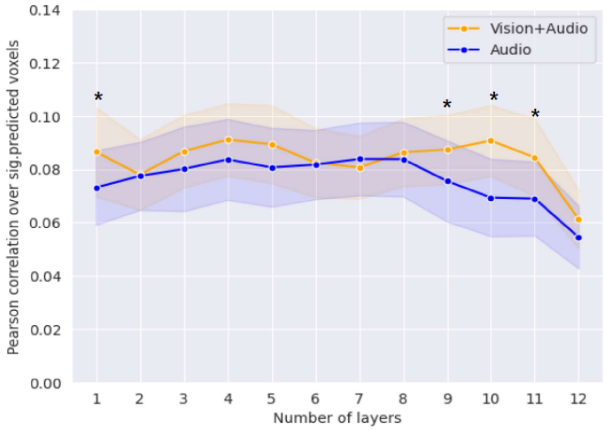

Figure 6: Pearson correlation of brain alignment over the significantly predicted voxels averaging over the language regions for all 12 layers of the joint encoder. Vision-language representations significantly improve the brain alignment over language-only representations at layer 1, 9-11.

7 RESULTS FOR BRAIN PLOTS

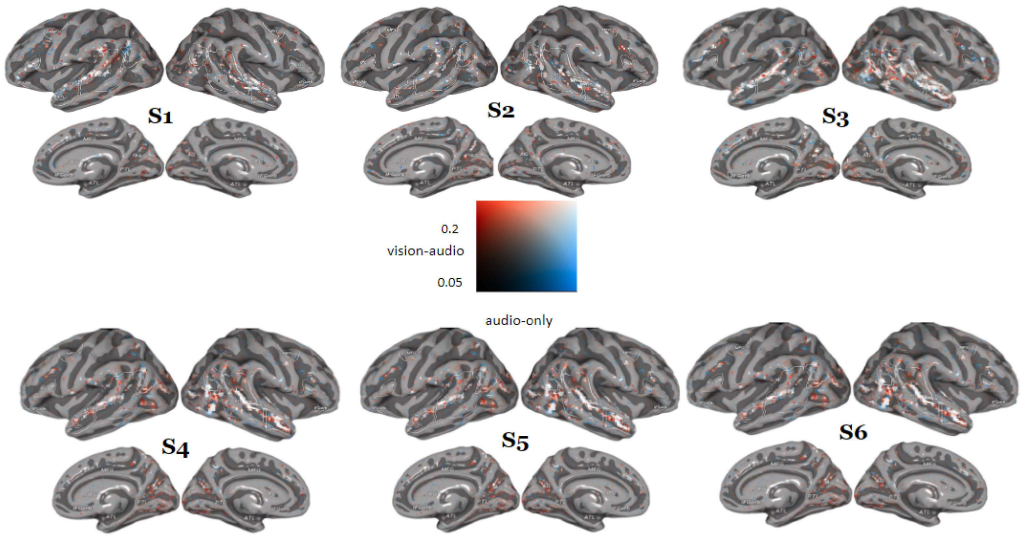

Figure 7: The contrast of voxel-based brain alignment between vision-language (audio) models and language (audio)-only models for six subjects at the representative layer 10. Voxels in red are better predicted by vision-language models than the language-only models. Voxels in blue are better predicted by language-only models than vision-language models. Vision-language representations improve alignment over language-only representations with language regions across six subjects.

8 MASKED LANGUAGE MODELING DETAILS

We extract the last-layer representation from the audio encoder implemented in MERLOT Reserve to obtain the true encoding of the MASK tokens. This is consistent with the pre-trained objective of the model, which must align the masked predictions with the independent encoding of the text or audio.

9 TVQA FINE-TUNING DETAILS

The entire TVQA dataset consists of 152,545 QA pairs from 21,793 video clips of popular American TV shows. Each question in the TVQA dataset has five options, only one of which is the correct answer. TVQA also provides annotations for a rough time region where the content of the question occurs (around 10 seconds). For the purpose of fine-tuning, 35 seconds of video centered around the provided time region are extracted for each question, in order to capture the full context for understanding what is happening. The model contextualizes five audio sequences, five text sequences as well as video frames for a given question in a joint transformer. Each audio/text sequence contains a question, an option (from five candidates), and a masked text (or audio) token followed by subtitles (or audio). The hidden representations of each candidate option are extracted from ten masked tokens. The model then scores the representations for each option through a linear projection layer and selects the option with the highest probability among each of the five audio sequences and the five text sequences. The joint prediction of audio and text is calculated based on the average softmax of two predictions, one from audio sequences, and another from text sequences. Finally, the model is optimized with softmax-cross entropy of the loss from the difference between the joint prediction and the ground truth. To extract the representations for our paper, we provide the model with five audio sequences, together with video frames for a given question in a joint transformer. We then pool the representations of the MASK token from the sequence that corresponds to the correct choice option. All questions are chosen from the validation set of the TVQA dataset.