



GALAXY: A Large-Scale Open-Domain Dataset for Multimodal Learning

Yihan Wu¹, Yichen Lu², Yijing Chen¹, Jiaqi Song², William Chen², Ruihua Song¹, Shinji Watanabe²

¹Gaoling Artificial Intelligence, Renmin University of China, China

²Carnegie Mellon University, USA

{yihanwu, rsong}@ruc.edu.cn, shinjiw@ieee.org

Abstract

Humans naturally use multimodal information, with vision, speech, and text working together to understand the world and solve problems. For artificial intelligence to achieve human-level capability, it must process multimodal information in a similar manner. However, there is a lack of large-scale open-domain datasets that support all three modalities—vision, speech, and text—with high-quality speech transcriptions. To address this gap, we introduce GALAXY¹, a large-scale, open-domain dataset designed for multimodal learning, containing 8,270 hours of videos, speech, and transcriptions across 16 diverse domains. We describe the data creation pipeline and provide detailed statistics and analyses of the dataset. Using multimodal speech recognition as a case study, we validate GALAXY’s effectiveness and evaluate baseline models’ performance across different data volumes and domains. The results highlight GALAXY’s potential as a valuable resource for advancing multimodal understanding.

Index Terms: speech recognition, multimodal understanding

1. Introduction

Humans inherently rely on multimodal information – vision, audio, and text – to solve problems collaboratively. In a similar manner, artificial intelligence systems require cross-modal or multimodal information to effectively perform tasks. Given that the performance of multimodal models is fundamentally constrained by the quality of the data and annotations used for training, there is a critical need for a high-quality, large-scale dataset containing visual, speech, and text modalities to facilitate the advancement of multimodal model development.

Many previous multimodal datasets focus on vision-text modalities [1–6] or speech-text modalities [7–11], with few providing high-quality, large-scale data across all three modalities—vision, speech, and text—simultaneously. While some datasets do include three modalities, they either focus on restricted domains or lack proper quality control. For instance, LRS2 [12] and LRS3-TED [13] focus on a restricted lip-reading scenario, which is not a wide setup. How2 [14] contains 300 hours of videos with speech and corresponding subtitles, but the videos are primarily sourced from the instructional domain, limiting the generalization of models. Ego4D [15] spans diverse scenarios but focuses solely on egocentric videos. HowTo100M [16] and YT-Temporal-180M [17] contain large amounts of multimodal data, but their text is derived by ASR models without any filtering or quality control.

To address these limitations, this paper presents GALAXY, a large-scale, open-domain multimodal dataset encompassing

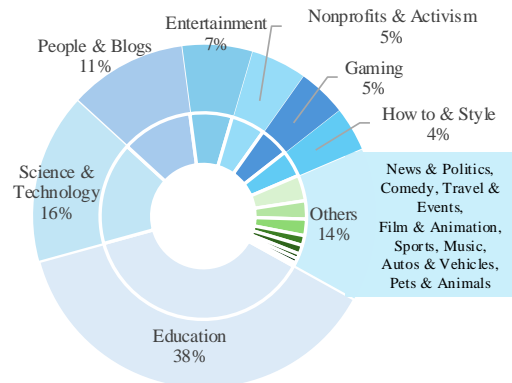


Figure 1: **Video domains in GALAXY dataset.** The outer circle displays the 7 most common scenarios (86% of the data). The inner portion highlights the remaining 14% of scenarios.

vision, speech, and text modalities. GALAXY is designed to meet the following key characteristics:

- **Multimodal.** GALAXY consists of videos paired with speech and transcriptions. The naturally paired multimodal information in GALAXY makes it an ideal resource for advancing multimodal development.
- **Large-scale.** The GALAXY dataset contains 8,270 hours of videos with more than 11M utterances, making it appropriate for both pretraining and downstream tasks.
- **Open-domain.** GALAXY covers a wide range of video domains from the real world, providing significant diversity to encourage models to learn from diverse scenarios.
- **High-quality.** GALAXY applies strict filtering to ensure a high correlation between speech and text, retaining only videos where the speech and text content align. This process minimizes the impact of data noise on model training.

To demonstrate GALAXY’s potential for multimodal understanding, we begin with multimodal automatic speech recognition (multimodal ASR) and plan to expand to various multimodal learning tasks, enhancing its utility. Multimodal ASR can be seen as a generalized audiovisual ASR task, where unconstrained visual information can be used to improve speech recognition accuracy. We choose multimodal ASR for two reasons. (1) It requires the model to understand both visual and speech inputs, necessitating the direct integration of these modalities. (2) Multimodal ASR involves learning the interleaved interactions between visual and speech modalities, making it a critical and foundational task for advancing multimodal learning, and serving as a basis for more complex tasks. Additionally, we construct a test set based on the GALAXY for the

¹<https://github.com/wyh2000/GALAXY>

Table 1: *Statistics of GALAXY dataset and its comparison with existing multimodal datasets.*

Dataset	Domain	# Videos	Hours (h)	Trans. quality control	Avg. Len(sec)	# Clips / Utterances	Per Clip Statistics
How2	Instruction	13,662	308	✓(Manual)	81.2	191,297	5.8 s & 20 words
HowTo100M	Instruction	1,221,000	134,472	✗(ASR only)	396.5	136,000,000	3.5 s
Ego4D	Ego	–	3,670	✓(Manual caption)	522.0	–	–
YT-Temporal-180M	Open	6,000,000	–	✗(ASR only)	–	180,000,000	–
GALAXY	Open	48,214	8,270	✓(ASR + filtering)	1216.8	11,807,029	4.8 s & 12 words

multimodal ASR task evaluation across different domains.

2. Related works

2.1. Multimodal datasets

With the development of multimodal learning, many large-scale datasets have been proposed to support model training [1,3,18]. Despite the rich content of these datasets, they primarily focus on visual-text understanding, neglecting the essential speech modality for multimodal understanding. Some datasets, like How2 [14] and Ego4D [15], include video, speech, and text. However, their videos are limited to instructional and ego-centric domains, restricting the generalizability of the models. HowTo100M [16] and YT-Temporal-180M [17] scale to 1M and 6M videos, respectively, but their text data are generated by ASR systems without quality control or filtering, potentially introducing noise during model training. To address these limitations, we propose the GALAXY dataset, a large-scale, open-domain dataset with high-quality text transcriptions. As shown in Table 1, GALAXY spans a broad range of domains, containing videos from 16 diverse categories, including education, entertainment, games, and more, as illustrated in Figure 1.

2.2. Multimodal speech recognition

Multimodal speech recognition [14, 19] aims to use visual and speech information as input, leveraging visual cues to enhance speech recognition accuracy. Unlike audiovisual ASR [20–22], which focuses solely on lip reading, multimodal ASR utilizes unrestricted visual information. This shows significant potential in scenarios like YouTube videos, online meetings, and live broadcasts, where visual data provides valuable contextual and content clues for enhancing speech recognition accuracy. Multimodal ASR is also a crucial task in multimodal learning [23,24]. It forms the foundation for understanding multimodal input, enabling more complex, cognition-aware tasks. Currently, multimodal ASR models can be classified into two main architectures: encoder-decoder-based and language model-based. Methods such as EVA [25], AVATAR [26], and AVFormer [19] use encoder-decoder architectures to incorporate visual information. Besides, some multimodal LLMs [23,24,27,28] leverage the capabilities of pre-trained large language models by incorporating speech and visual encoders to enable multimodal ASR. In this work, we establish baselines using both encoder-decoder and language model architectures to evaluate the performance of the GALAXY dataset with these two architectures.

3. GALAXY dataset

The GALAXY dataset consists of 48,214 videos, totaling 8,270 hours, along with corresponding speech and transcriptions. In this section, we will describe the dataset creation pipeline, which enables the construction of a large-scale, high-quality,

multimodal, open-domain dataset.²

3.1. Data collection

To construct the GALAXY dataset based on real-world videos across diverse video domains, we collect data from the YouTube platform. Specifically, we build GALAXY upon the YODAS dataset [8], a large-scale, multilingual speech dataset sourced from YouTube. In this work, we focus on the English subset of YODAS and directly utilize the speech and corresponding text from the dataset. We re-download the corresponding videos, and gather various types of metadata, including titles, video categories, video descriptions, like counts, view counts, and more. This data can be further used for multimodal understanding tasks, such as multimodal question answering, multimodal captioning, and multimodal dialogue.

3.2. Data filtering

As the collected raw videos contain considerable noise in speech-text alignment, directly using the low-quality data would degrade the performance of subsequent models. To address this issue, we filter the data based on speech-text alignment. Following YODAS [8], we use a pre-trained acoustic model [29] to assess speech-text alignment, retaining high alignment data only. The score is derived from the CTC loss [30], where a lower loss value indicates better alignment [31,32]. We first filter noisy videos at the video level using a threshold score of 3.0, then further refine the data by applying a sentence-level alignment threshold of 2.0. This process results in a high-quality dataset with well-aligned speech-text pairs. We solely use speech-text alignment for quality control in the training set and further incorporate video-text alignment for the high-quality test set, as described in Section 3.4.

3.3. Data statistics

We present key statistics of the GALAXY dataset in comparison with other popular multimodal vision-speech-text datasets in Table 1. Unlike How2 [14], HowTo100M [16], and Ego4D [15], GALAXY contains open-domain videos, ensuring greater diversity. Figure 1 illustrates the broad distribution of video domains within GALAXY, which includes real-world videos across 16 different domains, such as education³, entertainment, news, and more. In terms of video duration, each video has an average length of 1216.8 seconds, which is significantly longer than those in previous datasets. This highlights GALAXY’s potential for long-form multimodal understanding tasks. To support model training, we segment video-level data into utterances. Figure 2 shows the distribution of video and utterance lengths respectively. Both video length and utterance length exhibit long-tail distributions, with peaks around 3-5 minutes for

²We will release the dataset upon acceptance.

³The education domain is the largest because YODAS data only includes creative commons content.

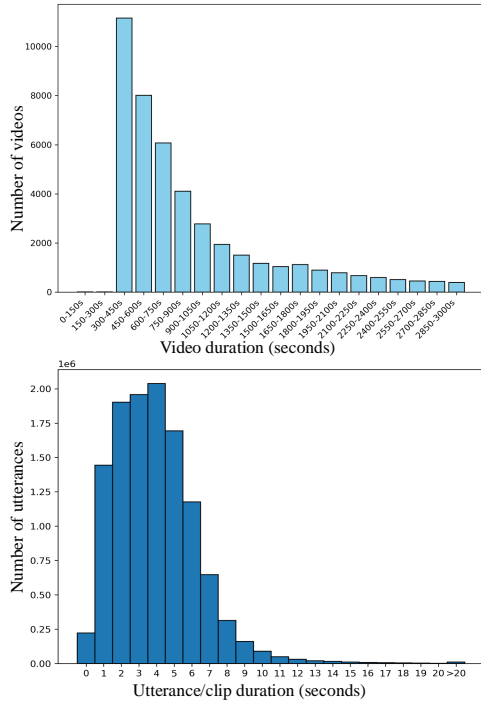


Figure 2: **Duration statistics.** Both video duration and utterance duration distribution show long-tail characteristics, with the highest concentration around 300–900 seconds for video length and 2–4 seconds for utterance length.

video length and 2–4 seconds for segment length.

3.4. Test set creation

To evaluate the performance of multimodal ASR in real-world scenarios across diverse domains, we split a test set from the GALAXY as a new evaluation benchmark. Our test set creation pipeline is driven by two objectives: (1) to include videos from diverse domains, and (2) to evaluate the contribution of the visual modality to speech recognition by selecting video segments with high visual-text correlation. Therefore, our pipeline consists of the following steps:

- **Step 1: Split subsets based on visual-text correlation.** To highlight challenging utterances where the visual modality can compensate for corrupted audio, we prioritize videos with a high visual-text correlation. This ensures that the visual information can effectively provide content clues in speech recognition, particularly in cases where the audio is noisy or ambiguous. More specifically, we use ImageBind [33] to compute the visual-text correlation score and retain only utterances with a score greater than 18.0.
- **Step 2: Balance videos from different domains.** To evaluate the performance across different video domains, we carefully constructed the test set by sampling data from 16 diverse domains based on Figure 1. This approach ensures a balanced representation of various content types, resulting in a total of 782 utterances. By constructing GALAXY test set, we aim to assess the generalizability and robustness of the multimodal ASR system across different real-world scenarios.

In summary, we create a high correlation test set across diverse video domains. As shown in Figure 3, the GALAXY test set has a much higher visual-text correlation than Ego4D [15]. When compared to the manually constructed VisSpeech dataset, GALAXY’s test set not only shows a similar average visual-text

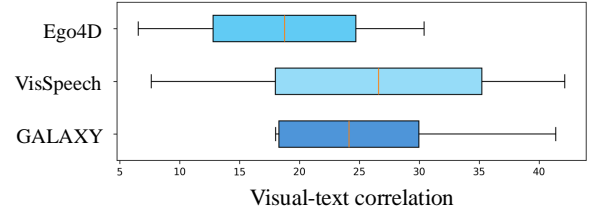


Figure 3: **Visual-text correlation of different test sets.** The GALAXY test set shows higher similarity, with a higher median visual-text correlation than the VisSpeech and Ego4D test set.

correlation but also exhibits a higher minimum value and comparable maximum correlation, making it a more comprehensive evaluation benchmark due to its broader range of video domains. However, VisSpeech only includes instructional videos, while the GALAXY test set covers multiple video domains, providing a more comprehensive evaluation.

4. Experimental settings

GALAXY provides a versatile resource for various tasks across vision, speech, and text modalities. In this work, we specifically focus on multimodal ASR tasks, exploring both encoder-decoder-based and language model-based architectures.

4.1. Models

- **OWSM-visual.** We use OWSM-visual [34] as the encoder-decoder-based model. OWSM-visual is built upon OWSM v3.1 [35], an open-source pre-trained speech recognition model that achieves robust performance on standard ASR benchmarks⁴. OWSM-visual uses CLIP-Large [36] as the visual encoder to process visual information, which is mapped to the speech token space via a linear layer. The visual and speech tokens are then concatenated and input into the model. We load the OWSM v3.1 pretrained weights and fine-tune the model with multimodal data. Unless otherwise specified, we fine-tune the model for 20 epochs with a batch size of 8.
- **VITA-1.5.** We use VITA-1.5 [27] as the language model-based method. VITA-1.5 is a multimodal model similar to GPT-4o, designed for complex video understanding by simultaneously processing vision and speech modality. VITA-1.5 incorporates InternViT-300M [18] as its visual encoder and 24 Transformer blocks as its speech encoder. A pre-trained Qwen2-7B [37] served as the language model backbone to receive multimodal information and generate appropriate predictions. We load the pretrained weights, freeze its vision and speech encoders, and fine-tune the language model and the speech encoder using LoRA with a LoRA rank of 64, and an alpha value of 16 for 2 epochs with a batch size of 8.

4.2. Evaluation

To build a comprehensive baseline, besides the GALAXY test set (as mentioned in Section 3.4), we employ other widely-used multimodal ASR test sets, including How2 test set [14], VisSpeech [26], and Ego4D audiovisual diarization test set [15]. We use word error rate (WER) as the evaluation metric for all experiments, with lower values indicating better performance.

⁴We opt for OWSM over Whisper due to potential data contamination concerns. Whisper’s training data, sourced through web crawling, might include our test set.

Table 2: **Results on different training sets.** WER is used as the evaluation metric, with lower values indicating better performance. * and † denote the best results with 180K training utterances for OWSM-visual and VITA-1.5, respectively, while bold values represent the overall best results.

Model	Training set	Domain	# Utterances	Test sets			
				GALAXY	How2	VisSpeech	Ego4D
OWSM	GALAXY (speech)	Open	180K	17.5	12.7	15.2	58.3
OWSM-visual	How2	Instruction	180K	28.3	9.2*	14.5	56.5
OWSM-visual	GALAXY	Open	180K	16.4*	11.9	14.0*	53.5*
OWSM-visual	GALAXY	Open	9M	15.9	11.1	13.9	52.4
VITA-1.5	How2	Instruction	180K	40.7	22.5	44.3	118.5
VITA-1.5	GALAXY	Open	180K	21.0†	20.4†	26.5†	117.2†

Table 3: **Comparison of OWSM-Visual performance across different domains.** We evaluate specific domains and open domains, using the same training data size (180K utterances) for all settings. Bold values represent the best results, and † denotes the second-best results.

Domain	GALAXY	How2	VisSpeech	Ego4D
Instruction	17.2†	11.5	14.4†	57.5
Education	17.3	11.5	14.9	67.6
Entertainment	17.4	13.2	14.9	52.8
Open	16.4	11.9†	14.0	53.5†

5. Results

5.1. Main results

As shown in Table 2, we use How2 and GALAXY as training sets to build baseline models, and evaluate them on four test sets respectively. Our observations are as follows. First, the OWSM-visual trained on the 9M GALAXY dataset outperforms the model trained on the 180K How2 dataset across most datasets, including the open-domain GALAXY and Ego4D test sets and out-domain VisSpeech test set. This improvement can be attributed to the richer data and greater domain diversity in GALAXY. Second, we evenly sample data from different domains within GALAXY, creating a subset with the same total data size (180K) as How2. Even with the same training data size, OWSM-visual trained on GALAXY achieves better performance on GALAXY, VisSpeech, and Ego4D test sets. It highlights the importance of domain diversity in training data. Third, the consistency of results on the GALAXY test set and the manually constructed VisSpeech test set demonstrates the effectiveness of our automatically constructed GALAXY test set. Finally, it is noteworthy that the performance of VITA-1.5 is inferior to that of OWSM-visual, which we attribute to the usage of the LoRA fine-tuning strategy. Besides, language model-based models typically require larger training datasets and more diverse training tasks, which we will explore in future work.

5.2. Effectiveness of data scale

To validate the importance of dataset size, we train OWSM-visual with training data of varying sizes, ranging from 10K to 9M utterances. The data are evenly sampled from 16 categories to eliminate domain-specific biases. As shown in Figure 4, the most significant reduction in WER occurs when the data size increases from 10K to 50K utterances. The results consistently show that increasing the number of utterances improves accuracy, highlighting the significance of large-scale training data in enhancing model performance.

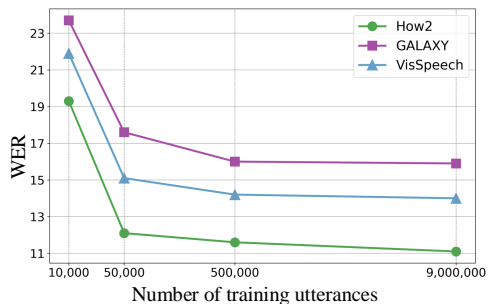


Figure 4: **WER on different training data size.** OWSM-visual’s performance on varying training data sizes from GALAXY, evaluated on GALAXY, How2, and VisSpeech test sets.

5.3. Effectiveness of data domain

To evaluate the impact of video domain diversity in training data on model generalization, we conduct experiments using both specific-domain and open-domain training subsets. Specifically, we sample 180K utterances from specific domains, including education, entertainment, and instruction. Also, we evenly sample 180K utterances from 16 domains as an open-domain training set. As shown in Table 3, we evaluate the OWSM-visual on test sets from various domains. The model trained on the entertainment domain performs best only on the Ego4D test set, with poor results on other datasets, likely due to the large proportion of entertainment data in the Ego4D test set. Similarly, models trained on the education and instruction domains perform well only on How2 but exhibit limited generalization to other domain test sets. With the same total training data size, training on open-domain data leads to more balanced performance across all test sets. It achieves the highest recognition accuracy on GALAXY and VisSpeech, and the second-best on How2 and Ego4D. This demonstrates the benefit of domain diversity in improving model generalization.

6. Conclusion

In this work, we introduce GALAXY, a large-scale multimodal dataset encompassing videos from diverse domains. We outline the data construction pipeline and provide preliminary analyses. Additionally, we provide baseline models based on this dataset on multimodal speech recognition, validating the effectiveness of GALAXY dataset. In the future, we will extend the GALAXY dataset to support multilingual and multi-task learning. We believe the GALAXY dataset can provide strong data support for several multimodal understanding tasks, including multimodal question answering, multimodal summarization, and beyond.

7. References

- [1] T. Chen *et al.*, “Panda-70m: Captioning 70m videos with multiple cross-modality teachers,” in *CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 13 320–13 331. [Online]. Available: <https://doi.org/10.1109/CVPR52733.2024.01265>
- [2] Y. Wang, *et al.*, “Internvid: A large-scale video-text dataset for multimodal understanding and generation,” in *ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=MLBdiWu4Fw>
- [3] H. Xue *et al.*, “Advancing high-resolution video-language representation with large-scale video transcriptions,” in *CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 5026–5035. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.00498>
- [4] Y. Sun *et al.*, “Long-form video-language pre-training with multi-modal temporal contrastive learning,” in *NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- [5] L. Chen *et al.*, “Sharegpt4video: Improving video understanding and generation with better captions,” in *NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- [6] R. Krishna *et al.*, “Dense-captioning events in videos,” in *ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 706–715. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.83>
- [7] J. Kahn *et al.*, “Libri-light: A benchmark for ASR with limited or no supervision,” in *ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 7669–7673. [Online]. Available: <https://doi.org/10.1109/ICASSP40776.2020.9052942>
- [8] X. Li *et al.*, “Yodas: Youtube-oriented dataset for audio and speech,” in *ASRU 2023, Taipei, Taiwan, December 16-20, 2023*. IEEE, 2023, pp. 1–8. [Online]. Available: <https://doi.org/10.1109/ASRU57964.2023.10389689>
- [9] V. Pratap *et al.*, “MLS: A large-scale multilingual dataset for speech research,” in *Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*. ISCA, 2020, pp. 2757–2761.
- [10] V. Pratap, A. Tjandra *et al.*, “Scaling speech technology to 1,000+ languages,” *J. Mach. Learn. Res.*, vol. 25, pp. 97:1–97:52, 2024.
- [11] R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *LREC 2020, Marseille, France, May 11-16, 2020*. European Language Resources Association, 2020, pp. 4218–4222.
- [12] J. S. Chung *et al.*, “Lip reading sentences in the wild,” in *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 3444–3453. [Online]. Available: <https://doi.org/10.1109/CVPR.2017.367>
- [13] T. Afouras *et al.*, “LRS3-TED: a large-scale dataset for visual speech recognition,” *CoRR*, vol. abs/1809.00496, 2018. [Online]. Available: <http://arxiv.org/abs/1809.00496>
- [14] R. Sanabria *et al.*, “How2: A large-scale dataset for multimodal language understanding,” *CoRR*, vol. abs/1811.00347, 2018. [Online]. Available: <http://arxiv.org/abs/1811.00347>
- [15] K. Grauman *et al.*, “Ego4d: Around the world in 3,000 hours of egocentric video,” in *CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 18 973–18 990. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01842>
- [16] A. Miech *et al.*, “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips,” in *ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2630–2640.
- [17] R. Zellers *et al.*, “MERLOT: multimodal neural script knowledge models,” in *NeurIPS 2021, December 6-14, 2021, virtual, 2021*, pp. 23 634–23 651.
- [18] Z. Chen *et al.*, “InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks,” in *CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 24 185–24 198. [Online]. Available: <https://doi.org/10.1109/CVPR52733.2024.02283>
- [19] P. H. Seo *et al.*, “Avformer: Injecting vision into frozen speech models for zero-shot AV-ASR,” in *CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 22 922–22 931.
- [20] T. Afouras *et al.*, “Deep audio-visual speech recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, 2022.
- [21] P. Ma *et al.*, “End-to-end audio-visual speech recognition with conformers,” in *ICASSP*, 2021, pp. 7613–7617.
- [22] J. S. Chung *et al.*, “Lip reading sentences in the wild,” in *CVPR*, 2017, pp. 3444–3453.
- [23] G. Sun *et al.*, “video-salmonn: Speech-enhanced audio-visual large language models,” in *ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=nYsh5GFIqX>
- [24] C. Fu *et al.*, “VITA: towards open-source interactive omni multimodal LLM,” *CoRR*, vol. abs/2408.05211, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2408.05211>
- [25] Y. Wu *et al.*, “Robust audiovisual speech recognition models with mixture-of-experts,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 43–48.
- [26] V. Gabeur *et al.*, “AVATAR: unconstrained audiovisual speech recognition,” in *Interspeech 2022, Incheon, Korea, September 18-22, 2022*. ISCA, 2022, pp. 2818–2822. [Online]. Available: <https://doi.org/10.21437/Interspeech.2022-776>
- [27] C. Fu *et al.*, “Vita-1.5: Towards gpt-4o level real-time vision and speech interaction,” *CoRR*, vol. abs/2501.01957, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2501.01957>
- [28] Y. Lu *et al.*, “Syneslm: A unified approach for audio-visual speech recognition and translation via language model and synthetic data,” in *Synthetic Data’s Transformative Role in Foundational Speech Models*, 2024, pp. 31–35.
- [29] X. Li *et al.*, “Universal phone recognition with a multilingual allophone system,” in *ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 8249–8253. [Online]. Available: <https://doi.org/10.1109/ICASSP40776.2020.9054362>
- [30] A. Graves *et al.*, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML 2006, Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, ser. ACM International Conference Proceeding Series, vol. 148. ACM, 2006, pp. 369–376. [Online]. Available: <https://doi.org/10.1145/1143844.1143891>
- [31] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *ICML 2014, Beijing, China, 21-26 June 2014*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 1764–1772.
- [32] L. Kürzinger *et al.*, “Ctc-segmentation of large corpora for german end-to-end speech recognition,” in *SPECOM 2020, St. Petersburg, Russia, October 7-9, 2020, Proceedings*, ser. Lecture Notes in Computer Science, vol. 12335. Springer, 2020, pp. 267–278. [Online]. Available: https://doi.org/10.1007/978-3-030-60276-5_27
- [33] R. Girdhar *et al.*, “Imagebind one embedding space to bind them all,” in *CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 15 180–15 190. [Online]. Available: <https://doi.org/10.1109/CVPR52729.2023.01457>
- [34] Y. Wu *et al.*, “Enhancing audiovisual speech recognition through bifocal preference optimization,” *CoRR*, vol. abs/2412.19005, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2412.19005>
- [35] Y. Peng *et al.*, “OWSM v3.1: Better and faster open whisper-style speech models based on e-branchformer,” *CoRR*, vol. abs/2401.16658, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2401.16658>
- [36] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139. PMLR, 2021, pp. 8748–8763.
- [37] A. Yang *et al.*, “Qwen2 technical report,” *CoRR*, vol. abs/2407.10671, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2407.10671>