Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Spoken question answering for visual queries

Nimrod Shabtay[1,2,*], *Zvi Kons*[1,*], *Avihu Dekel*[1], *Hagai Aronowitz*[1], *Ron Hoory*[1], *Assaf Arbelle*[1]

[1]IBM Research, Israel
[2]Tel-Aviv University, Israel

nimrod.shabtay@ibm.com, zvi@il.ibm.com

Abstract

Question answering (QA) systems are designed to answer natural language questions. Visual QA (VQA) and Spoken QA (SQA) systems extend the textual QA system to accept visual and spoken input respectively.

This work aims to create a system that enables user interaction through both speech and images. That is achieved through the fusion of text, speech, and image modalities to tackle the task of spoken VQA (SVQA). The resulting multi-modal model has textual, visual, and spoken inputs and can answer **spoken** questions on images.

Training and evaluating SVQA models requires a dataset for all three modalities, but no such dataset currently exists. We address this problem by synthesizing VQA datasets using two zero-shot TTS models. Our initial findings indicate that a model trained only with synthesized speech nearly reaches the performance of the upper-bounding model trained on textual QAs. In addition, we show that the choice of the TTS model has a minor impact on accuracy.

Index Terms: Large language models, Large multimodal models, spoken question answering, visual question answering, multimodal

1. Introduction

As the AI landscape continues to evolve, there is a growing shift from unimodal to multimodal tasks. Unimodal tasks are those that rely on a single type of input data, such as text, images, or audio, to perform a specific task, such as text generation, question answering, image classification, object detection, and automatic speech recognition (ASR). In contrast multimodal tasks process and integrate multiple types of input data, combining text with images (e.g., VQA [1], text-to-image synthesis [2]), or combining text with audio (e.g., SQA[3], and text-to-speech synthesis [4, 5]).

Large language models (LLMs) have become the primary tool for solving many text-based tasks, driven by the abundance of textual training data and the effort put into their training. As these models gain wider adoption, more attention is directed at the human-machine interface. Two crucial modalities for enhancing interaction and usability are speech and vision, i.e. the ability of the models to *listen* to the user's speech and to *view* relevant information in pictures. Speech enables more natural and intuitive communication while vision allows models to process a broader context leading to richer, more interactive and holistic AI systems.

In this paper we look at a model that can combine information from the three modalities of text, images and speech. Our focus is on the spoken visual question answering task (SVQA). In this task we give the model an audio with a spoken query. The query is about the visual information within a given image. The query may also include a text prompt with additional information (e.g. instructions and multiple-choice answers). The model is expected to provide a textual response.

Several approaches for augmenting an LLM with additional modalities have been explored. Due to the close relation between spoken and written text, one simple approach is to construct a pipeline that first converts the speech to text using an ASR model and then processes the text with an LLM. While this does not require any modifications to the LLM, it is clear that some audible information may be lost in the process and this is highly dependent on the accuracy of the ASR.

When dealing with visual information, however, constructing such a pipeline is far from optimal. To that end, models such as LLaVA [6] propose a modality alignment approach, where an encoded representation of the image, extracted from a pre-trained image encoder, is directly aligned to the LLM's input embedding space. We extend this approach to a Speech-Vision-Language model by jointly aligning two modality-specific encoders, Whisper [7] for speech and CLIP [8] for images, to the LLMs input embedding space.

The outline for our model's architecture is detailed in Section 3 starting from a baseline model that was pre-trained for VQA tasks. To incorporate speech capabilities, we first pre-train the speech modules on speech-only tasks before fine-tuning the entire model on spoken VQA (SVQA) tasks. This process requires suitable training and evaluation datasets, which include both human speech and text-to-speech (TTS)-generated speech. Section 4 provides details on the datasets used and discusses the limitations of using synthesized speech data, along with the methods we employed to mitigate these challenges.

In Section 5, we describe our two-phase training process: (1) separate pre-training of the speech and vision projectors and (2) fine-tuning the joint model. We perform an extensive ablation study by training several models with different combinations of datasets and examine the impact of various training choices.

We evaluate our models on several VQA and SVQA test sets, with the results presented in Section 6. Our model achieves an accuracy of 62% on SEED-Bench, despite the greater complexity of SVQA compared to VQA. This result highlights the potential of extending standard vision-language models to handle more than two modalities. Finally, in Section 7, we provide a discussion of our results and propose future research directions.

We summarize our contributions as follows:

1. We extend VQA models to accept spoken inputs.
2. We create new datasets for spoken visual question answering

[*]These authors contributed equally.

10.21437/Interspeech.2025-843

(SVQA).

3. We demonstrate the practical use of a zero-shot multi-speaker TTS system to synthesize QAs, overcoming previous challenges and preventing training-test speaker leakage.
4. We show that our models perform nearly as well as the textual upper bound, highlighting SVQA's potential in human-machine interaction.

2. Related work

Our work on SVQA extends two important tasks, VQA and SQA. Visual question answering (VQA) attempts to describe, locate, and reason regarding some visual input [9, 10, 11]. Several models that leverage powerful LLMs and vision encoders have been proposed [6, 12, 13, 14]. Spoken question answering attempts to answer a user's spoken question [15]. Several works address this challenge [3, 16].

Spoken visual question answering was tackled by several works, which leveraged a TTS system to synthesize the questions [17, 18, 19]. These works use a single-speaker TTS system. This may not generalize to unseen speakers, as their models have only seen a single speaker during training. Moreover, they test on data synthesized from the same TTS system, which does not expose this limitation. Furthermore, they formulate the problem as a classification rather than a generative task.

A recent work [20] uses cross-attention of audio/vision embeddings, resembling contrastive loss, making it more of a classifier than a generative model. Other works focus on SQA for the medical domain [21, 22]. Another related area of works that incorporate text, audio, and images is audio-visual question answering, which is a part of a more general task of video understanding [23, 24, 25, 26].

3. Spoken VQA model architecture

Our model is based on the LLaVA [6, 12] architecture. The LLaVA model extends a text-based, generative, large language model (LLM) for visual question answering (VQA) by allowing visual information input from images. This is done by first processing the images using an image encoder. The visual data from the encoder's output space is then aligned to the LLM input space using a projector module (Figure 1).

We extend this model to support spoken question answering (SQA) with additional speech input. This is achieved by processing the audio input with a speech encoder. The output of this encoder is aligned to the LLM using an additional projector module (Figure 1). With these two additional inputs, the model now supports spoken visual-questions answering (SVQA). For example, we can have a spoken query about visual information in an image. To support this type of interaction, the textual prompt may contain `<image>` and `<audio>` tags. After the textual tokenization and the token embeddings, these tags are replaced with the outputs of the image and audio projector, respectively.

In the experiments described below, our SVQA model is comprised of the following components:

- **LLM**: Vicuna-13b-v1.5 [27, 28]
- **Image encoder**: CLIP-ViT-Large-336px (305 M parameters) [8], with the input image resized to the CLIP input resolution (336x336).
- **Image projector**: 2 fully connected layers with 1024 hidden states and GeLU activation.
- **Speech encoder**: Whisper-large-v3 encoder (635 M parameters) [7]
- **Speech projector**: Our projector first downsamples the input by concatenating 10 consecutive frames into 1. This is followed by 2 fully connected layers with 4096 hidden states and ReLU activation.

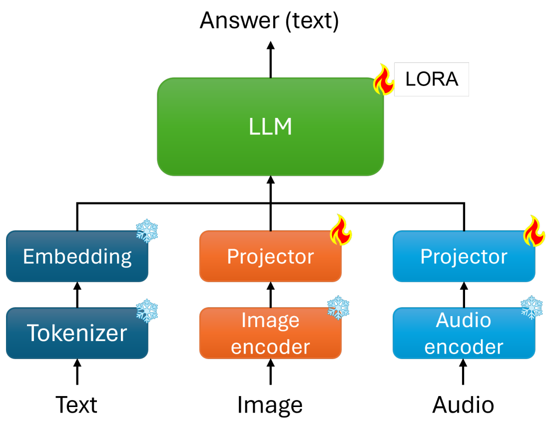

Figure 1: *SVQA model components. Frozen modules are marked with a snow icon and trainable modules with a fire icon.*

4. Datasets

4.1. Speech-only datasets

We pre-train the speech projector using the English subset of Multilingual LibriSpeech (MLS) [29]. The MLS speech samples are used for two different tasks. The first is ASR, where the LLM is presented with prompts such as `Transcribe speech to text. Input:<audio>` and the target is the audio transcript. We found that models trained only for ASR were only producing the transcripts of audio samples. The models would not address the content of the audio even when given direct instructions (e.g., `Answer the question in the following audio <audio>`).

To overcome this problem, we introduce a new task, which is based on the same audio samples. In this task, we first asked a textual language model to describe the sample's transcript (e.g., with the prompt `Describe the text. Text: {transcript}`). We then record the responses and use them as targets during training, where we use the same prompt but replace the transcript with the corresponding audio sample. This is similar to [30], which asks the LLM to respond or continue the textual transcript. However, they found that in many cases, the transcripts are short text fragments without proper context, and the LLM can only provide a generic answer without any meaningful connection to the text itself. By generating a description, we usually get an answer that is more closely related to the input text, thus improving the model's training.

4.2. Speech and images datasets

At the time of writing this article, we are unaware of any training or testing spoken-visual QA datasets. On the other hand, visual-textual datasets exist, such as LLaVA-1.5-Instruct [6], The Cauldron [31], and Cambrian-7M [32]. Consequently, we convert a textual LLaVA-1.5-Instruct training set into a spoken set by synthesizing the questions using zero-shot TTS systems [5].

Using synthetic TTS data for both training and testing might be considered problematic. The system can easily learn to over-fit the specific characteristics of the TTS output. For example, other synthesized datasets such as Spoken-SQuAD [33] or HeySQuAD [34] use one TTS with only one voice to synthesize the entire dataset. We overcome this limitation by employing several strategies. First, we use two different high-quality zero-shot TTS systems: StyleTTS2 [5] and F5-TTS [35].

Using these two systems, we perform cross-validation – train on samples generated using one TTS, and test on the samples generated using another TTS – to identify possible over-fitting. Moreover, we synthesize the speech with a large variety of voices and speaking styles, as both zero-shot systems receive as input an audio prompt and text. They synthesize the text into speech while trying to match the voice and the style of the audio prompt. We use random samples from the MLS dataset as audio prompts. This large variety of voices and styles helps us avoid over-fitting to a specific voice.

In addition, we use audio prompts from the MLS train and test sets for training and testing correspondingly. Because these sets are disjoint, our test also performs cross-validation with different speakers and styles selected at random, including 950 different speakers. This results in two training datasets that were based on the LLaVA-1.5 Instruct training set. Each one has 3.4M audio samples and a total duration of 5135 hours of StyleTTS2-generated speech and 4330 hours of F5-TTS-generated speech (StyleTTS2 speech is generally slower). A third mixture training dataset is constructed by randomly selecting audio file for each textual sample from either one of the above-mentioned TTS systems. The synthesis process takes approximately 2200 GPU hours on NVIDIA V100 processors.

We also convert three well-known VQA evaluation sets into SVQA test sets by synthesizing their textual questions into speech using the two TTS systems with 42 different speakers without speaker or sample overlap with the train set speakers. The test sets are SEED-Bench (Image subset) [9], MME [10], and DocVQA [11]. Only the question is synthesized, while the rest of the prompt, including any instructions or the multiple-choice answers, is kept in textual format. Each test set has two versions (one with each TTS). In addition, we extend SEED-Bench and created a spoken version including the multiple choices (see Sec. 6.2.)

5. Model training

The LLaVA-1.5 base model contains a trained Vicuna 13B LLM, a CLIP image encoder, and a trained image projector from [12]. The speech tower is composed of a trained Whisper encoder[1] and a speech projector.

In the first part of the training, we pre-train only the speech projector in a speech-only system (see Sec. 5.1). In the second part, we train the whole system with various combinations of frozen and trainable components (see Sec. 5.2).

5.1. Speech projector pre-training

The speech projector is pre-trained in a system that contains only the speech encoder, its projector, and the LLM. The weights of the encoder and the LLM are frozen; only the projector is tuned.

The projector is trained on two tasks: audio transcription and audio context description using the speech-only dataset described in 4.1. The training consists of one epoch over the two datasets and takes several days on four NVIDIA A100 GPUs.

For the vision projector we use the original LLaVA-1.5 projector pre-trained with CLIP [8] as the image encoder and Vicuna-13B [27] as the LLM.

5.2. Supervised fine-tuning

In our experiments we compare the following four models:
LLaVA-1.5-13B: This is the baseline tuned VQA LLaVA-1.5 evaluated on text-image datasets. We reproduced the results as reported in [12].
SVQA-baseline: This is the full SVQA system containing a Vicuna LLM, image tower and the speech tower. However, on this system the LLM and encoders are not fine-tuned for the task, and there is no joint vision-audio training. The image and speech projectors are each pre-trained separately.
SVQA-STTS2, **SVQA-F5**, **SVQA-Mix**: These are the SQVA-baseline systems that are fine-tuned with the corresponding StyleTTS2, F5-TTS, and mixed datasets described in 4.2. During this training, both image and speech encoders are frozen, while both projectors are allowed to be fine-tuned. In addition, the LLM is tuned using low-rank-adaptation (LoRA) with $r = 64$ and $\alpha = 16$.
SVQA-STTS2-FR: Same as **SVQA-STTS2**, but during the fine-tuning only the projectors are trained (LLM is frozen and no LoRA is used).

The models are trained over one epoch of the corresponding dataset. Each training takes about one week using four NVIDIA A100 GPUs.

6. Results

6.1. Spoken VQA

Table 1: ***Performance across VQA benchmarks (StyleTTS2 / F5-TTS).*** *Overall, we can see that the LoRA fine-tuned models perform better than the pre-trained and non-LoRA versions. The effect of the different TTS systems is minimal, and F5 shows a slightly better performance across all datasets.*

Model	SeedBench	MME	DocVQA$_{val}$
LLaVA-1.5-13B	68.2	1528	0.235
SVQA-baseline	27.4/28.9	494/457	0.015/0.009
SVQA-STTS2-FR	59.3/56.8	1049/911	0.161/0.147
SVQA-STTS2	61.4/59.2	1231/**1187**	0.166/0.145
SVQA-F5	**61.9**/60.3	**1256**/1173	**0.168**/0.153
SVQA-Mix	61.8/**60.5**	1239/1162	**0.171**/0.147

We evaluate our models using the *lmms-eval* [36, 37] evaluation platform, which has been widely adopted by the vision and language community. For SeedBench, we report the accuracy between the predicted answer and the ground-truth answer. For DocVQA, we report the average normalized Levenshtein similarity (ANLS) between the predicted and ground-truth answers. For MME, we follow LLaVA and report the sum of accuracy scores of the perception tasks. For all these metrics, higher results is better. We reproduce the results of LLaVA-1.5-13B as our baseline and evaluate our spoken-visual variants of SEED-Bench [9], MME [10], and DocVQA [11]. We provide samples from our SVQA datasets [2] and summarize our results in Ta-

[1] `https://huggingface.co/openai/whisper-large-v3`

[2] `https://ibm.biz/SvqaDemo`

ble 1[3].

Overall, the variants with LoRA fine-tuning are the most capable models, with the F5-TTS model slightly outperforming the STTS2-TTS on average. As expected, the pre-trained versions exhibited the lowest performance, while the non-LoRA variant performed worse than the LoRA-fine-tuned variants.

We also report the text-based VQA as an upper performance bound for our method. In general, we observe only small differences between the choice of TTS systems for the training sets, with no one system consistently outperforming the others across all tests. However, for the test sets, the StyleTTS2 systems yield better results, possibly due to producing slower, more intelligible speech. Given these small differences, we anticipate that similar results would be obtained when testing with natural human speech.

Our results demonstrate solid performance across various types of VQA, underscoring the potential of SVQA to open new avenues for more effective human-machine interaction.

6.2. The effect of a spoken post-prompt

Table 2: *SeedBench - Image accuracy of the full speech prompt (Full) vs. just the question as speech and the post-prompt as text (Question). Adding the post-prompt as part of the speech utterance consistently causes performance degradation.*

Model	StyleTTS2		F5-TTS	
	Question	Full	Question	Full
SVQA-STTS2	**61.4**	54.8	59.2	47.0
SVQA-F5	**61.9**	56.2	60.3	50.8

The textual version of the SeedBench-Image [9] has the format of a question followed by multiple-choice answers and instructions about the requested answers format. For the tests in Section 6.1, we synthesize only the question while the answers and prompts are kept in text format. In addition, we also create a second version of the test set where the whole text (question, answers, and post-prompt) is synthesized.

The results for the different sets are shown in Table 2. We can see that adding the post-prompt as part of the speech utterance performed worse consistently. We hypothesize that this is because putting the answers in this format is unnatural for speech. Also, the answers are usually very short, so small errors can easily lead to failures.

6.3. Word error rates

To understand the gaps between the models, we also examine their abilities in speech transcription. We feed the model with only speech and the appropriate textual prompt requesting a transcript and calculate the word-error-rate (WER) for the output. We do this for the natural human speech in the MLS test set and for the speech we synthesized for the SEED-Bench set with the two TTS systems. The results are presented in Table 3.

It is hard to extract useful insights from these results. The high error rates that we see in several cases are mostly caused by the failure of the model to follow the instructions in the prompt. Instead of providing a transcript of the speech, the model treats the content of the speech as a question and tries to answer it.

[3]A later analysis also included the results of the VQA model on the ASR transcript of the speech. These results will be presented elsewhere.

Table 3: *Transcription accuracy in WER for three different test sets using different models.*

Model	MLS	SeedBench	
		StyleTTS2	F5-TTS
SVQA-Baseline	7.4%	28.4%	86.0%
SVQA-STTS2-FR	10.0%	65.4%	108.9%
SVQA-STTS2	87.1%	115.4%	119.5%
SVQA-F5	84.6%	130.5%	128.9%
SVQA-Mix	85.6%	131.9%	134.2%

In the MLS case, most of the samples contain speech from text fragments, so this is less frequent.

From the results, we can see that training the projectors on the STTS2 data while the model is frozen (STTS2-FR) provides improvement to the STTS2 test set, but not to the F5. In all other models, there is a considerable increase in WER. This is most likely because the language models specialize in question answering and therefore try to do this instead of transcription.

7. Discussion

This paper presents our work extending a VQA model into spoken VQA by incorporating a speech encoder and a corresponding projector. A significant portion of the effort was spent on building both the speech and SVQA datasets. We demonstrate that, in the absence of natural speech datasets, synthetic speech generated via TTS can be effectively used.

Although our spoken models are not yet as accurate as their textual counterparts we have identified two possible causes. One possible explanation for the gap in the VQA performance is the fact that the original LLaVA model was fully fine-tuned for VQA as opposed to our LoRA fine-tuning. The other likely cause of this discrepancy may stem from errors introduced during speech synthesis (for training and testing) and during the conversion of speech back to text during inference. While the impact of using synthetic speech instead of natural speech is difficult to quantify, our cross-validation experiments using two different TTS systems yielded similar results. This suggests that the impact of synthetic speech is likely minimal.

As we demonstrate the concept of SVQA, we use LLaVA-1.5 as our base model and convert its original training data into a spoken training set. Newer models, incorporate stronger LLMs and additional enhancements that significantly improve overall performance. Our method could also benefit from synthesizing larger publicly available VQA datasets, such as Cambrian [32] and The Cauldron [31] to SVQA, to further enhance its effectiveness.

Future work is needed to investigate different directions for bridging the gap between the textual and spoken models. We might be able to improve the speech transcription accuracy by adding more datasets for the pre-training stage. These could be either natural speech or synthesized speech. Additional data may also help in the fine-tuning phase. For example, we could add the speech-only dataset and the textual VQA so that the model will not forget its basic abilities while training for SVQA.

8. References

[1] S. Antol, A. Agrawal, J. Lu *et al.*, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[2] R. Rombach, A. Blattmann, D. Lorenz *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[3] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L. shan Lee, "SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering," 2020. [Online]. Available: https://arxiv.org/abs/1910.11559

[4] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[5] Y. A. Li, C. Han, V. S. Raghavan *et al.*, "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," 2023. [Online]. Available: https://arxiv.org/abs/2306.07691

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[7] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[8] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[9] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "SEED-Bench: Benchmarking multimodal LLMs with generative comprehension," 2023. [Online]. Available: https://arxiv.org/abs/2307.16125

[10] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 11 2024. [Online]. Available: https://doi.org/10.1093/nsr/nwae403

[11] M. Mathew, D. Karatzas, R. Manmatha, and C. Jawahar, "DocVQA: A dataset for VQA on document images." 2021. [Online]. Available: https://arxiv.org/abs/2007.00398

[12] H. Liu, C. Li *et al.*, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[13] L. Karlinsky, A. Arbelle *et al.*, "Granite vision: a lightweight, open-source multimodal model for enterprise intelligence," 2025. [Online]. Available: https://arxiv.org/abs/2502.09927

[14] X. Dong, P. Zhang *et al.*, "InternLM-XComposer2: Mastering free-form text-image composition and comprehension in vision-language large model," 2024. [Online]. Available: https://arxiv.org/abs/2401.16420

[15] C.-H. Lee, S.-M. Wang, H.-C. Chang, and H.-Y. Lee, "ODSQA: Open-domain spoken question answering dataset," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 949–956.

[16] C. You, N. Chen, and Y. Zou, "Knowledge distillation for improved accuracy in spoken question answering," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7793–7797.

[17] T. Zhang, D. Dai, T. Tuytelaars, M.-F. Moens, and L. V. Gool, "Speech-based visual question answering," 2017. [Online]. Available: https://arxiv.org/abs/1705.00464

[18] A. P. Patil, A. ehera, P. Anusha *et al.*, "Speech enabled visual question answering using lstm and cnn with real time image capturing for assisting the visually impaired," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 2475–2480.

[19] F. Alasmary and S. Al-Ahmadi, "SBVQA 2.0: Robust end-to-end speech-based visual question answering for open-ended questions," *IEEE Access*, 2023.

[20] A. R. Chowdhury, T. Rajkhowa, and S. Sharma, "Towards multilingual spoken visual question answering system using cross-attention," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 9165–9175.

[21] K. Ramnath, L. Sari, M. Hasegawa-Johnson, and C. Yoo, "Worldly wise (WoW)-cross-lingual knowledge fusion for fact-based visual spoken-question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1908–1919.

[22] T. Rajkhowa, A. R. Chowdhury, S. Nagaonkar, and A. M. Tripathi, "TM-PATHVQA: 90000+ textless multilingual questions for medical visual question answering," 2024. [Online]. Available: https://arxiv.org/abs/2407.11383

[23] P. Yang, X. Wang, X. Duan *et al.*, "AVQA: A dataset for audio-visual question answering on videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3480–3491.

[24] G. Li, Y. Wei, Y. Tian *et al.*, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 108–19 118.

[25] G. Li, W. Hou, and D. Hu, "Progressive spatio-temporal perception for audio-visual question answering," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7808–7816.

[26] Z. Tang, J. Cho, Y. Nie, and M. Bansal, "TVLT: Textless vision-language transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9617–9632, 2022.

[27] W.-L. Chiang, Z. Li *et al.*, "Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[28] L. Zheng, W.-L. Chiang *et al.*, "Judging LLM-as-a-judge with MT-Bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[29] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, 2020, pp. 2757–2761.

[30] Y. Fathullah, C. Wu *et al.*, "Audiochatllama: Towards general-purpose speech abilities for llms," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 5522–5532.

[31] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?" 2024. [Online]. Available: https://arxiv.org/abs/2405.02246

[32] S. Tong, E. Brown *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs," 2024. [Online]. Available: https://arxiv.org/abs/2406.16860

[33] C.-H. Lee, S.-L. Wu, C.-L. Liu, and H.-y. Lee, "Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension," *Proc. Interspeech 2018*, pp. 3459–3463, 2018.

[34] Y. Wu, S. Rallabandi, R. Srinivasamurthy, P. P. Dakle, A. Gon, and P. Raghavan, "HeySQuAD: A spoken question answering dataset," 2024. [Online]. Available: https://arxiv.org/abs/2304.13689

[35] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," 2024. [Online]. Available: https://arxiv.org/abs/2410.06885

[36] K. Zhang, B. Li, P. Zhang *et al.*, "LMMs-Eval: Reality check on the evaluation of large multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2407.12772

[37] B. Li, P. Zhang *et al.*, "LMMs-Eval: Accelerating the development of large multimoal models," March 2024. [Online]. Available: https://github.com/EvolvingLMMs-Lab/lmms-eval