SVLA: A Unified Speech-Vision-Language Model for Multimodal Reasoning and Generation

Anonymous EMNLP submission

Abstract

Large Vision-Language Models have shown impressive capabilities in tasks such as image captioning, visual question answering, and cross-modal retrieval. However, there are still significant challenges that need to be addressed in order to fully unlock the potential of these models. First, integrating speech, text, and vision into a unified model is particularly difficult for tasks like Spoken Image Captioning and Spoken Visual Question Answering, where the interaction between these modalities introduces additional complexity. Second, existing speech generation approaches differ—some generate speech directly, while others use an intermediate text step—but their impact on fluency, coherence, and accuracy remains unexplored. To address these challenges, we propose **SVLA**, a unified **S**peech-**V**ision-**L**anguage **A**ssistant based on a decoder-only transformer architecture that seamlessly integrates multimodal inputs and outputs. We enhance model performance with a large-scale speech-text-image dataset containing 38.2 million examples and 64.1 hours of TTS-generated speech. Our approach advances multimodal understanding and generation, facilitating more effective integration of speech, text, and vision (http://github.com/vlm-svla/svla).

1 Introduction

Recent advances in Large Vision-Language Models (LVLMs) (OpenAI, 2023a; Li et al., 2023; Liu et al., 2024; Alayrac et al., 2022) mark a significant step toward multimodal AI, enabling models to interpret and reason over visual and textual inputs. However, these models remain constrained by their reliance on text-based instructions and lack native support for speech. Efforts such as LLaMA 3 (Touvron et al., 2023) and Qwen-Audio (Chu et al., 2023) have begun incorporating speech through modality-specific encoders, these models remain limited to speech perception and do not support speech generation.

Recent models (Zhang et al., 2023a; Zhan et al., 2024; Wu et al., 2024; Xu et al., 2025) address this limitation by introducing discrete speech representations, allowing language models to process speech as semantic tokens within a unified speech-text token space. This enables bidirectional modeling and multimodal integration. However, most systems remain focused on shallow tasks—such as text-to-speech (Veaux et al., 2017), image-to-music (Chowdhury et al., 2024), or basic audio captioning (Kim et al., 2019)—and struggle with more complex reasoning tasks like image captioning (IC) or visual question answering (VQA) involving spoken inputs or outputs. A key bottleneck is the absence of large-scale, richly aligned datasets spanning speech, text, and vision. Most models are trained on unimodal or bimodal data, limiting their ability to generalize to cognitively demanding, speech-enabled multimodal reasoning.

Furthermore, two prominent paradigms have emerged for enabling speech capabilities within multimodal systems (Zhang et al., 2023a; Nachmani et al., 2023): Cross-modal Instruction and Chain-of-Modality Instruction. Cross-modal Instruction directly maps inputs across modalities, such as converting an image to speech, offering efficiency but frequently compromising semantic coherence. In contrast, Chain-of-Modality Instruction involves first translating the input into textual form, performing reasoning over the generated text, and subsequently producing output in speech or text form. This method leverages the structured reasoning and planning capabilities inherent to large language models but may lead to verbosity. However, a comparative analysis of these paradigms under consistent tri-modal settings within LVLMs has not yet been conducted.

To address the above challenges, we introduce a large-scale synthetic trimodal dataset—combining text, image, and speech—to support instruction-following in both written and spoken formats. This

dataset integrates widely-used resources, including VQAv2 (Goyal et al., 2017), LAION-600M (Schuhmann et al., 2022), Visual Genome (Krishna et al., 2016), LibriHeavy (Kang et al., 2024), LibriSpeech (Panayotov et al., 2015), CommonVoice (Ardila et al., 2019), A-OKVQA (Schwenk et al., 2022), VizWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019), and COCO-Caption (Lin et al., 2014). To ensure natural-sounding and diverse speech data, we use a controllable text-to-speech (TTS) system that varies accent, speaking speed, and prosody. We further include noise and stylistic augmentation from MUSAN (Snyder et al., 2015) to improve robustness. Additionally, we implement a modality-switching instruction tuning strategy, enabling the model to flexibly process and respond in either text or speech.

Building on this foundation, we propose SVLA (Speech-Vision-Language Assistant), a unified, self-supervised multimodal model capable of reasoning over image, text, and speech inputs and generating outputs in either text or speech. SVLA features a hybrid fusion architecture: it applies early fusion between speech and text using discrete semantic units, allowing both to be modeled jointly in a shared language space (Zhan et al., 2024; Zhang et al., 2023b), and late fusion for visual input by integrating image embeddings from a pretrained encoder (Liu et al., 2024). This design supports complex tasks such as spoken VQA, speech-driven image captioning, and multimodal instruction following.

We also present the systematic evaluation comparing Cross-modal and Chain-of-Modality Instruction within a unified vision-language-speech framework. Our experiments span four controlled configurations—text-to-text, text-to-speech, speech-to-text, and speech-to-speech—using consistent instruction formats and shared example naming across tasks. Benchmarks such as VQAv2 and COCO-Captions are used to assess the trade-offs between fluency, coherence, and semantic quality under each setting.

Our contributions are as follows:

- We construct a large-scale, aligned speech-text-vision dataset from diverse benchmarks, supporting instruction-following in both textual and spoken forms.
- We propose SVLA, a unified tri-modal model that integrates early fusion (text-speech) and late fusion (vision) to enable robust multimodal reasoning and generation.
- We establish an evaluation framework that systematically compares Cross-modal and Chain-of-Modality paradigms across consistent input-output modality configurations, while also assessing robustness to variations in accent and speaking speed.

2 Related Works

Speech-Enabled Vision-Language Models: LVLMs (Yuan et al., 2021; Li et al., 2023; Liu et al., 2024; Chen et al., 2024) were initially designed for image-text reasoning, lacking native speech support. Early extensions added speech input and TTS output but remained limited in conversational expressiveness (OpenAI, 2023b; Dubey et al., 2024), with higher latency and limited contextual adaptability due to multi-stage pipelines. GPT-4o (OpenAI, 2024)addresses these issues with native speech generation for real-time, expressive interaction, though its closed-source nature has driven open efforts to replicate its capabilities. Models like NExT-GPT (Wu et al., 2024), CoDi-1/2 (Tang et al., 2023a, 2024), Unified-IO (Lu et al., 2024), and AnyGPT (Zhan et al., 2024) extend multimodal support via modality-specific encoders or projections, targeting perceptual tasks (e.g., image/audio/video generation, ASR, TTS). However, cross-modal reasoning (e.g., VQA, image captioning across text and speech modalities) remains underexplored. TMT (Kim et al., 2024) offers limited progress, enabling speech output from image captions but lacking joint multimodal reasoning.

Speech-Text-Vision Datasets: Most speech datasets used in multimodal learning—such as LibriSpeech (Panayotov et al., 2015), CommonVoice (Ardila et al., 2019), and GigaSpeech (Chen et al., 2021)—are designed for ASR or TTS and do not include vision. While datasets like SpeechCOCO (Havard et al., 2017) and SpokenCOCO (Hsu et al., 2020) add speech to image captioning, they remain limited in scale, task diversity, and interactivity. Others, like How2 (Sanabria et al., 2018), offer aligned speech and video, but focus on narrow instructional domains and do not support flexible input-output modality switching. Some recent works synthesize multimodal dialogues by using large language models to generate text-based

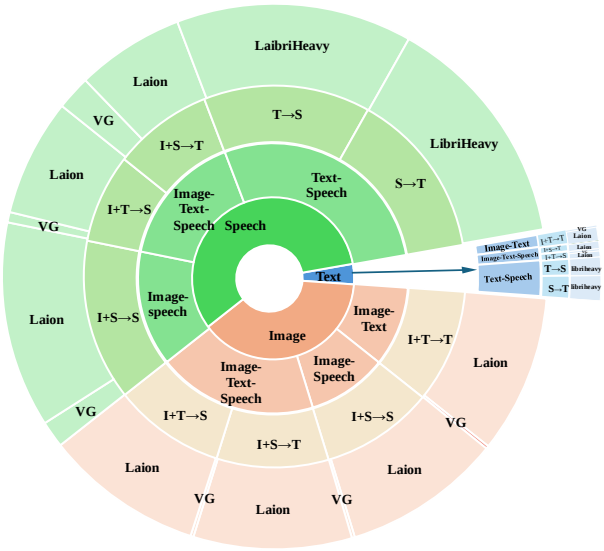


Figure 1: Pre-training data distribution across three modalities: Image, Text, and Speech.

prompts, which are paired with audio and visual content using tools like TTS and image generation models. While useful for data augmentation, these datasets often focus on perception tasks and lack cognitively demanding reasoning challenges such as VQA or multimodal instruction following.

Direct vs. Instruction-Based Speech Generation: Speech generation in multi-modal systems generally follows one of two paradigms: direct generation, where speech is produced end-to-end without textual intermediates, and instruction-based generation, where models first generate text and then synthesize speech from it. Instruction-based approaches—used in models like GPT-4o (OpenAI, 2024), AnyGPT (Zhan et al., 2024), NExT-GPT (Wu et al., 2024), and SpeechGPT (Zhang et al., 2023a)—offer modality consistency but often produce redundant or unnatural phrasing. In contrast, direct approaches aim to generate speech directly from inputs (e.g., Spectron (Nachmani et al., 2023)), allowing for more fluid prosody and conversational tone.

3 Data Generation

To construct a large-scale tri-modal dataset, we extend existing image-text corpora (Lin et al., 2014; Schuhmann et al., 2022) with corresponding spoken utterances. These datasets are selected for their abundance, diversity, and broad coverage of real-world visual and linguistic contexts, providing a strong foundation for scalable multimodal learning. We synthesize speech from text using the Melon-TTS model (Zhao et al., 2023), chosen for its controllability over prosody and speaker characteristics. To enhance diversity, we vary speaking rates from 0.7× to 1.3× and include multiple English accents, such as American, British, Indian, and Australian. Additionally, we introduce environmental background noise—e.g., rain, footsteps, and ambient sounds—from the MUSAN corpus (Snyder et al., 2015) to simulate realistic acoustic conditions. All audio is standardized to a 16 kHz sampling rate to ensure clarity and computational efficiency. This pipeline yields a rich and varied speech-text-image dataset suitable for both pre-training and fine-tuning multimodal models.

3.1 Pre-train dataset

Our pre-training strategy targets a unified speech-vision-language model capable of TTS, ASR, image captioning, and VQA. The training data includes: (1) 8M text-speech pairs from the publicly available LibriHeavy corpus (Kang et al., 2024) for speech generation and recognition, and (2) 6M image-text-speech triples, where speech is *synthetically generated* from image-caption pairs in LAION-COCO (Schuhmann et al., 2022) and question-answer pairs from Visual Genome (Krishna et al., 2016) to support vision-language reasoning. To enable modality switching, we use instruction-style prompts (e.g., "Answer this question in speech" or "Describe the image in text"), guiding the model to produce either spoken or written outputs. Speech outputs are capped at 5 seconds for training efficiency. Figure 1 shows the modality token distribution[1]. Additional details are provided in Appendix A.

3.2 Visual Instruction dataset

For supervised fine-tuning (SFT), we adopt a similar structure to pre-training but introduce multi-turn conversations to improve coherence and long-range context retention. The text-speech subset includes LibriSpeech (Panayotov et al., 2015) and Common-Voice (Ardila et al., 2019), covering diverse linguistic and command-oriented expressions. The vision-text-speech subset incorporates VQAv2, A-OKVQA (Schwenk et al., 2022), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), and COCO-Captions-2014, ensuring broad coverage of open-ended and visually grounded tasks. To enable consistent comparison across modalities, we

[1]Token counts: text via Qwen-2.5-1.5B (Yang et al., 2024), speech via SpeechTokenizer (Zhang et al., 2023b), and images estimated at 256 tokens each.

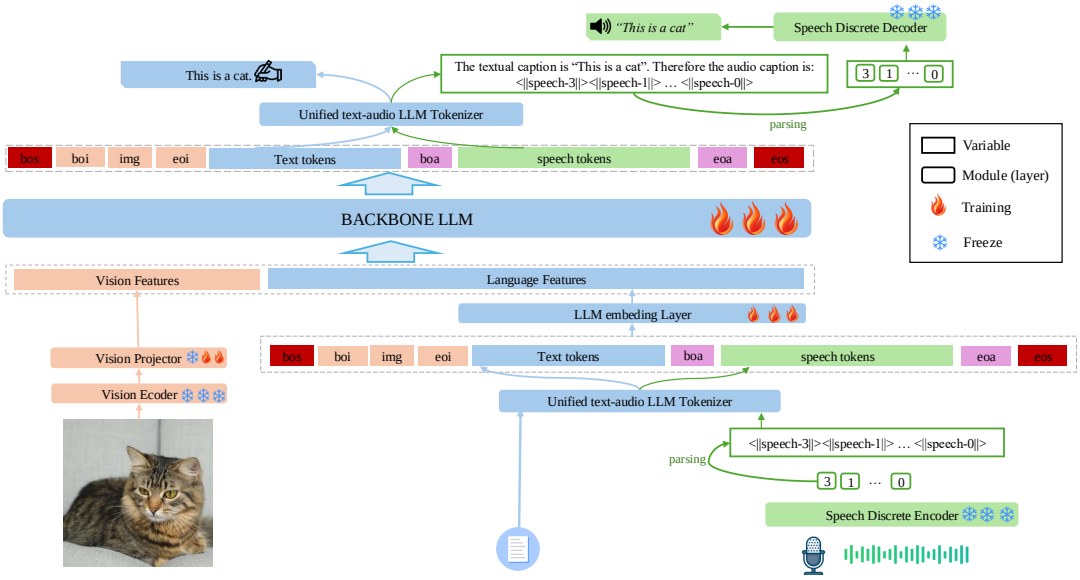


Figure 2: SVLA Architecture

design all tasks to support four input-output configurations: text-to-text, text-to-speech, speech-to-text, and speech-to-speech. Dialogue examples are constructed to fit within a 10-second speech limit and are paired with instruction prompts that guide the model to transition between text, speech, and vision as required. Additional details on the SFT dataset construction are provided in Appendix A.

4 Model Architecture

Our architecture is a hybrid of LLaVA (Liu et al., 2024) and AnyGPT (Zhan et al., 2024), combining visual grounding with speech processing via discrete tokens. It integrates speech and vision inputs into a unified language model, denoted as $f_\theta(x)$, where θ represents model parameters. The overall architecture is shown in Figure 2.

4.1 Vision Encoder

Given an input image X_v, we employ a pretrained VIT-based vision encoder, such as CLIP (Radford et al., 2021), denoted as $g(\cdot)$, to extract high-level visual representations:

$$Z_v = g(X_v), \quad Z_v \in \mathbb{R}^{n_v \times d_v}. \tag{1}$$

Here, Z_v is obtained from the final layer of the vision encoder, where n_v is the number of image patches. Each visual token in Z_v has a feature dimension of d_v.

Next, to align visual embeddings with the LLM's word embedding space, we apply a learnable linear projection layer W_v:

$$H_v = W_v Z_v, \quad H_v \in \mathbb{R}^{n_v \times d_h}. \tag{2}$$

Here, H_v represents the projected visual tokens, now residing in the same dimensional space d_h as the LLM's text embeddings, ensuring compatibility for multimodal fusion.

4.2 Speech Encoder and Tokenization

For speech input X_s, a pretrained speech encoder $s(\cdot)$ generates a sequence of discrete tokens:

$$Z_s = \{q_1, q_2, \ldots, q_{T_s}\}, \quad q_i \in \{1, 2, \ldots, d_s\},$$

where T_s is the number of tokens and d_s is the speech vocabulary size. Each token q_i is mapped to a text-like representation (e.g., $\langle\langle$speech-$i\rangle\rangle$) and added to the LLM's vocabulary, enabling text and speech to be handled uniformly by the tokenizer.

4.3 Multimodal Fusion

Given text tokens Z_t, speech tokens Z_s, and visual embeddings H_v, we construct the multimodal input sequence:

$$X = [\texttt{bos}, \texttt{boi}, \texttt{img}, \texttt{eoi}, Z_t, \texttt{boa}, Z_s, \texttt{eoa}, \texttt{eos}],$$

where $n = |X|$ is the sequence length. Special tokens mark modality boundaries and control flow.

After embedding, we replace the placeholder token `img` with H_v, yielding:

$$E' = [E_{\texttt{bos}}, E_{\texttt{boi}}, H_v, E_{\texttt{eoi}}, E_{Z_t}, E_{\texttt{boa}}, E_{Z_s}, E_{\texttt{eoa}}, E_{\texttt{eos}}],$$

with $E' \in \mathbb{R}^{n' \times d_h}$, where $n' = n_v + n - 1$. The LLM then processes E' for multimodal understanding.

Table 1: Performance of Text-Speech Tasks. We evaluate ASR on LibriSpeech (Panayotov et al., 2015) *test-clean* and TTS on VCTK (Veaux et al., 2017)

Model	Backbone	ASR LibriSpeech WER	TTS VCTK WER	TTS VCTK Similarity
Human-level	-	5.8	1.9	0.93
Wav2vec 2.0 (Papineni et al., 2002)	-	2.7	-	-
Whisper Large V3 (Radford et al., 2023)	-	**1.8**	-	-
VALL-E (Wang et al., 2023)	-	-	7.9	0.75
VILA (Fu et al., 2024)	Mixtral 8x7B (Jiang et al., 2024)	8.1	-	-
AnyGPT-7B	LLaMA-2-7B	8.5	8.5	**0.77**
MIO-Ins (Wang et al., 2024)	Yi-6B-Base (AI et al., 2024)	10.3	**4.2**	-
Qwen2.5-Omni-7B (Xu et al., 2025)	Qwen2.5-7B (AI et al., 2024)	**1.8**	-	-
SVLA-2B	Qwen-1.5B	10.2	21.7	0.65
SVLA-2B-Text-Ins	Qwen-1.5B	8.9	11.2	**0.72**

4.4 Speech Decoding

The multimodal LLM outputs a sequence of predicted tokens, from which the speech-related tokens Z_s' are extracted between the special tokens boa and eoa. A pretrained speech decoder $s^{-1}(\cdot)$ then transforms these discrete tokens back into speech waveforms:

$$X_s' = s^{-1}(Z_s'),$$

where Z_s' represents the predicted discrete speech token sequence, and X_s' is the resulting synthesized speech audio.

4.5 Instruction-Based Speech Generation

During pre-training (see Appendix A.2), the model learns to generate both text and speech tokens directly. In SFT, text responses are always generated directly, while speech outputs follow two alternative strategies: direct generation or instruction-based generation, which we compare experimentally. For simpler tasks like ASR and TTS, we apply lightweight prompts (e.g., *"The transcript of the given speech is:"* for ASR, and *"This is how your text sounds in speech:"* for TTS). For more complex vision-language tasks—such as image captioning and VQA—we adopt an instruction-based approach: the model first generates a textual response, which is then converted into speech using structured prompts like *"The textual caption is 'caption'. Therefore, the audio caption is:"*.

5 Experiments

This section details our experimental setup, the metrics used to assess performance, and the results achieved.

5.1 Implementation Details

We use Qwen2.5-1.5B as our backbone LLM and train it with PyTorch. The vision encoder is CLIP-Large-Patch14-336 (Radford et al., 2021), which produces 256 tokens per image. Speech data is handled by the SpeechTokenizer (Zhang et al., 2023b), which encodes each 1-second segment of audio into 50 discrete tokens. When higher speech fidelity is required, SoundStorm-SpeechTokenizer[2] extends this quantization approach to more nuanced tasks. For speech-input settings in image captioning and VQA, we use Melon-TTS to generate spoken questions or prompts, using the same configuration as in the training set, including a mix of speaking speeds and accents. More complementary details are provided in Appendix B.

5.2 Metrics

Text-Output tasks: For tasks where the model generates text outputs, we use Word Error Rate (WER) for ASR, CIDEr (Vedantam et al., 2015) for image captioning, and accuracy for VQA. These metrics provide a standardized evaluation framework for assessing performance across text-output tasks.

Speech-output tasks: We evaluate speech generation using two methods: We transcribe model-generated speech with Whisper Medium (Radford et al., 2023) and compare the resulting text to human references (as in text-output tasks). However, ASR models can exhibit biases that introduce transcription errors, thereby distorting the perceived quality of the generated speech. Therefore, we use WavLM-TDNN[3]to extract speech embeddings from both generated and reference speech and mea-

[2]https://github.com/ZhangXInFD/soundstorm-speechtokenizer

Table 2: Comparison of models on Image Captioning and VQA tasks. We evaluate Image Captioning on *COCO-Caption-2014*, *COCO-Caption-2017*, and *Flickr8k* datasets, while VQA performance is assessed on *VQAv2-val*, *OKVQA-test*, *GQA-test*, and *VizWiz* datasets. Modalities are denoted as I (Image), T (Text), and S (Speech). * indicates results on the test-dev set.

Model	Backbone	Input → Ouput	Image Captioning			VQA			
			COCO-2014-test	COCO-2017-test	Flickr8k	VQAv2-val	OKVQA-test	GQA-test	VizWiz
TMT (Kim et al., 2024)	-	I→T	108.7	–	79.7		–	–	–
TMT (Kim et al., 2024)	-	I→S	78.7	–	55.2		–	–	–
InstructBLIP (Liu et al., 2024)	Vicuna-7B	I+T→T	–	102.2	82.2	–	33.9	–	33.4
LLaVA (Liu et al., 2024)	LLaMA-2-7B	I+T→T	–	–	82.7	–	–	–	–
LLaVA-1.5 (Liu et al., 2024)	Vicuna-7B	I+T→T	–	–	–	78.5*	–	62.0	50.0
AnyGPT-7B (Tang et al., 2023a)	LLaMA-2-7B	I+T→T	107.5	–	–	–	–	–	–
CoDi (Tang et al., 2023b)	-	I+T→T	149.9	–	–	–	–	–	–
MIO-Ins (Wang et al., 2024)	Yi-6B-Base	I+T→T	120.4	–	–	65.5	39.9		53.5
Next-GPT-7B (Wu et al., 2024)	LLama-7B	I+T→T	158.3	124.9	84.5	66.7	52.1	–	48.4
SVLA-2B	Qwen-1.5B	I+T→T	120.0	117.8	61.4	68.7	45.4	53.3	57.7
	Qwen-1.5B	I+S→T	114.5	107.0	57.7	52.9	25.1	37.7	52.3
	Qwen-1.5B	I+T→S	2.0	2.2	1.7	4.0	4.0	3.7	0.0
	Qwen-1.5B	I+S→S	2.0	2.1	1.1	3.1	0.1	0.0	0.0
SVLA-2B-Text-Ins	Qwen-1.5B	I+T→T	120.2	117.0	67.7	69.7	47.4	52.7	58.0
	Qwen-1.5B	I+S→T	119.4	112.6	59.2	52.7	28.7	38.1	51.7
	Qwen-1.5B	I+T→S	64.7	53.36	49.4	37.5	11.6	29.6	29.8
	Qwen-1.5B	I+S→S	62.2	52.18	46.4	29.4	6.08	23.7	26.1

sure their similarity (via cosine similarity). This directly compares acoustic properties without relying on ASR.

5.3 Results

Text-Speech Performance: As shown in Table 1, SVLA-2B-Text-Ins outperforms SVLA-2B on both ASR and TTS, reducing ASR WER from 10.2 to 8.9 and TTS WER from 21.7 to 12 .2, while improving similarity from 0.65 to 0.72. Prompting with structured text (e.g., "This is the transcript:") boosts performance by providing clearer context. While both models lag behind specialized systems like Whisper and Wav2vec 2.0 (WER 2.7), SVLA-2B-Text-Ins is competitive with AnyGPT-7B. The results highlight a trade-off: direct speech generation is more natural, but instruction-based prompts yield greater clarity and accuracy.

Image Captioning Performance: From Table 2, instruction tuning in SVLA-2B-Text-Ins leads to minimal change in text-only captioning performance. For instance, on COCO-2014, the CIDEr score improves only slightly from 120.0 (SVLA-2B) to 120.2. However, in the speech captioning setting (I+S→S), instruction tuning results in a substantial gain: SVLA-2B scores only 2.0, while SVLA-2B-Text-Ins reaches 62.2. This highlights the effectiveness of structured prompts (e.g., *"The textual caption is ... Therefore, the audio caption is:"*) in guiding coherent speech generation. TMT (Kim et al., 2024), which performs I→S generation directly, achieves a CIDEr score of 78.7 on COCO-2014. While upper SVLA-2B-Text-Ins, it operates under a different paradigm—treating each modality independently—whereas SVLA supports unified multimodal reasoning across both text and speech outputs.

VQA Performance: Table 2 shows that SVLA-2B-Text-Ins performs competitively in the I+T→T VQA setting, despite using the smaller Qwen-1.5B backbone compared to 7B-scale models. While models like LLaVA-1.5 and Next-GPT-7B achieve strong performance on benchmarks such as VQAv2, OKVQA, and GQA, SVLA-2B-Text-Ins achieves comparable results, with a VQAv2 score of 69.7 and a VizWiz score of 58.0. These results demonstrate that a smaller, instruction-tuned model can rival or even surpass larger alternatives.

However, performance declines in the I+S→T setting, where the question is spoken. For example, SVLA-2B-Text-Ins drops to 52.7 on VQAv2 and 28.7 on OKVQA—approximately 15–20 points lower than in the I+T→T setting. This suggests that the model performs more effectively when the input question is provided in text rather than speech. Speech output accuracy is generally lower than text output across VQA tasks. In the I+T→S setting, SVLA-2B-Text-Ins achieves 37.5 on VQAv2 and 11.6 on OKVQA, whereas in I+S→S, the scores fall to 29.4 and 6.08, respectively. This drop highlights the difficulty of reasoning directly from speech inputs and generating accurate spoken responses. These results suggest that using text

[3] https://github.com/yangdongchao/UniAudio/blob/main/UniAudio/tools/evaluation/compute_similarity_vc.py

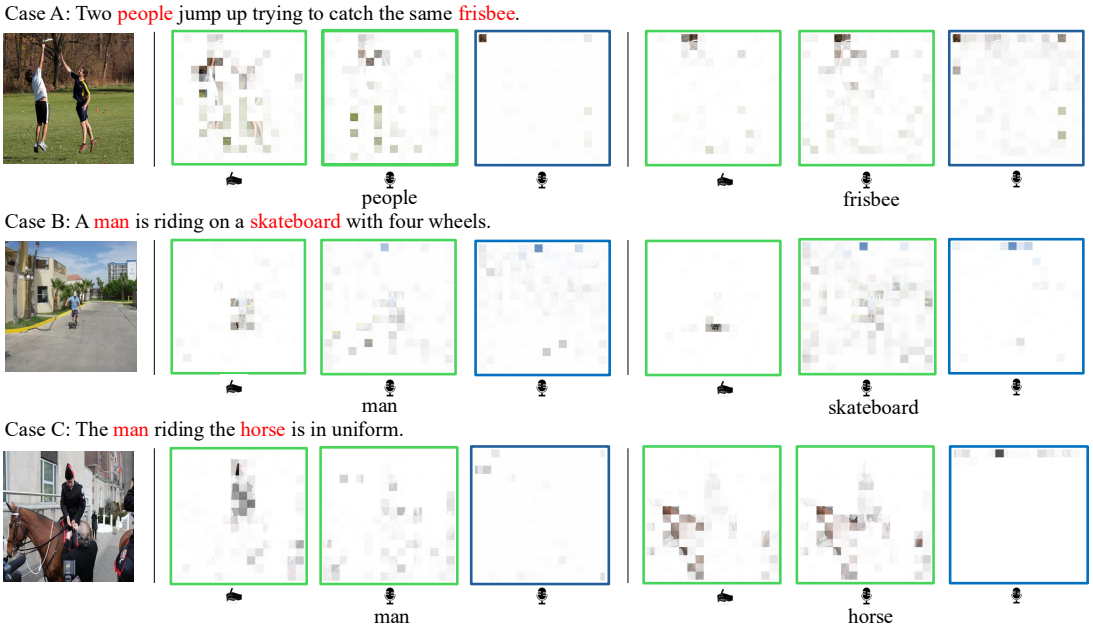
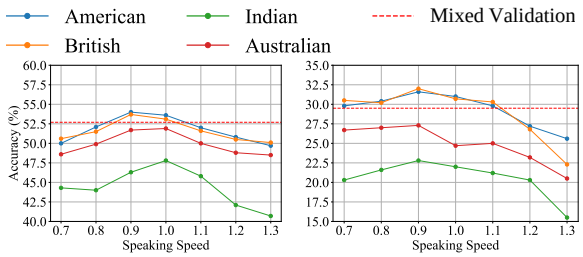


Figure 3: Visualization of attention maps comparing SVLA-2B's visual grounding accuracy with and without intermediate textual instructions during speech generation.

as an intermediate representation enhances semantic alignment, particularly for complex reasoning tasks. By contrast, SVLA-2B completely fails to handle VQA tasks.

5.4 Ablation Studies

Effect of Accent and Speaking Speed:



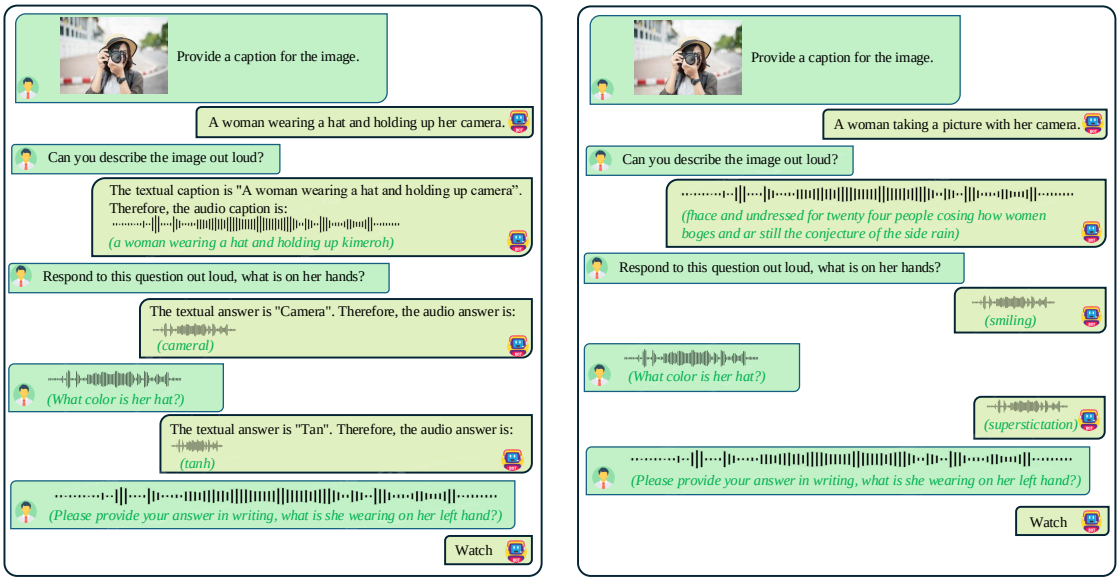
(a) I+T→S (b) I+S→S

Figure 4: Impact of accent and speaking speed on VQA performance. Both plots show accuracy on VQAv2-*val*.

We evaluate the impact of different accents and speaking speeds on VQA accuracy using the SVLA-2B-Text-Ins model on the VQAv2-*val* set. As show in table 4, in both the I+S→T and I+S→S settings, American and British accents yield higher performance, while Indian and Australian accents result in noticeably lower accuracy. Notably, despite being included in the training data, the Indian accent still underperforms, indicating potential challenges in generalization or speech variability. In terms of speaking speed, the model performs best around the default rate (1.0×), with accuracy dropping at both extremes. The decline is most pronounced at 1.3×, suggesting that faster speech reduces recognition and reasoning quality.

Where Do the Models Look in Images?: To examine how different speech generation strategies affect visual grounding, we visualize the model's attention maps in Figure 3. For each key word, we show three attention maps: the first green map is from the text output of SVLA-2B-Text-Ins, the second green map from its instruction-based speech output; and the blue map from SVLA-2B's direct speech output.The text output from SVLA-2B-Text-Ins exhibits the most focused and accurate attention, precisely grounding visual entities. Its speech output (green) generally retains meaningful grounding, showing acceptable attention consistency. In contrast, the speech output from SVLA-2B (blue) is often unfocused or misaligned, failing to attend to the relevant image regions. These results demonstrate that textual instructions play a critical role in guiding the model's visual attention. Without the intermediate text step, the model lacks semantic anchoring and often fails to locate the correct objects in the image, leading to degraded visual grounding during speech generation.

The Limits of ASR-Based Evaluation: ASR-based evaluation falls short in assessing speech output quality, as it often misinterprets minor phonetic variations as errors—even when the spoken response is semantically accurate. As shown in Fig-



(a) SVLA-2B-Text-Ins (b) SVLA-2B

Figure 5: Comparison of SVLA-2B-Text-Ins and SVLA-2B in Multimodal Image Captioning and VQA Responses.

ure 5a, the model's speech output closely matches the intended text, yet ASR transcribes words like *"camera"* as *"camaral"*, or *"tan"* as *"tanh"*. These subtle differences, while perceptually acceptable, are unfairly penalized because ASR prioritizes exact word-level matching over acoustic or semantic similarity. This limitation underscores the need for more robust evaluation metrics that go beyond transcription accuracy. In particular, speech-based VQA and captioning tasks would benefit from metrics that directly assess the fidelity of the generated speech waveform—capturing both semantic correctness and acoustic naturalness—without relying solely on error-prone intermediate transcriptions.

6 Ethical Considerations

We utilize publicly available datasets containing licensed image-text and speech-text pairs. All speech samples are either synthetic or derived from open corpora, explicitly excluding personal or sensitive data. While our datasets incorporate diverse accents and varying speaking rates to enhance representativeness, synthetic speech may still not capture the complete variability inherent in natural human speech. Observed performance disparities across different conditions highlight the necessity of ongoing research in fairness and robustness.

7 Limitations:

Our study has several limitations. First, the speech data is generated using a TTS model, which may lack the natural prosody and emotional variation of real human speech. Despite augmentations in accent, speed, and noise, the resulting speech may still be less diverse than natural input. Second, we use the Qwen2.5-1.5B backbone due to resource constraints, which limits model capacity. Third, the speech tokenizer introduces decoding errors, reducing intelligibility even when the underlying text is accurate.

8 Conclusion

In this work, we introduce SVLA, a unified Speech-Text-Vision Assistant capable of handling both language tasks (ASR, TTS) and vision-language reasoning tasks (image captioning, VQA). To support the community in building similar models, we also release a large-scale tri-modal dataset encompassing speech, text, and vision. Additionally, we analyze two settings for speech generation: directly producing spoken output and using a text prompt to guide speech synthesis. Our experiments show that while the model performs better with text outputs, speech outputs benefit from an instructive text prompt, yielding more coherent. In future work, we plan to incorporate real human speech, improve speech tokenization quality, and explore larger model backbones to better support nuanced prosody, robustness to speech variability, and high-fidelity speech generation.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, and 13 others. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. 2024. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, and 1 others. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

William Havard, Laurent Besacier, and Olivier Rosec. 2017. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. *arXiv preprint arXiv:1707.08435*.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. 2020. Text-free image-to-speech synthesis using learned segmental units. *arXiv preprint arXiv:2012.15454*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. 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 119–132.

Minsu Kim, Jee-weon Jung, Hyeongseop Rha, Soumi Maiti, Siddhant Arora, Xuankai Chang, Shinji Watanabe, and Yong Man Ro. 2024. Tmt: Tri-modal translation between speech, image, and text by processing different modalities as different languages. *arXiv preprint arXiv:2402.16021*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

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. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2023. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*.

OpenAI. 2023a. Chatgpt4. OpenAI API. Accessed: [13/17/2023].

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Gpt-4o.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Christoph Schuhmann, Andreas Köpf, Theo Coombes, Richard Vencu, Benjamin Trom, and Romain Beaumont. 2022. Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023a. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023b. Any-to-any generation via composable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, and 1 others. 2024. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv:2409.17692*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, and 1 others. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, and 1 others. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

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

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. Melotts: High-quality multi-lingual multi-accent text-to-speech.

A Training

A.1 Data Generation

Table 1 provides a detailed breakdown of the pre-training dataset, categorized into two stages: Stage 1, which consists solely of text-speech tasks, and Stage 2, which expands to include image-text-speech tasks. In Stage 1, the dataset comprises 2.0M samples from the Libriheavy corpus for TTS and ASR tasks, yielding a total of 1.5M text tokens, 7.4B speech tokens, and approximately 8.2K hours of speech. Stage 2 significantly scales up the dataset, incorporating Libriheavy (6M samples) alongside multimodal datasets such as Laion and VG, covering image-captioning (IC) and VQA tasks in both text and speech modalities. The total dataset spans 34.3M samples, 615.9M text tokens, 9.1B speech tokens, and 50.8K hours of speech, making it one of the most extensive speech-text-vision corpora for multimodal learning. Notably, the dataset supports a diverse set of multimodal tasks, including IC-TTT, IC-TTS, IC-STT, IC-STS, VQA-TTT, VQA-TTS, VQA-STT, and VQA-STS, ensuring broad coverage across different input-output combinations.

Table 2 presents the SFT dataset, covering both text-speech and image-text-speech tasks. The dataset includes 2.5M samples, with 308K image samples, 152M text tokens, and 920M speech tokens, totaling 33.9M seconds (5102 hours) of speech data. The text-speech tasks include 150K ASR samples from Librispeech and 388K TTS samples from CommandVoice, contributing 496 and 559 hours of speech, respectively. In the image-text-speech tasks, various VQA datasets (VQA, A-OKVQA, GQA, VizWiz) and COCO-Caption-2014 support text-based (VQA-TTT, IC-TTT), text-to-speech (VQA-TTS, IC-TTS), speech-to-text (VQA-STT, IC-STT), and speech-to-speech (VQA-STS, IC-STS) tasks. The dataset is diverse and well-balanced, ensuring broad multimodal coverage for fine-tuning models on speech, text, and vision-related tasks. Figure 1, Figure 2, Figure 3, and Figure 4 show the prompts of tasks used to train the models.

A.2 Training Strategy

Our model is trained in three sequential phases, progressively increasing multimodal complexity to enhance stability, efficiency, and modality integration.

Stage 1 - Text-Speech Pre-Training The model learns text-speech alignment through ASR and TTS on large-scale paired datasets, establishing a strong linguistic foundation before incorporating vision.

Stage 2 - Vision-Text-Speech Pre-Training Training expands to include vision-based tasks like VQA and image captioning alongside ASR and TTS, enabling the model to unify vision, text, and speech representations.

Stage 3 - Supervised Fine-Tuning The model is refined through supervised fine-tuning on benchmark-aligned datasets, focusing on multi-turn conversations, history retention, and multi-modal reasoning.

Dataset	Task	No. Sample	Image	Text Tokens	Speech Tokens	Speech Duration (s)	Speech Duration (h)
Pretrain Dataset - Stage 1							
Text-Speech tasks							
Libriheavy	TTS	1.0M	0	56.7M	730M	14.6M	4.1K
	ASR	1.0M	0	60.0M	736M	14.7M	4.1K
Total	-	2.0M	0	1.5M	7.4B	29.3M	8.2K
Pretrain Dataset - Stage 2							
Text-Speech tasks							
Libriheavy	TTS	3.0M	0	151.6M	2.2B	44.2M	12.3K
	ASR	3.0M	0	207.6M	2.2B	44.4M	12.3K
Image-Text-Speech tasks							
Laion	IC-TTT	5.8M	5.8M	111.1M	0	0	0
	IC-TTS	5.8M		48.4M	1.1B	22.0M	6.1K
	IC-STT	5.8M		63.1M	1.0B	20.2M	5.6K
	IC-STS	5.8M		0	1.9B	38.6M	10.7K
VG	VQA-TTT	1.3M	108K	12.1M	0	0	0
	VQA-TTS	1.3M		19.1M	97.0M	1.9M	538
	VQA-STT	1.3M		2.9M	321.0M	6.4M	1.8K
	VQA-STS	1.3M		0	249.6M	182.4M	1.4K
Total	-	34.3M	23.5M	615.9M	9.1B	182.8M	50.8K

Table 1: Statistics of the pretraining dataset

Dataset	Task	No. Sample	Image	Text Tokens	Speech Tokens	Speech Duration (s)	Speech Duration (h)
Text-Speech tasks							
Librispeech	ASR	150K	0	5.4M	89.2M	1.8M	496.0
CommandVoice	TTS	388K	0	4.8M	101.8M	2.0M	559.3
Image-Text-Speech tasks							
VQA	VQA-TTT	84K	83K	4.7M	0	0	0
	VQA-TTS	42K		3.0M	13M	260K	72
	VQA-STT	42K		282K	46.1M	921K	256
	VQA-STS	84K		0	86.0M	17.2M	476
A-OKVQA	VQA-TTT	50K	50K	520.0K	0	0	0
	VQA-TTS	25K		333K	1.4M	28K	8
	VQA-STT	25K		31K	5M	102K	28
	VQA-STS	50K		0	9.5M	190K	53
GQA	VQA-TTT	72K	72K	11.5M	0	0	0
	VQA-TTS	36K		6.7M	26.8M	536K	149
	VQA-STT	36K		528.5K	100.5M	2.0M	558
	VQA-STS	72K		0	168.0M	3.4M	933
VizWiz	VQA-TTT	20K	20K	780.5K	0	0	0
	VQA-TTS	10K		413K	883K	17.7K	5
	VQA-STT	10K		28.8K	5.7M	114.0K	31.7
	VQA-STS	20K		0	11.5M	230.7K	64.1
COCO-Caption-2014	IC-TTT	414K	83K	107.0M	0	0	0
	IC-TTS	212K		3.3M	39.4M	789K	219
	IC-STT	212K		2.4M	51.3M	1M	285
	IC-STS	414K		0	163.7M	3.3M	909
Total	-	2.5M	308K	152M	920M	33.9M	5102

Table 2: Statistics of the SFT dataset

ASR (Automatic Speech Recognition) Prompts:

- “Please convert this audio to text: "
- “Transcribe the following audio file, please: "
- “Can you convert this speech to text? "
- “Generate text from this audio recording: "
- “Please write out what’s being said in this audio: "
- “Turn this voice recording into text, please: "
- “Please create a transcript of this audio: "
- “Can you transcribe this audio? "
- “Convert this spoken content into written text: "
- “Please extract text from this speech: "
- “Transcribe the spoken words in this audio file: "
- “Create a written version of this audio: "
- “Convert the spoken words to text: "
- “Please generate a transcript from this recording: "
- “Transform the audio into a text document: "

Example Usage:

```
"Please convert this audio to text: <speech_start>{speech_tokens}<speech_end>"
```

Figure 1: Prompts for ASR Tasks.

TTS (Text-to-Speech) Prompts:

- “Please convert this text to speech: "
- “Turn this text into audio, please: "
- “Generate speech from this text: "
- “Please speak out this text: "
- “Convert these words to speech, please: "
- “Please make an audio version of this text: "
- “Can you read this text aloud: "
- “Transform this text into speech: "
- “Please give a voice to this text: "
- “Read out this text, please: "
- “Create a spoken version of this text: "
- “Convert the written text to speech: "
- “Turn the following words into sound: "
- “Provide an audio rendition of this text: "
- “Generate an audio file of these words: "

Example Usage:

```
"Turn this text into audio: {transcript}"
```

Figure 2: Prompts for TTS Tasks.

Captioning Prompts:
IC_TTT (Text-to-Text) and Caption_STS (Speech-to-Speech):

- “What do you see in the image?"
- “Explain what is shown in the picture."
- “Provide a caption for the image."
- “Describe the objects or people in the image."

Example Usage:

```
"<image>\nWhat do you see in the image?"
```

IC_TTS (Text-to-Speech):

- “Can you describe the image out loud?"
- “Read the description of the image aloud."
- “Turn the image caption into spoken words."
- “Provide a spoken description of the picture."

Example Usage:

```
"<image>\nCan you describe the image out loud?"
```

IC_STT (Speech-to-Text):

- “Write down what you see in the image."
- “Can you write a detailed description of the picture?"
- “Write a summary of the scene in the image."
- “Write down the elements present in the picture."

Example Usage:

```
"<image>\n Write down what you see in the image."
```

Figure 3: Prompts for Image Captioning Tasks.

VQA (Visual Question Answering) Prompts:
VQA_TTT (Text-to-Text) and VQA_STS (Speech-to-Speech):
Example Usage:

```
"<image>\n{question}\nAnswer the question using a single word or phrase."
```

VQA_STT (Speech-Text-to):

- “Please provide your answer in writing."
- “Respond to the question with a written explanation."
- “Answer this question using text."
- “Type your response to the question."
- “Write down your answer clearly."
- “Provide a detailed answer in text format."
- “Explain your response in written form."
- “Answer the question by typing a full response."

Example Usage:

```
"<image>\nPlease provide your answer in writing.\n{question}\nAnswer the quest
-ion using a single word or phrase."
```

VQA_TTS (Text-to-Speech):

- ‘Respond to this question out loud."
- ‘Please give your answer verbally."
- ‘Provide a spoken response to the question."
- ‘Answer this question using speech."
- ‘Explain your response in spoken form."

Example Usage:

```
"<image>\nRespond to this question out loud.\n{question}\nAnswer the quest
-ion using a single word or phrase."
```

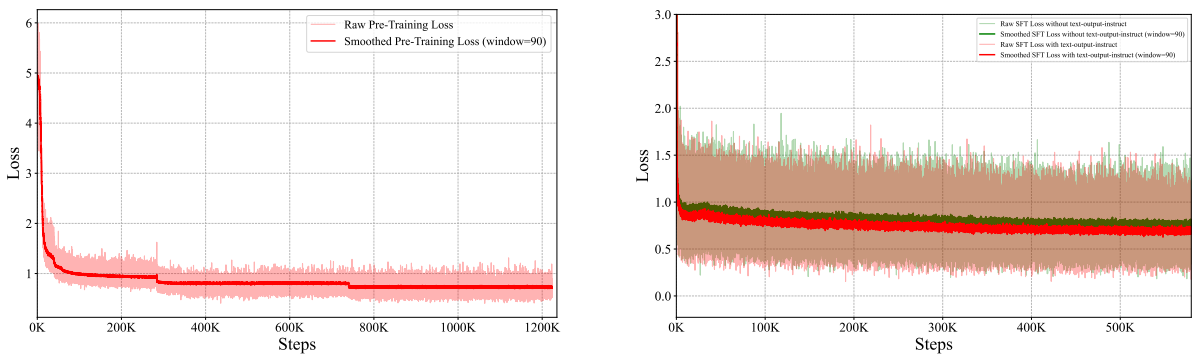
Figure 4: Prompts for VQA Tasks.

B Implementation Details

Figure 5 compares the loss curves for pre-training and supervised fine-tuning (SFT). Figure 5a plot shows the pre-training loss, where the raw loss exhibits high variance but progressively decreases, stabilizing after around 400K steps. The smoothed loss curve (window=90) highlights a consistent downward trend, indicating stable convergence. The 5b illustrates the SFT loss, comparing models with and without text-output instructions. While both configurations show a decreasing trend, the model trained with text-output instructions (red) achieves consistently lower loss than the version without instructions (green), suggesting that structured textual guidance improves fine-tuning efficiency and convergence. The variance in SFT loss remains higher compared to pre-training, reflecting the increased complexity of supervised instruction tuning.

Table 3: Implementation Details for Model, Pre-Training, and Supervised Fine-Tuning

Model and Hardware		
LLM	Qwen2.5-1.5B	
Vision Encoder	openai/clip-vit-large-patch14-336	
Speech Encoder	SpeechTokenizer (1s $\rightarrow$ 50 tokens),	
Speech Decoder	SoundStorm-SpeechTokenizer	
Frameworks	PyTorch, DeepSpeed	
Training Configurations		
	Pre-Training	**Supervised Fine-Tuning**
Batch Size	32	8
Epochs	1	2
Optimizer	Adam	Adam
Learning Rate	2×10^{-5}	1×10^{-5}
Warmup Ratio	0.03	0.02
LR Scheduler	Cosine	Cosine
Training Steps	1.2M	600K
Training Epochs	1	2
Maximum Tokens	4096	4096
GPUs	8 x NVIDIA H100	4 x NVIDIA H100
DeepSpeed Config	`zero2`	



(a) Pre-training Loss

(b) SFT loss

Figure 5: Comparison of Pre-training and SFT Loss Curves

C Evaluation Details

Generated Modality	Text	Speech
Beam size	5	1
Top-P	-	0.7
maximum	256	1024
Repetition Penalty	1.0	1.3

Table 4: Comparison of Text and Speech Generation Settings in evaluation.

ASR Prompt:

"Please convert this audio to text: <speech_start>{speech tokens}<speech_end>."

TTS Prompt:

"Please convert this text to speech: {transcript}."

IC_TTT and IC_STS Prompt:

"<image>\nrovide a caption for the image."

IC_TTS Prompts:

"<image>\nTurn the image caption into spoken words"

IC_STT Prompts:

"<image>\nWrite down what you see in the image."

VQA_TTT and VQA_STS Prompt:

"<image>\n{question}\nAnswer the question using a single word or phrase."

VQA_TTS Prompts:

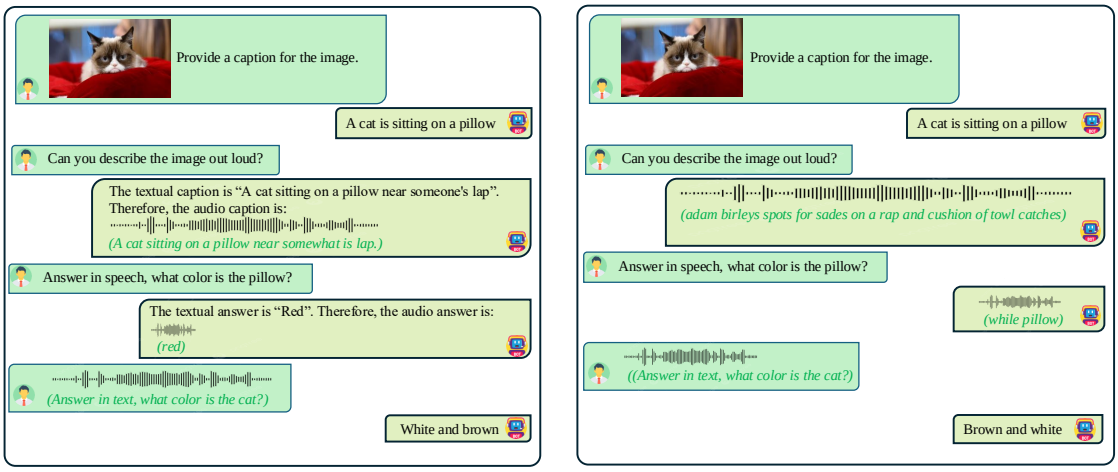
"<image>\nAnswer this question using speech.{question}\nAnswer the question using a single word or phrase."

VQA_STT Prompts:

"<image>\nAnswer this question using text.{question}\nAnswer the question using a single word or phrase."

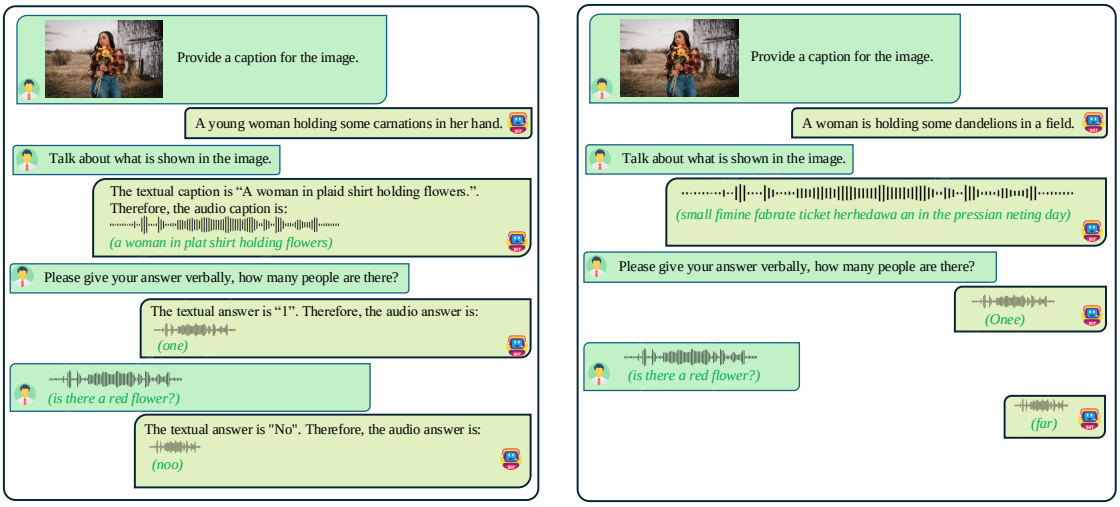
Figure 6: Evaluation Prompts for different tasks.

D Examples



(a) SVLA-2B-Text-Ins (b) SVLA-2B

Figure 7: Comparison of SVLA-2B-Text-Ins and SVLA-2B in Multimodal Image Captioning and VQA Responses (Example 2).



(a) SVLA-2B-Text-Ins (b) SVLA-2B

Figure 8: Comparison of SVLA-2B-Text-Ins and SVLA-2B in Multimodal Image Captioning and VQA Responses (Example 3).