

# MLLM-based Speech Recognition: When and How is Multimodality Beneficial?

Yiwen Guan, Viet Anh Trinh, Vivek Voleti, and Jacob Whitehill

**Abstract**—Recent advances in multi-modal large language models (MLLMs) have opened new possibilities for unified modeling of speech, text, images, and other modalities. Building on our prior work [1], this paper examines the conditions and model architectures under which multiple input modalities can improve automatic speech recognition (ASR) accuracy in noisy environments. Through experiments on synthetic and real-world data, we find that (1) harnessing more modalities usually improves ASR accuracy, as each modality provides complementary information, but the improvement depends on the amount of auditory noise. (2) Synchronized modalities (e.g., lip movements) are more useful at high noise levels whereas unsynchronized modalities (e.g., image context) are most helpful at moderate noise levels. (3) Higher-quality visual representations consistently improve ASR accuracy, highlighting the importance of developing more powerful visual encoders. (4) Mamba exhibits similar trends regarding the benefits of multimodality as do Transformers. (5) The input order of modalities as well as their weights in the loss function can significantly impact accuracy. These findings both offer practical insights and help to deepen our understanding of multi-modal speech recognition under challenging conditions.

**Index Terms**—Automatic speech recognition, Large language models, Multi-modal learning.

## I. INTRODUCTION

GIVEN the success of large language models (LLMs) for natural language processing as enabled by their reasoning and contextual understanding abilities, researchers are increasingly exploring how to develop multi-modal LLMs (MLLMs) to harness multiple input modalities, particularly in areas involving speech and vision [2–4]. The evolution of LLMs – specifically defined as large-scale autoregressive decoder-only models – has led to the emergence of LLM-based automatic speech recognition (ASR) systems in the past few years [5–7]. LLM-based ASR models typically adopt decoder-only architectures, taking either discrete units (e.g., extracted from HuBERT [8]) or continuous features (e.g., Log-Mel filterbank) as input, and producing text transcriptions as output. While conventional speech processing methods consider only the audio itself, MLLM-based speech recognizers jointly model and process multiple input modalities, such as subject-matter context or visual cues like images and the speaker’s lip movements [9–11]. By exploiting complementary multi-modal information, these systems can sometimes reach higher accuracy in challenging conditions: noisy environments [12, 13], ambiguous words, long-tail vocabularies, and accented speech [14–16]. For example, in noisy conditions, the scene context can resolve ambiguities between words like

“door” and “dough”, while lip movements help differentiate between “night” and “right”; or like the example in Fig. 2 that an image of a man in a fedora can help clarify the descriptive speech. They also benefit from LLM pre-training which endows them with linguistic information about which transcripts are more or less likely in a given context.

MLLM-based approaches hold significant promise for next-generation ASR systems but they also bring new challenges: While multiple input modalities may contain complementary information, they also increase the sequence length and complexity, making modeling more difficult. Each modality can be either beneficial or harmful in different situations that remain to be explored. Furthermore, it is still unclear what input formatting strategy – including the order of inputs, positional encodings, and modality-specific loss weights – is most effective in enabling the model to fully exploit multi-modal information. Moreover, Transformer-based models are inefficient when processing longer sequences due to their quadratic complexity attention mechanism, which sparks the research interest in faster architectures like Mamba [17]. These challenges also raise questions about where researchers should focus to achieve better ASR performance: improving LLM architectures that can handle longer and more complex sequences, or enhancing modality-specific encoders to provide higher-fidelity representations?

In this paper, we investigate when and how multiple input modalities (speech audio, visual cues, and lip movements) are beneficial in multi-modal ASR. We systematically examine the conditions and model architectures under which each modality can improve ASR accuracy in noisy environments. This article is an extended version of our earlier conference paper [1], in which we initially analyzed the modality impact based on a multi-modal language model named DMLM [5]. DMLM (described in Section III-B) is a discrete-token based unified model that processes multiple input modalities and performs various tasks with decoder-only backbone, including multi-modal ASR. While our previous work investigated when and how each modality helps, here we provide a more in-depth analysis, expanded experiments, and several new contributions, as outlined below:

- 1) Using both tightly-controlled synthetic as well as real-world datasets we examine the *conditional* accuracy benefit of multi-modal inputs to MLLM-based ASR systems as a function of the auditory noise level.
- 2) We assess how the quality and efficiency of visual encodings affects ASR accuracy in MLLM architectures.
- 3) We compare Transformer to Mamba as the MLLM backbone to assess the speed/accuracy trade-off as well as the difficulty of training.

Yiwen Guan, Viet Anh Trinh, Vivek Voleti, and Jacob Whitehill are all with Worcester Polytechnic Institute, MA 01609, US (e-mail: yguan2@wpi.edu; vtrinh@wpi.edu; vsvoleti@wpi.edu; jrwhitehill@wpi.edu).

- 4) We compare different input strategies – such as interleaving and swapping the modalities – to demonstrate their impact on MLLMs, thus suggesting new directions for optimizing multi-modal systems.

## II. RELATED WORK

### A. Multi-modal large language models

Recent advances in MLLMs have demonstrated impressive capabilities in integrating information from multiple input streams to understand and reason across multiple modalities. Among these works, models like LLaVA [18] and Qwen-Audio [19] focus on multi-modal comprehension tasks, such as answering questions grounded in visual scenes or spoken content, while other approaches, including NExT-GPT [3] and MM-Interleaved [20], extend unimodal LLMs to handle multi-modal generation tasks, such as rendering images or videos from textual descriptions or synthesizing speech from prompts.

A critical aspect of LLM design is its computational backbone, and in recent years researchers have explored replacing the most common Transformer architectures with more efficient ones. One notable alternative is the linear state-space model (SSM), such as Mamba [17] and Mamba-2 [21], which offers better scalability and efficiency than Transformers by avoiding the expensive attention mechanism. Cobra [22] addresses the efficiency bottleneck of Transformer-based MLLMs by adopting pre-trained Mamba models and performs  $3\times \sim 4\times$  faster than the most efficient SOTA Transformer methods. U-Mamba [23] leverages the powerful capability of SSMs in handling long sequences and proposes a CNN-SSM architecture to extract local features and long-range dependencies in biomedical image segmentation, which outperforms SOTA models in all tasks.

Understanding how MLLMs exploit multi-modal information is also crucial, so a growing body of work focuses on how different modalities interact within the model and how they contribute to overall performance [7, 9]. They highlight the importance of effective modality fusion, cross-modality alignment, and robustness to modality degradation. Some work examines the effect of different model sizes, visual encoders, and modality connectors in ablation study to compare their effectiveness [24]. Other research raises research questions on how to quantify modality utility and how to formalize why modalities are useful or harmful [25]. However, they do not conduct systematic experiments or quantitatively analyze the contribution of each modality and rarely discuss the importance of secondary modalities, especially in noise and across datasets. These areas are closely aligned with our goal of analyzing and quantifying the contribution of each modality in the context of speech recognition under noisy conditions.

We also note the rise of diffusion-based language models, such as Diffusion-LM and LLaDA [26, 27]. As our focus is on autoregressive models, we omit diffusion-based language models from our study.

### B. Speech & audio-visual systems

Visual speech recognition (VSR), often known as automatic lip-reading, aims to recognize the speech content with the

speaker’s lip movements. As a broader combination of ASR and VSR, audio-visual speech recognition (AVSR) integrates both auditory and visual data, such as lip movements or unconstrained images, to improve speech recognition performance. Previous work [13] has analyzed the robustness of ASR, VSR, and AVSR models under different input corruption types for each modality. However, they only consider synchronized lip motions as visual information and overlook unsynchronized information like static images. Our work takes into account unsynchronized visual inputs and different model architectures.

Recent research illustrates the effectiveness of LLMs for ASR, VSR, and AVSR [12, 19, 28–31]. The integration of LLMs enables stronger multi-modal perception and contextual reasoning, which is particularly useful in noisy or ambiguous conditions. In ASR, LLMs empower the system with stronger linguistic modeling, contextual adaptation, and generalization capabilities. In AVSR, initiatives like Llama-AVSR [12] incorporate powerful LLMs to more effectively combine auditory and visual cues under noisy conditions.

While conventional AVSR systems primarily rely on synchronized audio and lip motion streams, the adoption of LLMs opens up more flexible and generalizable architectures for multi-modal speech recognition. For instance, AVFormer [32] incorporates unconstrained visual frames, while [33] utilizes gestures as the visual modality, translating visual information into LLM-compatible representations for joint modeling. Zero-AVSR [34] leverages the pre-trained LLMs’ Roman-grapheme modeling capabilities to achieve zero-shot language recognition ability. Other approaches, including Llama-MTSK [35] and MMS-LLaMA [11], focus on optimizing audio-visual representations and introducing token compression techniques to reduce computational overhead without sacrificing accuracy. Despite these advancements, a fundamental question remains: how do individual modalities contribute to overall performance? A deeper understanding of modality contributions is essential for designing efficient, robust, and adaptive multi-modal systems for speech recognition.

## III. METHODOLOGY

This work aims to understand the conditions in which multiple input modalities can improve speech recognition accuracy in MLLMs. We consider scenarios in which multiple modalities supplementary to the speech are available, including both *synchronized* modalities such as lip movements (i.e., there is a temporal correspondence between the speech wave and the lip image sequence) as well as *unsynchronized* modalities such as a picture of what the person is talking about (e.g., a lecture slide or a document). Moreover, we explore how the amount of auditory and/or visual noise can impact the benefit of multimodality. Using both highly-controllable synthetic as well as real-world data, we systematically compare models with different modality combinations under varying noise conditions in order to quantify the accuracy benefits.

### A. Evaluation Metrics

Let  $\mathcal{M} = \{M_1, \dots, M_{|\mathcal{M}|}\}$  be the set of all modalities (e.g., image, speech, etc.);  $m$  refers to one specific modality in  $\mathcal{M}$ .

In our experiments we use three different metrics (WER, RB, and PS), which we report as percentages.

1) *Word Error Rate*: As a base metric to convey how accurate an ASR model is, we post-process the output transcriptions using the Whisper English text normalizer [36] and then calculate the word error rate (WER):

$$\text{WER} = \frac{n_S + n_D + n_I}{N} = \frac{n_S + n_D + n_I}{n_S + n_D + n_C} \quad (1)$$

where  $n_S$ ,  $n_D$ ,  $n_I$ ,  $n_C$ , and  $N$  represents the number of substitutions, deletions, insertions, correct words, and reference words, respectively. Lower WER is better.

2) *Relative Benefit of WER*: To evaluate how much the WER decreases by adding one (or more) additional modalities to the audio-to-text task, we calculate the relative benefit (RB):

$$\text{RB}(S) = \frac{\text{WER}(\{A\}) - \text{WER}(\{A\} \cup S)}{\text{WER}(\{A\})} \quad (2)$$

where  $A$  denotes audio and  $S \subseteq \mathcal{M} \setminus \{A\}$  stands for the set of additional modalities. RB can be used as an assessment of the extent to which the additional modalities help the existing model. Larger RB values imply greater benefit from adding the extra modalities. Negative values indicate that the extra modalities hurt rather than help.

3) *Perceptual Score of WER*: To evaluate the influence of each modality in a model that harnesses all modalities, we compute the perceptual score (PS) [37] for each modality  $m$ :

$$\text{PS}(m) = \frac{\text{WER}(\mathcal{M} \setminus \{m\} \cup \{\tilde{m}\}) - \text{WER}(\mathcal{M})}{\text{WER}(\mathcal{M})} \quad (3)$$

where  $\tilde{m}$  indicates that modality  $m$  of each example has been randomly swapped with another example's modality  $m$ . Larger PS indicates that  $m$  is more influential.

The relative benefit (RB) and perceptual score (PS) are subtly different: RB estimates the performance improvement of adding new modalities to a model with fewer modalities; this is useful when deciding whether to use an audio-only model versus a multi-modal model. In contrast, PS estimates the influence of each modality in a fully multi-modal model.

## B. Model Architecture and Pipeline

**Model Architecture**: Our experiments are performed on a Discrete Multi-modal Language Model (DMLM, depicted in Fig. 1) [5], a discrete token-based Transformer decoder-only model with an OPT-125M backbone [38] that was pre-trained on a variety of multimodal tasks including speech recognition, speech translation, image generation, and image captioning, using LibriSpeech-960, CVSS, CoVoST2, and COCO [39–42]. DMLM tokenizes the input data into discrete tokens using several frozen modality-specific encoders: we use Seamless [43] for audio, AV-HuBERT [44] for lip movements, and either DALL-E [45] or ViT [46] for images. The input tokens of each modality are concatenated to form an input sequence, then the entire sequence is fed into OPT to be processed like text tokens. We use absolute positional encodings that are trained and added element-wise to the embedded input tokens. This pipeline is similar to other works like SpeechGPT [28] which takes multi-modal inputs and generates multi-modal outputs.

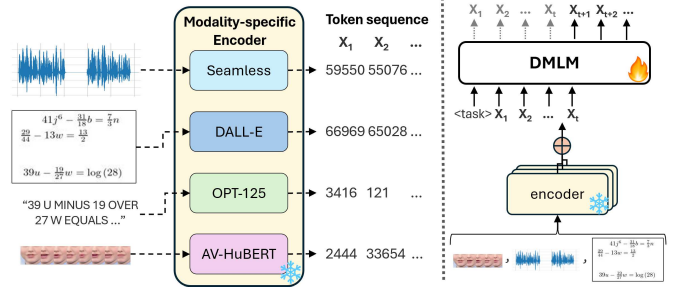


Fig. 1. An overview of Discrete Multimodal Language Model (DMLM). The modality inputs are first tokenized, and then passed to DMLM for processing. The  $\oplus$  in the figure represents modality vector concatenation.

As an alternative to OPT as the backbone, we also explored state space models using Mamba-130M.

For multimodal models, an important consideration is the positional encoding that helps the model to find relevant information. Although we experimented with different positional embedding strategies (see Supplementary Materials Section II), we found that absolute encodings worked the best.

**Training Details**: Following prior work, we feed the model with training data in the format: “< $m_1$  tokens>< $m_2$  tokens>...< $m_n$  tokens><text label (tokens)>”. The number of modalities ( $n$ ) and the order of modalities vary in each task. We trained each model with AdamW [47] with  $\beta=(0.9, 0.999)$ , weight decay of  $1e-4$ , learning rate of  $1e-6$  (unless specified). All experiments are conducted on a single NVIDIA A100 GPU with early stopping patience of 5.

**Inference Details**: The data are input in the format: “< $m_1$  tokens>< $m_2$  tokens>...< $m_n$  tokens>”. Then the model generates the corresponding text tokens autoregressively. To ensure determinism of the results, we set the temperature to 0 and use greedy decoding (no beam search) to generate tokens. Only the text part is used for evaluation.

## C. Modality-Weighted Loss Function

MLLMs trained autoregressively to predict the entire input sequence, which comprises multiple modalities, sometimes have hyperparameters ( $\lambda_S, \lambda_T, \lambda_I$ , etc.) controlling how much each token type (speech, text, image, respectively) influences the loss function. Previous work [5] found a strong benefit, in terms of both training stability and testing accuracy, in tuning these hyperparameters in a multi-modal loss of the form

$$\mathcal{L} = \sum_{m \in \mathcal{M}} \frac{\lambda_m}{n_m} \mathcal{L}_{CE}(m) \quad (4)$$

where  $\mathcal{L}_{CE}(m)$  is the cross-entropy loss for predicting tokens and  $n_m$  is the number of tokens in modality  $m$ . In particular, since  $n_m$  can differ significantly over modalities, judicious tuning of the  $\lambda$  can prevent any one modality from dominating the loss. More subtly, previous work [44, 48] has found that, in audio-visual speech recognition tasks, the audio signal can often dominate the model decisions because it is much easier for the model to associate the audio input with the lexical text output than lip reading. Hence, it is conceivable that a multi-modal ASR model may need a larger weight for non-speech

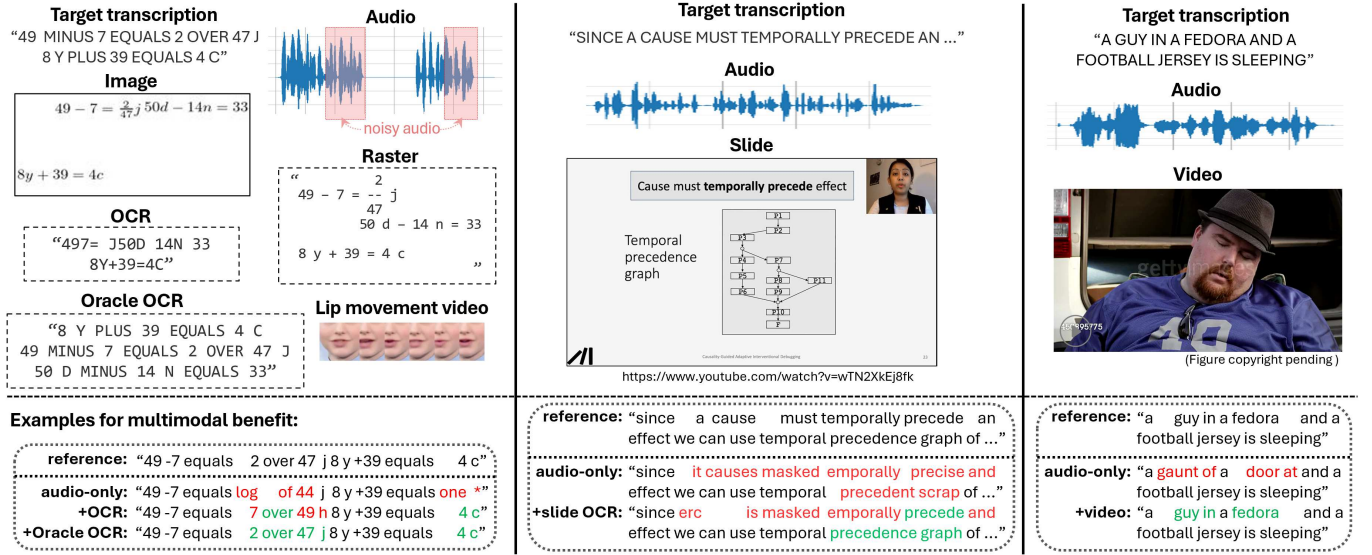


Fig. 2. Datasets used in our work. From left to right: Examples from the 3-Equations [1], SlideAVSR [49], and S-MiT [50]. For 3-Equations (left), we have different visual representations such as the image, OCR text, and raster representations. We provide one example for each dataset at the bottom that shows the help of incorporating multiple modalities in our experiments: "reference" is the ground truth transcription, "audio-only" is the output of an audio-only model ("audio→text"), and "+OCR" gives the output of the model trained on "OCR+audio→text" task.

modalities (image, lips) than speech in order to ensure better learning of image-to-text token correspondences.

In our experiments, we optimized the speech, text, and image loss weights using a grid search to minimize WER on a validation subset of the 3-Equations dataset (described below). This yielded an optimal hyperparameter configuration of  $\lambda_S=0.3$ ,  $\lambda_T=0.5$ , and  $\lambda_I=0.5$ . Our experiments revealed that tuning modality loss weights was very influential in both Transformer-based and Mamba models. See Section VII-C and Supplementary Materials (Section III).

#### IV. DATASETS

Our experiments use one synthetic dataset (3-Equations) and 2 real-world datasets (SlideAVSR; Spoken Moments in Time).

##### A. 3-Equations

We created a synthetic multi-modal speech recognition dataset called 3-Equations [1] to simulate scenarios in which multimodality is crucial to obtain accurate speech transcription. It consists of speech utterances, synchronized lip movements, and image context. In each utterance, a speaker reads 2 randomly selected mathematical equations that are displayed in an accompanying image containing 3 equations. We use the default male voice of pyttsx3 [51] to synthesize the voice. To each utterance, MUSAN noise [52] is added to the second half of each equation utterance so that only half of each utterance remains clean (see Fig. 2 left). This noise setting encourages the model to learn from both visual and auditory modalities: without visual information (lip movements, images), the model will fail in noise; without the audio, the model may not have enough information about which 2 of the 3 equations were actually spoken from the image. While our primary goal is a tightly-controlled dataset in which we can manipulate noise, visual representation, and other factors, 3-Equations is

also relevant to educational applications such as multi-modal learning assistants.

In order to study the impact on ASR accuracy of the visual representation quality, we created a "continuum" of visual modalities: no visual input, image encoding, optical character recognition (OCR), 2-D raster, and oracle-OCR text. For the image encoding, we used either DALL-E or ViT. For OCR, we applied EasyOCR [53] to each image. For oracle OCR, we concatenate the 3 equation sentences in each image, without filtering out the one that is not spoken. To similarly design a perfect 2-D representation that contains more spatial information than the OCR text, whilst still requiring conversion from math characters into English text, we created a 2-D text-based raster by placing characters from the equations onto a fixed-size text grid. With this new representation, the model is guaranteed to see the numbers and symbols in the equations, and it can potentially learn spatial information like image encoders. Hence, the continuum above can be considered as in an ascending order of visual representation quality. Fig. 2 (left) shows all visual modalities in 3-Equations.

Altogether, the dataset consists of 10,000 examples, 8,000 for training and 1,000 each for validation and testing. In total it contains 25.2 hours of synthesized speech, with 20.2 hours for training and 2.5 hours each for validation and testing.

##### B. SlideAVSR

SlideAVSR [49] is an audio-visual dataset of videos collected from YouTube in which the speakers explain scientific papers; the dataset provides manually transcribed speech, synchronized slides, and preprocessed OCR keywords (see Fig. 2, middle). The dataset contains many technical terms and the speakers have varying accents and speech intelligibility, making it difficult to transcribe accurately without the visual information (e.g., slides). It thus is a good choice to study



how multi-modal architectures can handle auditory noise in real-world scenarios.

Each utterance in SlideAVSR corresponds to one slide image with its processed OCR text. Some of the videos have talking faces, but since the coverage of talking face in this dataset is less than 50%, we omit the lip modality in our experiments. We also add MUSAN noise with different signal-to-noise ratios (SNRs) to the entire audio to simulate broader range of noise conditions. This dataset comprises 245 hours of audio data, with 195 hours for training, 20 hours for validation and 30 hours for testing.

### C. Spoken Moments in Time (S-MiT)

Spoken Moments in Time (S-MiT) [50] is a large-scale video caption dataset built from the Multi-Moments in Time [54]; it consists of 500k short videos depicting a broad range of different events, such as chef cooking or bird singing, along with audio recordings describing them (see Fig. 2, right). This dataset has a unique vocabulary size of over 50k, with high diversity and broad coverage examples.

While MLLMs with limited capacity may solely use static visual information gleaned from a single snapshot, stronger models can leverage more spatio-temporal information like motion. Using S-MiT, we can investigate whether models can gain the ability to harness such dynamics. Hence, we extracted 5 frames from each video uniformly over its duration. By comparing a visual modality consisting of all 5 frames to one containing only the first frame of each video, we can assess how much the motion is used beyond the static appearance.

Due to the input length limitation of the model we use in experiments, we subselect S-MiT to include examples whose audio duration is less than 5 seconds and video length less than 3 seconds, resulting in 56,385 examples for training, 692 for validation, and 146 for testing.

## V. EXPERIMENTS: NOISE & NUMBER OF MODALITIES

Multimodality can be a double-edged sword: With additional modalities, complementary information is provided, potentially leading to better accuracy. However, using more modalities increases the input length, which might increase the difficulty of finding relevant information and actually degrade performance. Moreover, even when multimodal inputs do increase ASR accuracy, their benefit might derive not from complementary information but from implicit regularization: training an MLLM to predict the next token for every modality prevents the model from dedicating all its capacity to minimizing the training loss on the ASR text tokens.

In our experiments, we investigated whether using more modalities generally leads to higher ASR accuracy, as well as the conditions under which each modality is most effective. We explore conditions including acoustic noise, visual noise, and sequence noise. We trained models with all combinations of modalities (e.g.,  $I+A \rightarrow T$ ,  $L+A \rightarrow T$ ,  $I+L+A \rightarrow T$ , etc.), then calculated the WER and RB of each model at different noise levels. Each model is fine-tuned on the 3-Equations training set containing a random mix of added MUSAN noise with an SNR in [20, 10, 5, 0, -5, -10, -20]. Then, each model

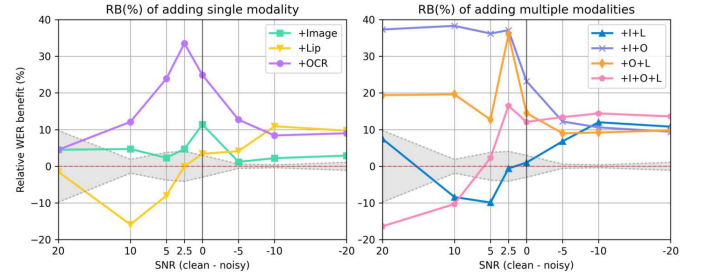


Fig. 3. RB (%) of adding image, OCR, or lip modalities on 3-Equations 2-noise test set. The gray shading part represents the statistical error range (between RB+ and RB-) from our shuffled experiments. Left: adding single modality; Right: adding multiple modalities.

was evaluated separately on an SNR-specific test set in [20, 10, 5, 2.5, 0, -5, -10, -20].

To ensure that any performance gap between models is statistically reliable, we estimate for each SNR value the variability in WER of the *same* model under a tiny perturbation (specifically, the order of examples across minibatches). This allows us to distinguish between meaningful WER differences versus just statistical anomalies. Specifically, the gray area in Fig. 3 between two dashed lines is the non-significant window due to just a mini-batch reshuffle. See Supplementary Materials (Section I) for more details.

Results on 3-Equations are shown in Table I and Fig. 3. As shown in the figure, most of the points on the curves are outside the non-significant window, indicating our results are non-trivial. We also calculated WERs for each non-audio modality: For  $I \rightarrow T$ ,  $L \rightarrow T$ ,  $O \rightarrow T$ , and  $O_{oracle} \rightarrow T$  they were 79.9, 41.4, 76.5, and 60.5, respectively, which illustrate how the audio signal itself is crucial to obtaining low WER. We interpret the results in the subsections below.

### A. Are more modalities always better for ASR?

Firstly, we examine the average performance, over all SNR levels, of adding more modalities. Compared to the audio-only baseline ( $A \rightarrow T$ ) in the first row of Table I, the average RB values of adding one modality of either image (I), lip (L), or OCR (O) are all positive. Multimodal inputs also performed much better than any of the non-audio single input modalities. When considering 3- or 4-modality combinations, both of them surpass the 2-modality models on average. The 4-modality model has the overall best average WER among all models, as shown in the last row in the table, while one 3-modality model has the highest RB.

Next, we analyzed the consistency of adding modalities across all noise levels. When introducing one modality, adding lips does not bring a consistent benefit in noise, whereas adding images or OCR does. For 3-modality or 4-modality combinations, we observe more consistent benefits in noise. The 3-modality models are better than their corresponding 2-modality models at most noise levels. The 4-modality model outperforms the 3-modality models for  $SNR < 0$  but not for  $SNR$  from 0 to 20. This may be attributed to the longer input sequences, from which it is difficult for the model to find useful information. We discuss more about sequence noise caused by input length in Section V-C. In general, we find evidence that multimodal ASR can benefit from at least 4

TABLE I

WERS AND RELATIVE BENEFITS OF ADDING MODALITIES TO THE TRANSFORMER-BASED MODEL ON 3-EQUATIONS, AT DIFFERENT NOISE LEVELS BASED ON SNR. A, I, O,  $O_{oracle}$ , L, R, AND T STAND FOR AUDIO, IMAGE, OCR, ORACLE OCR, LIP, RASTER, AND TEXT, RESPECTIVELY.

Input Modalities	20	10	5	2.5	0	-5	-10	-20	Average
	WER ↓ (RB ↑)								
A	0.34%	0.54%	1.07%	1.70%	2.97%	12.53%	25.96%	37.25%	10.3%
I + A	0.32% (+4.5%)	0.51% (+4.7%)	1.04% (+2.3%)	1.62% (+4.7%)	2.63% (+11.4%)	12.38% (+1.2%)	25.38% (+2.2%)	36.18% (+2.9%)	10.0% (+4.2%)
L + A	0.34% (-1.5%)	0.62% (-15.9%)	1.15% (-8.0%)	1.70% (0.0%)	2.87% (+3.4%)	12.01% (+4.1%)	23.14% (+10.9%)	33.64% (+9.7%)	9.4% (+0.3%)
O + A	0.32% (+4.5%)	0.47% (+12.1%)	0.81% (+23.9%)	1.13% (+33.5%)	2.23% (+24.9%)	10.94% (+12.7%)	23.78% (+8.4%)	33.91% (+9.0%)	9.2% (+16.1%)
R + A	0.4% (-19.4%)	0.67% (-25.2%)	1.19% (-11.7%)	1.91% (-12.4%)	3.31% (-11.4%)	13.58% (-8.4%)	26.87% (-3.5%)	37.81% (-1.5%)	10.7% (-11.7%)
$O_{oracle}$ + A	0.09% (+73.1%)	0.13% (+75.7%)	0.15% (+85.9%)	0.13% (+92.4%)	0.36% (+87.9%)	1.15% (+90.8%)	2.29% (+91.2%)	2.97% (+92.0%)	0.9% (+86.1%)
$O_{oracle,L}$ + A	0.30% (+10.4%)	0.23% (+57.0%)	0.37% (+65.3%)	0.37% (+78.2%)	0.94% (+68.4%)	5.19% (+58.6%)	10.51% (+59.5%)	15.38% (+58.7%)	4.2% (+57.0%)
I + L + A	0.31% (+7.5%)	0.58% (-8.4%)	1.17% (-9.9%)	1.71% (-0.6%)	2.94% (+1.0%)	11.67% (+6.8%)	22.85% (+12.0%)	33.22% (+10.8%)	9.3% (+2.4%)
I + O + A	0.21% (+37.3%)	0.33% (+38.3%)	0.68% (+36.2%)	1.07% (+37.1%)	2.28% (+23.2%)	11.00% (+12.2%)	23.22% (+10.6%)	33.74% (+9.4%)	9.1% (+25.5%)
O + L + A	0.27% (+19.4%)	0.43% (+19.6%)	0.93% (+12.7%)	1.09% (+35.9%)	2.54% (+14.5%)	11.40% (+9.0%)	23.57% (+9.2%)	33.61% (+9.8%)	9.2% (+16.3%)
I + O + L + A	0.39% (-16.4%)	0.59% (-10.3%)	1.04% (+2.3%)	1.42% (+16.5%)	2.61% (+12.1%)	10.85% (+13.4%)	22.21% (+14.4%)	32.18% (+13.6%)	8.9% (+5.7%)

TABLE II

WERS AND RELATIVE BENEFITS OF ADDING MODALITIES ON SLIDEAVSR AT DIFFERENT NOISE LEVELS. THE OCR WORDS ARE RANKED BY FQ RANKER [49], AND K INDICATES THE MAXIMUM OCR WORD COUNT.

Inputs	+∞	10	0	-10	Average
	WER ↓ (RB ↑)				
A	33.8%	42.5%	44.8%	70.6%	47.9%
$O_{All}$ + A	31.4% (+7.0%)	37.2% (+12.4%)	47.7% (-6.4%)	75.9% (-7.6%)	48.0% (+1.4%)
$O_{K=30}$ + A	30.5% (+9.7%)	35.7% (+15.8%)	46.9% (-4.7%)	75.5% (-6.9%)	47.2% (+3.5%)
$O_{K=10}$ + A	30.6% (+9.5%)	34.6% (+18.5%)	46.2% (-3.1%)	77.9% (-10.3%)	47.3% (+3.6%)
I + A	28.5% (+15.5%)	34.3% (+19.3%)	46.1% (-13.6%)	70.8% (-0.2%)	46.1% (+5.3%)

TABLE III

WERS AND RB OF ADDING MODALITIES ON S-MIT TEST SET. V REFERS TO VIDEO MODALITY, AND I REFERS TO IMAGE.

Inputs	Vcodec	Visual Data	WER (RB ↑)
A	-	-	28.51%
V + A	ViT	5 frames	26.26% (+7.9%)
I + A	ViT	1 <sup>st</sup> frame	26.71% (+6.3%)
I + A	DALL-E	1 <sup>st</sup> frame	25.97% (+8.9%)

modalities as inputs, but the benefit varies with the noise level and thus how much complementary information is useful.

We found similar trends on SlideAVSR and S-MiT. On SlideAVSR, we use the OCR words and the first video frames as images because the videos in the dataset are about explaining slides. As shown in Table II, adding all OCR words ( $O_{All} + A \rightarrow T$ ) and image ( $I + A \rightarrow T$ ) has an average RB of +1.4% and +5.3%, respectively. On S-MiT, we extract video frames to be processed by visual encoders such as ViT and DALL-E. As shown in Table III, the model achieves consistent improvements over the audio-only baseline when extra visual information is included. These demonstrate that the addition of supplementary visual modalities improves the overall recognition performance across datasets.

#### B. Acoustic noise: when is each modality most helpful?

Modalities synchronized with audio (e.g., lip movement) could be effective in different conditions than unsynchronized ones (e.g. image). As shown in Fig. 3, for models with 2 modality inputs, the benefit of unsynchronized modalities follows a trend of increasing first and then decreasing, reaching a “sweet spot” in the middle. In contrast, the benefit of synchronized modalities grows stronger as the noise increases, similar to previous studies [55]. This discrepancy is because of how each visual modality aligns with the primary modality – audio: when the audio is too noisy, the unsynchronized modalities (image, OCR) cannot establish a correspondence with the audio, and the visual information is useless.

When adding more than one modality, the model with additional image and lip information ( $I + L + A$ ) displays

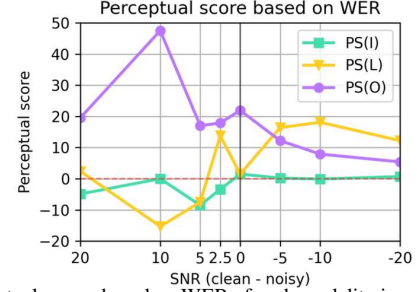


Fig. 4. Perceptual scores based on WER of each modality in models fine-tuned on 3-Equations. PS is computed on the 4-modality model ( $I + O + L + A$ ).

a similar trend as  $L + A$  model; the  $O + L + A$  model, on the other hand, is similar to  $O + A$  model. This may indicate that the lip and OCR are more prominent modalities in the model decision. Meanwhile, the 4-modality model exhibits a trend similar to  $O + A$  when there is less noise, and a trend similar to  $L + A$  when there is more noise, having a sweet spot at SNR=2.5dB. This also supports the view that OCR and lip dominate the performance.

We also compute each modality’s Perceptual Score (PS) to better understand their contribution and effectiveness across noise levels. While we mostly find a similar PS trend to RB results, the fact that the PS(I) curve is near 0 for most SNR values suggests that, although adding the image information can improve ASR accuracy, the benefit might have more to do with regularization than complementary information. See Fig. 4 and Supplementary Materials (Section IV) for more details.

We observe some similar phenomena on SlideAVSR, as shown in Table II. For all three experiments that add OCR words and one experiment that add images, they all outperform the audio-only baseline at low noise levels (+∞ and 10dB), but underperform the baseline at higher noise levels. They exhibit a sweet spot in the middle at 10dB, and this follows our conclusion that unsynchronized modalities need to establish a reliable correspondence with the audio to have an effect.

#### C. Do irrelevant inputs (sequence noise) impact accuracy?

We investigated how the proportion of irrelevant information in the input sequence – which we call *sequence noise* – can impact ASR accuracy. Recall that, in the 3-Equations dataset, each example contains 3 written equations, but only 2 of them are actually spoken. Therefore, the OCR modality inherently only contains around  $\frac{2}{3}$  relevant information, and the model needs to learn – by combining the OCR input with auditory

TABLE IV  
WERS AND RELATIVE BENEFITS OF ADDING MODALITIES TO THE MAMBA-BASED MODEL ON 3-EQUATIONS, AT DIFFERENT NOISE LEVELS.

Inputs	20	10	5	2.5	0	-5	-10	-20	Average
	WER ↓ (RB ↑)								
<i>A</i>	1.02%	1.37%	2.03%	2.50%	4.34%	15.13%	28.45%	38.18%	11.6%
<i>I + A</i>	0.91% (+10.8%)	1.34% (+2.2%)	2.02% (+0.5%)	2.76% (-10.4%)	4.68% (-7.8%)	15.54% (-2.7%)	28.97% (-1.8%)	39.10% (-2.4%)	11.9% (-1.5%)
<i>L + A</i>	1.31% (-28.4%)	1.65% (-20.4%)	2.48% (-22.2%)	3.12% (-24.8%)	4.94% (-13.8%)	17.99% (-18.9%)	31.88% (-12.1%)	42.71% (-11.9%)	13.3% (-19.1%)
<i>O + A</i>	0.66% (+35.3%)	0.82% (+40.1%)	1.30% (+36.0%)	1.80% (+28.0%)	3.29% (+24.2%)	12.31% (+18.6%)	24.63% (+13.4%)	33.44% (+12.4%)	9.8% (+26.0%)

information about what was actually said – how to locate useful information from the input sequence.

To create a scenario with higher sequence noise, we added to each example 7 extra irrelevant oracle OCR sentences that are randomly selected from other examples in addition to the original 3 sentences, resulting in a dataset with around  $\frac{1}{5}$  relevant information. We then compared ASR accuracy of the two models (one with less, and one with more sequence noise). Results are shown in Table I ( $O_{oracle} + A$  and  $O_{oracle,10} + A$ ). We observe that including more sequence noise leads to much worse accuracy, with the overall average RB dropping from +86% to +57%. In particular, the RB gap between the two models increases as the level of acoustic noise increases. These results may also explain why the 4-modality model is worse than the 3-modality or even 2-modality models under some conditions: adding less informative modalities introduces a longer sequence while reducing the proportion of relevant information in the sequence.

In SlideAVSR, we similarly explore the influence of sequence noise using OCR words. A sample slide in SlideAVSR can contain hundreds of OCR words, while only a small fraction of them are mentioned in the speech. According to our experiments on 3-Equations, it is conceivable that this information overload could harm performance. To evaluate this impact, we use FQ Ranker [49] to filter the OCR words based on the frequency of word occurrences in Wikipedia, setting the maximum word count ( $K$ ) to 10 and 30. As shown in Table II ( $O_{All}$ ,  $O_{K=30}$ ,  $O_{K=10}$ ), although the performance of adding OCR is worse than the audio-only model at some noise levels, its overall RB increases as we filter more stringently.

#### D. How do different visual representations impact accuracy?

Next we explore the impact of the visual representation quality on ASR accuracy. As shown in Fig. 2, there are four image-based visual modalities in the 3-Equations dataset: image encoding, OCR text, oracle OCR text, and raster representation. These modalities can be placed along a spectrum from *raw* to *abstract* [25], where the image encoding is raw, and the OCR texts are more abstract as they present information extracted from images. As summarized in Table I, the average WER of 2-modality models combining visual modalities follows:  $A$  (10.3%) >  $I + A$  (10.0%) >  $O + A$  (9.2%) >  $O_{oracle} + A$  (0.9%). This can be attributed to the visual representations becoming more abstract and informative, thus reducing the visual noise level, making it easier for the model to utilize. Therefore, better visual representation with less visual noise can lead to better supplementary performance. We notice that adding raster has even worse accuracy than adding image, we suppose this is because: (1) the raster representation has a much longer token sequence (384 for image vs. 864 for

raster) that hinders the model from finding useful information, as discussed in Section V-C, and (2) the model still struggles in handling 2-D representations, even when it is as perfect as rasters. This suggests that more powerful positional encoding schemes than an absolute encoder might be necessary for 2-D representations.

Beyond the variations in visual representations, the capacity of the visual encoder itself may further influence the effective quality, or SNR, of the extracted visual features. We therefore investigate the impact of visual encoders by comparing models utilizing ViT and DALL-E on the S-MiT, the results are included in Table III. From the results in the last two rows, DALL-E, with a relative benefit of +8.9%, performs better than ViT on this task. This may be attributed to their distinct training objectives: ViT excels at extracting structural details, whereas DALL-E produces richer semantic features more aligned with language, thus facilitating multi-modal integration in our task. We also note that  $V + A$  outperforms  $I + A$  when both use ViT as encoder, which indicates that DMLM is capable of learning from spatio-temporal information.

#### VI. EXPERIMENTS: TRANSFORMERS VS. MAMBA

Although Transformer-based MLLMs have achieved impressive performance, their quadratic-complexity attention is inefficient in terms of both latency and memory usage, especially for multimodal input streams which tend to be long. As an alternative method with fast inference speed and competitive accuracy, linear sequential state space models such as Mamba [17] are becoming more powerful and are growing in popularity for MLLMs in particular [22].

To explore whether Mamba-based MLLMs exhibit some of the same trends regarding multimodality as Transformers do, we conduct similar 2-modality experiments ( $I + A$ ,  $L + A$ ,  $O + A$ ) to compare a pretrained Mamba-130M to OPT-125M as the MLLM backbone. Results of the Mamba model on 3-Equations are summarized in Table IV. By adding lip movements and OCR, we reach the same conclusion as OPT, that the modality synchronized with audio has increasing RB when there's more noise, while the unsynchronized modality has the greatest RB in moderate noise. At the same time, we also notice that all experiments that are conducted on Mamba perform worse than those on OPT. The RB of adding an image has a decreasing trend as the noise level goes up, which is different from the trend on OPT. We test the audio-only model and the  $I + A$  model on clean audio (SNR=+∞), where their WERs are 0.6% and 0.94%, respectively, resulting an RB of -10.6% when adding  $I$ . This confirms that the general trend of adding images on Mamba is the same as OPT, but the sweet spot may be more architecture-dependent.

As a fundamental advantage of leveraging Mamba models, we want to examine the time cost of Mamba-based and



Transformer-based multi-modal models. With our experimental configurations on 3-Equations, training Mamba for  $A \rightarrow T$  takes about 197 seconds per epoch (sec/ep), while an OPT model of similar size takes 383 sec/ep. With more input modalities, the length of the input sequence increases, which increases the time cost. In particular, for an  $I + A \rightarrow T$  model, the training time of OPT grows to 305 sec/ep (55% increase), while OPT it increased to 643 sec/ep (68% increase). With three modalities ( $I + L + A \rightarrow T$ ), the time cost further rises to 313 sec/ep (59% total increase) and 837 sec/ep (219% total increase), respectively. These indicate that Mamba-based models in general are more efficient, and are less affected by the increase in length.

On the other hand, our practical experiences with Mamba suggest it is not as stable as a Transformer-based model, since it seems to be sensitive to hyperparameters. For example, varying the learning rate for the  $A \rightarrow T$  Mamba from  $1e-3$  to  $1e-4$  substantially increased the WER from 13.7% to 20.4%, but a comparable change in learning rate for OPT changed its accuracy only very slightly from 10.3% to 10.6%. Furthermore, when adapting the model to accept multi-modal inputs, we find that different hyperparameters have to be used for different modality combinations to achieve reasonable performance. This may indicate that Mamba-based model may have high specialization but relatively low generalizability if the hyperparameters are optimal. We also experiment with different sizes of OPT and Mamba models, and observe that they exhibit different patterns when scaling up the model size. See details in Supplementary Materials (Section V).

## VII. EXPERIMENTS: INPUT FORMAT AND LOSS WEIGHTS

With MLLMs, practical questions arise of (1) how to format the input stream so that the model can find what it needs, and (2) how much to weigh each modality in the loss function.

### A. Multi-modal inputs: interleaved or blocked?

Interleaved inputs are common in vision-language tasks, as this format is ubiquitous in web-crawled data. Previous works have examined the benefit of using interleaved data, and find it endows MLLMs with in-context few-shot learning capabilities [20, 56]. However, there has been little exploration of interleaving modalities when both modalities contain dense and synchronized information. We thus compared interleaved to non-interleaved (blocked) input formats with  $V + A$  task. Specifically, we conducted experiments on the S-MiT by interleaving the 5 video frames evenly with audio. I.e., if the speech contains 300 audio tokens, the final input sequence will be like: “<Frames [0]><AudioTokens [0:60]><Frames [1]><AudioTokens [60:120]>...<Frames [4]><AudioTokens [240:300]>”. Using ViT as visual encoder to tokenize video frames, the model achieves a WER of 28.4% in the interleaving manner, which is +0.4% RB to the audio-only model. Although this model is better than the audio-only model, its performance is still much worse than the non-interleaved (blocked) input.

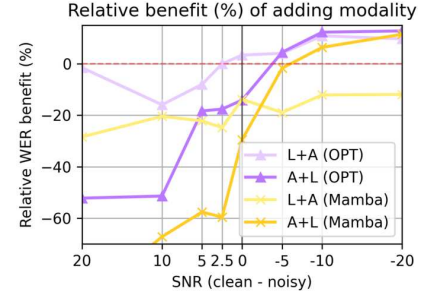


Fig. 5. RB of audio-first or audio-later inputs on OPT and Mamba model.

TABLE V  
EVALUATE WER(%) ON THE 3-EQUATIONS OF ALL NOISE LEVELS (SNRS) WITH DIFFERENT MODALITY LOSS WEIGHT CONFIGURATIONS. LOSS WEIGHT IS DISPLAYED IN THE ORDER OF (SPEECH, TEXT, IMAGE). ONLY TEXT WEIGHT IS NON-ZERO IN THE SECOND EXPERIMENT.

Task	Loss	20	10	5	2.5	0	-5	-10	-20
$A \rightarrow T$	(0.3,0.5,0.5)	0.3%	0.5%	1.1%	1.7%	3.0%	12.5%	26.0%	37.0%
$A \rightarrow T$	(0.0,0.5,0.0)	0.3%	0.6%	1.2%	1.8%	3.3%	13.5%	27.0%	38.4%

### B. In what order should the different modalities be input?

Previous work studying multi-modal inputs observed order sensitivity in MLLMs, and found that prompt orders and modality orders can significantly affect results. Some demonstrate that MLLMs have a preference for the beginning and end of contexts, and placing important content in these positions can enhance performance [57]. Others examine the impact of modality sequencing in prompt context, and found the impact is context-dependent and related to task complexity [58].

We explored the impact of switching modality input orders in 2-modality models using either Transformer or Mamba as backbone. Specifically, we compared the accuracy of a model  $L + A \rightarrow T$  to  $A + L \rightarrow T$ , i.e., the order of the two inputs (audio, lips) are swapped. Results are shown in Figure 5. They indicate that RB of the audio-first models keeps increasing as the noise level increases (lower SNR) and eventually outperforms audio-later models, and that audio-first is more beneficial in noisy situations. We found similar trends for other 2-modality models.

We speculate that the results are due to an interaction between three factors: (1) the absolute positional encoder that spans the entire input sequence; (2) primacy/recency effects of how Transformers sometimes prefer to attend to tokens near the start/end (respectively) of the input [57]; and (3) how the audio noise in 3-Equations was added to the *second* half of each utterance. With the  $A + L \rightarrow T$  model, the clean part (first half) of the first utterance is at a known location (i.e., the start of the input sequence) so that the model can easily find it via the absolute positional encoding. In contrast, with the  $L + A \rightarrow T$  model, the audio begins only after a *variable-length* lip sequence, and hence the clean part of the first utterance is difficult for the model to find. Moreover, the most recent part of the audio is the noisy half and thus, especially at low SNR, not very useful. See Supplementary Materials (Section VI) for more explanation.

### C. How to weigh the different multimodal loss components?

During the training stage of a Transformer-based model, the loss function is used to predict the next token as the



autoregressive objective. Hence, a modality-weighted loss can be viewed as how well we want the model to learn to generate tokens of that modality. Since in the ASR tasks we only care about the transcription, which is merely text generation, we wonder if setting the loss weights of other modalities to zero can lead to good results with faster convergence. However, experimental results in Table V show that this is not the case. This may be because the model needs to learn the internal logic or structure of that modality, and such tasks can be viewed as an intrinsic multitask learning setting (e.g., the speech-to-speech next token prediction task contributes to the speech-to-text task). This suggests that, even for simple speech-to-text generation tasks on a pretrained MLLM, learning from modalities other than text remains important.

### VIII. CONCLUSION AND FUTURE RESEARCH

This work provides an in-depth investigation into how multiple input modalities impact speech recognition accuracy in MLLMs. Our experiments suggest that: (1) Fusing more modalities usually enhances performance, but the benefit of each modality depends on the auditory noise. Specifically, synchronized modalities become more influential at high noise levels, whereas unsynchronized modalities provide the greatest benefit at moderate (“sweet spot”) noise levels. (2) Adding irrelevant information into the input sequence decreases accuracy. (3) Visual representations with higher-quality and generated from better visual encoders consistently improve accuracy. Also, the performance gap between raster and oracle OCR (both are essentially perfect representations) suggests that MLLMs with absolute positional encoders struggle to handle 2-D representations. (4) Mamba shows similar trends as Transformers with faster speed, but are less stable to train. (5) Different input formats may vary in effectiveness, suggesting new directions for optimizing multi-modal input formatting.

Our results suggest some potential research directions to improve multimodal ASR: (1) Develop more powerful positional encodings that help the model to collate information across modalities and also to handle 2-D visual representations; (2) Develop stronger visual encoders that can extract more detailed (e.g., text content) from images; (3) Explore more recent state space models that might obtain the “best of both worlds” of fast inference and high accuracy.

**Acknowledgement:** This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805, and also from an NSF CAREER grant #2046505. The opinions expressed are those of the authors and do not represent views of the NSF.

### REFERENCES

- [1] Y. Guan, V. A. Trinh, V. Voleti, and J. Whitehill, “Multimodal speech transformer decoders: When do multiple modalities improve accuracy?” *arXiv:2409.09221*, 2024.
- [2] G. Comanici, E. Bieber, M. Schaekermann *et al.*, “Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities,” *arXiv:2507.06261*, 2025.
- [3] S. Wu, H. Fei, L. Qu *et al.*, “Next-gpt: Any-to-any multimodal llm,” in *Forty-first ICML*, 2024.
- [4] B. Liao, H. Tao, Q. Zhang *et al.*, “Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation,” *arXiv:2502.13145*, 2025.
- [5] V. A. Trinh, R. Southwell, Y. Guan *et al.*, “Discrete multimodal transformers with a pretrained large language model for mixed-supervision speech processing,” *arXiv:2406.06582*, 2024.
- [6] X. Chang, B. Yan, K. Choi *et al.*, “Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study,” in *ICASSP*, 2024.
- [7] Z. Ma, G. Yang, Y. Yang *et al.*, “Speech recognition meets large language model: Benchmarking, models, and exploration,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai *et al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, 2021.
- [9] J. Wang, H. Jiang, Y. Liu *et al.*, “A comprehensive review of multimodal large language models: Performance and challenges across different tasks,” *CoRR*, 2024.
- [10] F. Chen, M. Han, H. Zhao *et al.*, “X-llm: Bootstrapping advanced large language models by treating multimodalities as foreign languages,” *arXiv:2305.04160*, 2023.
- [11] J. H. Yeo, H. Rha, S. J. Park, and Y. M. Ro, “Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens,” *arXiv:2503.11315*, 2025.
- [12] U. Cappellazzo, M. Kim, H. Chen *et al.*, “Large language models are strong audio-visual speech recognition learners,” in *ICASSP*, 2025.
- [13] J. Hong, M. Kim, J. Choi, and Y. M. Ro, “Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring,” in *CVPR*, 2023.
- [14] Y. Hu, C. Chen, C.-h. H. Yang *et al.*, “Large language models are efficient learners of noise-robust speech recognition,” in *ICLR*, 2024.
- [15] V. Gabeur, P. H. Seo, A. Nagrani *et al.*, “Avatar: Unconstrained audiovisual speech recognition,” in *INTER-SPEECH*, 2022.
- [16] B. Mu, X. Wan, N. Zheng *et al.*, “Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition,” *IEEE Signal Processing Letters*, 2024.
- [17] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in *First Conference on Language Modeling*, 2023.
- [18] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” in *CVPR*, 2024.
- [19] Y. Chu, J. Xu, X. Zhou *et al.*, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv:2311.07919*, 2023.
- [20] C. Tian, X. Zhu, Y. Xiong *et al.*, “Mm-interleaved: Interleaved image-text generative modeling via multimodal feature synchronizer,” *CoRR*, 2024.
- [21] T. Dao and A. Gu, “Transformers are ssms: Generalized

- models and efficient algorithms through structured state space duality,” in *ICML*, 2024.
- [22] H. Zhao, M. Zhang, W. Zhao *et al.*, “Cobra: Extending mamba to multi-modal large language model for efficient inference,” in *Proceedings of the AAAI Conference*, 2025.
- [23] J. Ma, F. Li, and B. Wang, “U-mamba: Enhancing long-range dependency for biomedical image segmentation,” *arXiv:2401.04722*, 2024.
- [24] Y. Qiao, Z. Yu, L. Guo *et al.*, “VI-mamba: Exploring state space models for multimodal learning,” *CoRR*, 2024.
- [25] P. P. Liang, A. Zadeh, and L.-P. Morency, “Foundations & trends in multimodal machine learning: Principles, challenges, and open questions,” *ACM Computing Surveys*, 2024.
- [26] X. Li, J. Thickstun, I. Gulrajani *et al.*, “Diffusion-lm improves controllable text generation,” *Advances in neural information processing systems*, 2022.
- [27] S. Nie, F. Zhu, Z. You *et al.*, “Large language diffusion models,” in *ICLR 2025 Workshop on Deep Generative Model in Machine Learning*.
- [28] D. Zhang, S. Li, X. Zhang *et al.*, “Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,” in *Findings of the Association for Computational Linguistics: EMNLP*, 2023.
- [29] X. Zhang, Q. Zhang, H. Liu *et al.*, “Mamba in speech: Towards an alternative to self-attention,” *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
- [30] J. H. Yeo, S. Han, M. Kim, and Y. M. Ro, “Where visual speech meets language: Vsp-llm framework for efficient and context-aware visual speech processing,” in *EMNLP*. Association for Computational Linguistics, 2024.
- [31] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen *et al.*, “Audiopalm: A large language model that can speak and listen,” *arXiv:2306.12925*, 2021.
- [32] P. H. Seo, A. Nagrani, and C. Schmid, “Avformer: Injecting vision into frozen speech models for zero-shot av-asr,” in *CVPR*, 2023.
- [33] S. Kim, D. Lee, B. Stark, and J. Han, “Gesture-aware zero-shot speech recognition for patients with language disorders,” in *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*, 2025.
- [34] J. H. Yeo, M. Kim, C. W. Kim *et al.*, “Zero-avsr: Zero-shot audio-visual speech recognition with llms by learning language-agnostic speech representations,” *arXiv:2503.06273*, 2025.
- [35] U. Cappellazzo, M. Kim, and S. Petridis, “Adaptive audio-visual speech recognition via matryoshka-based multimodal llms,” *arXiv:2503.06362*, 2025.
- [36] A. Radford, J. W. Kim, T. Xu *et al.*, “Robust speech recognition via large-scale weak supervision,” in *ICML*, 2023.
- [37] I. Gat, I. Schwartz, and A. Schwing, “Perceptual score: What data modalities does your model perceive?” *Advances in Neural Information Processing Systems*, 2021.
- [38] S. Zhang, S. Roller, N. Goyal *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv:2205.01068*, 2022.
- [39] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*, 2015.
- [40] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, “Cvss corpus and massively multilingual speech-to-speech translation,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022.
- [41] C. Wang, A. Wu, and J. Pino, “Covost 2 and massively multilingual speech-to-text translation,” *arXiv:2007.10310*, 2020.
- [42] T.-Y. Lin, M. Maire, S. Belongie *et al.*, “Microsoft coco: Common objects in context,” in *Computer vision—ECCV*. Springer, 2014.
- [43] L. Barrault, Y.-A. Chung, M. C. Meglioli *et al.*, “Seamlessm4t: Massively multilingual & multimodal machine translation,” *arXiv:2308.11596*, 2023.
- [44] B. Shi, W.-N. Hsu, K. Lakhota, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *ICLR*, 2022.
- [45] A. Ramesh, M. Pavlov, G. Goh *et al.*, “Zero-shot text-to-image generation,” in *ICML*, 2021.
- [46] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *ICLR*, 2020.
- [47] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2017.
- [48] T. Afouras, J. S. Chung, A. Senior *et al.*, “Deep audio-visual speech recognition,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [49] H. Wang, S. Kurita, S. Shimizu, and D. Kawahara, “Slideavr: A dataset of paper explanation videos for audio-visual speech recognition,” in *Workshop on Advances in Language and Vision Research*, 2024.
- [50] M. Monfort, S. Jin, A. Liu *et al.*, “Spoken moments: Learning joint audio-visual representations from video descriptions,” in *CVPR*, 2021.
- [51] “pyttsx3: Offline text to speech (tts) converter for python,” <https://github.com/nateshmbhat/pyttsx3>.
- [52] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” *arXiv:1510.08484*, 2015.
- [53] “Easyocr: Ready-to-use ocr with 80+ supported languages,” <https://github.com/JaidedAI/EasyOCR>.
- [54] M. Monfort, B. Pan, K. Ramakrishnan *et al.*, “Multi-moments in time: Learning and interpreting models for multi-action video understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [55] P. Ma, A. Haliassos, A. Fernandez-Lopez *et al.*, “Auto-avsr: Audio-visual speech recognition with automatic labels,” in *ICASSP*, 2023.
- [56] J.-B. Alayrac, J. Donahue, P. Luc *et al.*, “Flamingo: a visual language model for few-shot learning,” *Advances in neural information processing systems*, 2022.
- [57] Z. Tan, X. Chu, W. Li, and T. Mo, “Order matters: Exploring order sensitivity in multimodal large language models,” *arXiv:2410.16983*, 2024.
- [58] G. Wardle and T. Sušnjak, “Image first or text first? optimising the sequencing of modalities in large language model prompting and reasoning tasks,” *Big Data and Cognitive Computing*, 2025.