Do Slides Help? Multi-modal Context for Automatic Transcription of Conference Talks

Anonymous ACL submission

Abstract

State-of-the-art (SOTA) Automatic Speech Recognition (ASR) systems primarily rely on acoustic information while disregarding additional multi-modal context. However, visual information are essential in disambiguation and adaptation. While most work focus on speaker images to handle noise conditions, this work also focuses on integrating presentation slides for the use cases of scientific presentation.

In a first step, we create a benchmark for multi-modal presentation including an automatic analysis of transcribing domain-specific terminology. Next, we explore methods for augmenting speech models with multi-modal information. We mitigate the lack of datasets with accompanying slides by a suitable approach of data augmentation. Finally, we train a model using the augmented dataset, resulting in a relative reduction in word error rate of approximately 34%, across all words and 35%, for domain-specific terms compared to the baseline model.

1 Introduction

Automatic Speech Recognition (ASR) like many other NLP tasks are currently solved by using pre-trained models rather than learning models from scratch (Han et al., 2021). Although modern ASR systems have an overall similar to human performance on general data yet one important challenge remains in accurately transcribing specialized vocabulary for example, in academic settings. Figure 1 illustrates a challenge for current ASR systems. A system relying on only audio is not able to correctly transcribe the domain-specific terms Kenya-Birth and Kenya Rwandan (in red).

As conference talks and lectures often include presentation slides, humans can correctly identify these words by using this additional context. Therefore, we propose to integrate visual context (slides) into existing state-of-the-art ASR system and enable them to also exploit this context. As shown on the right side of Figure 1, the final model is able to properly transcribe these words as Kinyabert and Kinyarwanda (in blue) when the correct words are presented to the model in the additional information provided from the accompanying slides of the talk.

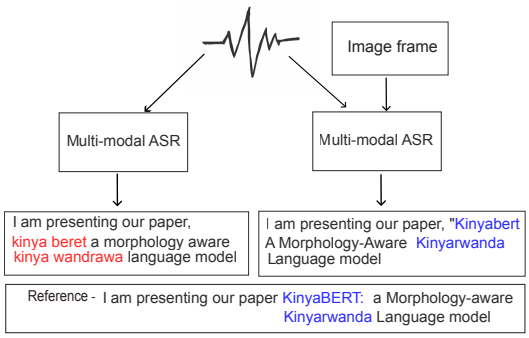

Figure 1: An example of ASR transcription before and after using multi-modal input. Left: ASR baseline makes mistakes (in red) for multiple words. Right: Model correctly transcribes words (in blue) using multi-modal inputs.

In a first step, we extend an existing benchmark, the ACL dataset (Salesky et al., 2023) with additional slide context, as well as a, target automatic evaluation for domain-specific terms to evaluate this assumption. Furthermore, we verify our assumption that these terms are challenging for SOTA models like Whisper (Radford et al., 2023), Phi-4-multimodal (Abouelenin et al., 2025) and SALMONN (Tang et al., 2023).

When integrating visual context into ASR models to handle domain-specific words, we want to keep the strong SOTA performance of current large-scale models. Therefore, we focus on approach that can add this ability to existing models. One interesting aspect of current models is their ability to handle zero-shot task. To this end, we first propose a zero-shot integration that is already able to exploit the visual context.

In a second step, we investigate methods to train the model to better integrate the contextual infor-

mation. This gives rise to the challenge that we need dedicated training data for this scenario. We address this problem by using large language models (LLMs) to augment ASR training data with presentation slides.

The primary contributions of this paper are:

- Analysing the ability of ASR to transcribe domain-specific words, particularly from scientific talks.
- Integration of multi-modal information into existing pre-trained models.
- Application of training approaches with augmented data to improve the transcription on domain-specific terms.

2 Related Work

There are multiple work where model performance is improved by additional information integration. In this regard, (Maergner et al., 2012) create a lecture specific vocabulary, based on the content of the related documents of the lectures. Construction of a vocabulary with relevant content improves the model performance and results in a reduced word error rate of up to 25 percent.

Additionally, combining modalities for the improvement of ASR is also considered in the literature. Starting from Hidden Markov model for speech recognition and manually created features represented visual components, combining modalities were also considered for the task of establishing relation between words and non-linguistic context (Fleischman and Roy, 2008) to compensate data deficiency. Later extraction of visual feature from videos using deep learning architectures was incorporated into ASR models on open-domain videos (Miao and Metze, 2016). These approaches are extended with SOTA sequence to sequence model (Gupta et al., 2017) which helped to extract relevant context information from the videos for ASR. (Sun et al., 2022) proposes using words from slides and presents GNN encoding using tree-RNN for contextual speech recognition. In addition, (Huber and Waibel, 2025) performs a technique of continuous learning of new words in ASR from slides.

Automatic speech recognition has made a significant progress in recent years by generating accurate transcriptions. Whisper (Radford et al., 2023) has made it possible to generate better transcriptions on unseen datasets. However, transcribing domain-specific datasets or low resource datasets, abbreviations, disfluencies still posses challenge for the SOTA ASR models(Ma et al., 2023). Many approaches focus on fusing audio and visual modalities to address challenges such as proper name transcription, error correction, noisy environments, and multi-modal context (Peng et al., 2023),(Kumar et al., 2023).

In recent work, the integration of presentation slides into Multi-modal ASR has gained attention due to the potential benefits of leveraging visual information to improve transcription. The SLIDESPEECH dataset (Wang et al., 2024b) a large scale audio-visual corpus enriched with slides is created from online conference videos. However only a part of their dataset is transcribed and synchronized with the slides. In a previous work, (Yang et al., 2024) creates a multi-modal-assisted LLM-based ASR model, and uses SLIDESPEECH dataset along with its accessible keywords provided with the dataset to enhance the ASR performance. In contrast to this paper, we explore a strategy to augment existing domain-specific speech-only datasets with images of slides, to enhance model performance on domain-specific vocabulary. Unlike (Yang et al., 2024), we further demonstrate that incorporating images rather than textual context yields additional improvements in ASR performance. Similar to the SLIDESPEECH dataset (Wang et al., 2024a) creates a dataset SlideAVSR, using scientific paper explanation videos. They propose a FQ ranker in this work which helps to select words based on their frequency to be used as prompts. In contrast, we focus on words unique to specifically scientific domain by removing all words commonly existing in a general dataset.

Methods of data augmentation has been proposed to create synthetic data with variations of audio and visual modality for the purpose of enhanced speech recognition (Oneață and Cucu, 2022). In this work, we augment an existing speech-only dataset and enrich them with visual modality for the purpose of multi-modal ASR. (Chen et al., 2024), (Wang et al., 2024a) present a multi-modal academic dataset for audio-visual recognition and understanding tasks. Both datasets requires manual annotation, which is both time consuming and expensive, making such an approach to large data collection non-feasible. In contrast, we show that ASR model performances can be improved when trained through an automatically augmented dataset. While most of the conference videos available are in English, our data augmentation allows utilization of datasets in other languages addition

to english.

3 Multi-modal Scientific Presentation Benchmark

In this section we analyze three baseline models on the ability to transcribe on domain-specific words. The models are evaluated using an evaluation dataset. We describe the dataset in Section 3.1 and give details of model performance on the dataset in Section 3.4.

3.1 Benchmark

For evaluating the model performances we use the ACL 60/60 dataset (Salesky et al., 2023). This dataset consists of a development (*dev*) and evaluation (*eval*) data each with audio recordings and manual transcripts of technical presentations from ACL 2022 conference. Both the dev and eval sets consist of five recordings each. Each of these datasets has a duration of approximately one hour. The dataset consists of manually created aligned text and audio segments which we consider for our task.

3.2 Metric

The traditional Word Error Rate (WER) metric is employed to evaluate the performance of ASR models, assigning equal weight to all words in the transcript. In addition to WER, this study places particular emphasis on the ASR performance for words that are frequently encountered within scientific domains. These words are referred to as domain-specific words, and the term special words is used interchangeably throughout this paper. In this work, we define a domain specific-word as words that does not occure in the general domain corpus (in most experiements this is the Must-C (Di Gangi et al., 2019) corpus) We measure the quality of the domain-specific words with respect to the reference and the hypothesis similar to recall and precision. First, we investigate how many domain-specific words in the reference are missed or wrongly transcribed by the model, by aggregating the deletion and the substitution counts, and dividing it by the total occurrences of domain-specific words in the manual transcript.

In this paper, we calculate a reference-centric WER metric $\text{WER}_{t_{ref}}$.

$$\text{WER}_{t_{ref}} = \frac{|\text{substituted+deleted}|}{|\text{recognized +substituted+deleted}|}$$

Next, we calculate the $\text{WER}_{t_{hyp}}$ to evaluate how many domain-specific words in the model's output are incorrectly transcribed.

$$\text{WER}_{t_{hyp}} = \frac{|\text{substituted+inserted}|}{|\text{recognized +substituted+inserted}|}$$

3.3 Baseline

To study the ability of ASR models to transcribe domain-specific words we use the models, Whisper, SALMONN and Phi-4-multimodal.

Whisper: Whisper is a transformer-based encoder-decoder model developed by OpenAI for ASR and translation tasks (Radford et al., 2023). Trained on approximately 680k hours of web-sourced speech data, it encodes input audio into features, which are then processed by the decoder to generate transcriptions using positional encoding and prior outputs. In this work, we use the Whisper Large V2 model.

SALMONN: The SALMONN model, developed by Tsinghua University and ByteDance (Tang et al., 2023), extends LLMs such as Vicuna (Chiang et al., 2023) to directly process and understand general audio inputs, enabling strong performance on various speech and audio tasks. It integrates outputs from Whisper (Radford et al., 2023) and BEATs (Chen et al., 2022) encoders using a window-level Q-Former module (Zhang et al., 2024), producing augmented audio tokens aligned with the LLM's internal representations. In this work, we use the SALMONN 13B v1 model.

Phi-4-multimodal: Phi-4-multimodal (referred to as Phi) is a 5.6B-parameter, instruction-tuned multi-modal transformer developed by Microsoft. It supports unified processing of text, image, and audio inputs for vision-language, vision-speech, and speech-language tasks, with a context length of up to 128K tokens. The model employs 32 transformer layers with Grouped Query Attention (GQA) (Ainslie et al., 2023) for efficient long-context handling. Vision and audio features are projected into the text embedding space using two-layer MLPs. Phi achieves strong performance across multilingual and multi-modal benchmarks.

3.4 Analysis

We evaluate the models on their ability to transcribe the ACL dataset specifically on domain-specific words.

Table 1 gives the statistics on the domain-specific words extracted from the dataset with this approach. The count of total special words in the ACL dev dataset is 333 of which 130 are unique. Similarly, there are in total 276 special words in the ACL eval dataset of which 115 are unique.

Table 1: Statistics of domain-specific words

			Whisper		SALMONN		Phi	
Data	**Total special words**	**Unique special words**	**Times recognised**	**Times not recognised**	**Times recognised**	**Times not recognised**	**Times recognised**	**Times not recognised**
ACL dev	333	130	251	82	204	129	244	89
ACL eval	276	115	150	126	116	160	150	126

Table 2: WER, $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ for Whisper, SALMONN and Phi.

Model	**ACL dev**			**ACL eval**		
	WER	**$\text{WER}_{t_{ref}}$**	**$\text{WER}_{t_{hyp}}$**	**WER**	**$\text{WER}_{t_{ref}}$**	**$\text{WER}_{t_{hyp}}$**
Whisper	8.81	**24.62**	**20.57**	**13.45**	**45.65**	**44.03**
SALMONN	17.42	38.44	37.31	20.31	57.97	57.04
Phi	**7.01**	26.73	25.38	18.58	45.65	44.65

The results of the model performance on the ACL dataset are summarized in Table 2. We find that for all models, the word error rate ($\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$) on domain-specific words is significantly higher compared to WER on all words. Whisper makes approximately three times more mistakes on ACL dev and eval datasets. Similar results can be also observed for SALMONN and Phi models which implies that all models consistently make more mistakes while transcribing domain-specific words. Additionally, since $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ are similar, there appears to be no specific problem with over or under-generating domain-specific words.

We also present the number of times the models are able to recognize the special words. Columns *Times recognized* and *Times not recognized* of Table 1 show the details of how many of the domain-specific words are recognized and not recognized by Whisper, Phi and SALMONN models respectively. We find that Whisper identifies the highest number of domain-specific words on both the ACL dev and ACL eval sets compared to all other models. Notably, Phi matches Whisper's performance in recognizing domain-specific words on the ACL eval set. Whereas the overall results demonstrate that the domain-specific words pose a difficult challenge for state-of-the-art ASR systems. This motivates the integration of additional context like presentation slides. The following section describes our approach of additional context extraction and integration to models.

4 Multi-modal Context Extraction and Integration

Our analysis on Section 3.4 shows that the current ASR models make up to three times more mistakes while transcribing domain-specific words.

Based on this analysis, we propose a multi-modal context extraction and integration system. We build our system on top of an existing ASR model and enrich it through multi-modal information. We propose both a cascaded approach and an end-to-end approach to incorporate additional information into the model. In both cases, we focus on ASR systems based on multi-modal foundation models to allow an easy integration of additional context. Figure 2 provides an overview of both approaches.

In the cascaded approach, we represent the important domain-specific terms explicitly as words and provide these words to the ASR system. In a first step, we obtain text from extracted images. In a second step, we apply additional filtering on these words. Finally, these words are presented as context to the ASR system.

One disadvantage of this approach is that only the text from the slide is represented and that we can be harmed by cascading errors. Therefore, we also investigate the direct integration of the image in an end-to-end fashion. In this case, the image is provided directly as additional context to the multimodal ASR system.

The following section provides the details on our approach to text extraction from images and integration into models.

4.1 Image Frame Extraction

To obtain the relevant context, we begin with the corresponding video recordings of the scientific talks of the ACL dataset and extract aligned image frames (denoted by 1 in Figure 2). Since presentation video recordings are not usually accompanied by their respective slides, we extract frames directly from the recordings. Our audio segments are less than 30 seconds, therefore we assume that while demonstrating the content of a particular segment, the presenter uses only one single slide.

For each of the audio files, we use the available audio segments, with their durations and offset timestamps relative to the full recording. This information is used to align segments with the original video and extract a single frame from the midpoint of each video segment. The images are then directly integrated into the end-to-end models or processed to extract the specific vocabulary for the

cascaded approach.

4.2 Text Extraction

In the second component, (denoted by 2 in Figure 2) we perform text extraction on the obtained frames from the previous step (Section 4.1). To perform this task, we use LLaVA-NeXT (Liu et al., 2024) (referred to as Llava in rest of this paper), due to its ability of better visual reasoning and optical character recognition (OCR) capability. We provide the model with previously extracted image frames and a suitable prompt as input (explained in Appendix 9), to generate information for each provided frame.

The Llava method results in a large number of extracted texts, which needs to be filtered further (denoted by 3 in Figure 2). The primary motivation behind this is to obtain only domain-specific words. To this end, we filter the extracted text by removing all common words. This is done by discarding all words present in a general presentation dataset (Di Gangi et al., 2019), resulting in a collection of only domain-specific words.

4.3 Context Integration

The extracted information is then provided to an existing multi-modal ASR model (denoted by 4 in Figure 2). Such ASR systems include an LLM which can be prompted with text to perform the required transcription task. In this work, we focus on improving ASR performance by integrating the context as part of such prompts.

In particular, we use the additional information to enrich the input to SALMONN and Phi model. By default, there exists text prompts used in these models that provides instruction (explained in Appendix 9) about the task to be performed. We modify the default text prompt with the information extracted from the previous step (Section 4.2).

5 Data Augmentation

ASR systems with integrated LLMs can be prompted in a zero-shot manner. Existing work (Wei et al., 2021) has shown that compared to zero-shot, fine-tuning of models can be useful to achieve further improvements. To this end, we first perform a zero-shot prompting and further enhance the capability of the ASR model to generate accurate transcriptions by incorporating and training with additional information.

Enhancing ASR using visual modality, a dataset comprising both visual (e.g. images or slides) and

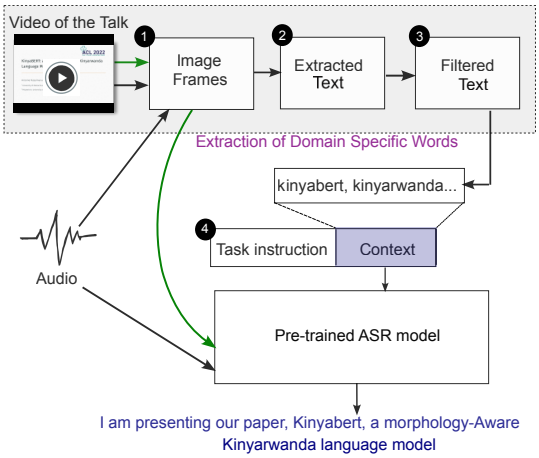

Figure 2: Overview of our two approaches. The green arrows represent the end-to-end approach.

speech data is essential. To address the lack of required relevant multi-modal domain-specific data, this work synthesizes a dataset by data augmentation. For our purpose, we augment images to an existing dataset where we generate images that corresponds to presentation slides. This generated image is then added to the dataset lacking inherent similar multi-modal content. This novel strategy of automatically generating and augmenting a visual modality allows us to use any existing speech dataset while also supporting domain-specific training.

5.1 Generation of Presentation Slides

In this approach, we generate presentation slides for existing speech content through a series of steps. First, we segment the speech transcript into smaller textual units, selecting a chunk size of eight sentences. Our choice of chunk size results in approximately 15–20 slides for a 20–30 minutes speech, ensuring an allocation of 60–90 seconds of speech per slide.

Next, we employ LLaMA 3 to generate LaTeX code for these text chunks. We guide LLaMA 3 with a pair of instructions consisting of a high level system prompt and a more task specific prompt to generate latex code based on the text chunks (explained in Appendix 9). In the final stage, we convert the generated LaTeX code into images. This involves first compiling the LaTeX code into PDFs and subsequently extracting images from the generated PDF files. We adopt a methodology where images are generated from PDFs rather than directly utilizing the PDFs, as such resources are often unavailable in standard datasets. Conversely, presentation videos are typically accessible, which allows us to extract time-aligned slides correspond-

ing to the speech, as described in Section 4.1.

5.2 Text Extraction

After obtaining the images from the generated slides, we follow the approach of text extraction by Llava as described in Section 4.2. Since the target dataset for information augmentation is a general purpose dataset, we apply a separate filtration strategy on the extracted text, differing from the one used for the ACL datasets. For each talk, we retain only the words that are relevant to the talk by discarding words that also appear in all other talks within the dataset. We consider such words that are unique to each talk to be the domain-specific words for that particular talk.

6 Experimental Setup and Results

This section provides details on our experimental setup in Section 6.1 and information about the dataset used for training is included in Section 6.2.

6.1 Experimental Setup

We adopt two models, SALMONN 13B v1, and Phi-4-multimodal to perform our experiments. For extracting text from the images with LLaVA-NeXT, we use llava-v1.6-mistral-7b model which uses CLIP-ViT-L-336px (Radford et al., 2021) as image encoder and LLaMa (Touvron et al., 2023) for language understanding. We provide the model with an image as well as a suitable prompt to generate the text from the image.

For generation of slides we use LLaMa 3 (Dubey et al., 2024) to create latex code and use the python library *subprocess* to execute the shell commands *pdflatex* and *pdftoppm* respectively to generate latex code to PDF and image.

6.2 Dataset

For training the ASR model, we use MuST-C (Multilingual Speech Translation Corpus) (Di Gangi et al., 2019) which is primarily designed as a speech translation dataset. The dataset consists of around 400 hours of audio recordings from English TED Talks speech, transcription and translated transcripts in multiple languages, which are applicable to train model for speech recognition and speech translation tasks.

Since MuST-C does not contain any visual modality, we augment it with the generated images as described in Section 5. Based on the text extraction and filtration approach described in Section 5.2, we obtain 16,830 domain-specific words for 2551 talks present in the dataset.

7 Results

In this section we first analyse the quality of the text extracted using Llava and Phi in Section 7.1. Next, we describe the zero-shot performances of the model on the extracted text presented in Section 7.2 and finally we compare the zero-shot performance of the model to a model fine-tuned using the additional information elucidated in Section 7.3.

7.1 Quality of the extracted text

We perform an analysis to check the quality of the extracted text from the images using Llava and Phi models. This assessment is essential, as the extracted text is intended to support the model's transcription of domain-specific terms. For this, we compare the special words that are present in the reference text to the extracted text. Table 3 summarizes this result. We find that both the Llava and the Phi model produces a large number of unique special words of which 62% and 66% are common to the special words present in the reference of ACL dev and 52% is common to the special words in reference of ACL eval dataset.

We also measure the performance of the ASR models on the Llava and Phi extracted words. The considered models for our qualitative analysis are SALMONN, Phi and Phi+image (Phi trained to perform ASR with image) shown as separate columns in Table 3. As an example, consider the ACL dev dataset where Phi extracted text contains 86 unique special words common to the reference. These 86 words are present in total 260 times in the dataset. The results presented for each ASR models show the number of times out of 260, it has been recognized and not recognized. Consider the results for SALMONN which is able to recognize the Phi extracted special words 164 times but fails for 96 times. Similar ASR model performance results are shown in Table 3 for the Llava extracted text and the reference.

7.2 Zero-shot performance of the ASR model on the extracted data

We evaluate the zero-shot performance of SALMONN and Phi models providing the extracted domain-specific words as prompts and compare it to the model without any additional prompts.

Table 4 shows the results of these experiments. It includes our experiments with two models in four configurations. The first configuration referred to

Table 3: Statistics of domain-specific words extracted using Llava, Phi models and counts of special words recognized and not-recognized by SALMONN, Phi and Phi+image (Phi trained to perform ASR with image).

				SALMONN		Phi		Phi + image	
Dataset	**Text source**	**Unique special words**	**Common with ref**	**Times recognised**	**Times not recognised**	**Times recognised**	**Times not recognised**	**Times recognised**	**Times not recognised**
ACL dev	ref	130	-	204	129	244	89	278	55
	Phi	321	86	164	96	193	67	218	42
	Llava	367	81	173	96	204	65	231	38
ACL eval	ref	115	-	116	160	150	126	179	97
	Phi	645	60	77	108	103	82	124	61
	Llava	669	60	73	107	95	85	125	55

Table 4: WER, $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ scores using context words from Llava, Phi and reference for SALMONN and Phi zero-shot approaches.

Model	**ACL dev**			**ACL eval**		
	WER	**$\text{WER}_{t_{ref}}$**	**$\text{WER}_{t_{hyp}}$**	**WER**	**$\text{WER}_{t_{ref}}$**	**$\text{WER}_{t_{hyp}}$**
SALMONN	17.42	38.44	37.31	20.31	57.97	57.04
+ LlaVA prompts	**10.31**	28.62	28.09	**16.54**	**48.33**	**47.75**
+ Phi prompts	15.36	**27.69**	**27.13**	28.08	58.38	57.92
+ Ref prompts	10.93	17.12	20.66	14.09	35.87	34.93
Phi	7.01	26.73	25.38	18.58	45.65	44.03
+ LlaVA prompts	**6.95**	21.18	20.0	18.29	38.9	38.20
+ Phi prompts	7.05	**20.38**	**19.46**	**15.62**	**38.38**	**37.36**
+ Ref prompts	7.01	18.02	14.95	12.30	37.68	36.06

Table 5: WER, $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ scores of different setup using SALMONN and Phi.

Model	**ACL dev**			**ACL eval**		
	WER	**$\text{WER}_{t_{ref}}$**	**$\text{WER}_{t_{hyp}}$**	**WER**	**$\text{WER}_{t_{ref}}$**	**$\text{WER}_{t_{hyp}}$**
SALMONN						
Zero-shot	17.42	38.44	37.31	20.31	57.97	57.04
Zero-shot Llava	10.31	28.62	28.09	16.54	**48.33**	47.75
Fine-tuned	10.9	30.33	25.48	15.74	51.45	50.0
Fine-tune with Llava	**10.24**	**19.33**	**17.80**	**14.85**	48.89	**46.82**
Fine-tuned with ref	9.67	10.51	8.31	14.63	29.35	26.13
Phi						
Zero-shot	7.01	26.73	25.38	18.58	45.65	44.03
Zero-shot Llava	**6.95**	21.18	20.0	18.29	38.9	38.20
Fine-tuned	9.03	22.30	20.83	13.99	40.22	39.33
Fine-tune with Llava	8.81	17.41	15.85	13.66	35.0	33.52
Fine-tune with image	8.70	**14.13**	**13.48**	**12.23**	**30.56**	**30.17**
Fine-tuned with ref	6.73	16.22	14.68	18.70	41.30	40.22

as base configuration is the models without any additional prompts shown in first and fifth row of the table. The remaining three configurations considers model with additional context using Llava, Phi and from the reference text. We conduct experiment using the special words from reference to show the model performance in the best possible configuration.

We find that the model configurations containing additional context outperforms the base configuration. For the SALMONN model the configuration containing Llava context outperforms the base configuration by 26% and 25% on ACL dev and 17% and 16% on ACL eval on $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ respectively. For the Phi model the configuration with additional context extracted from Phi achieves the best results. It outperforms the base configuration by 24% and 23 % on ACL dev and by 16% and 15% on ACL eval on $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ respectively.

The SALMONN configuration with Phi context performs poorly on ACL eval in comparison to the base configuration. In contrast, we find consistent improvements over the base configuration for the models when special words obtained from Llava are considered. As a result, for further experiments presented in the paper, we only consider special words from Llava.

7.3 Fine-tuning performance using augmented data

For this experiment, our goal is to check if the performance of the ASR models can be improved further by fine-tuning compared to zero-shot performance. To this end, we fine-tune SALMONN and Phi using the augmented dataset obtained in Section 5 and compare it to additional setups described below. Table 5 summarizes the results of our experiment.

The upper part of the table illustrates the SALMONN specific setups and their corresponding results while the lower part contains the Phi specific details. The following provides details on the setups of our experiment that corresponds to Table 5.

Zero-shot Llava: The model with additional context using Llava (Section 7.2).

Fine-tuned: The model fine-tuned without any additional context using the configurations used by the model authors i.e., no changes are made to the task description. (Section 7.2).

Fine-tuned with Llava: The model fine-tuned with additional context words from Llava (default setup).

Fine-tuned with image: The model (only done for Phi since it accepts image as input) fine-tuned with image instead of additional text as context.

Fine-tuned with ref: The model fine-tuned with context obtained as special words from transcripts (best possible setup).

For the setups mentioned above that uses additional context words, we modify the model's task description with additional special words and change the instruction to consider the special words

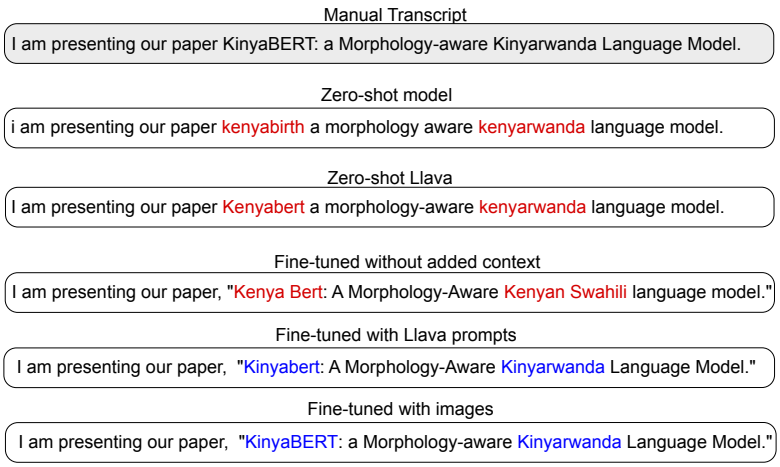

Figure 3: Example of transcriptions generated by different models with respect to the manual transcript. The figure shows that the best possible transcript is generated while fine-tuning the ASR model with llava prompts and image.

while transcribing (explained in Appendix 9). Additionally, we make sure that during extraction of special words as outlined in Section 5, there exists no overlap between special words from training and evaluation datasets.

As illustrated in Table 5, both SALMONN and Phi models improve its overall performance when fine-tuned with Llava context words over fine-tuned with no context words. For the SALMONN setups, fine-tuning with Llava words achieves the best possible scores across both the datasets. We observe, similar results for the Phi setups with additional context words. We conduct additional experiments with Phi using image instead of extracted words as addition context. We find this to be our best possible overall setup for Phi, even outperforming the setup containing context words from reference. This improvements can be attributed to the fact that in addition to text in the slides, the included figures, plots and tables also contribute to the model performance.

We perform a significance test by using matched-pair test for error counts for two hypothesis 1) transcripts from model using only speech 2) transcripts from model using speech and additional context. We find a p-value of less than 0.001 showing the significance of our results.

Figure 3, shows an example prediction by the Phi model with each setup described earlier. Considering both the Zero-shot model and the Fine-tuned model without context words, we find that the models make mistakes on both words *KinyaBERT* and *Kinyarwanda*. The zero-shot with Llava model improves but is unable to transcribe correctly. Whereas the Fine-tuned model with LLava generates the correct transcription likely due to its acquired ability to incorporate from the additional information. Finally, the model trained with images not only accurately transcribes the content but also preserves the textual formatting as it appears in the presentation slide. As illustrated by the above example, our experiments show encouraging results in improving existing ASR performance either using context words or images.

To be used for ASR of scientific talks, the approach requires minimal additional effort to setup. An example setup comprises of a system to generate images from slides of a presenter which is directly utilized by the ASR models for improved transcription.

8 Conclusion and Future work

Current ASR systems exhibit challenges in accurately transcribing domain-specific words. This limitation hinders their effectiveness in various applications. We present an analysis of the model performance on transcribing domain-specific words to demonstrate this. This paper investigates the potential of augmenting ASR models with information extracted from slides to improve performance. We explore the use of visual information extracted from video recordings of slides as prompts. When trained with additional context, the model develops ability to generate better transcription on domain-specific words. This shows the effectiveness of multi-modal information in enhancing ASR performance.

The results presented in Section 7 highlight the potential for further advancements. We find that integrating image as an additional input improves ASR performances for Phi and as future work propose to investigate on SOTA ASR uni-model performances on such end-to-end approaches.

Limitations

While our augmented data approach proves effective and results in significant improvements in model performance, it is not without limitations, presenting opportunities for future research.

In our work we consider slides to extract domain-specific words that can be used as additional information for context integrated ASR. Slides often contains summarized, bullet-pointed information which may lead to omit domain-specific words to some extend which may effect the models ability to recognize them correctly. Speakers often elaborate the slides with their own words introducing mismatch between speech and the slide content which also creates similar problem.

Apart from that, the ASR model in this work integrates a pre-trained LLM. LLMs are heavily dependent on the quality and diversity of their training data. Although we achieve improved model performance with our augmented data there remains further scope of improvement. When integrating additional information to the LLM, it may fail to effectively combine these sources of information, leading to misaligned predictions for some cases. Incorporating LLMs into the ASR pipeline for context integration introduces substantial computational overhead, which can slow down the processing time.

On the other the LLM might misinterpret the contextual information for the speech and lead to produce incorrect transcription.

Our experiment involving image integration into the existing ASR model is limited to the Phi-4-multimodal model. Further comprehensive studies are required to draw conclusive insights into model performance under such configuration.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M^3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. *arXiv preprint arXiv:2403.14168*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Michael Fleischman and Deb Roy. 2008. Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of ACL-08: HLT*, pages 121–129.

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. 2017. Visual features for context-aware speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5020–5024. IEEE.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Christian Huber and Alexander Waibel. 2025. Continuously learning new words in automatic speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Vanya Bannihatti Kumar, Shanbo Cheng, Ningxin Peng, and Yuchen Zhang. 2023. Visual information matters for asr error correction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Rao Ma, Mengjie Qian, Mark JF Gales, and Kate M Knill. 2023. Adapting an asr foundation model for spoken language assessment. *arXiv preprint arXiv:2307.09378*.

Paul Maergner, Alex Waibel, and Ian Lane. 2012. Unsupervised vocabulary selection for real-time speech recognition of lectures. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4417–4420. IEEE.

Yajie Miao and Florian Metze. 2016. Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, page 3.

Dan Oneață and Horia Cucu. 2022. Improving multimodal speech recognition by data augmentation and speech representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4579–4588.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. *arXiv preprint arXiv:2305.11095*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 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.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Guangzhi Sun, Chao Zhang, and Philip C Woodland. 2022. Tree-constrained pointer generator with graph neural network encodings for contextual speech recognition. *arXiv preprint arXiv:2207.00857*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hao Wang, Shuhei Kurita, Shuichiro Shimizu, and Daisuke Kawahara. 2024a. Slideavsr: A dataset of paper explanation videos for audio-visual speech recognition. *arXiv preprint arXiv:2401.09759*.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li. 2024b. Slidespeech: A large scale slide-enriched audio-visual corpus. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11076–11080. IEEE.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Guanrou Yang, Ziyang Ma, Fan Yu, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Mala-asr: Multimedia-assisted llm-based asr. *arXiv preprint arXiv:2406.05839*.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

9 Appendix

Textual Context Integration to SALMONN We instruct SALMONN by providing text prompts to Vicuna that ask questions about the processed audio. The LLM then responds with textual answers based on its understanding. The model is trained for various speech related tasks with suitable prompt structure, as follows

```
USER: [Auditory Tokens] Can you transcribe the
    speech into a written format? \n ASSISTANT:
```

Here, *[Auditory Tokens]*are the output tokens of the window-level QFormer, followed by user prompts in the form of questions with respect to the task performed by the model on the given audio.

Our extracted domain-specific terms from accompanying slides are included in prompts with the following structure

```
USER: [Auditory Tokens] Please can you
    transcribe the speech referring to the
    following tokens wherever needed:
    kinyarwanda, kinyabert, nlp, pre-trained,
    ...? \n ASSISTANT:
```

Here, domain-specific words like *Kinyarwanda, Kinyabert, NLP, and pre-trained* are included in the user prompt. The overall prompt is designed to emphasize both these special words and the task itself.

Context Integration to Phi Depending on the input required for training Phi-4-multimodal modal, we construct its prompt format.

Format for Speech-Language with special words:

```
user_message = {
    "role": "user",
    "content": "<|audio_1|>\n" + Can you
         transcribe the given speech referring to
          the following words wherever needed
         #### kinyarwanda, kinyabert, nlp, pre-
         trained, ...?
}
```

Format for Speech-image-Language:

```
user_message = {
    "role": "user",
    "content": "<|image_1|>\n<|audio_1|>\n" +
    Can you transcribe the given speech?
}
```

Model Instruction for Text Extraction To exhibit LLaVa-Next models OCR quality an extract text from slides we provide the model with an image and a suitable text prompt. the structure of the instruction is given as follow:

```
"[INST] <image>\nUSER: Extract the text from the
     sides? [/INST]"
```

the <*image*> tag is replaced with the image input for LLaVa-Next following with the user prompt. The instruction should always start with the *[INST]* tag and end with *[/INST]* tag.

Model Instruction for Data Augmentation For creating the multi-modal context for data augmentation, we use LLaMa 3 and guide it with a pair of instructions consisting of a high level system prompt and a more task specific prompt to generate latex code based on text chunks. This consists of a system prompt and a user prompt as follows:

```
{"role": "system", "content": "you are a
    presenter who wants to inform and inspire"},

{"role": "user", "content": generate one
    presentation slide with the main points and
    concepts in latex, from the following text:<
    chunk>}
```

The *chunk* in the user prompt is replaced by the parts of talk for which we want to generate the latex code.