

# Exploring the Potential of Multimodal LLM with Knowledge-Intensive Multimodal ASR

Minghan Wang<sup>1</sup>, Yuxia Wang<sup>2</sup>, Thuy-Trang Vu<sup>1</sup>,  
Ehsan Shareghi<sup>1</sup>, Gholamreza Haffari<sup>1</sup>

<sup>1</sup>Department of Data Science & AI, Monash University

<sup>2</sup>MBZUAI

{minghan.wang, trang.vu1, ehsan.shareghi, gholamreza.haffari}@monash.edu  
yuxia.wang@mbzuai.ac.ae

## Abstract

Recent advancements in multimodal large language models (MLLMs) have made significant progress in integrating information across various modalities, yet real-world applications in educational and scientific domains remain challenging. This paper introduces the Multimodal Scientific ASR (MS-ASR) task, which focuses on transcribing scientific conference videos by leveraging visual information from slides to enhance the accuracy of technical terminologies. Realized that traditional metrics like WER fall short in assessing performance accurately, prompting the proposal of severity-aware WER (SWER) that considers the content type and severity of ASR errors. We propose the Scientific Vision Augmented ASR (SciVASR) framework as a baseline method, enabling MLLMs to improve transcript quality through post-editing. Evaluations of state-of-the-art MLLMs, including GPT-4o, show a 45% improvement over speech-only baselines, highlighting the importance of multimodal information integration.<sup>1</sup>

## 1 Introduction

Recent years have witnessed remarkable progress in multimodal large language models (MLLMs), with GPT-4o achieving impressive performance on tasks involving the integration of information from multiple data modalities such as text, images, speech and video (OpenAI et al., 2024). However, many real-world applications, particularly in the educational and scientific domains, present significant challenges that push the boundaries of current multimodal capabilities (Gupta et al., 2023; Benedetto et al., 2024).

One such challenging task is multimodal automatic speech recognition (ASR) from knowledge-intensive presentation videos, which involves not

<sup>1</sup>Our code is available at <https://github.com/yuriak/MS-ASR>

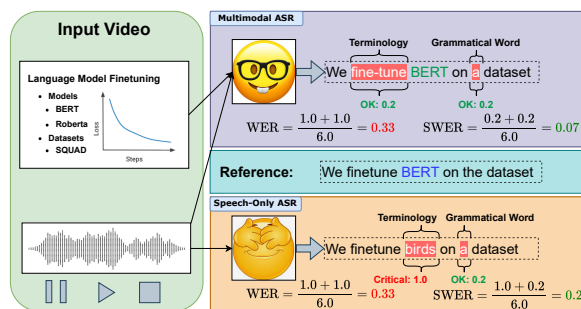


Figure 1: This figure illustrates the importance of visual information in accurately recognizing the terminology *BERT*. It also introduces our proposed evaluation metrics Severity-aware WER (SWER) by calibrating WER with the severity of ASR errors.

only transcribing spoken content but also understanding and integrating visual information from images or videos (Oneata and Cucu, 2022; Chang et al., 2023). Presentation videos, often found in scientific conferences or educational contexts, contain dense, knowledge-rich content spanning various topics and concepts. Effective recognition and comprehension of such content are crucial for applications such as automatic transcript generation (Angrave et al., 2020), simultaneous translation (Wang et al., 2024b,a) in conference venue, lecture summarization (Benedetto et al., 2024) and accessibility tools for researchers and learners (McCarron, 2021). ASR on such videos presents several challenges, including keeping pace with the rapid evolution of technical domain knowledge and handling the variability in speaker accents, proficiency in English, and speech fluency (Cettolo et al., 2004). Mitigating these challenges requires ASR models to effectively extract and incorporate information from the speech and presentation slides in the video recordings (see Figure 1).

In this paper, we investigate whether the current multimodal systems can successfully generate transcripts for presentation videos. We propose the

task of transcribing scientific conference presentation recordings, namely Multimodal Scientific ASR (MS-ASR). Unlike conventional multimodal ASR tasks that focus on motion-rich scenarios, MS-ASR focuses on knowledge-intensive content, requiring accurate transcription of domain-specific terminology using visual information. Standard evaluation metrics like word error rate (WER) fail to accurately assess these specialized terms, leading to biased performance evaluations. To address this, we propose a more nuanced evaluation protocol: severity-aware WER (SWER) for the task, which calibrates WER based on the severity of ASR errors, aligning system performance more closely with human evaluation.

Solving the proposed MS-ASR task requires proficiency in several areas, including speech recognition, visual information extraction and integrating them to post-edit the transcript. To this end, we propose a zero-shot inference framework named Scientific Vision Augmented ASR (SciVASR) that leverages MLLMs to collaboratively produce contextually accurate transcripts. We further explore the potential of the current SoTA MLLM, such as GPT-4o, for end-to-end post-editing the transcript. Experimental results demonstrate that MLLMs can effectively improve the quality of the transcript compared to the speech-only ASR baseline, where visual information plays a key role for all MLLMs including GPT-4o to obtain the improvements. We found that speech-only ASR suffers from high term errors (298) and 43.04% critical errors, adding visual information not only reduces 152 term errors but also 28% critical errors (Section 6). Human evaluation confirmed that remaining errors are negligible to affect the understanding on transcripts.

The contributions of this paper are threefold.

- We introduce the Multimodal Scientific ASR (MS-ASR) task and evaluate both open- and proprietary SoTA MLLMs with our proposed zero-shot inference framework Sci-VASR.
- We propose a more fine-grained ASR evaluation framework and severity-aware WER (SWER) metric. It can systematically assess transcription accuracy across multiple dimensions, closely reflecting human evaluation.
- Our experiments demonstrate that the proposed task effectively evaluates MLLMs' multimodal capabilities, with GPT-4o performing

the best, achieving a 45% improvement in SWER over the speech-only baseline.

## 2 Multimodal Scientific ASR

### 2.1 Task Formulation

Given a presentation video consisting of a sequence of visual frames and a corresponding speech audio sequence, the goal is to generate an accurate transcript of the spoken content in the video. The visual frames typically contain presentation slides with text, images, diagrams and other visual aids relevant to the presented topic. The speech audio comprises the spoken explanations and narrations from the presenter, which are closely coupled with the visual content displayed on the slides (Figure 1). The knowledge-intensive and technical nature of the presentation content requires the ASR model the ability to recognize domain-specific terminologies and effectively integrate information from both visual and speech modalities. Unlike conventional ASR task, which operates at sentence-level utterances over well-split speech segments, we treat each video as a whole and generate a presentation-level transcript for the entire video. This task formulation aims to evaluate the accuracy of the ASR system on long videos and also aligns better with the real-world application where pre-segmented utterances are not typically available.

Standard ASR evaluation metrics, such as WER, can be used to evaluate model performance on this task. In our preliminary experiments, we found that WER does not align well with human evaluations and fails to adequately measure the model's ability to recognize technical terminologies (§4.3). Therefore, we propose severity-aware evaluation protocol that calibrates WER with the severity of ASR errors, providing a more nuanced and accurate assessment of model performance (§5).

### 2.2 ACL60/60 Dataset

We utilize the ACL 60/60 dataset comprising video recordings of accepted papers at ACL 2022 (Salesky et al., 2023). Each recording lasts 10-15 minutes, with a speaker presenting their paper in English. It was primarily used for speech translation evaluation but contains high-quality human-annotated ASR transcripts and terminologies, making it a suitable dataset for our research.

This dataset comprises 10 video recordings, in total of around 120 minutes, with 5 videos split as the development set and 5 as the test set. These videos

ID	Gender	L1	Country	Track	Length	# Slides	DS ↓
268	F	Spanish	Spain	Resources and Evaluation	00:11:35	18	0.853
367	M	Kinyarwanda	Rwanda	Theme: Language Diversity (Best Paper)	00:11:35	18	0.824
590	M	Polish	Poland	Machine Learning for NLP	00:12:17	24	0.756
110	M	-	-	Dialogue and Interactive Systems	00:12:09	27	0.683
117	F	Marathi	India	Question Answering	00:09:37	14	0.831
410	M	Chinese	China	NLP Applications	00:12:03	20	0.657
468	M	-	Belgium	Resources and Evaluation	00:12:02	16	0.891
567	F	Romanian	Romania	Language Grounding, Speech and Multimodality	00:09:22	16	0.937
597	M	Japanese	Japan	NLP Applications	00:14:02	20	0.490
111	M	Hebrew	Israel	NLP Applications	00:11:53	15	0.571

Table 1: Meta information (video ID, gender, native language (L1), country, track and video length are from (Salesky et al., 2023), #Slides is counted manually) and difficulty score (DS) of the dataset where lower scores indicate higher difficulty.

are uniformly sampled based on the speaker’s nationality, accent, gender, English proficiency, and topic, as shown in Table 1. We also estimate the difficulty of each video by calculating the ratio of the conditional perplexity of the baseline speech generated by the text-to-speech system and the real speech. The detailed calculation of the difficulty estimation is in Appendix A.

Compared to other multimodal ASR datasets like How2 (Sanabria et al., 2018), the ACL60/60 dataset differs due to its greater sample-wise length (presentation level vs utterance level), more specific domain coverage (scientific domain in NLP vs daily life), and higher knowledge density, making it an ideal benchmark for the proposed task.

### 3 Scientific Vision-Augmented ASR

Due to the nature of the long, knowledge-rich, and un-segmented videos, to effectively tackle the task of transcribing presentation videos, a multimodal ASR must excel at four subtasks: (i) *speech recognition* to accurately transcribe the speech audio, (ii) *video segmentation* to identify slide boundaries, (iii) *visual information extraction* from each slide, and (iv) integrating the visual information to *post-edit* the initial speech recognition transcript. In this section, we present our approach to address these subtasks, illustrated in Figure 2.

#### 3.1 Speech Recognition

We utilize state-of-the-art ASR models, such as Whisper-base (Radford et al., 2023) to transcribe the video’s speech into an initial transcript along with corresponding timestamps. While this provides a reasonable starting point, the transcript may contain errors, especially in technical terminology, due to its knowledge-intensive nature.

#### 3.2 Video Segmentation

In contrast to motion-rich videos, where each frame contributes to the depiction of a continuous movement, presentation recordings are typically composed of static scenes with slides presented by the speaker. To reduce the complexity and computational cost, we segment the video into a sequence of scenes, with one slide representing the visual context of each scene.

Specifically, we first apply the Kernel Temporal Segmentation (KTS) algorithm (Potapov et al., 2014; Afham et al., 2023) on the video frame features extracted by a pretrained image encoder, e.g., CLIP (Radford et al., 2021). KTS derives the optimal segmentation policy by identifying the boundaries (change points) of each scene from the video frames. The output of KTS is a sequence of consecutive time spans in seconds, with each span indicating the start and end times of a scene.

We find that the predicted scene boundaries align well with the moment when the speaker moves to the next slide. Therefore, for each scene, we sample a frame image at 90% of the time span before the speaker moves to the next slide<sup>2</sup>. This aims to ensure that the sampled frame contains the complete textual content of the slide even if there were animations. Processing only 10-30 scene images instead of all video frames significantly reduces computational demands without losing key information. The scene time spans also guide the segmentation of transcripts corresponding to each slide. Ultimately, we obtain a sequence of scenes  $\mathcal{S} = [S_1, \dots, S_N]$ , where  $N$  is the number of scenes. Each scene is represented by a tuple  $S_i = (\mathcal{I}_i, \mathcal{T}_i)$  with the slide image  $\mathcal{I}_i$  and the corresponding tran-

<sup>2</sup>The number of sampled slides from the predicted scenes has an average difference of less than 1.2 pages compared to manually counted number.

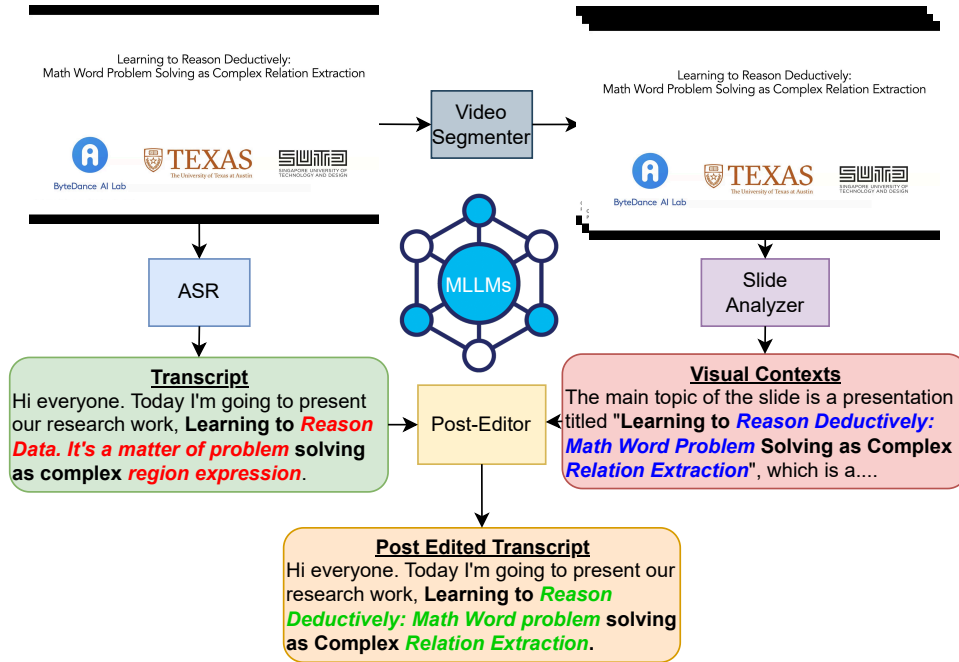


Figure 2: **The architecture of SciVASR.** Baseline ASR model transcribes a presentation recording into a transcript. **Video segmenter** splits video frames into a sequence of scenes with each presenting a slide. **Slide analyzer** applies multimodal LLMs to extract key information from scene images into visual contexts represented by text. **Post-editor** instruct a text-LLM to leverage visual contexts to refine the ASR transcript.

scripts  $\mathcal{T}_i$ , where  $i$  is the scene index.

### 3.3 Slide Analysis

Visual contexts are extracted from the segmented scene images and transformed into natural language. These textual representations of scenes are further used as references for transcript refinement.

**Visual Context Extraction** A text-LLM and a vision-LLM collaborate to extract the required visual contexts from the slide images. The necessity of using LLMs in both modalities instead of only a vision-LLM lies in the fact that many state-of-the-art vision-LLMs cannot handle complex reasoning tasks, particularly when dealing with false-premise or unanswerable questions (Wang et al., 2023a). (See Figure 4 for example) In contrast, text-LLMs are superior in reasoning and are more robust in these situations. Combining both types of LLMs offers a positive complementary effect.

Specifically, we predefine eight questions covering aspects such as layout, title, bullet points, figures, and tables in the slide that are used to prompt the vision-LLM to extract relevant information (see questions in Table 6 of Appendix B). These answers are often granular, duplicative, excessively lengthy, and even contain misinformation. Therefore, we regard them as raw contexts, and fur-

ther leverage a text-LLM to cross-validate, correct inaccuracies caused by the vision-LLM hallucinations, and summarize the lengthy raw contexts into a concise summary, referred to as the scene-level context  $C_i$ . This summary is added to the scene tuple  $S_i = (\mathcal{I}_i, \mathcal{T}_i, C_i)$ .

**Visual Context Condensation** For certain scenes, information on the current slide might not suffice to correct errors in the corresponding transcripts, as the relevant cues could appear in earlier slides. Therefore, we use the text-LLM to further condense scene-level summaries into an overall presentation-level summary to share temporal contexts across scenes. This is achieved by prompting the text-LLM to generate a comprehensive summary based on the concatenated scene-level summaries, denoted as  $\mathcal{P}$ . Both scene-level and presentation-level summaries will be passed to post-edit the initial transcripts.

### 3.4 Post-editing

In the post-editing stage, we instruct the text-LLM to utilize the extracted visual context  $C_i$  and  $\mathcal{P}$  to correct errors in the ASR transcript  $\mathcal{T}_i$  (See prompts in Figure 5 of Appendix B). The output of the post-editor for each scene is denoted as  $\hat{\mathcal{T}}_i$ .  $\sum_{i=1}^N \hat{\mathcal{T}}_i$  constitutes the final output.

**End-to-end Vision Post-editing** As GPT-4o has the capability of solving complex vision-language tasks. We set a special step only for GPT-4o to prompt it post-edit the transcript given the slide image in an end-to-end manner. In this setting slide analysis and post-editing are merged into one step, performed by GPT-4o directly.

## 4 Experiments

### 4.1 Models

Models used in our experiments involve three types across multiple modalities.

**ASR Models** We experiment with Whisper-base and Whisper-large-v3 (Radford et al., 2023) as speech recognition models. To challenge the effectiveness of our proposed method, we use the inferior Whisper-base as the baseline ASR whose outputs demands more complex refinement.

**Vision-LLMs** We evaluate both open-sourced model CogAgent-VQA (Hong et al., 2023; Wang et al., 2023c) given the strengths in GUI and textual-related VQA tasks, and closed-sourced model GPT-4o (OpenAI et al., 2024).

**Text-LLMs** Based on the observation that models with less than 13B parameters often fail to follow complex instructions and exhibit severe hallucinations, we apply open-sourced Llama3-70B-Instruct (AI@Meta, 2024) and closed-sourced GPT-4o<sup>3</sup> in our experiments for effective error correction.

### 4.2 Experimental Settings and Metrics

We experiment in four settings:

- **ASR-Only** - an ASR-only baseline using Whisper-base and Whisper-large-v3.
- **Text-Only Post-editing (Text-PE)** employs the text-LLM to post-edit the transcripts from the ASR baselines without using any visual information. Post-editing is performed on scene-level transcripts.
- **Visual-Augmented Post-editing (Vision-PE)** is the full pipeline of our proposed method presented in Section §3. It performs post-editing with the visual information. In the open-source model setup, Llama3-70B-Instruct

and CogAgent-VQA are used. In the proprietary model setup, GPT-4o deal with both visual information extraction and post-editing.

- **End-to-end Vision Post-editing (E2E-Vision-PE)** where GPT-4o directly correct the transcript with the slide image without the slide analysis step, as described in §3.4.

**Evaluation Metrics** Following prior work, we use Word Error Rate (WER) to evaluate the quality of ASR outputs (Salesky et al., 2023). We additionally compute terminology recall given that the dataset provides annotated terminologies in the ground-truth transcript. We also conduct a human evaluation by comparing the raw ASR transcript, the text-only post-edited version, and vision-augmented post-edits. Specifically, the annotators (introduced shortly) are asked to determine which version is better given the ground-truth transcript, without revealing the system they are evaluating. Both automatic metrics and human evaluation are conducted at the scene level.<sup>4</sup>

**Annotators** To guarantee the annotation quality, instead of employing crowd-sourcing annotators, we perform in-house labeling by four annotators who are Master’s and PhD students and postdocs and are familiar with ASR. Two annotators as a group are responsible for 3-4 videos. Annotators first independently finish individual annotations and then consolidate their results with the group partner. In consolidation, partners discuss their disagreements until reaching a consensus.

### 4.3 Results and Analysis

The results with four configuration groups are presented in Table 2. A negative correlation (-0.636 in spearman- $\rho$ ) is found between the Whisper-base WER and the difficulty score in Table 1, indicating that the harder the video, the higher the WER. In terms of Term-Recall, compared with the baseline Whisper-base, Whisper-large-v3 performs better, which is within our expectations. **Text-PE** using only the text-LLM for post-editing also improves the term recall, and including visual context (**Vision-PE**) brings further improvements across all videos. GPT-4o consistently outperforms open-sourced combinations for both text- and vision-PE,

<sup>4</sup>Original dataset provides the utterance level ground truth, we re-segment both references and hypotheses for each video using <https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz>

<sup>3</sup>Model signature: gpt-4o-2024-05-13

VID	ASR only				Text-PE				Vision-PE				E2E-Vision-PE	
	Whisper-base		Whisper-large		Open		GPT-4o		Open		GPT-4o		GPT-4o	
	WER↓	TR↑	WER↓	TR↑	WER↓	TR↑	WER↓	TR↑	WER↓	TR↑	WER↓	TR↑	WER↓	TR↑
110	14.41	86.06	<b>0.35</b>	4.96	<b>0.70</b>	1.82	-1.23	7.88	<b>1.17</b>	6.06	<b>-2.52</b>	<b>9.70</b>	-1.41	9.70
117	11.16	89.11	<b>4.31</b>	1.75	<b>1.84</b>	<b>-0.40</b>	-1.46	2.02	<b>5.31</b>	0.40	<b>-1.73</b>	4.84	-0.76	<b>5.65</b>
268	18.58	85.92	<b>5.80</b>	2.96	<b>1.03</b>	4.37	-1.57	6.80	<b>1.35</b>	7.77	<b>-1.79</b>	<b>11.65</b>	<b>0.38</b>	11.65
367	17.99	76.44	-1.21	3.89	-2.04	4.89	-3.42	8.44	-1.89	9.78	<b>-4.23</b>	9.33	-3.06	<b>10.22</b>
590	14.44	95.73	<b>5.19</b>	<b>-2.12</b>	<b>0.91</b>	0.86	-0.99	0.86	<b>0.33</b>	0.86	<b>-1.07</b>	<b>1.71</b>	<b>2.56</b>	<b>-3.42</b>
111*	33.12	67.23	-0.87	7.15	-0.89	9.60	-1.13	2.82	-4.13	16.95	<b>-5.34</b>	<b>18.08</b>	-1.05	16.38
468*	19.81	84.30	<b>4.32</b>	7.14	-1.35	8.14	-3.40	8.72	-0.18	11.63	<b>-5.39</b>	11.05	-4.28	<b>12.79</b>
410*	22.99	77.94	<b>-2.19</b>	2.41	<b>1.72</b>	2.21	-1.55	6.62	<b>0.92</b>	11.77	-1.49	<b>13.97</b>	-1.26	13.97
567*	9.73	83.87	<b>4.90</b>	<b>-2.82</b>	<b>1.39</b>	1.94	-0.85	4.52	<b>0.23</b>	7.10	-0.85	<b>10.32</b>	<b>-1.70</b>	9.68
597*	28.58	71.84	<b>0.51</b>	9.17	-1.65	6.90	-3.15	8.05	-5.44	16.09	<b>-8.02</b>	<b>18.97</b>	-4.37	18.97
AVG	19.08	81.85	<b>2.11</b>	<b>3.45</b>	<b>0.17</b>	4.03	-1.88	5.67	-0.23	8.84	<b>-3.24</b>	<b>10.96</b>	-1.49	10.56

Table 2: This table presents the full settings of our experiment showing the result in WER and Term-Recall (TR). Results are organized into 4 groups with various configurations (introduced in §4.2). Videos with \* are from the test set. For clarity, we show the relative improvements compared to the Whisper-base with red and green distinguishing worse or better performance. Best results are highlighted with bold.

the end-to-end setting of GPT-4o obtains comparable performance to its pipelined version. This exhibits the effectiveness of incorporating visual information.

**Why does post-editing increase WER?** Despite the significant improvements in term recall, WER doesn’t reveal consistent improvements. Open-sourced PE methods even increase the WER in 6 out of 10 videos. The discrepancy and the counter-intuition motivate us to investigate why the improvements aren’t reflected on WER. By analyzing the model outputs, we find that while post-editing accurately corrects a certain amount (roughly 40%) of explicit homophone errors, WER still judges them as errors due to the tiny differences with ground truth text in spelling, e.g., *present* vs. *presents*, and *fine-tuning* vs. *finetuning*. Additionally, acceptable minor rephrasing caused by the post-editing also increases WER, despite the improvements in fluency. These negligible errors are penalized similarly in WER compared with severe errors like recognizing *BERT* as *birds*, making **WER fail to reflect the true quality of the transcript**, significantly deviating from human judgments as shown in Figure 3. Post-edited transcripts are consistently voted to be better than the baseline transcript, and **Vision-PE** is more preferred than **Text-PE** results. To address this, we propose a severity-aware ASR evaluation protocol in §5.

## 5 Severity-Aware Evaluation Protocol

When evaluating the transcripts, the annotators focus more on the correctness of keywords, termi-

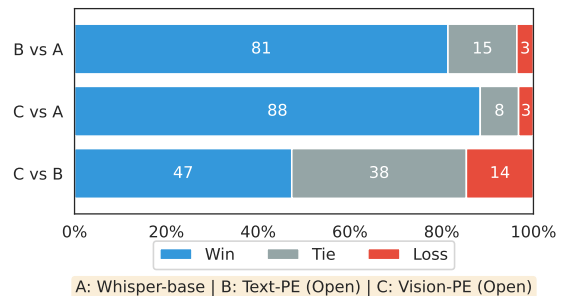


Figure 3: The win, tie and lose ratio between different settings by human annotations, based on Whisper-base and SciVASR in open-source setup.

nologies, and numbers, which could impact the meaning of the speech, and care less about minor errors in casing, tense, spelling and grammar. This inspires us to propose the severity-aware evaluation protocol, imitating human evaluations.

### 5.1 Breakdown of WER

WER is computed by counting the number of required operations (insertion, omission, substitution –  $I, O, S$ ) to transform the hypothesis into the reference and then normalizing it by the reference length  $N$ . Each operation relates to a mismatch in the hypothesis. Formally, WER is calculated as  $\frac{I+O+S}{N}$  where it assigns equal weights to all operations, without considering the varying importance and the semantic impact of the different types of mismatched content. This makes the phenomenon that a transcript with ten errors on general words like “the” or “a” has the same WER as a transcript

with ten errors in terminologies. We mitigate it by re-weighting mismatches according to the content type and error severity.

## 5.2 ASR Error Taxonomy

Inspired by previous studies in NMT that define nuanced error taxonomy and explainable metrics for evaluating the translation hypothesis (Guerreiro et al., 2023; Sharou and Specia, 2022; Freitag et al., 2021), we propose a taxonomy for ASR errors from two perspectives: content type and severity level of mismatches. The difference between ASR and NMT error taxonomy is discussed in Appendix E.

**What is a mismatch pair?** Given a pair of (hypothesis, reference), where the hypothesis is the predicted transcript. A mismatch is a text span differing between hypothesis and reference at the same position, identified by the edit-distance algorithm.

**Content Type** The content type reflects the nature of the recognized terms in conveying the intended meaning of the spoken language. We classify the mismatched content into six types including terminology, numerical data, named entities, grammatical words, disfluencies & fillers and general words. We label the content type of each mismatch in the hypothesis by its corresponding span in the reference text – *ground-truth content*. Table 3 shows detailed definitions and corresponding examples. Categorizing mismatches by the content type help interpret the nature of each error and capture the full distribution of errors.

**Severity Level** The content type of a mismatch is not sufficient to determine whether it leads to a severe recognition mistake. For example, transcribing the terminology “finetuning” as “fine-tuning” is tolerable due to a minor formatting change. Therefore, we further categorize each mismatch into three levels of severity based on its impact on understanding the text’s meaning, including ok, minor and critical, see examples in Table 3.

## 5.3 Severity-Aware WER

Based on the error taxonomy, we introduce the Severity-aware WER (SWER) — WER calibrated by the error severity. It distinguishes different types of mismatches, making the evaluation of transcripts more accurate and better aligned with human evaluation. Computation of SWER requires three steps.

**Pair and highlight mismatches** using the three types of operations for both reference and hypothesis transcripts. We implement this by first using Levenshtein distance to identify the required operations given the transcript reference and hypothesis. Then, three types of parentheses are used for highlighting: “[ ]” for substitutions, “< >” for insertions, and “{ }” for omissions. The highlighted references and hypotheses are presented in two paragraphs, as the input of the next step: annotating mismatches using the error taxonomy.

**Annotate content type and error severity of mismatch pairs** This can be performed by either human annotators or automatic methods such as LLM-based categorization. We used both in our experiments. Note that annotations in this step demand understanding the semantics of the text and making full use of the context. LLMs are recommended leveraging its powerful language understanding capability.

**Calculate SWER** After acquiring the content type and error severity of each mismatch, we can re-weight the importance of each mismatch during the computation of WER. We define SWER as

$$\text{SWER} = \frac{\sum_{i=1}^K w_i}{N}$$

where  $N$  is the reference length,  $K$  is the number of mismatches which is same as the total number of operations required in WER, and  $w_i$  is the weight corresponding to the severity level of each mismatch. In our experiment, we simply define the mapping of severity levels to weights as  $w_{critical} = 1.0$ ,  $w_{minor} = 0.6$ , and  $w_{ok} = 0.2$ . The rationale is to down-weight less severe mismatches and place greater emphasis on critical errors. We also tried the setting of (1.0, 0.5, 0.1) and found that (1.0, 0.6, 0.2) gives a better correlation with human evaluations on the dev set.

## 5.4 How reliable is LLM evaluation?

We prompt an LLM to label each mismatch with its content type category and severity level as shown in Figure 7 (see detailed prompt in Figure 6 in Appendix C). GPT-4o<sup>5</sup> is used in our experiments.

Compared with human annotators who follow the same prompt as guideline to annotate the errors in transcripts produced by ASR-only (Whisper-base results), the LLM performs well

<sup>5</sup>Model signature: gpt-4o-2024-05-13

Names	Definition	Example
CONTENT TYPE CATEGORY		
<b>Terminology (TERM)</b>	Specialized words or phrases used extensively within a specific field or industry.	“Neural networks” in a tech discussion
<b>Numerical Data (NUM)</b>	Numbers, dates, quantities, or statistical data that provide specific, quantifiable details.	“3.14”, “thirty percent” or “3 out of 4”
<b>Named Entities (NE)</b>	Names of people, places and organizations.	“Albert Einstein” as a person’s name
<b>Grammatical Words (GRAM)</b>	Words that primarily serve a grammatical function in a sentence, including articles, prepositions, conjunctions, and auxiliary verbs.	“the” in “the cat”
<b>Disfluencies and Fillers (DISF)</b>	Words or phrases that reflect speech disfluency, including fillers, repetitions, or corrections.	“uh”, “um”, or repeated words due to stuttering
<b>General Words (GEN)</b>	Common words that are not specialized terms, named entities, grammatical words, disfluencies, or fillers.	“very” in “a very big dog”
SEVERITY LEVEL		
<b>OK (OK)</b>	No significant impact on the meaning.	REF: “We try to um try to do” HYP: “We try to do”
<b>MINOR (MIN)</b>	Slight impact on the meaning, but the main point is still understandable.	REF: “this is the number” HYP: “it is the number”
<b>CRITICAL (CRI)</b>	Significant or complete change or distortion of the original meaning.	REF: “finetune BERT” HYP: “finetune birds”

Table 3: The categories of content type and severity level of ASR errors with abbreviations, definitions and examples.

Content Type	Precision	Recall	F1
<b>TERM</b>	98.34	91.93	95.02
<b>NUM</b>	98.21	98.21	98.21
<b>NE</b>	70.00	100.00	82.35
<b>GRAM</b>	91.78	97.81	94.70
<b>DISF</b>	91.43	67.37	77.58
<b>GEN</b>	93.82	96.72	95.25
<b>Micro Average</b>	94.04	93.59	93.52
Severity Level	Pearson	Spearman	
	93.45	93.45	

Table 4: The performance of LLM-based evaluator on Whisper-base transcripts of all videos, against gold labels annotated by humans. Pearson and Spearman correlation coefficients are calculated with three levels of severity converted into integers (0, 1, 2).

in the content type classification task, achieving an F1 score of 93% as shown in Table 4. We also convert the severity levels into integers (0,1,2) and compute the Pearson and Spearman correlations between humans and GPT-4o, obtaining significant positive correlations that are more than 0.93. This indicates that GPT-4o is capable of accurately classifying mismatches into the correct categories and assigning reasonable severity levels.

## 6 Re-Evaluating SciVASR

In this section, we re-evaluate our proposed framework using the fine-grained metrics.

**Results and Analysis** As shown in Table 5, we present metrics including term recall, WER and SWER for overall evaluation. Additionally, we calculate the content-type-wise severity scores for all mismatches in the test set. The weighted score sums up the severity weight of mismatches belonging to this category:  $\sum_{i=1}^{K_c} w_i$ , unweighted score is simply the number of mismatches of the category:  $K_c$ . There are four major findings.

**Parameterized knowledge of Text-LLM helps** Compared with ASR baselines, text-only post-editing without visual information reduces both SWER and the content-wise severity scores to some extent. SWER decreases from 5.0 (**Whisper-base**) to 4.76 in open-sourced setting and 4.08 in close-sourced setting (**Text-PE**), indicating that parameterized knowledge in text-LLM is helpful in correcting potential errors.

**Visual context is critical for correcting terminology, named entity and numbers** Incorporating visual information brings a significant reduction in the severity scores for TERM (-152), NE (-38) and NUM (-5) in the close-sourced group (GPT-4o **Vision-PE** vs **Whisper-base**). SWER reaches 2.75 from 5.0. This suggests that visual modalities are highly effective in correcting these types of errors.

**Evaluation with SWER is reliable** SWER can accurately reflect the relationship between the severity of errors and the final scores. This is con-



Settings	Content-wise Severity Score (weighted/unweighted)↓						Metrics		
	TERM	NUM	NE	GRAM	DISF	GEN	WER ↓	SWER ↓	Term-Recall ↑
<i>ASR Baselines</i>									
Whisper-base	271.4/298	45.9/111	68.2/69	42.2/147	<b>14.6/69</b>	349.0/535	<b>19.08</b>	5.00	81.85
Whisper-large-v3	<b>185.7/237</b>	<b>44.8/110</b>	<b>58.5/66</b>	<b>39.2/117</b>	22.6/81	<b>336.7/590</b>	21.19	<b>4.62</b>	<b>85.29</b>
<i>SciVASR with Open-sourced LLMs (Llama3-70B-Instruct + CogAgent-VQA)</i>									
Text-PE	221.6/268	42.6/101	62.6/65	<b>44.4/176</b>	14.6/63	330.8/546	19.25	4.76	85.88
Vision-PE	<b>116.0/184</b>	<b>37.6/111</b>	<b>39.2/44</b>	44.8/174	<b>13.6/64</b>	<b>277.8/585</b>	<b>18.85</b>	<b>3.44</b>	<b>90.68</b>
<i>SciVASR with Proprietary LLMs (GPT-4o)</i>									
Text-PE	203.2/236	41.6/106	57.2/63	42.4/160	<b>15.4/73</b>	267.4/474	17.21	4.08	87.52
Vision-PE	<b>71.2/146</b>	<b>37.0/105</b>	18.2/31	<b>39.8/167</b>	16.0/72	<b>215.0/441</b>	<b>15.84</b>	<b>2.75</b>	<b>92.81</b>
E2E-Vision-PE	78.0/160	39.8/107	<b>16.4/18</b>	43.6/178	15.8/71	236.0/484	17.59	2.81	92.40

Table 5: Aggregated evaluation results of all settings on both dev and set set. The content-wise severity score with the weighted at left and the unweighted at right for each content type. The weighted score sums up the severity weight of mismatches in this category, and unweighted score is simply the number of mismatches of the category.

sistent with the trend presented by the metric of term-recall, a higher correlation of SWER and difficulty score in Table 1 is also obtained for the Whisper-base result (-0.927 in spearman- $\rho$ ), highlighting the reliability of SWER. In contrast, WER cannot accurately capture this relationship.

**Close-sourced setup is superior than the open-sourced one** Though open-sourced PE implementation improves ASR performance, close-sourced GPT-4o still outperforms them, achieving the highest accuracy. However, this is not purely attributed to the powerful capability of GPT-4o, but the pipeline strategy as well, since the GPT-4o end-to-end post-editing setting is inferior to the pipeline method. This suggests that despite the strong visual understanding capabilities of GPT-4o, the complexity of this task poses challenges. Our proposed pipeline approach provides an effective solution.

These findings underscore the importance of using visual context to improve ASR performance and validate the efficacy of SWER as a more reliable evaluation metric compared to WER.

## 7 Conclusion

This paper introduces the MS-ASR task, addressing the challenge of transcribing knowledge-intensive scientific presentations by integrating visual and speech modalities. We identified the inefficacy of WER and proposed the SWER metric to better evaluate ASR performance by considering the content type and error severity. Through our inference framework – SciVASR, we demonstrate that state-of-the-art MLLMs, particularly GPT-4o can significantly enhance transcript accuracy by

leveraging visual information, improving by 45% over speech-only baselines. The experiments underscore the importance of multimodal integration for accurate ASR in complex, real-world scenarios. Our future works focus on expanding benchmark datasets and refining evaluation protocols to further enhance the applicability of the task.

## Limitation

We summarize the limitations of our work in the following three aspects:

**Benchmark Dataset** The proposed task would benefit from the inclusion of more datasets as benchmarks, as it currently relies on a single dataset. Expanding the range of datasets will provide a more comprehensive evaluation of the system’s performance and enhance its robustness across different domains and types of presentation videos.

**Evaluation Protocol** As the evaluation framework leverages the capabilities of LLMs, it should be used alongside conventional metrics to ensure better reliability. Combining new and traditional evaluation methods will provide a more balanced and accurate assessment of the ASR system’s effectiveness, addressing potential biases that might arise from relying solely on LLM-based metrics.

**Latency and Error Propagation** Moreover, the latency and error propagation inherent in our zero-shot framework may pose challenges for practical applications. The time required for processing and the potential for errors to compound through multiple stages can affect real-time usability and overall

accuracy. Future work will focus on optimizing these aspects to improve the framework’s practicality and reliability.

## Ethical Statement

In conducting this research, we have adhered to ethical standards to ensure integrity and transparency. All data used in our experiments were obtained from publicly available sources or through collaborations with researchers who provided consent for the use of their materials. We have taken steps to anonymize any potentially sensitive information to protect the privacy of individuals involved in the datasets. Our methods were designed to respect intellectual property rights, and we have appropriately cited all works that have informed our study. Furthermore, our research aims to contribute to the broader scientific community by providing insights and methodologies that can enhance automated speech recognition (ASR) systems, particularly in knowledge-intensive domains. We remain committed to the responsible use of AI technologies, ensuring that our findings are used to benefit educational and scientific advancements while minimizing any potential negative impacts.

## Acknowledgments

The authors are grateful to the anonymous reviewers for their helpful comments and corrections. The first author was supported by an Australian Government Research Training Program (RTP) Scholarship. This work is supported by the ARC Future Fellowship FT190100039.

## References

Mohamed Afham, Satya Narayan Shukla, Omid Pour-saeed, Pengchuan Zhang, Ashish Shah, and Sernam Lim. 2023. [Revisiting kernel temporal segmentation as an adaptive tokenizer for long-form video understanding](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 1181–1186. IEEE.

AI@Meta. 2024. [Llama 3 model card](#).

Lawrence Angrave, Zhilin Zhang, Genevieve Henricks, and Chirantan Mahipal. 2020. Who benefits? positive learner outcomes from behavioral analytics of online lecture video viewing using classtranscribe. In *Proceedings of the 51st acm technical symposium on computer science education*, pages 1193–1199.

Irene Benedetto, Moreno La Quatra, Luca Cagliero, Lorenzo Canale, and Laura Farinetti. 2024. [Abstractive video lecture summarization: applications and](#)

[future prospects](#). *Education and Information Technologies*, 29(3):2951–2971.

- Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loïc Barrault, and Florian Metze. 2019. [Multimodal grounding for sequence-to-sequence speech recognition](#).
- M. Cettolo, F. Brugnara, and M. Federico. 2004. [Advances in the automatic transcription of lectures](#). In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–769.
- Allen Chang, Xiaoyuan Zhu, Aarav Monga, Seoho Ahn, Tejas Srinivasan, and Jesse Thomason. 2023. [Multimodal Speech Recognition for Language-Guided Embodied Agents](#). In *Proc. INTERSPEECH 2023*, pages 1608–1612.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. [The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation](#).
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2022. [Avatar: Unconstrained audiovisual speech recognition](#).
- Shahram Ghorbani, Yashesh Gaur, Yu Shi, and Jinyu Li. 2020. [Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations](#).
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. [xcomet: Transparent machine translation evaluation through fine-grained error detection](#).
- Anchit Gupta, CV Jawahar, Makarand Tapaswi, et al. 2023. [Unsupervised audio-visual lecture segmentation](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5232–5241.
- Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. [Cogagent: A visual language model for gui agents](#).
- Yash Jain, David Chan, Pranav Dheram, Aparna Khare, Olabanji Shonibare, Venkatesh Ravichandran, and Shalini Ghosh. 2024. [Multi-stage multi-modal pre-training for automatic speech recognition](#).

- Seonmin Koo, Chanjun Park, Jinsung Kim, Jaehyung Seo, Sugyeong Eo, Hyeonseok Moon, and Heuseok Lim. 2024. [Toward practical automatic speech recognition and post-processing: a call for explainable error benchmark guideline.](#)
- Liz McCarron. 2021. Creating accessible videos: Captions and transcripts. *Communications of the Association for Information Systems*, 48(1):19.
- Dan Oneata and Horia Cucu. 2022. [Improving multimodal speech recognition by data augmentation and speech representations.](#)
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. 2024. [Gpt-4 technical report.](#)
- Danila Potapov, Matthijs Douze, Zaïd Harchaoui, and Cordelia Schmid. 2014. [Category-specific video summarization.](#) In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 540–555. Springer.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision.](#)
- 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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona T. 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@ACL 2023, Toronto, Canada (in-person and online), 13-14 July, 2023*, pages 62–78. Association for Computational Linguistics.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. [How2: A large-scale dataset for multimodal language understanding.](#)
- Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2023. [Avformer: Injecting vision into frozen speech models for zero-shot av-asr.](#)
- Khetam Al Sharou and Lucia Specia. 2022. [A taxonomy and study of critical errors in machine translation.](#) In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.
- Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. [Audio-visual llm for video understanding.](#)
- Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. [Fine-grained audio-visual joint representations for multimodal large language models.](#)
- Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. 2024. [Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue.](#)
- Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. [Evaluation and analysis of hallucination in large vision-language models.](#) *arXiv preprint arXiv:2308.15126*.
- Minghan Wang, Jiaxin Guo, Yimeng Chen, Chang Su, Min Zhang, Shimin Tao, and Hao Yang. 2021. [Make the blind translator see the world: A novel transfer learning solution for multimodal machine translation.](#) In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 139–149, Virtual. Association for Machine Translation in the Americas.
- Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. 2023b. [Whislu: End-to-end spoken language understanding with whisper.](#) In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 770–774. ISCA.
- Minghan Wang, Thuy-Trang Vu, Yuxia Wang, Ehsan Shareghi, and Gholamreza Haffari. 2024a. [Conversational simulmt: Efficient simultaneous translation with large language models.](#) *Preprint*, arXiv:2402.10552.
- Minghan Wang, Jinming Zhao, Thuy-Trang Vu, Fatemeh Shiri, Ehsan Shareghi, and Gholamreza Haffari. 2024b. [Simultaneous machine translation with large language models.](#) *Preprint*, arXiv:2309.06706.
- Wei Han Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023c. [Cogvlm: Visual expert for pretrained language models.](#)
- Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. [Instructscore: Explainable text generation evaluation with finegrained feedback.](#)
- Hang Zhang, Xin Li, and Lidong Bing. 2023. [Video-LLaMA: An instruction-tuned audio-visual language model for video understanding.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

## Appendix

### A Presentation Difficulty Estimation

The difficulty of transcription mainly lies in the accent, fluency, and pronunciation of a speech. To eliminate the difficulty stemming from the content, we compare the speech with a standard TTS generation as follows:

- Segment the first 30-second speech chunk from each video and obtain the ground-truth transcript for the chunk.
- Use *OpenAI TTS API* to generate a baseline speech with a standard accent, where the gender of the voice is set as the original speakers’.
- Compute the conditional perplexity (CPPL) of the baseline speech (TTS generated) and the real speech by the Whisper-base model, obtaining  $CPPL_{\text{bsl}}$  and  $CPPL_{\text{spch}}$ .
- Calculate the difficulty score (DS) as  $CPPL_{\text{bsl}}/CPPL_{\text{spch}}$  reported in Table 1. The smaller, the harder.

### B Details about Implementation

Table 6 lists the questions used for visual information extraction. In Figure 4, we show the example of how vision-LLM could hallucinate.

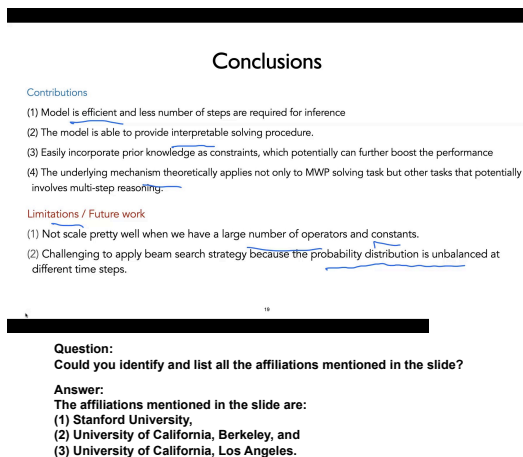


Figure 4: An example showing that vision-LLM could answer with hallucination when the question has false-premise. Image is sampled from video 410, CogAgent-VQA is used for inference.

### C Details about LLM-based Evaluation

Figure 6 present the annotation guideline for our proposed evaluation protocol, and Figure 7 present

the annotated result with the LLM. The performance of LLM-based evaluation is presented in Table 4.

### D Additional Experimental Results

We statistic the severity in three configurations in Table 7.

### E Related Work

#### E.1 Incorporating Visual Modality in ASR

Oneata and Cucu (2022); Ghorbani et al. (2020); Gabeur et al. (2022); Caglayan et al. (2019); Wang et al. (2021, 2023b) propose various approaches to effectively incorporate visual modalities into ASR tasks. These methods typically involve encoding image and speech inputs through visual and audio encoders into dense representations, which are then fused using a modality fusion component such as a transformer encoder or lightweight adapters. This approach ensures that representations of both modalities are mapped into a shared space during training, allowing the decoder to leverage them to predict transcripts in an end-to-end manner during inference.

Despite their success, end-to-end approaches often face significant challenges. One major issue is the scarcity of high-quality triplet data (image, speech, text), which makes it difficult to train models from scratch. To address this, various studies (Jain et al., 2024; Seo et al., 2023) explore leveraging transfer learning on modalities to better utilize unimodal data or introducing an additional stage for self-supervised pretraining to learn the shared representation before fine-tuning the model on the ASR task.

Moreover, existing works are often trained and evaluated on content-less scenarios such as the How2 dataset (Sanabria et al., 2018). In these scenarios, vision models only need to provide coarse features, such as object classes, to sufficiently reinforce speech features. However, real-world ASR systems are often used in content-rich scenarios, which are more challenging and require vision models to process more complex content such as text, tables, and figures in slides.

#### E.2 Multimodal ASR in the LLM era

Recent advancements in multimodal large language models (LLMs), such as Video-LLaMA (Zhang et al., 2023), Audio-Visual LLM (Shu et al., 2023), and AVicuna (Tang et al., 2024), have demonstrated

---

### Questions for vision-LLMs to perform visual-context extraction

---

- "Please describe the layout of the slide."
  - "What is the main topic or headline of the slide?",
  - "Can you list the key points or bullet points presented on the slide?",
  - "Are there any important dates, statistics, or quantitative data mentioned? Please summarize.",
  - "Could you identify and list all the affiliations mentioned in the slide?",
  - "Are there any visual aids (e.g., charts, graphs, images, diagrams or tables) on the slide? What information do they convey?",
  - "Does the slide include any quotes, citations, or references to other works? Please detail.",
  - "Please provide a holistic summary that encapsulates the entire content and context of the slide.",
- 

Table 6: The predefined questions for vision-LLMs.

**Task Introduction**

You are tasked with post-editing the ASR transcript of a presentation of an academic paper. Focus on rectifying inaccuracies introduced by ASR (Automatic Speech Recognition) errors within the current page slide's transcript. Utilize the comprehensive presentation summary and the individual slide summaries as references for your corrections. If the transcript is error-free, simply submit it as is. Your edits should strictly address corrections without adding extraneous content such as explanations or notes. Aim for minimal alterations; restrict your changes solely to essential corrections (e.g. terminologies, names, entities) without rewording the original sentences in the transcript. Just return the fixed ASR transcript without any other explanations. DO NOT return any other irrelevant content!

**Overall Presentation Summary**

{overall\_summary}

**Page-level Slides Summary**

{slide\_summary}

**Transcripts**

{transcript}

Figure 5: This figure shows the prompt for post-editing with text-LLM.

Severity →	CRITICAL	MINOR	OK	Total
Whisper-base	529 / 43.04%	176 / 14.32%	524 / 42.64%	<u>1229</u>
Text-PE (GPT-4o)	328 / 26.69%	185 / 15.05%	599 / 48.74%	1112
Vision-PE (GPT-4o)	178 / 14.48%	181 / 14.73%	603 / 49.06%	962

Table 7: This table presents the absolute error number for each severity level as well as the percentage of it normalized by the baseline error number (underlined).

the ability to leverage both audio and video frames as inputs. However, these models primarily focus on video understanding tasks, paying less attention to speech input. Unlike natural sounds, speech signals often carry more detailed information in a video, making them more challenging for multimodal LLMs to process effectively. While Sun et al. (2023) have explored this area, their work also centers on content-less scenarios such as the How2 dataset (Sanabria et al., 2018), leaving the potential of these models in content-rich environments largely unexplored.

### E.3 LLM-based Quality Evaluation

The reliance on WER for evaluating ASR systems has been increasingly criticized for its inadequacy in reflecting true system performance, mirroring issues in machine translation (MT) where metrics like BLEU fail to capture the full quality spectrum. Sharou and Specia (2022) address this by categorizing critical translation errors that could lead to severe user consequences, suggesting a nuanced approach to error analysis. Similarly, Guerreiro et al. (2023) propose a transparent evaluation framework for machine translation, emphasizing fine-grained error detection could be potentially adapted for ASR. Freitag et al. (2021) also highlight the need for multi-dimensional quality metrics that can provide a more holistic view of translation quality, underscoring the need for ASR evaluations to consider context and practical usability.

The difference between ASR and NMT error taxonomy relies on the interpretation of error, for NMT, errors are often referred to as mistranslation,

## Role Introduction

You are a professional ASR transcript error annotator. Your task is to identify and categorize mismatches between the ground-truth text and the ASR-generated transcript.

## Input Data Format

The data you will review is presented in two sections:

- **### Ground-Truth:** This is the original, correct version of the text.
- **### ASR-Transcript:** This is the ASR-generated text that may contain mismatches.

## Mismatch Symbols

In both transcripts and ground-truth, mismatches are highlighted using specific symbols:

- "**<>**": **Insertion** - Text that the ASR system added erroneously.
- "**[]**": **Substitution** - Text that the ASR system altered incorrectly.
- "**∅**": **Omission** - Text that the ASR system failed to include.

## Mismatch Categories

You will categorize each mismatch into one of the following types, based on its content and the context:

- **Terminology:** Specialized terms used in specific domains. These terms have specific meanings and are often used in a particular context. Example: "Neural networks" in a tech discussion, "Quantum entanglement" in physics, "BERT" or "GPT" in an introduction of NLP models.
- **Numerical Data:** Numbers, quantities, or statistical data. Example: "3.14", "thirty percent" or "3 out of 4" when discussing experimental results.
- **Named Entities:** Names of people, places and organizations. These are unique identifiers that do not change across different contexts. Example: "Google" in a business case study, "Albert Einstein" as a person's name.
- **Grammatical Words:** Words that primarily serve a grammatical function in a sentence, including articles, prepositions, conjunctions, and auxiliary verbs. Example: "and" in "cats and dogs", "the" in "the cat".
- **Disfluencies and Fillers:** Words or phrases that reflect speech disfluency, including fillers, repetitions, or corrections. Example: "uh", "um", "actually", or repeated words due to stuttering.
- **General Words:** Common words that are not specialized terms, named entities, grammatical words, disfluencies, or fillers. Example: "increase" in "the company saw a significant increase in profits", "very" in "a very big dog".

Note: The distinction between these categories should always consider the context of the content, not just the meaning of the text within the brackets. Additionally, classification should be based on the corresponding words in the Ground-Truth text, not the erroneous words in the ASR-Transcript.

## Severity Annotation

For each mismatch, assess the severity of the impact on the meaning of the text, categorized into one of three levels:

- **OK:** No significant impact on the meaning. These are typically minor grammatical errors or slight pronunciation issues that do not alter the understanding of the sentence. For example, minor grammatical errors, tense, voice, singular/plural errors, conversions between numerical and text formats (e.g., "0" vs "zero"), missing colloquial words, missing or inserted punctuation, case conversion, and minor rephrasing that does not change the original meaning. Example: "present" instead of "presents" due to unclear pronunciation (Ground-Truth: "The speaker [presents] the results."; ASR-Transcript: "The speaker [present] the results.")
- **MINOR:** Slight impact on the meaning, but the main point is still understandable. These errors may cause slight confusion or require minor inferencing to understand the original intent. For example, non-critical homophone errors and spelling mistakes of general words. Example: "their" instead of "there" (Ground-Truth: "[There] were several experiments conducted."; ASR-Transcript: "[Their] were several experiments conducted.")
- **CRITICAL:** Significant or complete change or distortion of the original meaning. These errors alter the key message or cause major confusion, potentially leading to misunderstanding of the content. For example, critical homophone errors and serious misspellings or misrecognition of key terms and names, misrecognition of numerical values (e.g. 90 vs 19), and large omissions or rephrasings that deviate from the original meaning. Example: "social lightwork" instead of "neural network" (Ground-Truth: "The [neural network] showed high accuracy."; ASR-Transcript: "The [social lightwork] showed high accuracy.")

Note: The determination of severity requires careful consideration of the context.

## Return Format

Return the annotation results as a JSON object with a key named "mismatches", the value of it should be a JSON list. Each object in the list should correspond to a mismatch in the text and include the following information:

```
```json
{
  "mismatch_type": "insertion | substitution | omission",
  "mismatch": "For insertions: inserted string | For substitutions: incorrect text | For omissions: empty string",
  "correct_form": "For insertions: empty string | For substitutions: correct text | For omissions: omitted text",
  "mismatch_content_type": "One of the defined category names",
  "severity": "OK | MINOR | CRITICAL",
}
```
```

Figure 6: This figure shows the guideline used for automatic evaluation.

```
REF: We {um} finetune [BERT] on <> downstream [tasks].
HYP: We {} finetune [birds] on <the> downstream [task].
LLM_response = [
  {"mtype": "O", "mismatch": "", "correct": "um", "ctype": "DISF", "severity": "OK"},
  {"mtype": "S", "mismatch": "birds", "correct": "BERT", "ctype": "TERM", "severity": "CR"},
  {"mtype": "I", "mismatch": "the", "correct": "", "ctype": "GRAM", "severity": "OK"},
  {"mtype": "S", "mismatch": "task", "correct": "tasks", "ctype": "GEN", "severity": "MIN"}
]
```

Figure 7: An example of how to leverage LLM to annotate ASR errors based on the taxonomy in Table 3. We shorten the mismatch type of omission, substitution and insertion with O, S and I.

hallucination or deletions deviate from the source sentence (Sharou and Specia, 2022), while ASR errors are referred to mismatches to the ground-truth transcript due to misrecognition. The pursuit of a comprehensive taxonomy of ASR errors is exemplified in Koo et al. (2024), they propose an Error Explainable Benchmark (EEB) that categorizes errors at both the speech and text levels, offering a granular understanding of model shortcomings. While this taxonomy is a significant step towards a more detailed diagnosis of ASR systems, its extensive nature may pose challenges in practical implementation due to the complexity of distinguishing between numerous error types.

LLMs have demonstrated their utility in evaluating the quality of generative text, moving beyond surface-level metrics to provide deeper insights into the quality of machine-generated content. Xu et al. (2023) introduces a metric that generates not only a score but also a diagnostic report, identifying specific errors and their severity. Similarly, Fernandes et al. (2023) presents AUTOMQM, a prompting technique that leverages LLMs to identify and categorize translation errors, aligning with the Multidimensional Quality Metrics (MQM) framework. These approaches underscore the potential of LLMs to offer interpretable and detailed assessments of translation quality, which can be instrumental in refining ASR systems and enhancing user satisfaction.