Augmenting Lecture Information via Redundancy-Filtered LLM Slide Descriptions for Blind and Low-Vision Students

Katharina Anderer[1,2] **and Jan Niehues**[1] **and Karin Müller**[1] **and Matthias Wölfel**[2]
[1]Karlsruhe Institute of Technology
[2]University of Applied Sciences Karlsruhe
{katharina.anderer, jan.niehues, karin.mueller}@kit.edu
matthias.woelfel@h-ka

Abstract

Slides in video lectures often contain information that is inaccessible to blind and low-vision students unless explicitly described. Although previous studies have addressed this accessibility gap, existing solutions often fail to align visual content with spoken narration, leading to redundant or omitted details. This paper presents an automated method for generating slide descriptions to enhance lecture content. A large language model refines them to eliminate redundancy with respect to spoken narration. Experimental results demonstrate that the system can effectively filter the information already mentioned by the lecturer, reaching word count compressions of up to 65%, while preserving important information. Our findings suggest that careful prompt design can directly balance efficiency with content preservation.

1 Introduction

Visual slides serve as essential support in academic lectures, scientific talks, and professional presentations to engage the audience. However, for blind and low-vision (BLV) students, these slides often pose a barrier rather than a support. Since visual information is often explained inadequately or not at all (Lundgard and Satyanarayan, 2021; Anderer et al., 2024), BLV students have to spend a lot of time deciphering information on a slide or rely exclusively on auditory information (Wilkens, 2025).

Online learning, hybrid learning formats and lecture recordings that contain visual slides are becoming increasingly popular (Pabst et al., 2023). Accessibility needs to be ensured for these formats. In particular, lecture recordings have great potential to support BLV students not only for revision, but by making visual content more accessible with AI-based methods such as vision-language models (VLMs).

In recent years, VLMs have made great progress in terms of quality and reliability of image descriptions, although hallucinations remain a major challenge (Li et al., 2025b). Despite these advancements, there is a research gap on how to distinguish between visual information that is already present by audio and information that needs to be made accessible to BLV individuals. Preventing information overload is especially relevant in academic settings, where BLV students often report that they require more time than their peers to rehearse lecture material (Butler et al., 2017; Armstrong and Murray, 2007; Frank et al., 2014).

In this work, we present a technical framework for automatically generating efficient slide descriptions for a current video frame of a lecture recording and for filtering redundancies with respect to the spoken narration. To evaluate whether the generated slide descriptions convey essential content, we apply Visual Question Answering (VQA) as a proxy for information coverage. This proxy follows the logic that if a compressed text description allows an LLM (or a student) to answer specific questions about a slide's content, relevant information is covered by the description. With our research questions, we address the following:

RQ1 To what extent do slide descriptions improve visual question-answering accuracy across distinct question categories (text, graphs, tables, formulas) compared to transcripts alone?

RQ2 Can LLM-driven redundancy filtering significantly compress description word count while preserving relevant information?

RQ3 How do variations in prompt design and LLM selection influence the balance between information compression and performance?

To address these questions, we utilize the MCIF benchmark (Papi et al., 2025), a multi-modal benchmark dataset containing question-answering pairs of different scientific talks based on slides. The

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 401–411
October 17-21, 2026 ©2026 Association for Computational Linguistics

MCIF dataset provides a solid foundation for simple visual questions such as text extraction, while it does not cover a large variety of different and complex visual questions. Because understanding complex visual elements is often the steepest barrier for BLV students, evaluating our framework required different visual question types. Therefore, we designed a custom dataset based on the same lecture material as the MCIF, but covering more complex visual questions and including formula and table extraction, as well as graph-related questions, which are crucial in academic settings [1].

To filter out information already present in the transcript, we use an LLM-based filtering approach and compare different prompt strategies and LLM models (Gemma 3-12B vs. Mistral-Small-4). By measuring how well the LLMs answer the custom questions when supplied with either uncompressed or compressed slide contexts, we explicitly test whether information is lost during the efficiency optimization step.

We build our pipeline upon the open source project "Lecture Translator" (LT) (Huber et al., 2023), utilizing its core infrastructure that already aligns the lecture transcript with a lecture video. Our framework builds on the pipeline by adding a slide description module to automatically generate efficient slide descriptions.

Our contributions are two-fold: (i) Slide description and filtering: An LLM-driven approach to generate descriptions and to remove transcript-redundant information from slide descriptions based on prompt constraints (RQ3). (ii) Evaluation dataset and pipeline: A custom dataset and evaluation pipeline to measure downstream visual Q&A accuracy across distinct question categories compared to spoken narration alone (RQ1) and to test information preservation and efficiency (RQ2).

2 Related Work

This section will first introduce work aimed at improving slide accessibility, followed by work that optimizes the efficiency of assistive systems. Finally, we describe general redundancy filtering approaches in natural language processing (NLP).

2.1 Assistive Systems for Slides

Recent research has focused on identifying and closing the information gap between visual slide content and auditory descriptions during lectures (Peng et al., 2021b,a; Park and Lim, 2022; Anderer et al., 2024; Zhang et al., 2023). Already in 2021, Peng et al. (2021b) introduced a word-level analysis to highlight slide elements left unaddressed by the lecturer. They extracted slide elements and aligned them with the presenter's speech, providing audio notifications for slide changes and unaddressed elements. For image descriptions, they relied on the Microsoft Cognitive Services Vision API, which generates captions. Although this offers a rough overview of slide content, it cannot capture all relevant elements of the slide. Park and Lim (2022) presented a system that integrates slide content directly into audio by converting slide information to text and then to audio for commentary during lectures. However, this approach limits user control over information density and lacks interactive Q&A capabilities. Zhang et al. (2023) developed a system to read and author slides, enabling interaction through gestures or voice input for Google Slides. Although this improves accessibility for Google Slides, it fails to generalize to other formats, which are common in higher education. To make lecture recordings more accessible to BLV students, Anderer et al. (2025) developed an application that allows users to upload a video and query a chatbot about its visual content. In addition to this Q&A component, their application automatically generates chapters and enables quick navigation to a video position by chapters or querying the chatbot about a topic. BLV students in their study explicitly wished for an overview of the current slide, so that repetitive questions can be reduced. Therefore, with our work, we tackle a concrete user need that was identified in their paper.

2.2 Assistive Systems that Optimize Efficiency

Cognitive load theory (Sweller, 1994) posits that redundancy increases the extraneous cognitive load, which is the mental effort caused by the complexity of the material. Reducing redundancies should therefore be a key design goal. Applying this principle to navigation, Li et al. (2025a) proposed a reasoning framework to avoid redundant utterances by making outputs more concise and information-dense for a walking assistant. They also introduced an "awareness discriminator" to assess the "scene novelty" and potential risks to avoid repetitive notifications.

Similarly, Zhao et al. (2025) designed an agent

[1] The dataset and result files are available on GitHub

with core principles and goals to help BLV with daily tasks and navigation. Their study with 15 BLV participants revealed that "inefficient verbosity" was a primary frustration when interacting with current AI assistants. To make output short and non-redundant, they enforce the agent by critical self-assessment and by a brevity constraint, limiting the output of the LLM to a maximum of 128 tokens. However, hard token constraints might fail to preserve relevant information and are not flexible enough to explain a complex graph adequately.

In the context of digital content, Peng et al. (2023) identified overlapping text and visuals in subsequent slides to find hierarchical groups that provide structural navigation for BLV. This helped in efficient navigation through hierarchical anchors, which was evaluated by, for instance, a reduced time to navigate to a slide.

Very recently, Haque and Hong (2026) presented an approach to filter redundancy in speech within video lecture recordings, such as fillers, self-corrections, or long pauses and removed them from the final audio stream. They also added slide descriptions to the edited audio stream, but limited it to a short description consisting of the slide number, title, and a brief description of visual content.

While these works address efficiency with regard to navigation and presentation structures, research focusing on efficient slide descriptions for BLV in a lecture context is lacking.

2.3 Redundancy Filtering and Evaluation

Minimizing redundant information while preserving semantically relevant context has long been a fundamental challenge in NLP. Recent work has increasingly utilized LLMs to filter redundancies, such as for filtering repetitive reports (Lin et al., 2026) or eliminating data redundancy via structured prompt engineering (Guang et al., 2024). Specifically, Guang et al. (2024) used structured prompt constraints such as explicitly instructing the LLM what to discard and controlling information density by consolidating overlapping statements into one. They demonstrated that downstream task performance remained stable when inputting the compressed instead of the original text.

Evaluating summarization quality and measuring information redundancy has traditionally relied on lexical overlap metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004). While computationally efficient, these metrics do not capture semantic similarity and might fail when alternative phrasing is used. To address this limitation, dense sentence embeddings are often leveraged to measure semantic similarity in high-dimensional vector spaces such as BERTscore (Zhang et al., 2020) or E5 embeddings (Wang et al., 2024).

In this work, we adapt these metrics to a novel task: evaluating automated redundancy-filtering for slide descriptions conditioned on spoken narration.

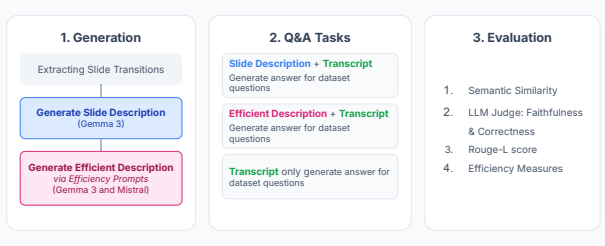


Figure 1: Generation and Evaluation Framework

3 Generation of Slide Descriptions

In this section, we will describe our method for generating slide descriptions. We will detail how slide transitions are detected and slide descriptions are automatically generated by a VLM.

3.1 Slide Description Module

The Slide Description Module generates slide descriptions for each slide presented in a lecture video. Therefore, slide transitions within a lecture video need to be detected beforehand. The next subsections detail the slide transition detection, as well as the automatic slide description generation and the generation of efficient descriptions. An overview of the generation framework is also detailed in Fig. 1 on the left side.

3.1.1 Slide transition detection

To detect the single lecture slides within the video, we applied an algorithm for slide transition detection. In a first step, slide transitions are detected with the help of the image transformer SwiftFormer (Shaker et al., 2023) that generates high-dimensional visual embeddings for video frames sampled at 1 Hz. This rate captures rapid transitions while a maximum detection delay of one second is negligible for user feedback and a high amount of redundant computations are avoided. We calculate the cosine similarity between embeddings of consecutive frames to identify potential transitions. The final output is a list of timestamps that mark the slide transition.

3.1.2 Automated slide description generation.

After the slide transitions are calculated, in a second step, a VLM is used to generate an alternative text description for every slide. For our experiments, Gemma 3-12B, an open source model from Google (Gemma Team, 2025) with 12 billion parameters, is used.

We iteratively adapted slide description generation prompts to evaluate how structure and detail levels affect information coverage. The first prompt for generating descriptions asks about a description structured into "main content", "text content" and "visual information" (P1), while the second prompt asks specifically about formulas, exact table data, and detailed descriptions of graphs including axes, labels and legends (P2). The exact prompts can be found in the appendix A. The descriptions coming from this step will be labeled as "Full-P1" and "Full-P2" respectively, as they are produced without a filtering-redundancy step, yet. The next section will describe how redundancy-filtered descriptions were derived based on the full descriptions.

3.1.3 Filtering Redundancy

In a third step, an LLM is instructed to rewrite generated slide descriptions, removing information already present in the transcript. For Gemma, a redundancy-reduction prompt (v1) had already been integrated into the pipeline. Descriptions generated with P1 and P2 were thus reduced by using the following prompt. This filtering step operates purely on text: the model receives only the previously generated slide description and the transcript as input (see Prompt v1 below).

> **Prompt v1.** Transcript of what the speaker said during this slide:
>
> `{transcript_context}`
>
> Visual description of the slide:
>
> `{analysis_result_without_transcript}`
>
> Rewrite the visual description to remove information the speaker already explained in detail. Keep information (diagrams, table numbers, formulas, etc.) as long as the speaker did not talk about it precisely.

The Gemma-generated and redundancy filtered descriptions will be referred to as "Eff-P1" or "Eff-P2", so that one can distinguish whether the filtering was based on prompt P1 or P2. For Mistral (using version mistral-small-2603 via API), redundancy filtering was always based on descriptions generated by P2. In addition to redundancy-filtering by prompt v1, we tested two other versions that balance information coverage and efficiency differently to test how this can impact the descriptions: **v2** explicitly notes to omit information fully conveyed already. In terms of information retention, it is similar to v1, but it explicitly adds to prioritize information retention instead of sacrificing when in doubt. **v3** utilizes a strict "REMOVE/KEEP framing" to explicitly preserve diagrams, axes, table data, and formulas unless explicitly described. In this way, it concentrates more on information coverage, in contrast to v1 and v2. The exact prompts can be found in the appendix A.

We will refer to the descriptions generated from these descriptions as "Eff" later on, and add a prefix for the model used and a suffix for the prompt version used, e.g. "Mistral-Eff-v2".

4 Evaluation of the Slide Descriptions

The core contribution of this work lies in balancing information efficiency with information preservation. For systematic validation, we designed an evaluation pipeline to test whether compressing slide descriptions degrades downstream Q&A comprehension. We deploy an LLM-based Q&A paradigm that treats the generated descriptions as LLM context to answer visual questions.

First, we introduce the baseline MCIF benchmark (Papi et al., 2025) and detail the curation of our custom dataset. Second, we formalize the evaluation tasks, where unfiltered and filtered slide descriptions, as well as only the transcribed speech of the lecturer, serve as contexts for the utilized LLM (mistral-small-2603 via API) to answer questions of the custom dataset. Finally, we describe the metrics used to evaluate LLM answers. Together, this pipeline provides a technical lens for measuring whether information is preserved in an automated way. The tasks and the evaluation metrics are also shown in Fig.1 as an overview.

4.1 Datasets

4.1.1 MCIF benchmark

The MCIF (Multimodal Crosslingual Instruction Following) benchmark was introduced by Papi et al. (2025). It is a human-annotated benchmark consisting of 220 question-answering pairs, from which 49 are only answerable from the visuals shown on the video, 129 are answerable by both audio and video,

17 by audio-only, and 25 questions are not answerable. In its original intention, the benchmark compares different multi-modal large language models to analyze how different modalities impact their performances (Papi et al., 2025).

In the MCIF dataset, we analyzed the "video-only" questions — those answerable only through visual capabilities. We found that about a third of these questions are about the author affiliation questions, and another third of the questions ask about the number of authors or the speaker's name. Most of these questions can be answered by simple text extraction and do not require complex visual extraction such as table or formula extraction, or graph related questions. Therefore, we decided to build our own custom dataset to include more complex visual questions that are more relevant to BLV students, which is described next.

4.1.2 Creating own customized dataset

To ensure our evaluation framework directly reflects the actual information-seeking behaviors and lived experiences of BLV students, we grounded our dataset design in authentic user interactions. Therefore, we analyzed inquiries submitted during a three-month testing phase with 8 BLV students. During this time, they were able to use our platform LectureAssistant (Anderer et al., 2025), where they could upload their own videos and engage with a chatbot to ask visual and non-visual questions about the lecture (Anderer et al., 2026). We systematically categorized these inquiries based on the underlying visual or textual content types required to answer them, such as graphs, tables, or formulas. We then calculated the percentages for these categories by applying a regex-based algorithm—for instance, categorizing queries that contained terms like "describe" or "explain" alongside "image," "graph," or "figure" as graph-related queries. The highest percentage of questions concerned text extraction from slides (36%), followed by graph-related inquiries (25%). Other visual questions included formula extraction (2%) and table/numeric extraction (4%). Remaining queries were non-visual, relating to summaries, video navigation, or lecture concepts. As these questions were primarily driven by one highly active user, this distribution should not be seen as a general representation but rather as a qualitative proxy. Ultimately, student inquiries are dependent on lecture style, topic, or prior knowledge.

Taking the above mentioned categories into account, we created a custom dataset including in total 53 questions. These involve questions about slide text extraction (32%), diagram or graph-related inquiries (45%), formula extraction (6%), and table/numeric extraction (17%). To measure the direct incremental benefit of slide descriptions over a transcript-only baseline, we did not exclude questions answerable by the transcript alone.

4.2 Evaluation Tasks

We defined three tasks to evaluate how different slide description strategies impact LLM (mistral-small-2603) Q&A performance. By comparing the LLM's answer using a **Transcript-only** baseline against its performance when supplemented with **Full Descriptions** (non-filtered) versus **Efficient Descriptions** (filtered), we can quantify information gain and verify if crucial data is preserved during compression. This framework allows us to iteratively optimize LLM prompts for maximum information retention at minimum word count.

4.3 Evaluation Metrics

To evaluate how effectively slide descriptions aided question-answering, we used four performance metrics: BERTScore for semantic similarity compared to the ground truth, ROUGE-L for lexical overlap, LLM-correctness for answer accuracy, and LLM-faithfulness to ensure answers were rooted in the context. Additionally, we measured efficiency using two metrics: the average word count to assess word count reduction, and a transcript-conditioned uniqueness score to measure information novelty.

4.3.1 Semantic and lexical similarity

We compare model answers to the MCIF human-annotated reference answers using BERTScore F1 (Zhang et al., 2020) for semantic similarity and ROUGE-L (Lin, 2004) for lexical overlap (rouge-score v0.1.2), which measures overlap via longest common subsequences.

4.3.2 LLM as a judge

An LLM additionally rates correctness (correct=1, partially correct=0.5, false=0) and faithfulness (whether answer claims are grounded in the given context) of each model answer. Faithfulness ratings require forwarding the full context (slide description and transcript) to the judge; correctness ratings do not. Mistral-small-4 served as the primary judge; we validated reliability against ministral-14b-2512 (same model family) as well as gpt-5.4

(different family).

4.3.3 Efficiency measures

For the different slide descriptions, we analyzed both the reduction in redundant information and the decrease in overall description word count. While the latter simply measures the average word count of the slide descriptions, the former evaluates the degree of novel information each sentence in a slide description introduces relative to the lecture transcript. To quantify this, every sentence from both the slide descriptions and the transcript is encoded into a high-dimensional vector using BAAI/bge-large-en-v1.5 [2], which was chosen for its state-of-the-art performance on the MTEB leaderboard. For each description sentence, the transcript sentence with the highest cosine similarity s is then searched. Taking the residual norm $\sqrt{1-s^2}$ gives the sine, which translates to the uniqueness of the sentence. Therefore, a high score of the residual norm indicates that novel information is introduced in the slide descriptions compared to the transcript.

5 Results

This section details the comparative performance of the Full, Efficient, and Transcript-only tasks across different slide description strategies. Specifically, we compare standard baseline descriptions generated by Gemma 3 (**Gemma-Full-P1** and **Gemma-Full-P2**) against efficient descriptions generated by Gemma 3 (**Gemma-Eff-P1** and **Gemma-Eff-P2**) and efficient descriptions generated by Mistral-Small (**Mistral-Eff-v1**: original rewriting, **Mistral-Eff-v2**: detailed visual, **Mistral-Eff-v3**: explicit REMOVE/KEEP framing), alongside a **Transcript-only** baseline. Note that for Mistral, only efficient descriptions were generated based on Gemma-generated full descriptions.

The analysis proceeds along our three research questions. Before, interrater reliability scores are reported.

5.1 Interrater reliability

The interrater reliability between the different LLM judges was substantial for the two models from the Mistral family, mistral-small and ministral-14b (linear-weighted Cohen's $\kappa = 0.778$, overall Spearman $\rho = 0.848$, $n = 419$), as well as between mistral-small and gpt-5.4 ($\kappa = 0.772$, $\rho = 0.850$, $n = 419$). The choice of the model family did

[2]https://huggingface.co/BAAI/bge-large-en-v1.5

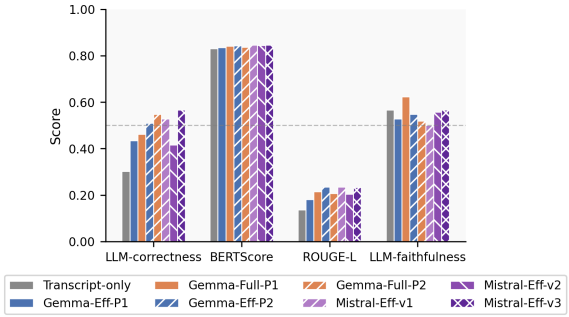


Figure 2: Evaluation with different metrics

therefore not materially change the reliability estimate.

While linear-weighted Cohen's κ assesses the level of agreement while comparing the absolute scoring and penalizing large discrepancies across the ordinal scale, Spearman's ρ measures the strength and direction of the rank correlation between the two judges' ratings, thereby looking at the relative consistency.

For clarity, all LLM-based scores reported in the remainder of this paper were obtained using mistral-small-latest (version mistral-small-2506) as the judge.

To assess validity against human judgment, the first author rated a blinded stratified subsample of 80 items (10 per condition), stratified by condition and question type. The human rater saw neither the condition nor the automatic scores. Agreement with the LLM judge scores was again substantial: for human vs. gpt-5.4 ($\kappa = 0.764$, $\rho = 0.8344$) and for human vs. mistral ($\kappa = 0.698$, $\rho = 0.798$).

5.2 RQ1: Do the Slide Descriptions Help

Comparing the **Transcript-Only** condition with all description-augmented conditions yields a clear trend: LLM-correctness scores improve by 0.13 to 0.26 points across conditions. This improvement is statistically significant under a Wilcoxon signed-rank test with False Discovery Rate (FDR) correction ($\alpha = 0.05$). Fig. 2 illustrates the average metric scores across the eight conditions.

In contrast, the trends for BERTScore F1 are less pronounced. The average BERTScore for the transcript-only condition (0.8297) was lower than all others, and statistically significant for all except for the comparison to **Gemma-Eff-P1** and **Gemma-Full-P2**. The mean differences ranged from -0.0058 to -0.0164, which seem to be rather small, although statistically significant for all re-

maining conditions. For ROUGE-L, differences between the transcript-only and alternative conditions were all significant, except for **Gemma-Eff-P1** and **Gemma-Full-P1**. This suggests that both the transcript and prompt 1 descriptions had low lexical overlap with the reference answers. However, BERTScore may not be sensitive enough to discriminate between small semantic differences.

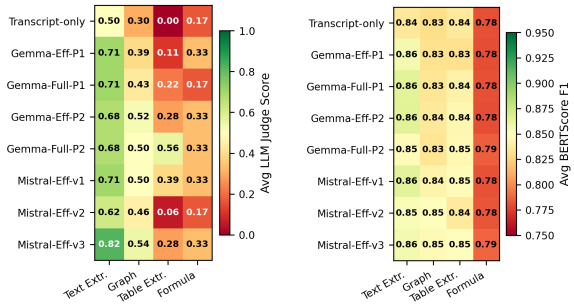
No statistically significant differences were observed for LLM-faithfulness scores, either between the transcript-only and experimental conditions or across the description-augmented conditions themselves.

Since **Gemma-Full-P1** and **Gemma-Full-P2** are generated from the same slides prior to any redundancy filtering, we directly compared the two to test whether the more exhaustive P2 prompt yields better downstream performance than the more concise P1 prompt. LLM-correctness was on average higher for **Gemma-Full-P2** than for **Gemma-Full-P1** (0.5472 vs. 0.4623), but this difference did not survive FDR correction. Faithfulness showed the opposite pattern, with **Gemma-Full-P1** scoring higher on average than **Gemma-Full-P2** (0.6226 vs. 0.5189); this difference was likewise not significant after correction (raw $p = 0.049$, adjusted $p = 0.634$). BERTScore and ROUGE-L differences between the two prompts were smaller still and non-significant. Descriptively, the by-question-type breakdown in Fig. 3 shows a numerically larger gap favoring Gemma-Full-P2 for graph, table, and formula questions than for text extraction, though we did not test this pattern statistically given the small number of questions per type.

Fig. 3 plots the metric score comparison for the LLM-correctness and BERTScore by question-answer types. The scores are on average highest for text extraction, followed by graph-related questions. Formula and table extraction performed worst, which is expected as these questions were the most complex and detailed.

5.3 RQ2: Is Information Preserved?

Direct comparison between efficient and standard baseline descriptions (**Gemma-Eff-P1** vs. **Gemma-Full-P1** and **Gemma-Eff-P2** vs. **Gemma-Full-P2**) reveals no statistically significant degradation in LLM-correctness via Wilcoxon testing. Mean LLM-correctness scores are marginally higher for the non-efficient variants. Furthermore, **Mistral-Eff-v1** and **Mistral-Eff-v3** matched the performance of the **Gemma-Full-P2** condition, although **Mistral-Eff-v2** did not.



(a) LLM correctness score by QA type

(b) BERTScore F1 by QA type

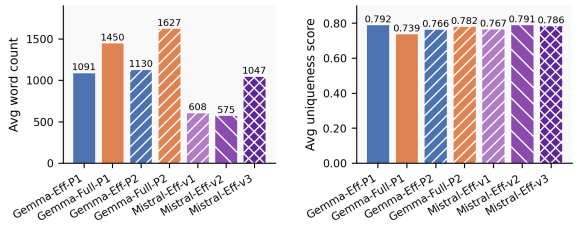
Figure 3: Scores by QA type.

For BERTScore and ROUGE-L, significant differences emerge when comparing **Mistral-Eff-v1** and **Mistral-Eff-v3** against the non-efficient baseline **Gemma-Full-P2**, but with very small differences between -0.0045–0.0048. This might indicate differences of description style between the two LLM models, rather than real semantic differences.

The two efficiency metrics, plotted in Fig. 4 show that the average word count of the descriptions was drastically reduced under the efficient conditions. The largest reduction was induced by **Mistral-Eff-v2**, having an average word count of 575 compared to the **Gemma-Full-P2** condition with an average word count of 1677 (compression ratio = 0.35). **Gemma-Eff-P1** and **Gemma-Eff-P2** both had a reduction of about 400 words on average per lecture.

Conversely, **Mistral-Eff-v3** yielded a word count of 1047 (compression ratio = 0.64). This is explainable because the v3 prompt explicitly prioritized information preservation by instructing the model to retain data when in doubt.

Regarding transcript-conditioned uniqueness scores, differences between conditions are subtle but informative (Fig. 4b). **Gemma-Eff-P1** achieved a score of 0.792, indicating a higher density of novel information relative to the transcript than **Gemma-Full-P1** (0.739). However, **Gemma-Full-P2** scored higher than **Gemma-Eff-P2**, implying that the compression step in **Gemma-Eff-P2** may have inadvertently discarded some non-redundant details. Comparing the three MS conditions, **Mistral-Eff-v2** has the highest uniqueness score among them, confirming effective redundancy elimination, albeit at the cost of downstream performance on the Q&A. Because the uniqueness scores were calculated per-sentence and were un-



(a) Average Word Count (b) Uniqueness Score

Figure 4: Efficiency metrics.

paired across conditions, we performed a Mann-Whitney U test using FDR p-value adjustments. The test indicated that, while these score variances are unlikely to be products of random chance, the absolute differences remain small.

In summary, the efficient conditions did not lose information significantly, and some efficient conditions performed equally well as the non-efficient conditions.

5.4 RQ3: Does Rewriting the Prompt Matter?

Our iterative prompt engineering began with **Mistral-Eff-v1**. For v1 and v2, the LLM focused heavily on information density reduction, as shown by their lower word counts. Checking the performance of **Mistral-Eff-v2** by Q&A type (Fig. 3a) revealed that the model excessively stripped structural data from tables and formulas. This caused its formula extraction score to fall to the **Transcript-only** baseline (0.17) and its table extraction score to drop near zero (0.06 vs. 0.0 for baseline).

By contrast, **Mistral-Eff-v3** compressed descriptions more moderately (compression ratio of 0.64) and performed on par with the uncompressed **Gemma-Full-P2**. Thus, prompt engineering can function as a tool for balancing compression factors against information preservation.

5.5 Failure Cases

A detailed inspection of the generated answers that we did revealed three primary failure types for answering questions: (1) **context not clear**, where information in the context was missing or vague, leading all models to deviate from the reference answer (e.g., Q4 on network layers); (2) **extraction errors**, where table, graph or color information extraction was imprecise (e.g., Q6, Q27, Q34), an issue likely mitigable with more powerful VLMs; and (3) **question ambiguity**, where unspecific phrasing of the dataset question led to misinterpretation (e.g., Q18 on a circled label, Q44 on dialog lines).

6 Discussion

In this work, we proposed a technical framework to automatically generate efficient, transcript-conditioned slide descriptions for online video lectures, as well as an evaluation pipeline to analyze information coverage and efficiency. Our empirical evaluation demonstrated that the generated descriptions significantly enhanced the evaluation model's capability to correctly answer complex, visual questions. These results provide strong evidence that automated slide descriptions successfully capture and supply visual context that is missing from lecture audio alone. Furthermore, the results show that our evaluation framework can be used to carefully balance efficiency and information preservation. Our efficiency metrics revealed how different prompts compressed information density and preserved information to different degrees. This provides actionable insights into how LLMs must be instructed to preserve information. For instance, a performance drop was observed in formula and table extraction when the model was instructed to prioritize text reduction (**Mistral-Eff-v2**). However, allowing the automated system to skip this kind of information may be a design choice in practice, provided that users could dynamically retrieve these specific details via an interactive VLM chatbot component.

In the following sections, we outline the challenges for curating a qualitative QA dataset for lectures, evaluate the broader dimensions of text efficiency, and reflect on the limitations of purely automated metrics.

6.1 Crafting Academic Datasets

A challenge in designing high-quality Q&A pairs is the requirement of sufficient domain knowledge. A pedagogically meaningful description of an image requires sufficient context to address not only basic descriptions but also the intention of the information. Furthermore, we observed that several MCIF presentation slides lacked adequate axis labels or legends, hindering data extraction for both human annotators and VLMs. This highlights a critical prerequisite for automated accessibility: the source material itself must be clear, well-structured, and properly annotated. This benefits all students, not only BLV students.

6.2 Efficiency measures

Reducing redundancy is key to enabling BLV students to review lecture materials efficiently and maintain a study pace equivalent to that of their sighted peers. In addition, interactive question-answering should remain available for specific inquiries, but slide descriptions should minimize the need for students to ask recurring questions about individual slides.

In future research, other aspects of efficiency should also be considered such as structuring text layouts into bulleted or nested segments to allow users of screen-readers to skip and scan content.

Finally, aggressive text compression carries the risk of degrading narrative flow, possibly making the text harder to mentally synthesize. Future research must look at human evaluation to test how condensed texts impact user cognitive load, which was beyond the scope of our technical evaluation.

7 Limitations

One limitation of the study is the use of relatively small, open-source VLM and LLM architectures for both context generation and question answering. More powerful models would likely yield superior information retrieval capabilities, particularly for complex visuals. Furthermore, while the custom dataset was tailored to the target group, the scalability of this approach depends on a systematic method to generate such questions in real-world applications. Future research should look at how to automate this. The experimental factors—namely, the two models and the different prompt versions—were not fully crossed, as not all prompt versions were evaluated with Gemma. This could be a limitation because prompt effects may vary across models. However, given the already substantial computational cost of the experiments and the limited additional insight expected from testing every combination, we gave greater weight to sustainable resource use than to the limited expected gain in knowledge. In addition, we did not explicitly set sampling configurations such as temperature or random seed. Since both the Mistral API's provider defaults and the local Gemma model are configured to enable stochastic sampling, results might slightly differ across runs, which limits the reproducibility of our results. Finally, this work presents a technical evaluation. While automated pipelines provide a valid baseline for system evaluation, human-centered testing is indispensable for understanding user preferences and behavioral metrics, such as cognitive load and time-to-answer metrics.

8 Conclusion

Driven by the goal of making visual lecture materials accessible and efficient for BLV students, this work shows that automated slide descriptions significantly improve Q&A accuracy, and that LLM-based redundancy filtering eliminates transcript-redundant information. Careful prompt design can balance efficiency and information preservation. Future work should investigate human-centric aspects of efficiency, such as hierarchical structuring and time-to-answer metrics.

9 Acknowledgment

We acknowledge the use of generative AI for grammatical text corrections (Gemini, Writefull) and for coding assistance (Claude Code).

References

Katharina Anderer, Karin Müller, Lukas Strobel, Matthias Wölfel, Jan Niehues, and Kathrin Gerling. 2025. Making lecture videos accessible for students who are blind or have low vision through ai-assisted navigation and visual question answering. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15.

Katharina Anderer, Karin Müller, and Matthias Wölfel. 2026. Lessons learned for long-term accessibility studies: a case study with blind and low vision students on a prototype for accessing video lectures. In *International Conference on Computers Helping People with Special Needs*, pages 148–156. Springer.

Katharina Anderer, Matthias Wölfel, and Jan Niehues. 2024. Identifying the information gap for visually impaired students during lecture talks. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 168–173. IEEE.

Helen Armstrong and Iain Murray. 2007. Remote and local delivery of cisco education for the vision-impaired. In *Proceedings of the 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, ITiCSE '07, page 78–81, New York, NY, USA. Association for Computing Machinery.

Matthew Butler, Leona Holloway, Kim Marriott, and Cagatay Goncu. 2017. Understanding the graphical challenges faced by vision-impaired students in australian universities. *Higher Education Research & Development*, 36(1):59–72.

Helen Frank, Mike McLinden, and Graeme Douglas. 2014. Investigating the learning experiences of student physiotherapists with visual impairments: An exploratory study. *British Journal of Visual Impairment*, 32(3):223–235.

Gemma Team. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Jin Guang, Jing Chao, Zong Tianqi, Chen Siya, Cui Chaoyuan, and Fan Jun. 2024. Data redundancy elimination and noise processing via large language model prompt engineering. In *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*, pages 818–825. IEEE.

Azizul Haque and Jonggi Hong. 2026. Enhancing slide presentation accessibility for blind and low-vision audiences through delay-buffered editing. In *Proceedings of the 2026 Designing Interactive Systems Conference*, DIS '26, page 3488–3508, New York, NY, USA. Association for Computing Machinery.

Christian Huber, Tu Anh Dinh, Carlos Mullov, Ngoc-Quan Pham, Thai Binh Nguyen, Fabian Retkowski, Stefan Constantin, Enes Ugan, Danni Liu, Zhaolin Li, Sai Koneru, Jan Niehues, and Alexander Waibel. 2023. End-to-end evaluation for low-latency simultaneous speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Ed.: Yansong Feng, Els Lefever*, page 12–20. Association for Computational Linguistics (ACL).

Chongyang Li, Zhiqiang Yuan, Jiapei Zhang, Ying Deng, Hanbo Bi, Zexi Jia, Xiaoyue Duan, Peixiang Luo, and Jinchao Zhang. 2025a. Less redundancy: Boosting practicality of vision language model in walking assistants. *Preprint*, arXiv:2508.16070.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025b. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yen-Chun Lin, Pei-Hua Tsai, Yu-Hao Weng, Hsin-Lun Chiu, Chu-Yun Ma, and Yung-Ju Chang. 2026. " there were too many to check, so i just added one": Using an llm-powered agent to reduce redundant reports in crowdsourced reporting. In *Proceedings of the 2026 Designing Interactive Systems Conference*, pages 1223–1244.

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083.

David M. Pabst, Vincent J. Taverna, Samantha E. Wismer, Lee A. Dosse, David Schmidt, and Matthew M. Barry. 2023. Student engagement with pre-recorded lecture videos in a flipped-class format. *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–5.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2025. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *Preprint*, arXiv:2507.19634.

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

Dong-Yeon Park and Soon-Bum Lim. 2022. Video screen commentary system supporting online learning of visually impaired students. In *International Conference on Computers Helping People with Special Needs*, pages 360–368. Springer.

Yi-Hao Peng, Jeffrey P Bigham, and Amy Pavel. 2021a. Slidecho: Flexible non-visual exploration of presentation videos. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–12.

Yi-Hao Peng, Peggy Chi, Anjuli Kannan, Meredith Ringel Morris, and Irfan Essa. 2023. Slide gestalt: Automatic structure extraction in slide decks for non-visual access. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. 2021b. Say it all: Feedback for improving non-visual presentation accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. *Preprint*, arXiv:2303.15446.

John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Leevke Wilkens. 2025. Using videos in higher education: Experiences of students with visual impairment. *British Journal of Visual Impairment*, 0(0).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Zhuohao Zhang, Gene SH Kim, and Jacob O Wobbrock. 2023. Developing and deploying a real-world solution for accessible slide reading and authoring for blind users. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15.

Yi Zhao, Siqi Wang, Qiqun Geng, Erxin Yu, and Jing Li. 2025. "less is more": Reducing cognitive load and task drift in real-time multimodal assistive agents for the visually impaired. *Preprint*, arXiv:2511.00945.

A Appendix

A.1 Prompts to Generate Descriptions

Prompt 1: "Analyze this lecture slide and provide a clear, but efficient description that would help someone understand what this slide is about and what visual information is there. Structure it into 'Main content', 'text content' and 'visual information'."

Prompt 2: "You are analyzing a slide from an academic lecture video. Extract and reproduce every piece of information on the slide with full detail: Title (exact text), Text content (bullet points, labels verbatim), Formulas (unicode/LaTeX), Tables (exact values, headers, units), Figures/graphs (axes, legends, trends), and other visual elements. Be exhaustive. Preserve exact wording and numbers."

A.2 Prompts to Filter Redundancies

v1: "Rewrite the visual description to remove information the speaker already explained in detail. Keep information (diagrams, table numbers, formulas, etc.) as long as the speaker did not talk about it precisely."

v2: "Rewrite the slide description so it contains only information a reader cannot derive from the transcript alone. If fully conveyed, omit it. When in doubt, keep it—never sacrifice visual information."

v3: "Rewrite the slide description [...] REMOVE only statements that the speaker already conveyed fully [...] KEEP everything else—especially: diagrams, chart axes, table data, formulas, and citations."