# EduFusion: A Multimodal RAG System based on ERNIE 4.5 VLM for Traceable Lecture Summarization

## Basic Information


• Track: Multimodal
• Participant Name:
Lunshuo Tian, Master of Information Technology, University of New Southern Wales
Qinjin Ji, Master of Electronic Information, Fudan University
• Contact Email:
lunshuo.tian@student.unsw.edu.au
jqj1202@126.com

• Baidu AI Studio UID: 18303603, 18315078


## 1. Synopsis

This project proposes the development of **EduFusion, an intelligent lecture assistant**, designed to overcome the critical challenges of **knowledge reliability and depth-of-field limitations** often encountered in automated lecture summarization and quiz generation.

EduFusion is structured around a cutting-edge multimodal framework built upon the **ERNIE-4.5 (28B) Vision-Language Model**. To handle diverse input modalities, the system integrates advanced perception tools, specifically **PaddleOCR-VL (0.9B)** and **PaddleSpeech**, which are crucial for converting raw data streams (including lecture speech, slides, blackboard notes, etc.) into the structured, multimodal inputs required by the ERNIE model.

The output of EduFusion includes highly reliable **structural summaries, segmented session records, and customized quizzes**, ultimately enhancing learning efficiency and knowledge retention.

## 2. Meaningful application

**The initial motivation for proposing EduFusion stems from addressing a pervasive challenge facing all students:** efficiently assimilating knowledge when attendance is infeasible, and subsequently, managing the overwhelming information volume during high-stakes periods like finals week. **Rapid and accurate knowledge acquisition**—across numerous subjects and dense concepts—remains a major pain point.

The fundamental goal of EduFusion is to serve this universal need: to empower every student, including our own team members, to quickly and precisely grasp core lecture content. This demand for efficiency and accuracy is truly global.

As our exploration deepened, we recognized that the application potential of EduFusion extends far beyond the general student body. Other key beneficiaries include:

- **Educators:** Teachers can utilize EduFusion to comprehensively review their teaching content, allowing for data-driven curriculum optimization and highly relevant quiz generation.
- **Inclusivity Groups:** Learners with **hearing impairments** or those who are **non-native speakers** can significantly lower their learning barriers by accessing high-fidelity transcripts and structured summaries.

Ultimately, the widespread adoption of EduFusion promises a future where **remote and asynchronous learning is no longer a challenge**, democratizing access to education globally.

# 3. Related work

This study is positioned at the intersection of advanced Vision-Language Models (VLMs)[1], Retrieval-Augmented Generation (RAG)[2] architectures, and educational technology (EdTech)[3]. Our work specifically contributes to the field by addressing limitations in existing VLM summarization and RAG-based quiz generation systems.

## 3.1 VLM Application and Limitations in Multimodal Presentation Summarization

Recent research has explored the use of Vision-Language Models for processing complex multimodal documents. For instance, **Gigant et al. (2025)**, in their work titled *"Summarization of Multimodal Presentations with Vision-Language Models: Study of the Effect of Modalities and Structure"*[6], systematically evaluated VLM performance in summarizing presentations by combining audio transcripts and visual slides. Their analysis utilized detailed ablation studies to quantify the contribution of various modalities (speech, OCR text, image) to the final summary quality.

However, this line of research, which primarily focuses on **abstractive summarization**, presents two critical limitations that our **Edufusion RAG architecture** is designed to overcome. Firstly, these models operate in a **closed domain**, which inherently compromises the **authority and factual accuracy** of the output when dealing with external knowledge or requiring strict source verification. In contrast, our system embeds **ERNIE 4.5 VL 28B** within a **RAG framework**, enabling real-time retrieval from the authoritative **MIT 6.867 course knowledge base**, thereby fundamentally guaranteeing the factual correctness and **knowledge source traceability** necessary for rigorous academic settings. Secondly, their

focus remains on simple generative summarization, overlooking the **bidirectional requirements** of a functional teaching assistant system.

### 3.2 RAG-Based Adaptive Quiz Generation and the Need for Multimodal Coverage

The application of RAG to automate educational content creation is a significant advancement in EdTech. **Sreekanth and Dehbozorgi (2024)**, in their research *"Enhancing Engineering Education Through LLM-Driven Adaptive Quiz Generation: A RAG-Based Approach"*[5], demonstrated the feasibility of using an LLM-driven RAG system for generating adaptive quiz questions, successfully enabling the precise assessment of knowledge in an engineering domain.

While that work validated the RAG paradigm for the crucial pedagogical function of **quiz generation**, it is primarily limited by **unimodal knowledge coverage** and scope. Their systems predominantly rely on retrieving information from **purely textual materials**, rendering them incapable of handling the **multimodal information carriers** ubiquitous in modern teaching, such as complex diagram figures, abstract mathematical derivation charts, or handwritten notes. Our research directly addresses this gap: Edufusion employs **ERNIE 4.5 VL 28B** alongside **multimodal augmentation processing** to extract semantic information synchronously from **images, formulas, and text**. This capability allows our system to generate quiz questions that specifically test **multimodal knowledge points**. Furthermore, by reinforcing **ERNIE 4.5 VLM's Chain-of-Thought (CoT)** [4] through fine-tuning, our architecture ensures the generation of structured questions involving **complex derivation processes** and **deep conceptual analysis**, providing a significant competitive advantage in terms of teaching generality and knowledge depth.

## 4. Preliminary work

### 4.1 Testing Datasets

The foundation of this study is the construction of a multimodal dataset that covers diverse teaching scenarios to ensure the generalizability of the Edufusion system. We have completed the initial round of test data collection, acquiring two sets of high-quality raw video datasets from the domain of machine learning:

Firstly, one set comprises **21 segments** of offline lecture videos [9] with an average duration of approximately **50 minutes** and a high resolution 1280 * 720. These videos capture the professor's blackboard writing, where the **handwritten notes** constitute a significant source of visual information, as Figure 1(a). Secondly, a set of **14 segments** of online lecture videos, averaging approximately **80 minutes** in length, where the primary information carrier is the **slides** [8], as Figure 1(b). By collecting these two distinctly different recording scenarios (offline whiteboard writing versus online slides), we aim to enable Edufusion to exhibit robust performance across a broad spectrum of educational data.

(a) (b)

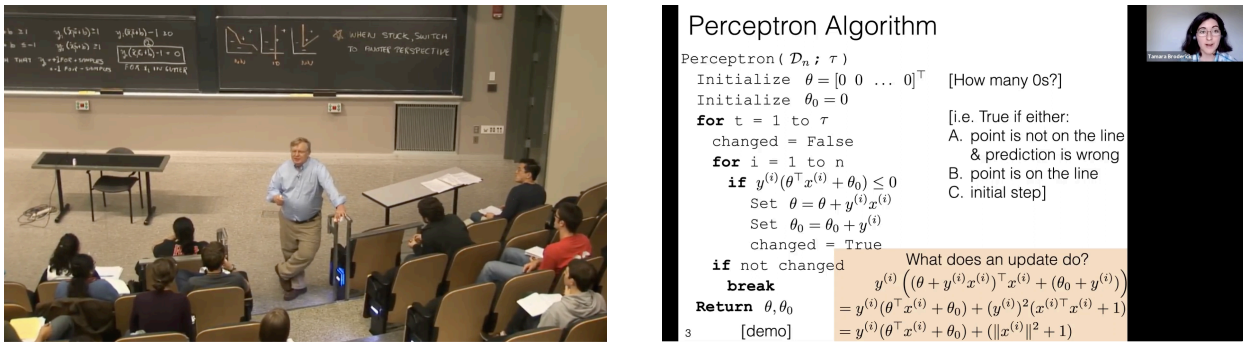


Figure 1. Examples of different source of information, (a) Offline lecture (b) Online lecture

## 4.2 Cross-Modal Information Extraction and API Integration

Building upon the data acquisition, we proceeded with the extraction and preprocessing of key information. We successfully configured and deployed the Baidu Paddle model ecosystem, including **PaddleOCR-VL 0.9B**, **PaddleSpeech**, and the core generation model **ERNIE-4.5 VL-28B**. By calling the **PaddleSpeech API**, we performed high-fidelity **transcripts** on the audio streams of both datasets. Concurrently, utilizing the **PaddleOCR-VL 0.9B API**, we successfully extracted effective visual information from the video frames, encompassing both blackboard writing and slides, shown as Figure 2.

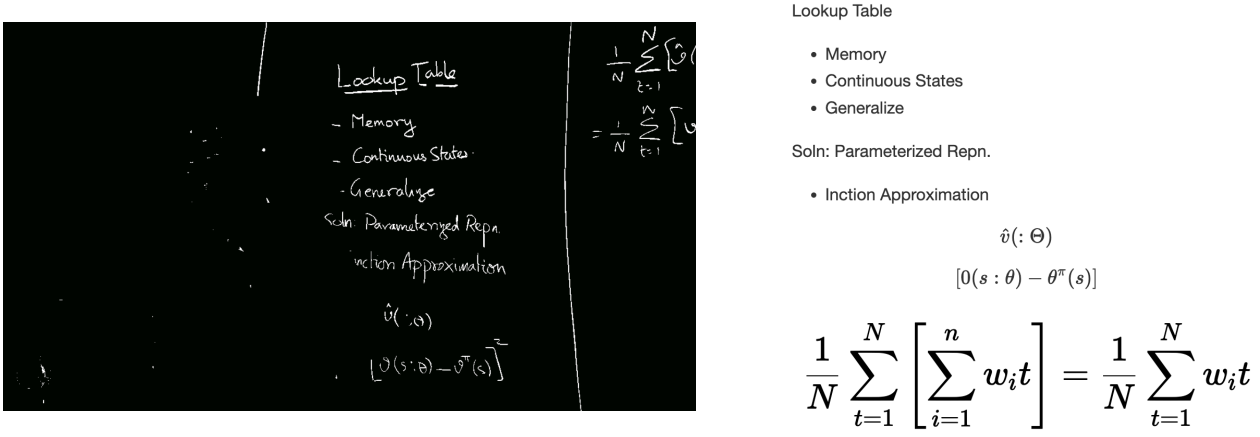


Figure 2. Example of hand-writing blackboard notes converted by PaddleOCR-VL[7]

## 4.3 Interleaved Image-text Representation

To effectively link the multi-channel inputs and enhance the model's contextual semantic understanding, we adopted an **Interleaved slides-transcript representation structure** [3] to process the data input stream. As shown (please refer to Figure 3, the knowledge application model flowchart), this representation method tightly associates and interleaves the speech transcription text with the visual content (OCR text from slides or blackboard writing) corresponding to the respective timestamps. We posit that this approach can effectively

connect **multiple channels of input** (verbal explanations and visual supplementary materials), thereby generating a richer **contextual semantics** within the input sequence, ensuring the model's comprehensive mastery of the lecture content, and significantly improving its cross-modal reasoning accuracy. This meticulous refinement of data representation is a vital prerequisite for enabling the system to handle the complexity inherent in multimodal samples.


Figure 3. Flow chart of preprocessing the multimodal data stream[6]

# 5. Methodology

## 5.1 Dataset

In Section 4, we completed the collection and preprocessing of the test dataset. Our subsequent plan focuses on constructing the RAG knowledge base for the topic covered by this dataset: the field of machine learning. Following preliminary research, we selected the Massachusetts Institute of Technology (MIT) OpenCourseWare **6.867 Machine Learning (Fall 2006) website**[10] as the authoritative knowledge source for building the machine learning knowledge base. The course materials are characterized by their rigorous content, clear structure, and comprehensive coverage of classic core machine learning theories. The data acquisition process primarily targeted three core resource types: **Lecture Notes**, **Problem Sets**, and **Exams** along with their standard solutions.

The initially established dataset rooted in a single authoritative domain, serves as the critical **initial trial for system validation and model tuning**. However, to ensure the ultimate **generalizability** of the model and robustness of the evaluation results across diverse educational contexts, we plan to significantly expand the dataset in subsequent phases. Moving forward, the final dataset is expected to exceed **100 multimodal lecture recordings**, intentionally broadening the scope beyond core Machine Learning to include subjects such as

mathematics, computer science, and physics. This expansion, encompassing diverse visual and auditory classroom characteristics, will be essential for validating the resilience of the Edufusion system.

### 5.1.1 Construction of the RAG Knowledge Base

The construction of the RAG knowledge base relies primarily on the core **Lecture Notes** as input. We adopt a layered processing strategy to address information modalities of varying complexity.

For common textual descriptions, clearly structured mathematical formulas, and intuitive flowcharts or data figures, we primarily utilize the **PaddleOCR-VL model** for efficient recognition and transcription. Leveraging its strong performance in document understanding and vision-language tasks, this model ensures that the vast majority of text, the LaTeX format conversion of formulas, and the structured description of figures are accurately extracted, laying a solid foundation for the RAG system's basic text retrieval layer.

However, when confronted with highly abstract images in the machine learning domain (such as the geometric representation of high-dimensional data spaces or abstract logic diagrams of complex algorithms) or obscure theoretical derivation figures, traditional OCR or limited VLM models struggle to accurately capture and extract the deep semantics. These abstract slides or diagrams often represent critical bottlenecks in understanding core theories.

Therefore, for this subset of abstract visual content that cannot be effectively semanticized by the foundational model, we designate it as input for the higher-capability **ERNIE 4.5 28B VL model**.

### 5.1.2 LoRA Fine-Tuning Dataset

To enhance the expressive capability of **Edufusion** in the teaching scenario, we utilized the structurally extracted **Problem Sets and Exam data** to construct an **Instruction Tuning dataset**.

A knowledge block from the lecture notes is set as the input context, with its corresponding practice question or exam question serving as the target output. This instruction tuning enables the model to learn the inverse mapping: converting an authoritative concept into a test question that adheres to pedagogical standards (such as multiple-choice or short-answer formats).

The standard solution steps from the assignments and exams are used as the ideal output for the model. By instructing the model to "provide a detailed mathematical derivation process," the fine-tuning reinforces **ERNIE 4.5 VLM's logical Chain-of-Thought (CoT)** [4], enabling it to generate professionally detailed and logically rigorous solution steps when faced with complex derivation problems.

Ultimately, the meticulously processed MIT 6.867 dataset will simultaneously serve the RAG system's precise knowledge retrieval (ensuring **factual accuracy**) and the model's instruction fine-tuning (ensuring **generation quality and pedagogical format**).

The currently described dataset focusing on the **Machine Learning** domain serves as an initial trial for system validation and model tuning. As the project progresses, we plan to significantly expand the dataset to include a broader range of subjects, teaching styles, and classroom environments. This will ensure greater generalizability of the model and robustness of evaluation results. The final dataset is expected to exceed 100 multimodal lecture recordings, covering disciplines such as mathematics, computer science, and physics, with diverse visual and auditory characteristics.

## 5.2 Models Architecture

The knowledge application stage is central to this system's ability to perform intelligent question answering and content generation. As depicted in Figure 4, the core of this phase is the coupling of the structured knowledge base with the high-performance Vision-Language Model **ERNIE 4.5 VL 28B**, forming a robust RAG framework. This architecture is specifically designed to overcome the issues of knowledge obsolescence and "hallucination" common in traditional generative models, thereby ensuring the accuracy and authority of the output.

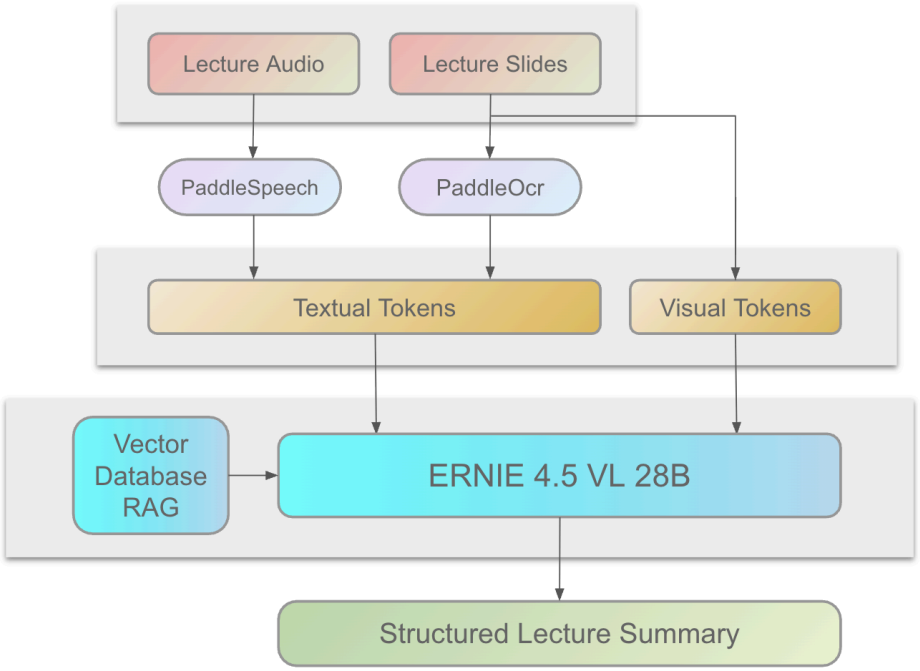


Figure 4. Models architecture of the proposed Edufusion

### 5.2.1 Multimodal input

The system is initiated by two primary raw forms of instructional content: **Lecture Audio** and **Lecture Slides**. The lecture audio is processed through the **PaddleSpeech** module for high-fidelity speech recognition, converting it into manageable **Textual Tokens**.

Simultaneously, the lecture slides pass through the **PaddleOCR** pipeline to extract high-fidelity Textual Tokens, with the original image data directly providing **Visual Tokens**. This dual-stream input strategy ensures comprehensive knowledge coverage, accommodating both the details of verbal explanations and the precision of visual materials.

### 5.2.2 RAG

During the user query stage, the system first activates the retrieval module. The user query is encoded and used to retrieve pre-stored knowledge chunks from the **Vector Database RAG**. This step ensures that only the knowledge context highly relevant in semantics to the current query is recalled.

Subsequently, the retrieved knowledge chunks, along with the raw Textual Tokens and Visual Tokens, collectively form the augmented context, which is then fed into the core generative model, **ERNIE 4.5 VL 28B**. This process follows the classic RAG paradigm, where the retrieval system provides the generative model with real-time, external, and authoritative "open-book" knowledge. **ERNIE 4.5 VL 28B**, leveraging its powerful multimodal heterogeneous MoE architecture, is capable of simultaneously processing and deeply fusing textual and visual information, making it particularly well-suited for handling the complex, visually-rich reasoning tasks prevalent in machine learning courses.

### 5.2.3 Generated Content

**ERNIE 4.5 VL 28B** undertakes the core generation tasks of the RAG system. Upon receiving the precise context provided by the retrieval mechanism, its capabilities are extended into two key pedagogical application areas:

Firstly, the model is responsible for **Structured Lecture Summaries and In-Depth Answers**. Based on the retrieved knowledge chunks and raw multimodal tokens, it performs high-level semantic reasoning and content synthesis. For complex queries posed by students, **ERNIE 4.5 VL 28B** outputs a **Structured Lecture Summary**, which includes key concept summaries with hierarchical headings, detailed derivation steps for complex formulas, and examples of theoretical applications. This process not only ensures the completeness of the information and academic rigor but also strictly attaches knowledge provenance to the output.

Secondly, the model accomplishes **High-Quality Teaching Test Question Generation (Quiz Generation)**. Benefitting from the instruction fine-tuning conducted in Section 4.2.3 using exam and assignment data, **ERNIE 4.5 VL 28B** is empowered with the role of a professional "Question Creator." When a user issues a command to generate test questions, the model is capable of converting the retrieved topic-specific knowledge chunks into pedagogical test questions. It strictly adheres to the format standards learned during fine-tuning, outputting structured questions, such as multiple-choice, fill-in-the-blank, or short-answer types, while simultaneously providing the standard answer and detailed solution steps.

## 5.3 Evaluation

### 5.3.1 Quantitative Evaluation

To comprehensively measure the effectiveness of the constructed RAG knowledge base and the **ERNIE 4.5 VL 28B** model, we strictly applied the multi-dimensional set of evaluation metrics defined in Section 3.4. This evaluation aims to quantify the system's performance across **knowledge recall accuracy, fluency of generated text, and efficiency of multimodal information fusion**. The evaluation baseline is established using standard reference answers (Reference Summaries) and test sets generated from the structured extraction of assignment and exam data within the MIT 6.867 course materials.

Furthermore, to assess the **linguistic quality** of the generated text, we employed the **reference-free metric GRUEN (G)**[11]. Since this metric relies solely on the generated text itself, it effectively evaluates the ability of **ERNIE 4.5 VL 28B** to produce fluent, coherent natural language answers and summaries that adhere to high academic standards while maintaining its domain-specific expertise.

Most crucially, we utilized the multimodality-adapted **Importance-based Relevance score (IbR)**[6] to quantify the model's capacity to capture key information across different input modalities. Given the complexity of the RAG knowledge base sources, we refined IbR into three modality-specific versions: **IbR_transcript**, **IbR_ocr**, and **IbR_overall**. These metrics precisely reveal the reliance and balance of **ERNIE 4.5 VL 28B** on oral details and visual-textual information during knowledge recall, by analyzing the coverage of important words in the summary originating from the speech transcription text and the OCR-extracted text. **IbR_overall** provides a comprehensive assessment of the model's effectiveness in extracting the main ideas from the entire multimodal source document, demonstrating the scaling advantage of the RAG framework when processing long documents.

### 5.3.2 Human Evaluation

Acknowledging the potential limitations in relevance of automated metrics, particularly when evaluating structured summaries and quiz generation tasks for long documents and multimodal inputs, we designed a detailed **Double-Blind Human Evaluation** experiment. This experiment is intended to capture subtle quality differences that are difficult to quantify and directly validate the efficacy of our proposed RAG and multimodal enhancement collaborative architecture in a practical teaching application scenario.

We recruited ten students from Computer Science or related backgrounds who have taken machine learning courses to serve as evaluators. This group, representing the primary target users of the system, can provide the most authentic feedback on pedagogical content. Evaluators were asked to independently score the system's generated structured summaries and **teaching test questions (Quiz)** on a 1 to 10 scale, where a higher score indicates better quality.

The human evaluation primarily focused on the following three core metrics:

1. **Relevance:** Measures the semantic agreement between the generated content and the user's original query or specified knowledge point.
2. **Correctness:** Assesses the **factual accuracy** and **logical rigor** of the generated content, especially the accurate coverage of complex mathematical derivations and code examples.
3. **Pedagogical Value:** Evaluates the **clarity, organizational structure**, and **practical utility** of the generated content as learning material; for the Quiz, it assesses the **precision of the learning objective** and the **reasonableness of the difficulty level**.

To isolate and verify the gains provided by the RAG mechanism and multimodal input, we divided the evaluation targets into three comparison groups:

1. **Baseline Model (Baseline):** Generates content using only the **ERNIE 4.5 VL 28B** model, **without access to the RAG knowledge base** and **without multimodal enhanced input**.
2. **RAG-Augmented Model (RAG-Only):** Uses **ERNIE 4.5 VL 28B** with the RAG knowledge base, but **only utilizes purely textual Chunks**, excluding VLM-processed multimodal enhancement information.
3. **Final Model (Final Model):** Uses **ERNIE 4.5 VL 28B**, accesses the RAG knowledge base, and includes **all VLM-processed multimodal enhanced Chunks** (as shown in Figure 2).

Through this comparative experimental design, we can perform detailed statistical analysis of the scoring results. We anticipate that this human evaluation will unequivocally demonstrate that the **Final Model**, enhanced by the RAG mechanism and multimodal input processing, will significantly outperform the other two comparison groups across the three core metrics of Relevance, Correctness, and Pedagogical Value, thereby strongly supporting the superiority of the proposed dual-model collaborative architecture.

## 6. Project Plan

| Phase | Description | Outcome | Duration |
|---|---|---|---|
| Data Acquisition & RAG Foundation | Data collection; Initial processing via PaddleOCR-VL; Multimodal augmentation using ERNIE 4.5 VL 28B; RAG knowledge base and retrieval pipeline setup. | Structured RAG Knowledge Base; Integrated Retrieval Functionality. | Oct 23 - Oct 30 |
| Model Fine-Tuning & Core Features | Instruction fine-tuning of ERNIE 4.5 VL 28B using assignment/exam data; Implementation of structured summarization and Quiz generation capabilities. | Fine-tuned **ERNIE 4.5 VL 28B Model**; Core API Interface. | Oct 31 - Nov 6 |

| | | | |
|---|---|---|---|
| Evaluation & Web UI Development | **Quantitative Evaluation** (R1, R2, G, IbR); **Web UI Interface development**, integration of the RAG-VLM API, and implementation of front-end interaction. | All Quantitative Results; Working **Web UI Prototype**. | Nov 7 - Nov 16 |
| Human Evaluation & Final Deliverables | Execution of double-blind human evaluation experiments; Data organization and analysis; **Technical Report (Paper) writing**; Production and editing of the **Demo Presentation Video** | Final Evaluation Data; Complete Technical Report Draft; **Demo Video**. | Nov 17 - Nov 23 |

# 7. OpenSource Contributions

## 7.1 Dateset Contribution

**To foster reproducibility and accelerate research in educational AI, we will release the curated multimodal lecture dataset constructed in this project.** This dataset includes offline blackboard-writing lectures and online slide-based lectures, along with synchronized ASR transcripts and OCR-extracted text. We aim to fill a critical gap in existing open-source resources, where authentic multimodal classroom scenarios are largely underrepresented. The dataset will provide significant value for future research in lecture summarization, multimodal reasoning, and educational technology.

## 7.2 Paddle Toolkit Innovation

Our project leverages and extends the Paddle ecosystem, particularly **PaddleOCR-VL** and **PaddleSpeech**, in innovative ways:

For **PaddleOCR-VL**, we adapted its pipeline to handle challenging classroom conditions such as handwritten mathematical formulas and low-contrast blackboard notes, improving robustness in real-world educational settings.

For **PaddleSpeech**, we integrated high-fidelity speech recognition with multimodal timestamp alignment, enabling precise synchronization between spoken explanations and visual materials.

## 7.3 Community Sharing

To maximize impact, **we will publish the EduFusion Demo Pipeline on AtomGit and Github**, including **data preprocessing scripts** for multimodal lecture inputs, **RAG knowledge base** construction workflow, **LoRA fine-tuning configurations** for **ERNIE 4.5 VL 28B** and other necessary code and files.

# 8. References


[1] Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *46*(8), 5625-5644.
[2] Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A Survey on RAG with LLMs. *Procedia computer science*, *246*, 3781-3790.
[3] Donahoe, B., Rickard, D., Holden, H., Blackwell, K., & Caukin, N. (2019). Using EdTech to enhance learning. *International Journal of the Whole Child*, *4*(2), 57-63.
[4] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
[5] Gopi, S., Sreekanth, D., & Dehbozorgi, N. (2024, October). Enhancing Engineering Education Through LLM-Driven Adaptive Quiz Generation: A RAG-Based Approach. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1-8). IEEE.
[6] Gigant, T., Guinaudeau, C., & Dufaux, F. (2025). Summarization of Multimodal Presentations with Vision-Language Models: Study of the Effect of Modalities and Structure. *arXiv preprint arXiv:2504.10049*.
[7] Ashwin, S. S., Yogesh, B. R., Rutvik, B., & Jayashree, R. (2021). Summarization of video lectures. In *Artificial Intelligence and Speech Technology* (pp. 149-158). CRC Press.
[8] https://tamarabroderick.com/ml.html
[9] https://ocw.mit.edu/courses/6-034-artificial-intelligence-fall-2010/
[10] https://ocw.mit.edu/courses/6-867-machine-learning-fall-2006/
[11] Zhu, W., & Bhat, S. (2020). GRUEN for evaluating linguistic quality of generated text. *arXiv preprint arXiv:2010.02498*.