VidExplainAgent: A Multimodal RAG System for Accessible STEM Video Education

Syed Ali Haider
Dartmouth College
Hanover, NH, USA
syed.ali.haider.gr@dartmouth.edu

Abstract

We present VidExplainAgent, a multimodal RAG system designed to make STEM video content accessible to blind and low vision (BLV) learners. The system uses a vision-language model for visual description, a vector database for indexing, and an interactive text-to-speech interface. We evaluated the system using a two-tiered approach (component and end-to-end RAG), achieving strong performance on a quantum physics video. Key results include a `BERTScore F1` *of* 0.588 *for visual descriptions and an* 87.5% *overall pass rate for question-answering, featuring perfect* 100% *retrieval accuracy and* 90% *answer faithfulness. These findings show the system's potential to significantly improve STEM accessibility.*

Keywords: Accessibility, Multimodal AI, RAG Systems, Vision-Language Models, Educational Technology, Evaluation Metrics

1. Introduction

1.1. Motivation

Blind and low vision (BLV) learners face significant barriers when accessing STEM educational content, particularly video lectures that rely heavily on visual elements such as diagrams, equations, animations, and demonstrations. While audio descriptions exist for entertainment media, they are rarely available for educational content, and manual creation is time-consuming and expensive. This accessibility gap prevents BLV learners from fully engaging with cutting-edge STEM education resources.

1.2. Contributions

We present VidExplainAgent, a system that automatically makes STEM videos accessible through:

1. **Multimodal Visual Description Generation**: Automatic extraction and description of visual elements using state-of-the-art vision-language models
2. **Interactive RAG Pipeline**: Question-answering system allowing learners to query specific aspects of video content
3. **Comprehensive Evaluation Framework**: Two-tiered evaluation methodology combining traditional NLP metrics with RAG-specific assessments
4. **Production-Ready Implementation**: Full-stack application with accessibility-first design principles

Our evaluation on a quantum physics educational video demonstrates strong performance across both semantic understanding (BERTScore F1: 0.588) and question-answering capabilities (87.5% pass rate), with particular strength in retrieval accuracy (100%) and answer faithfulness (90%).

2. Related Work

2.1. Vision-Language Models for Accessibility

Recent advances in vision-language models (VLMs) have enabled automatic image and video captioning [6], [11]. However, most work focuses on general-domain content rather than specialized educational material. Our work extends these capabilities to STEM education, handling complex visualizations such as equations, diagrams, and animations.

2.2. Retrieval-Augmented Generation

RAG systems [5] have shown promising results in question-answering tasks by combining retrieval mechanisms with large language models. While RAG has been applied to text-based educational content [4], its application to multimodal educational videos remains underexplored. We demonstrate that RAG can effectively handle temporal video content with rich visual semantics.

2.3. Accessibility in Education

Prior work on educational accessibility for BLV learners has focused on static materials [1] or manual annotation [3]. Automated systems have been limited to simple image captioning [8]. VidExplainAgent addresses the gap in automated, comprehensive video accessibility for complex STEM content.

2.4. Evaluation of Generative Systems

Traditional NLP metrics (BLEU [9], ROUGE [7]) have limitations for evaluating semantic understanding in generative systems. Recent work introduces semantic similarity metrics (BERTScore [13]) and RAG-specific evaluation frameworks (RAGAS [2]). We employ both traditional and modern metrics for comprehensive evaluation.

3. System Architecture

3.1. Overview

VidExplainAgent consists of three main components: (1) an ingestion pipeline for processing and indexing videos, (2) a retrieval and generation pipeline for answering queries, and (3) a user interface designed for accessibility. Figures 1 and 2 illustrate the complete system architecture. Appendix C shows the product's inference.

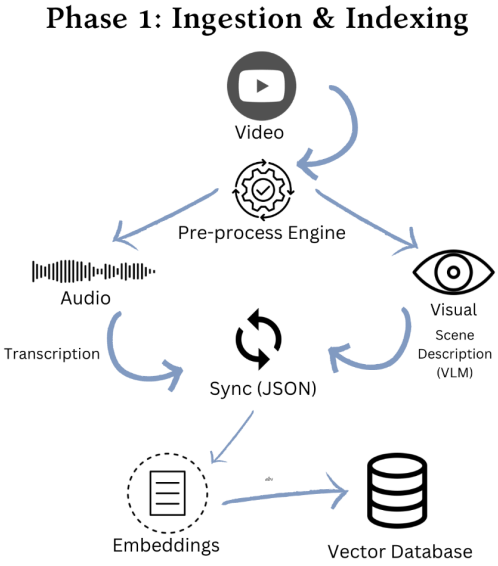

Figure 1. The VidExplainAgent Ingestion Pipeline. A source video is processed in parallel: audio is transcribed (ASR) and visual frames are analyzed (VLM). The resulting data is synchronized into a unified multimodal representation and indexed in a vector database.

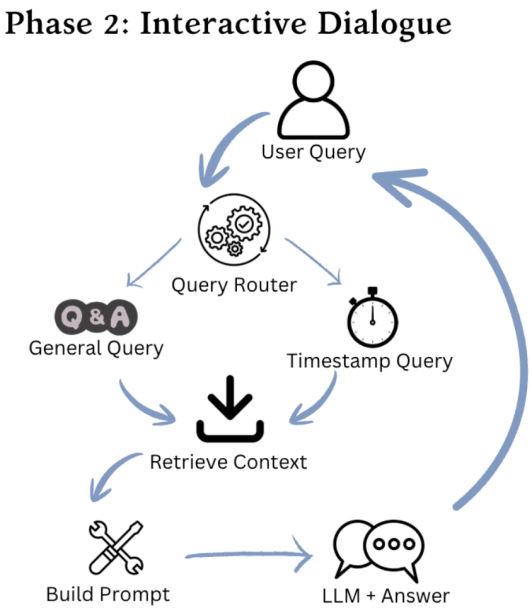

Figure 2. The RAG Retrieval and Generation Pipeline. A user query is routed for either semantic or timestamp-based search. The retrieved context is used to construct a detailed prompt, which the LLM uses to generate the final answer.

3.2. Ingestion Pipeline

3.2.1. Video Processing and Temporal Segmentation

The system accepts YouTube URLs or direct video uploads. Videos are downloaded and processed into temporal segments to balance granularity with context. Our approach uses natural scene boundaries and content transitions rather than fixed-interval sampling.

3.2.2. Multimodal Analysis with VLMs

The VidExplainAgent employs Gemini 2.5 Flash for multimodal content extraction. Each video segment is analyzed using a carefully engineered prompt (see Appendix A.1) that instructs the Vision-Language Model (VLM) to extract several distinct features: Visual Descriptions, Transcript Alignment, Cognitive Summaries, Technical Details (including equations and code), Speaker Information, and Educational Context (such as difficulty and prerequisites).

3.2.3. Robust JSON Parsing

We implement a robust JSON parsing pipeline with multiple fix strategies (e.g., removal of markdown, trailing comma correction) to handle stochastic VLM outputs, achieving a $> 95\%$ successful parsing rate.

3.2.4. Vector Database Indexing

Processed segments are embedded using 'BAAI/bge-large-en-v1.5'[12] and indexed in ChromaDB. Each document

stores textual content and rich metadata (timestamps, concepts, etc.).

3.3. Retrieval and Generation Pipeline

User queries are embedded using the same bge-large model [12]. The vector database, ChromaDB, then performs an approximate nearest neighbor search, typically retrieving the top-k segments (where $k = 5$) using cosine similarity. Retrieved segments are then assembled into a rich context. The subsequent RAG prompt (see Appendix A.2) is constructed to incorporate principles like progressive disclosure, cognitive load management, and specific accessibility guidelines. Finally, Gemini 2.5 Flash generates the answers, equipped with reliability features such as exponential backoff for API failures, response validation, and precise timestamp attribution. The generated text is subsequently synthesized into speech using dual options: macOS Native (for speed) or Gemini TTS (for high quality).

4. Evaluation Methodology

4.1. Dataset Preparation

For the dataset preparation, we selected a challenging test case, "Wave-Particle Duality" by Perimeter Institute [10], due to its conceptually dense, highly visual content on quantum mechanics and short duration. The video also doesn't talk directly about any of the visualizations in the narration so it serves as a perfect case study. To enable our two-tiered evaluation, two expert annotators then created the necessary ground truth resources. Specifically, we generated Visual Descriptions for 23 segments, which are frame-accurate descriptions following audio description standards, and prepared Question-Answer Pairs (20 total) that include diverse questions spanning factual recall, visual detail, and conceptual understanding.

4.2. Component-Level Evaluation

Our comprehensive evaluation is structured across two tiers. For the component-level evaluation, we assess the quality of generated visual descriptions against human ground truth using three complementary metrics: BLEU [9] measures lexical precision via n-gram overlap; ROUGE-L [7] assesses sequence similarity; and BERTScore [13] captures semantic similarity using contextual embeddings. For the end-to-end RAG evaluation, we utilize the RAGAS framework [2] to systematically test the full question-answering pipeline. This framework utilizes four key metrics to ensure robustness: Context Relevance (whether retrieved context is relevant), Answer Faithfulness (whether the answer is grounded in the context, preventing hallucination), Answer Relevancy (whether the answer addresses the question), and Answer Correctness (whether the answer is factually accurate). This RAGAS assessment is executed using an LLM-as-Judge methodology, where GPT-4o-mini serves as the evaluator, employing metric-specific prompts and deterministic (temperature=0.0) binary (pass/fail) classification for highly consistent results.

Table 1. Component-Level Metrics for Visual Description Quality (n=10)

Metric	Mean	Median	Std	Min	Max	Target
N-gram Overlap (BLEU)						
BLEU-1	0.210	0.250	0.100	0.070	0.317	>0.40
BLEU-4	**0.045**	0.019	0.049	0.004	0.140	>0.30
Sequence Similarity (ROUGE-L)						
F1-Score	**0.248**	0.232	0.124	0.077	0.389	>0.40
Semantic Similarity (BERTScore)						
Precision	0.546	0.556	0.103	0.413	0.704	>0.85
Recall	0.640	0.675	0.096	0.512	0.770	>0.85
F1-Score	**0.588**	0.606	0.099	0.458	0.720	>0.85

4.3. Experimental Setup

The system was executed on an Apple Macbook Pro M4 CPU with 16GB RAM, leveraging a software stack consisting of Python 3.13, FastAPI, Next.js 14, and ChromaDB. The key models utilized in the implementation were: the VLM/Generator (`Gemini 2.5 Flash`), the Embedding model (`BAAI/bge-large-en-v1.5`), and the LLM Judge (`GPT-4o-mini`).

5. Results

5.1. Component-Level Performance

The Component-Level Performance of the system, presented in Table 1, was assessed using 10 generated visual descriptions compared against human ground truth. Analysis of the results shows the system achieves a strong `BERTScore F1` of 0.588, which indicates solid semantic understanding of the visual content. The high recall value (0.640) confirms that the system successfully captures 64% of the semantic information present in the human annotations. A key finding is the notable disparity between the low `BLEU` score (0.045) and the high `BERTScore` (0.588): this demonstrates that the system conveys accurate meaning using vocabulary different from the annotators, a characteristic acceptable for generative accessibility tools.

5.2. RAG System Performance

The end-to-end RAG evaluation results, summarized in Table 2, were obtained across 20 question-and-answer pairs. Analysis of the results shows the system achieved an 87.5% overall pass rate. Most notably, the retrieval component was perfect, demonstrating 100% Context Relevance, which

validates the effectiveness of our embedding and search strategy. Answer Faithfulness was high at 90%, indicating minimal hallucination and establishing trust crucial for an educational tool. While the 75% Answer Correctness is acceptable, it highlights the primary area requiring improvement in factual precision. Overall, 14 of the 20 questions (70%) passed all four metrics perfectly (a complete per-question breakdown is available in Appendix B).

Table 2. RAG System Evaluation Results (RAGAS Framework)

Metric	Pass	Fail	Pass Rate	Target
Context Relevance	**20**	0	**100%**	>80%
Answer Faithfulness	18	2	**90%**	>90%
Answer Relevancy	17	3	85%	>80%
Answer Correctness	15	5	75%	>70%
Overall Pass Rate	-	-	**87.5%**	>75%

6. Discussion

Our evaluation highlights three key findings. First, the system's generated descriptions prioritize *meaning* over *exact wording*, as shown by a high `BERTScore` (semantic match) despite a low `BLEU` score (lexical match). This suggests semantic-based metrics are more appropriate for accessibility. Second, the system achieved perfect 100% Context Relevance, meaning it consistently retrieved the correct information for every query, providing a reliable foundation. Finally, a gap exists between high 90% Faithfulness (staying true to the source) and lower 75% Correctness (factual accuracy), indicating the system sometimes misunderstands or incompletely synthesizes the information it retrieves.

The system's strengths lie in its practical implementation. It was built with an *Accessibility-First* design, integrating features like voice input/output, screen reader compatibility, and partial keyboard navigation. It also demonstrates *Robust Engineering* through reliable JSON parsing and API error handling (exponential backoff), making it suitable for real-world use. The rigorous two-tiered evaluation provides a nuanced view of its performance, and the system proved capable of handling complex subject matter, as shown by its strong results on the quantum physics test video.

Several limitations guide future improvements. The *Temporal Granularity* is coarser than human annotation (11 system segments vs. 23 human segments), which may cause the system to miss rapid visual transitions. The most significant limitation is in *Factual Correctsness* (75%), especially when questions require multi-step reasoning. Finally, the *Dataset Scale* is currently limited to a single video, which means the strong results may not yet be generalizable to all types of STEM content.

7. Resources

The following resources accompany this paper:

- GitHub Repository
- Demo

References

[1] E. Brady, M. R. Morris, and J. P. Bigham, "Visual challenges in the everyday lives of blind people," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2013.

[2] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, *Ragas: Automated evaluation of retrieval augmented generation*, 2025. arXiv: 2309.15217 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.15217.

[3] C. Gleason, P. Gara, A. Z. C., J. P. Bigham, et al., "Making memes accessible," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2020.

[4] Z. Jiang, F. F. Liu, J. Koch, S. Liu, et al., "Active retrieval augmented generation," *arXiv preprint arXiv:2305.06983*, 2023.

[5] P. Lewis et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks*, 2021. arXiv: 2005.11401 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.11401.

[6] J. Li, D. Li, R. Saville, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning (ICML)*, 2023.

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[8] H. MacLeod, A. J. P., J. P. Bigham, et al., "Understanding blind people's experiences with computer-generated captions of social media images," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2017.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[10] Perimeter Institute for Theoretical Physics, *Wave-Particle Duality*, YouTube video, Video length: 3:32. Accessed: [Date Accessed], 2019. [Online]. Available: https://youtu.be/DfQH3o6dKss?si=gRNcDb_Dronj080W.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Messa, I. Sutskever, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[12] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, *C-pack: Packaged resources to advance general chinese embedding*, 2023. arXiv: 2309.07597 [cs.CL].

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr.

A. Prompt Templates

A.1. VLM Extraction Prompt

```
You are analyzing a video segment to
create comprehensive audio descriptions
for blind and low-vision learners.
Extract the following
structured information:

1. COGNITIVE SUMMARY: High-level
conceptual understanding
(2-3 sentences)

2. VISUAL DESCRIPTION: Detailed spatial
and visual information
   - Describe all visual elements on screen
   - Include spatial relationships
   - Describe animations, transitions, movements
   - Note text overlays, equations, diagrams

3. SPEAKER INFORMATION (if applicable):
   - Name, role, visual appearance

4. TECHNICAL DETAILS:
   - Exact notation for equations, formulas
   - Code syntax, diagram structures

5. EDUCATIONAL CONTEXT:
   - Difficulty level: beginner/intermediate/adv
   - Key concepts introduced
   - Prerequisites needed

Output as valid JSON with double quotes and
no trailing commas.
```

A.2. RAG Generation Prompt

```
You are an expert educator creating accessible
explanations for blind and low-vision STEM
learners. Answer the question using ONLY
information from the provided video contexts.

QUESTION: {query}

RETRIEVED CONTEXTS:
{contexts}

GUIDELINES:
1. Progressive Disclosure: Start with overview,
   then details, then technical.
2. Cognitive Load Management: Break complex
   ideas into chunks, define jargon.
3. Grounding: Base all claims on provided
   contexts, cite timestamps.
4. Accessibility: Describe visual elements
   verbally, avoid demonstratives.

Generate a clear, accurate, accessible answer.
```

B. Evaluation Details

B.1. Per-Question RAGAS Results

Table 3 shows the pass/fail status for all 20 questions across the four RAGAS metrics.

C. Product Interface

The figure 3 and 4 below show the UI/UX of the product.

Table 3. Per-Question RAGAS Results (Pass/Fail)

ID	Question Type	Difficulty	Context	Faithful	Relevant	Correct	Pass Rate
q1	Visual Detail	Easy	✓	✓	×	✓	75%
q2	Factual	Easy	✓	✓	✓	✓	100%
q3	Visual Detail	Medium	✓	✓	✓	✓	100%
q4	Visual Detail	Easy	✓	✓	✓	✓	100%
q5	Visual Detail	Medium	✓	✓	✓	✓	100%
q6	Conceptual	Medium	✓	✓	✓	✓	100%
q7	Visual Detail	Medium	✓	✓	✓	✓	100%
q8	Conceptual	Medium	✓	✓	✓	✓	100%
q9	Visual Detail	Medium	✓	✓	✓	✓	100%
q10	Conceptual	Hard	✓	✓	✓	✓	100%
q11	Conceptual	Hard	✓	✓	×	×	50%
q12	Factual	Easy	✓	✓	×	✓	75%
q13	Visual Detail	Medium	✓	✓	✓	✓	100%
q14	Conceptual	Medium	✓	✓	✓	✓	100%
q15	Factual	Easy	✓	✓	✓	✓	100%
q16	Visual Detail	Hard	✓	×	✓	×	50%
q17	Factual	Easy	✓	✓	✓	✓	100%
q18	Factual	Easy	✓	✓	✓	✓	100%
q19	Conceptual	Medium	✓	✓	✓	×	75%
q20	Factual	Easy	✓	×	✓	×	50%

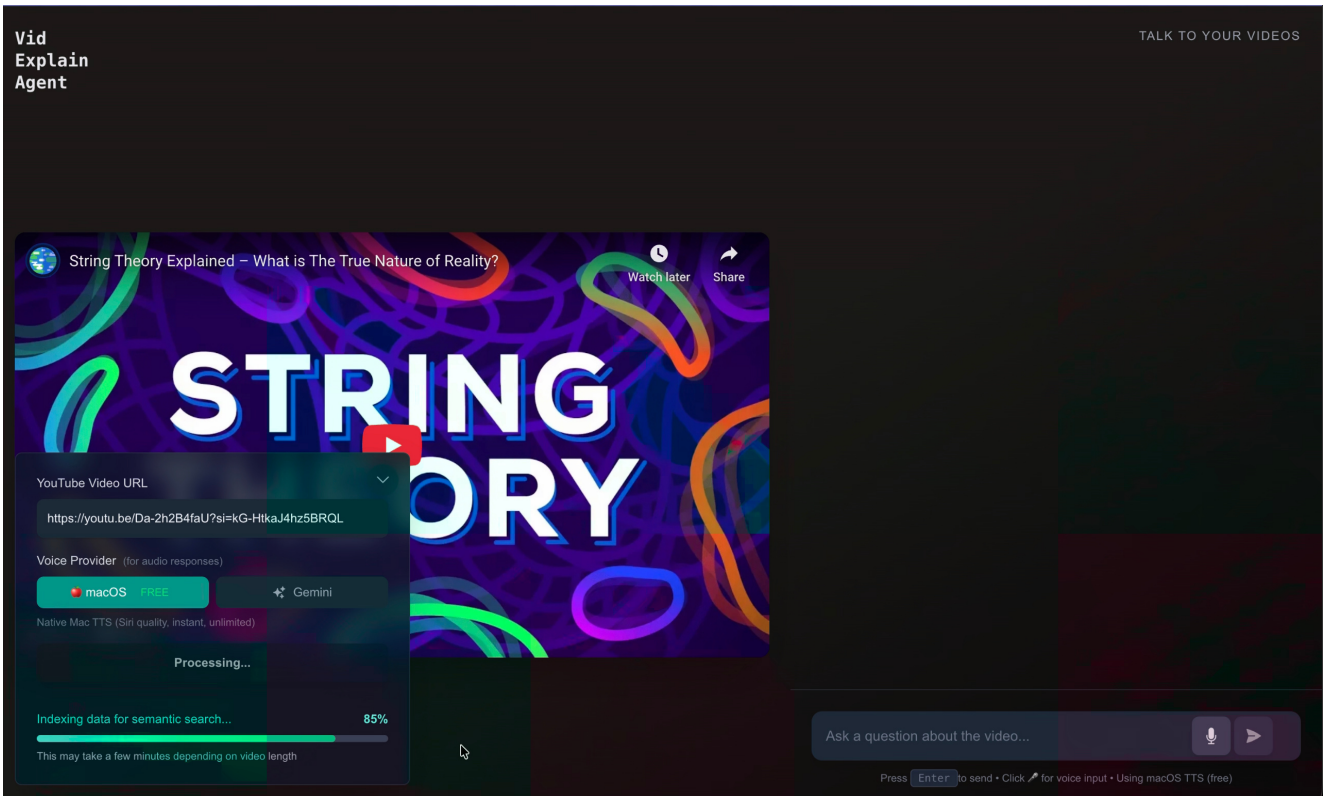

Figure 3. Uploading URL phase

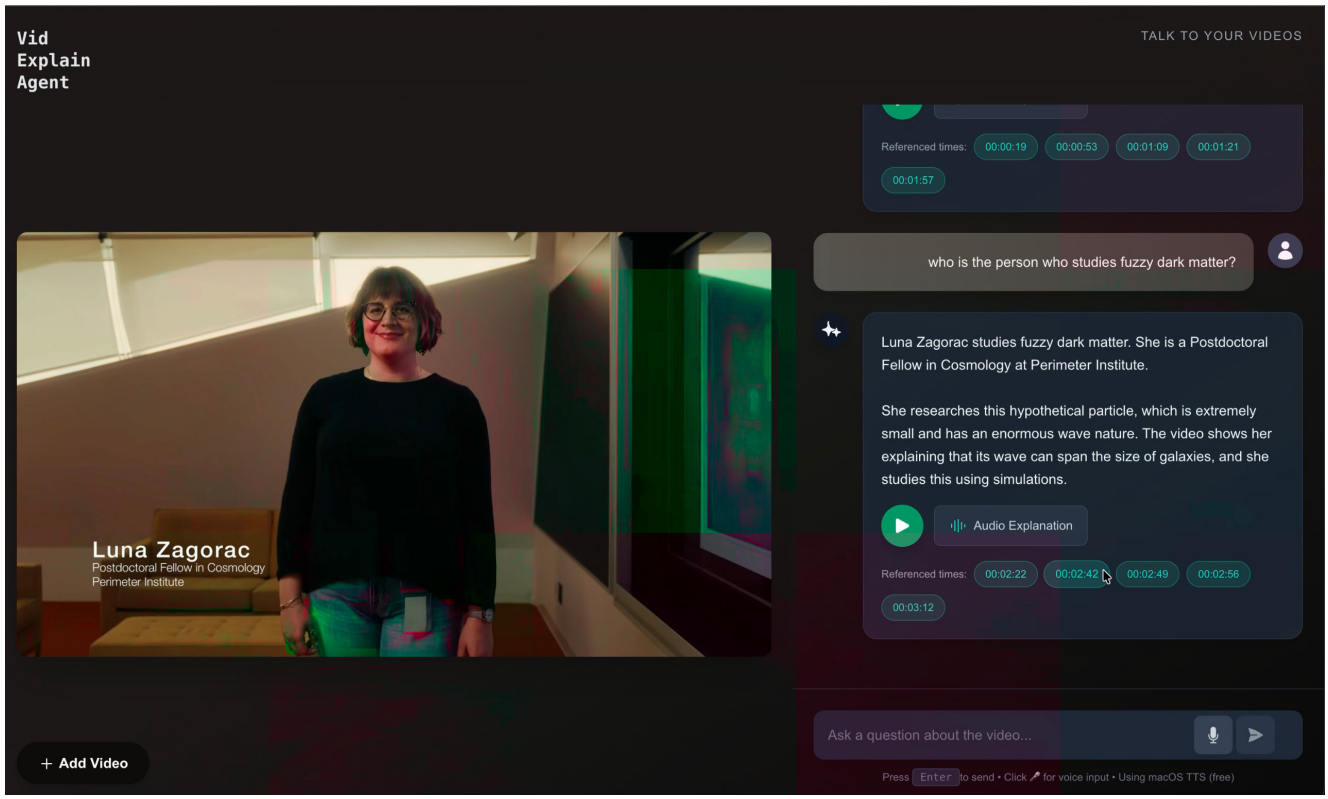

Figure 4. User watching video and asking questions side by side