Multimodal Methods for Educational Applications

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Megha Mariam K. M
2023701006
megha.km@research.iiit.ac.in

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
June 2026

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Multimodal Methods for Educational Applications"** by **Megha Mariam K. M** , has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: C. V. Jawahar

Acknowledgements

My journey through this Master's program has been shaped by far more than research papers, deadlines, and experiments. It has been a journey of learning how to think, communicate, struggle through uncertainty, and grow alongside remarkable people. Above all, I thank God for guiding me through every stage of this journey and for giving me the strength, opportunities, and people who made it meaningful.

First and foremost, I would like to express my sincere gratitude to my advisor, **Prof. C. V. Jawahar**, for giving me the opportunity to be part of this research environment and for profoundly shaping the way I approach research and learning. One of the most valuable lessons I learned from him was the importance of stepping back to understand the bigger picture rather than getting lost in isolated details. Over time, I realized that many of our discussions were not just about research problems, but about learning how to think clearly, ask better questions, and communicate ideas effectively. His feedback was always honest, direct, and purposeful, often arriving exactly when I needed to improve. I am especially grateful for the care and effort he invested in teaching me how to structure ideas, write research papers, and present work in a way that others can follow and appreciate. Watching the clarity and depth with which he approaches problems has been both inspiring and transformative for me.

I am also grateful to **Prof. Vineeth N. Balasubramanian** for his calm and thoughtful mentorship throughout my work. His technical feedback consistently pushed my research to become sharper and more rigorous, while his patient and encouraging nature made discussions both engaging and insightful. I deeply value the guidance, support, and perspective he provided throughout this journey.

I feel incredibly fortunate to have worked with **Dr. Aditya Arun** and **Dr. Zakaria Laskar**. Research becomes far less intimidating when surrounded by people who are approachable, encouraging, and genuinely willing to help, and both of them created exactly that environment. I truly appreciate the patience with which they answered even my simplest questions and the openness with which they discussed ideas, experiments, and problems.

I would also like to thank **Prof. Pawan Kumar**, **Prof. Makarand Tapaswi**, and **Prof. Anoop M. Namboodiri**. Their courses did far more than introduce technical concepts—they sparked my curiosity about Computer Vision and generative models and shaped the direction of my research interests. Many of the ideas and questions that later stayed with me during research began in their classrooms.

A special thanks to my classmates and coursework friends, especially **Sahithi**, **Swaroopa**, **Siddharth**, **Gaurang**, **Varghese** and **Amal** for making the coursework phase feel less overwhelming and far more enjoyable. Long assignments, deadlines, and exam preparations became easier because they

were shared with people who could laugh through the stressful moments together. Some of my fondest memories from this journey came from those everyday moments of collective struggle, discussion, and support.

I am deeply grateful to all my lab mates at CVIT—**Dr. Ajoy, Janak, Snigdha, Lalitha, Varun, Saumya, Saket, Sourvik, Shaon, Harsh, Devika**, and many others—for adding warmth and positivity to this journey. Whether through brief conversations, shared moments, or simply their presence, they made the overall experience at CVIT feel welcoming and memorable.

I would also like to thank the CVIT staff, especially **Ram, Varun, the LACES team, Vidya, Cecilia, Aradhana, and Nagaraju**, for their constant support and willingness to help whenever needed. Their efforts behind the scenes and readiness to assist, even during stressful situations or odd hours, made many things much easier throughout this journey.

Finally, I owe my deepest gratitude to my family, whose support remained constant through every high and low of this journey. To my **Appa and Amma**, thank you for your unwavering belief in me and for the countless sacrifices you made so that I could pursue my goals freely and confidently. Your trust gave me the courage to keep moving forward, even during moments of self-doubt. To my sisters, **Deby and Proby**, thank you for being a constant source of encouragement and strength throughout this journey. Your confidence in me, your support during difficult moments, and your encouragement to pursue this Master's in the first place played a huge role in bringing me here. I could not have done this without all of you.

Lastly, I would like to thank everyone I had the opportunity to meet and interact with during this journey. Sometimes it is the small conversations, unexpected encouragement, shared frustrations, or moments of kindness that stay with us the longest. In ways both big and small, every person I encountered contributed to making this experience worthwhile.

Abstract

Recent advances in large language models (LLMs), vision–language models (VLMs), and text-to-video (T2V) systems have sparked growing excitement about the future of education. Imagine a learning experience where complex ideas can be explained through synchronized text, visuals, narration, and video; where educational content can be generated on demand; and where learning becomes more interactive, accessible, and personalized. While these possibilities are increasingly within reach, an important question remains: do current multimodal AI systems truly understand and connect information across different modalities in ways that meaningfully support learning? This thesis explores that question through the development of benchmarks and evaluation frameworks for multimodal educational AI.

A familiar challenge for learners is following presentation slides while simultaneously listening to a speaker. Slides are often dense, explanations move quickly, and learners can easily lose track of where to focus their attention. To address this, we explore a method that automatically highlights slide regions relevant to the ongoing narration, helping guide learners toward the most important visual information in real time. Supporting this effort, we introduce a dataset designed to evaluate how effectively models can align spoken language with slide content. The dataset consists of 14 presentation videos and 150 slides, where each slide is paired with its corresponding audio segment and transcript. It captures a wide range of visual and textual elements—including figures, equations, tables, and diverse layouts—making it a challenging and realistic benchmark for multimodal alignment.

We then turn to physics education, where visual intuition is often essential for understanding abstract concepts, yet creating high-quality educational animations remains time-consuming and labor-intensive. Motivated by recent advances in T2V generation, we investigate whether these models can create videos that are not only visually realistic, but also pedagogically meaningful. To this end, we introduce *PhyEduVideo*, a benchmark designed to evaluate the ability of T2V models to generate simple, accurate, and educational visualizations for physics concepts. Each concept is decomposed into fine-grained teaching points and paired with carefully crafted prompts intended to elicit clear visual explanations. Beyond visual quality and motion smoothness, we evaluate whether the generated videos actually convey the underlying physics correctly. Our findings reveal a striking gap: while current models often produce visually appealing outputs, their conceptual understanding of physics remains inconsistent.

Finally, we address a broader challenge in scientific communication: knowledge is often scattered across research papers, slides, presentation videos, and explanatory videos. Although these modalities frequently convey the same core ideas, they provide complementary perspectives and rarely contain ex-

plicit links between one another. To bridge this gap, we introduce the *Multimodal Conference Dataset (MCD)*, the first benchmark to integrate all of these modalities from the same set of research works. Using MCD, we evaluate both embedding-based approaches and vision–language models on their ability to discover fine-grained correspondences across formats. Our experiments show that while VLMs are generally robust, they still struggle with precise alignment, whereas embedding-based methods effectively capture text–visual relationships but tend to isolate equations and symbolic content into separate regions of the embedding space. These findings expose important limitations in current multimodal systems while also highlighting promising directions for future research in multimodal scientific understanding.

Together, these contributions move toward educational AI systems that are not only visually impressive, but also accurate, aligned, and pedagogically meaningful—ultimately enabling richer, more connected, and more effective learning experiences.

Contents

List of Figures

List of Tables

List of Related Publications

[P1] Megha Mariam K. M, C. V. Jawahar, **"Attend to what I say: Highlighting relevant content on slides"**, in International Conference on Document Analysis and Recognition (ICDAR), 2025.

[P2] Megha Mariam K. M, Aditya Arun, Zakaria Laskar, C. V. Jawahar, **"PhyEduVideo: A Benchmark for Evaluating Text-to-Video Models for Physics Education"**, in Winter Conference on Applications of Computer Vision (WACV), 2026.

[P3] Megha Mariam K. M, Vineeth N. Balasubramanian, C. V. Jawahar, **"Unifying Scientific Communication: Fine-Grained Correspondence Across Scientific Media"**, in Conference on Computer Vision and Pattern Recognition Finding (CVPRF), 2026.

Chapter 1

Introduction

Artificial intelligence (AI) is rapidly transforming education, changing the way learning content is created, delivered, and experienced. Recent advances in large language models (LLMs), vision–language models (VLMs), and text-to-video (T2V) systems have enabled automated content generation, personalized tutoring, and rich multimodal explanations that extend far beyond traditional learning methods. These technologies have the potential to make education more accessible, adaptive, and engaging by tailoring learning experiences to individual needs and presenting information through multiple forms of media.

At the same time, learning today is no longer limited to static text alone. Modern educational experiences increasingly combine speech, visuals, videos, and interactive elements, each playing an important role in how knowledge is communicated and understood. Text provides structure and precision, while visuals and videos make learning more engaging, help build intuition, and make complex ideas easier to grasp. Bringing these different modalities together effectively is essential for creating meaningful and impactful learning experiences.

However, despite this exciting progress, important challenges still remain. Current AI systems often struggle to connect and reason across multiple modalities, limiting their ability to align information across formats, generate pedagogically meaningful content, and support smooth interaction between different forms of knowledge. As these systems become more integrated into education, it is increasingly important not only to build more powerful models, but also to understand how well they actually perform on multimodal educational tasks.

In this thesis, we explore these challenges through three complementary directions. First, we study how spoken content can be aligned with presentation slides to improve presentation understanding and guide learner attention. Second, we investigate the ability of text-to-video models to generate educationally meaningful videos for physics concepts. Finally, we examine fine-grained correspondences across multiple scientific communication formats, including papers, slides, and videos. Together, these efforts aim to advance the development of more effective, reliable, and pedagogically meaningful multimodal AI systems for education.

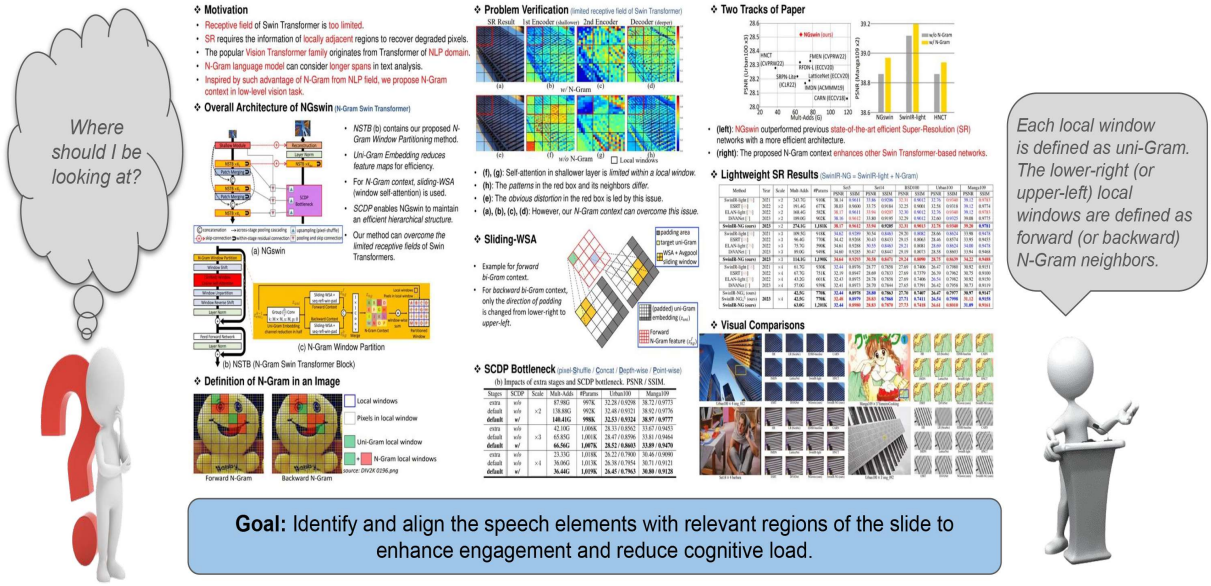


Figure 1.1: This illustration highlights the challenge audiences face during presentations in simultaneously locating relevant content on slides while paying attention to the speaker. By synchronizing spoken information with corresponding visual elements, cognitive load is reduced, engagement is improved, and comprehension and retention of key insights are enhanced.

1.1 Linking Spoken Content to Slides

Understanding a multimedia presentation often requires juggling multiple streams of information simultaneously(Figure 1.1). As a speaker explains concepts, the audience must navigate slides in real time to locate relevant visual elements. Although all the information is present, identifying what is important at a given moment can be challenging. This challenge becomes more pronounced as the speaker moves across topics, requiring listeners to quickly map spoken references to specific regions on the slide. Since speech is transient, by the time the audience identifies the appropriate visual cue, the explanation may have already progressed, resulting in missed connections and fragmented understanding.

This disconnect forces learners to split their attention between listening and visual search, increasing cognitive effort and reducing comprehension. Over time, divided attention can lead to information loss, reduced engagement, and a diminished learning experience. A natural solution is to highlight relevant slide regions as the speaker talks. However, this is not a trivial task. The difficulty increases when speakers refer to multiple elements simultaneously or fail to visually emphasize key content in sync with their narration. In such cases, the lack of coordination between modalities places an additional cognitive burden on the audience. Prior work on cognitive load in multimedia learning shows that increased working memory demand can negatively affect learning outcomes [1]. Furthermore, individual differences in processing multimodal information can intensify this challenge when the presentation format does not align with a learner's preferred style [2].

Addressing this problem requires a deeper understanding of the multimodal alignment. While prior research has explored alignment across modalities such as movies and scripts, speech and transcripts, and books and images [3], [4], these approaches typically operate at a coarse level. In contrast, presentation understanding requires fine-grained alignment between spoken content and specific visual regions within slides. Achieving this alignment is essential for creating a seamless and intuitive learning experience.

To this end, we introduce a curated collection of conference presentations to study the problem in detail. The collection includes video segments with time-aligned speech, transcripts, slide layouts and textual content. We explore methods for aligning the speaker's narration with both textual and visual components of the slides, enabling precise identification of relevant content as it is discussed. We believe that advancing research in this direction will not only improve presentation comprehension but also contribute to broader progress in multimodal document understanding with applications in education, accessibility, and beyond.

1.2 Content Generation and Evaluation

The role of AI in education has expanded rapidly in recent years, moving from generating textual explanations to supporting interactive tutoring and, increasingly, multimodal learning experiences [5], [6], [7], [8]. Systems such as Khan Academy's integration with GPT-4 [9] and tools like Socratic by Google [10] illustrate the growing promise of AI-assisted education. However, these systems remain largely centered around text-based interactions, offering limited support for rich visual or video-based learning.

Simultaneously, research on intelligent tutoring systems (ITS) has made significant progress in delivering personalized instruction and adaptive feedback. However,, much of this work is still confined to structured or text-driven interfaces. In contrast, recent advances in text-to-video (T2V) models [11], [12], [13], [14], [15], [16], [17], [18], [19], [20] have opened up new possibilities for generating dynamic visual content directly from natural language. These models are now capable of producing visually compelling and coherent videos, suggesting a powerful opportunity for their application in educational domain.

Despite this progress, a key question remains largely unanswered: Can such models generate videos that genuinely support learning rather than merely appearing realistic? This question is particularly important in domains such as physics, where visual intuition and dynamic representations play a central role in understanding fundamental concepts. If effective, T2V models could significantly reduce the effort required to create high-quality educational content while making such resources more accessible.

To systematically investigate this potential, we introduce *PhyEduVideo*, the first benchmark designed to evaluate the ability of T2V models to generate pedagogically meaningful videos for physics education. Unlike prior benchmarks, which primarily focus on visual quality or physical plausibility, our approach emphasizes educational effectiveness. Specifically, we ground the evaluation in well-defined

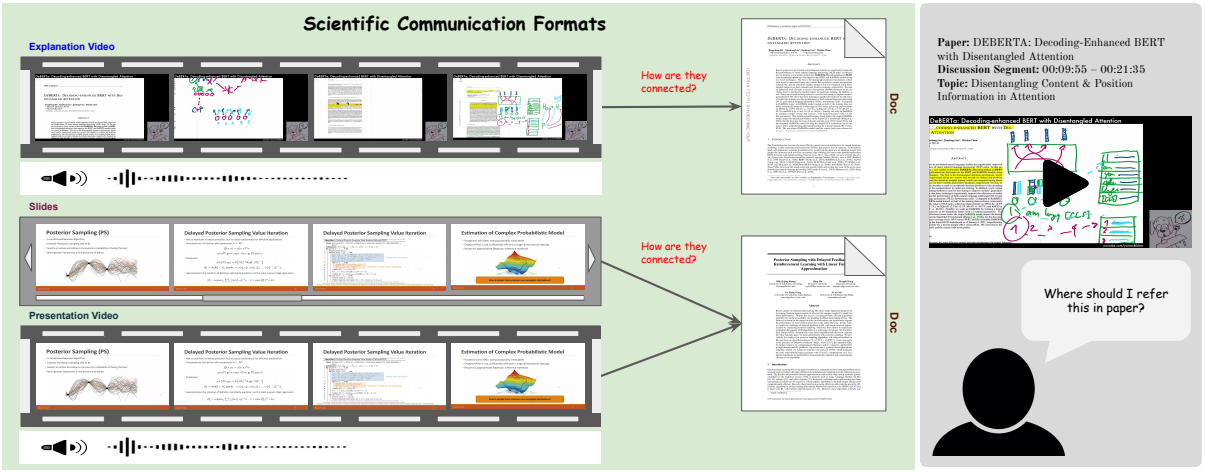


Figure 1.2: **Scientific communication formats and their interconnections:** The figure illustrates how research knowledge is represented and shared across multiple formats, including research papers (Docs), presentation slides, conference videos and explanation videos. These formats are not isolated; rather, strong semantic and structural connections exist among them. Slides often summarize and visualize key insights from papers, presentation videos provide verbal and contextual elaboration, and explanation videos further distill the content for broader understanding. Together, they form a coherent, interconnected network of scientific communication that captures complementary aspects of the same underlying research.

physics concepts and decompose each concept into a set of fine-grained teaching points that ensure comprehensive coverage of the topic.

This structured design enables a more meaningful assessment of whether the generated videos support conceptual understanding rather than simply producing visually plausible outputs. The *PhyEduVideo* benchmark consists of 205 prompts spanning 60 physics concepts, with each concept broken down into 1–5 teaching points aligned with specific instructional goals. The prompts, ranging from 16 to 45 words, are carefully designed to elicit clear and targeted visual explanations.

Our evaluation reveals important trends. Among the models studied, Wan2.1 achieves the strongest overall performance, followed by PhyT2V. Performance varies across domains; areas such as Mechanics, Fluids, and Optics show relatively higher accuracy, while Electromagnetism and Thermodynamics remain more challenging. These findings highlight both the current capabilities of T2V models and the gaps that must be addressed to make them truly effective educational tools.

1.3 Multimodal Correspondence Across Scientific Formats

Scientific communication has evolved far beyond traditional research papers into a rich, multi-format ecosystem. Today, a single research work is often presented through several complementary mediums—formal manuscripts, presentation slides, recorded talks, and explanatory videos. Each format contributes something unique to the overall understanding of the work. Research papers provide tech-

nical depth and rigor, carefully detailing methods, experiments, and results. Slides condense the main ideas into concise and visually engaging summaries, while presentation videos add the researcher's voice, emphasis, and narrative flow. Explanatory videos further improve accessibility by transforming complex concepts into more intuitive and widely understandable explanations. Together, these modalities create a more complete, layered, and engaging representation of scientific knowledge.

Yet, despite conveying the same underlying ideas, these formats largely remain disconnected from one another (Figure 1.2). The relationships between them—such as how a figure in a paper relates to a specific slide [21], or how a spoken explanation corresponds to an equation—are rarely made explicit [21], [22], [23], [24], [25], [26]. As a result, scientific information becomes fragmented across modalities, with valuable insights in one format remaining isolated from others. This fragmentation makes it difficult to smoothly navigate, compare, or fully understand research across its different representations. The impact is broad: students lose opportunities to learn through complementary resources, researchers face challenges in synthesizing information efficiently, and automated systems struggle to organize and interpret scientific knowledge in a coherent way.

Addressing this challenge requires moving beyond treating each modality separately and instead building structured connections across formats. Establishing fine-grained correspondences between modalities can fundamentally change how users interact with scientific content. For example, a concept in a research paper could be directly linked to its visual explanation in a slide, or spoken narration from a presentation could guide users to the most relevant section of a document. Such connections can greatly improve accessibility, comprehension, and exploration of scientific knowledge, while also laying the groundwork for intelligent systems capable of reasoning across multiple forms of communication.

Motivated by this vision, we introduce the *Multimodal Conference Dataset (MCD)*, a unified collection that brings together research papers, slides, presentation videos, and explanatory videos from the same set of works. Using this dataset, we evaluate six models spanning both embedding-based and vision–language approaches across multiple retrieval settings, including explanatory video-to-paper, slide-to-paper, and presentation video-to-paper retrieval. Our evaluation covers diverse paper elements such as paragraphs, figures, equations, and algorithms. Through this analysis, we find that vision–language models demonstrate strong cross-modal robustness but still struggle with precise fine-grained alignment. In contrast, embedding-based models effectively capture text–visual relationships, yet tend to isolate equations and symbolic content into distinct clusters within the embedding space. While the models show promising retrieval capabilities overall, these findings reveal important gaps in achieving a truly unified understanding of scientific content across modalities, highlighting key directions for future research in multimodal scientific reasoning.

1.4 Contributions

In this thesis, we introduce evaluation frameworks for multimodal AI in education, addressing speech–slide alignment, physics-oriented text-to-video generation, and cross-modal knowledge link-

ing, along with a systematic analysis of the strengths and limitations of the current models. Our core contributions are as follows -

- We address the problem of aligning spoken content with slides by proposing a framework that highlights relevant regions of a slide based on the speaker's narration, supported by annotations for fine-grained speech–slide correspondence.
- To evaluate this task, we curate a collection of 14 presentation videos comprising 150 slides, where each slide is paired with its corresponding audio segment and transcript. The slides contain diverse elements, including tables, figures, and dense text.
- To assess the capability of text-to-video (T2V) models in supporting physics intuition, we introduce *PhyEduVideo*, a benchmark designed to evaluate the generation of simple, accurate, and pedagogically meaningful visualizations.
- We propose a structured evaluation framework grounded in pedagogical units (teaching points), enabling fine-grained assessment of how effectively generated videos convey core conceptual ideas.
- We provide empirical insights into the strengths and limitations of current T2V models for instructional video generation, showing that while they produce visually coherent outputs, they often lack physical commonsense and struggle with semantic alignment.
- To study cross-format scientific communication, we introduce the *Multimodal Conference Dataset (MCD)*, which integrates research papers, slides, presentation videos, and explanatory videos from the same works. We analyze how well existing models capture fine-grained correspondences across these modalities, highlighting their key strengths and limitations in achieving robust cross-format understanding.

1.5 Organization of Thesis

This thesis is organized into six chapters, each addressing a distinct aspect of multimodal understanding while collectively contributing to a unified framework for aligning, generating, and evaluating information across modalities.

Chapter 1 introduces the motivation and scope of the thesis. It outlines three primary research directions—linking spoken content to slides, content generation and evaluation for educational videos, and multimodal correspondence across scientific formats—followed by a summary of key contributions and the overall structure of the thesis.

Chapter 2 reviews relevant prior work, covering human interaction with documents, understanding of slide-based content, advances in text-to-video models, evaluation benchmarks for generative systems, cross-modal retrieval, and the broader role of AI in scientific research.

Chapter 3 presents the first contribution, focusing on aligning spoken narration with visual regions in presentation slides. It describes the problem formulation, dataset construction, preprocessing, alignment techniques, evaluation metrics, and both qualitative and quantitative analyses, along with insights from a user study.

Chapter 4 introduces *PhyEduVideo*, a benchmark for evaluating text-to-video models in the context of physics education. This chapter details the benchmark construction, statistical analysis, proposed evaluation metrics, human evaluation studies, experimental results, and a discussion of the limitations and future directions of current T2V systems.

Chapter 5 presents the *Multimodal Conference Dataset (MCD)*, addressing fine-grained correspondence across research papers, slides, and videos. It includes dataset construction, preprocessing, annotation strategies, problem formulation, evaluation setup, and an in-depth performance analysis across different model families and traversal settings.

Chapter 6 concludes the thesis by summarizing the key findings across all contributions and outlining potential directions for future research in multimodal alignment, generation, and scientific understanding.

Chapter 2

Prior Work

Prior work lies at the intersection of multimodal understanding and AI-driven educational technologies, with a central focus on how information is created, interpreted, and utilized across diverse formats. This research spans a broad spectrum of problems, beginning with an investigation into how humans interact with complex documents in environments where text, visuals, and structural elements must be processed together. Such interactions are rarely linear, requiring users to continuously shift attention between modalities to construct meaning. This observation naturally motivates the study of structured visual artifacts such as slide-based documents, where effective navigation and interpretation play critical roles in comprehension.

The rapid evolution of generative models, including GAN and diffusion-based approaches, has significantly expanded the scope of multimedia content creation. These models have enabled the synthesis of increasingly realistic and semantically aligned outputs across modalities. Building on these advancements, recent efforts have focused on exploring the capabilities of modern generative systems, particularly text-to-video (T2V) models, for producing rich and dynamic educational media. Importantly, the emphasis has not been limited to generation alone but also extends to the systematic evaluation of such models across multiple dimensions, including fidelity, alignment, and pedagogical relevance.

In parallel, research has explored the broader role of AI in supporting the research ecosystem, particularly in transforming and navigating scientific content. Applications in this field include generating posters from research papers, producing explanatory videos from textual descriptions, and identifying relevant segments within large documents. These tools aim to reduce the cognitive and temporal burdens associated with content creation and consumption. Collectively, this body of work reflects a sustained effort toward designing intelligent systems that can effectively understand, generate, and organize multimodal information in meaningful and impactful ways.

2.1 Human Document Interaction

Human interaction with documents has evolved significantly, extending far beyond traditional modes of reading static text. In contemporary settings, individuals increasingly "consume" information through multimodal formats such as educational videos, slide presentations, and interactive media [27]. These

modern documents inherently integrate multiple modalities, including speech, text, graphics, and annotations, requiring users to process and synthesize heterogeneous information streams simultaneously. From a document understanding perspective, this necessitates the development of systems capable of jointly reasoning across modalities rather than treating them in isolation from one another.

To facilitate such processing, Optical Character Recognition (OCR) techniques are employed to extract textual content from visual layouts, while Automatic Speech Recognition (ASR) systems convert spoken language into time-aligned transcripts [28], [29]. Recent advancements, particularly in models such as Whisper, have significantly improved multilingual speech recognition and robustness across diverse acoustic conditions. However, effectively integrating visual and auditory signals remains challenging because these modalities often exhibit temporal and semantic misalignment. Consequently, research on audio–visual learning and multimodal fusion has gained increasing attention, focusing on aligning and integrating information across different streams [30], [31], [32], [33].

Despite these advancements, real-world scenarios such as technical presentations introduce additional complexities that are not fully addressed by current systems. Domain-specific terminology, variations in speaker accents, and environmental noise can adversely affect ASR performance, while visually dense and heterogeneous slide layouts pose challenges for OCR systems. To mitigate these limitations, recent approaches have explored domain adaptation strategies and post-processing techniques that incorporate lexicons and leverage large language models for contextual correction [34], [35], [36]. These challenges underscore the need for more robust multimodal understanding frameworks that can seamlessly integrate diverse information sources while maintaining semantic coherence.

2.2 Understanding Slide Documents

Slide-based presentations play a pivotal role in education, communication, and knowledge dissemination, serving as complementary medium to spoken explanations. Therefore, understanding the structure and content of slides is a critical task in document analysis. Early efforts in this domain have been shaped by benchmark competitions such as the Page Segmentation Competition [37] and the ICDAR 2017 RDCL challenge [38], which established foundational problems in layout analysis and document understanding. Recently, the ICDAR 2023 competition on structured text extraction [39] emphasized the challenges associated with extracting meaningful textual information from complex layouts, particularly under zero and few-shot conditions.

A central aspect of slide understanding is layout analysis, which involves identifying and segmenting distinct regions, such as text blocks, images, tables, and charts, while preserving their semantic relationships. Deep learning-based approaches have significantly advanced this area by modeling layout as a structured prediction problem. For instance, Yang and Hsu (2020) [40] proposed SLLD, a convolutional neural network-based approach that formulates slide segmentation as an object detection task, achieving improved performance over traditional methods. Building upon this, Yang and Hsu (2022) [41] intro-

duced TRDLU, a transformer-based framework that leverages encoder–decoder architectures to achieve more accurate classification and representation of slide components.

Beyond static digital slides, real-world applications have extended these techniques to dynamic environments. Systems such as WiSe [42] enable segmentation and analysis from photographs of slides captured in live settings such as classrooms and conferences. Additionally, information retrieval techniques have been applied to enhance accessibility, with OCR-based slide retrieval methods achieving performance comparable to traditional document search systems [43]. Video-based frameworks further enrich this domain by dynamically analyzing the slide content over time. The SlideVQA dataset [44], which includes over 2,600 slide decks and 52,000 images, supports complex reasoning tasks that combine visual understanding with question answering. Although recent document VQA models that integrate evidence selection and reasoning have surpassed earlier approaches, they still fall short of human-level comprehension. Therefore, continued advancements in multimodal learning and dataset development are essential to bridge this gap and achieve more robust slide understanding systems. Here is a refined version with improved flow, stronger connections, and expanded content while preserving all citations and structure:

2.3 Text-to-Video Models

The emergence of text-to-video (T2V) models marks a significant milestone in multimodal generative modeling, enabling the synthesis of dynamic and temporally coherent visual content directly from textual description. This paradigm extends beyond static image generation by incorporating motion, temporal dependencies, and narrative consistency, thereby opening new possibilities for content creation. Early approaches in this domain, such as MoCoGAN [45] and TGAN [46], introduced foundational techniques for disentangling motion and content and modeling spatiotemporal dynamics. However, these methods were constrained by limitations in scalability, motion realism, and alignment with complex textual prompts [45], [46], which restricted their applicability in real-world scenarios.

These challenges motivated a transition toward diffusion-based approaches [47], [48], which generate videos through iterative denoising processes and have demonstrated substantial improvements in visual fidelity and temporal consistency. Prominent diffusion-based systems, including ModelScope [48], VideoCrafter [11], [12], CogVideo [13], AnimateDiff [49], and Text2Video-Zero [50], have significantly advanced the field. Furthermore, large-scale models such as Imagen Video [47] and Make-A-Video [51] have pushed the boundaries of resolution, realism, and semantic alignment, demonstrating the benefits of scaling in generative modeling. Despite these advancements, many diffusion-based architectures rely on convolutional UNet backbones [48], which are limited in capturing long-range temporal dependencies and global contextual relationships across frames.

To overcome these limitations, recent research has increasingly shifted toward Diffusion Transformers (DiTs), which leverage self-attention mechanisms to model global spatial–temporal interactions more effectively [14], [18]. By capturing dependencies across both space and time, these architec-

tures enable improved coherence and consistency in generated videos. State-of-the-art systems such as Sora [18], CogVideoX [14], Hunyuan [15], Wan2.1 [16], Pika [19], Lumiere [20], and Kling [17] exemplify this transition, reflecting a broader trend toward enhanced scalability, controllability, and semantic alignment in T2V modeling. This evolution—from GAN-based methods to diffusion models and subsequently to transformer-based architectures [18], [45], [48]—highlights a continuous effort to address the core challenges of video generation.

In addition to architectural innovations, complementary research directions have focused on improving the realism, controllability, and applicability of T2V systems. Multimodal pretraining strategies and retrieval-augmented generation techniques have enhanced contextual grounding and text–video alignment [12], [48]. Moreover, physics-aware conditioning has emerged as a promising direction for ensuring that generated videos adhere to physical laws and exhibit plausible interactions. For instance, PhyT2V [52] integrates large language model guidance with simulation-based priors to more accurately capture physical dynamics. Such approaches are particularly valuable in educational contexts, where accurate and interpretable representations of scientific phenomena are essential. By enabling intuitive visualization of abstract concepts, T2V systems hold the potential to significantly transform educational content creation and delivery. Motivated by these opportunities, recent work has emphasized the importance of systematically evaluating T2V models in instructional settings to better understand their capabilities, limitations, and suitability for real-world educational applications.

2.4 Evaluation Benchmarks for Text-to-Video Models

As text-to-video models continue to advance in terms of visual quality, temporal coherence, and semantic alignment, their evaluation has largely relied on general-purpose benchmarking frameworks. VBench [53] introduced a structured and hierarchical evaluation protocol spanning key dimensions such as prompt adherence, spatial consistency, and temporal coherence. This framework was further extended in VBench++ [54], which improved coverage and flexibility across evaluation scenarios, and later in VBench 2.0 [55], which incorporated more challenging aspects such as commonsense reasoning, physical plausibility, and aesthetic quality. Alongside these developments, complementary benchmarks have explored specific dimensions of generation quality, including compositional generalization (e.g., T2VCompBench), motion understanding, and temporal dynamics, as in DEVIL [56]. Collectively, these efforts have established a strong foundation for systematic and large-scale evaluation of text-to-video systems.

In parallel, a number of benchmarks have focused more explicitly on physical reasoning and adherence to physical laws. VideoPhy [57] introduced structured evaluation metrics such as Semantic Adherence (SA) and Physical Commonsense (PC) to assess whether generated videos align with basic physical intuition. This line of work was extended in VideoPhy2 [58], which further introduced a Physical Rules (PR) dimension to capture more fine-grained violations of physical consistency. Similarly, Physics-IQ [59] emphasized real-world physical reasoning through naturally occurring video scenarios,

while PhyGenBench [60] broadened evaluation coverage across multiple domains including mechanics, thermodynamics, and optics, leveraging simulation-based probes and large language model-assisted evaluation. Together, these benchmarks represent significant progress toward assessing physical realism in generated videos, although their primary focus remains on perceptual and physical plausibility rather than instructional effectiveness.

Despite these advancements, most existing benchmarks prioritize whether generated videos are realistic or physically plausible, rather than whether they are pedagogically meaningful. In other words, they primarily evaluate *how real the video appears* rather than *how well it communicates or teaches a concept*. To bridge this gap, we introduce the first benchmark specifically designed for physics education. In our formulation, each concept is decomposed into fine-grained teaching objectives, enabling a more structured assessment of whether generated videos faithfully convey core ideas in a clear and coherent manner. This shift reframes evaluation from surface-level realism toward instructional utility, providing a complementary perspective to existing benchmarks and moving T2V evaluation closer to real-world educational applications.

2.5 Cross-Modal Retrieval

The challenges associated with multimodal understanding and generation naturally extend to retrieval tasks, where the objective is to identify semantically aligned information across different modalities. Cross-modal retrieval enables users to query in one modality, such as text, and retrieve relevant content in another, such as images, audio, or video. This capability is particularly important in modern information systems, where data is inherently multimodal and interconnected. Cross-modal retrieval has been widely studied in general domains, including image–text retrieval [61], [62], [63], video–text alignment [64], [65], and vision–language understanding [66], [67]. However, many of these approaches primarily focus on broad visual–linguistic relationships and may not adequately capture the structured, semantically dense, and domain-specific nature of educational and scientific content.

Recent methods such as E5-V [68] and GME [69] have leveraged multimodal large language models to learn unified embedding spaces that support scalable retrieval across modalities. E5-V adapts multimodal large language models to generate universal embeddings using prompt-based representations, enabling flexible retrieval across text, image, audio, and video modalities [68]. Similarly, GME fine-tunes multimodal retrievers on large fused-modal datasets to support single-modal, cross-modal, and fused-modal retrieval within a unified framework [69]. By aligning representations from different modalities into a common semantic space, these approaches facilitate seamless retrieval and interaction across heterogeneous data sources.

Despite these advances, educational and research content presents unique challenges that extend beyond general-purpose retrieval tasks. Such content is often highly structured and semantically rich, combining precise textual explanations, abstract figures, equations, slides, and spoken presentations. Effective retrieval in this setting requires accurate alignment across modalities while preserving contextual

and pedagogical relationships. Earlier work explored pairwise alignment strategies, such as linking presentation slides with research papers [22], [23], later improving performance through the incorporation of visual cues and sequential modeling [22]. Subsequent research expanded toward aligning presentation videos with slides [24], [25], [26], integrating speech transcripts, visual frames, and temporal synchronization mechanisms [70], [71]. Datasets such as the Google I/O collection [71] further enabled research into multimodal synchronization between spoken explanations and presentation materials. In parallel, fine-grained figure–text associations were also explored to better connect visual elements with corresponding textual descriptions [21].

While prior work highlights the importance of multimodal alignment in educational settings, most approaches remain limited to pairwise correspondence between modalities. In practice, educational content spans interconnected representations such as papers, slides, videos, figures, and transcripts, all contributing to a unified understanding of the material. Addressing this requires integrated cross-modal retrieval frameworks capable of modeling relationships across multiple modalities simultaneously. Recent efforts toward unified benchmarks and multimodal alignment systems aim to support seamless navigation across these representations, improving accessibility and enabling deeper multimodal understanding.

2.6 AI for Research

Advances in multimodal understanding, generation, and retrieval are collectively driving a broader transformation in the research ecosystem, where artificial intelligence is becoming an integral component of the knowledge creation and dissemination process. Traditionally, the production of research artifacts—including papers, presentations, and visual summaries—has been a manual and time-intensive endeavor, requiring significant effort to translate ideas across formats. However, recent developments in AI have enabled the automation and augmentation of these processes, thereby improving efficiency, scalability, and accessibility [72], [73], [74], [75], [76]. AI-driven systems now assist in summarizing complex papers, generating multimodal representations such as slides, videos, and posters, and supporting interaction across textual, visual, and spoken modalities. This convergence of AI and scholarly communication is redefining how scientific knowledge is created, communicated, and consumed.

Emerging systems now support the automatic generation of slides [72], [77], [78], [79], videos [74], [76], and posters [73] directly from research papers, effectively reducing the effort required to communicate ideas across different modalities. These systems leverage advances in natural language processing, computer vision, and generative modeling to produce coherent and contextually relevant outputs. Slide generation models [72], [77], [78] identify key sections, figures, and equations to create structured presentation decks, while video generation systems [74], [76] synthesize narrated visual explanations from scientific content. Poster generation frameworks [73] further enable concise and visually engaging summaries of research contributions. Beyond static outputs, systems such as Paper2Agent [75] transform research papers into interactive AI agents capable of explaining methods, reasoning about findings, and

engaging in dialogue with users, thereby enabling more dynamic and intuitive interaction with scientific content.

More broadly, these developments reflect a unifying trend in which AI is bridging gaps between modalities and redefining how knowledge is created, communicated, and consumed. By integrating multimodal generation, evaluation, retrieval, interaction, and alignment, current research is moving toward a more interconnected and intelligent ecosystem for education and scientific discovery. This vision emphasizes not only efficiency and scalability but also interpretability, accessibility, and contextual understanding, ultimately contributing to more effective knowledge sharing and learning at scale.

Collectively, our research aims to systematically evaluate the capabilities of current multimodal models in educational and research-oriented settings, while identifying critical gaps in reasoning, reliability, evaluation, and real-world usability. Through studies involving documents, slides, videos, retrieval systems, and research-assistance frameworks, we seek to advance the development of more robust, interpretable, and practically useful AI systems that can effectively support knowledge-intensive and human-centered applications.

Chapter 3

Highlighting relevant content on slides

In this chapter, we discuss the problem of highlighting relevant content on presentation slides based on the corresponding audio. The objective is to establish meaningful alignment between spoken narration and visual elements, enabling precise identification of important regions in the slides that correspond to segments of the accompanying speech.

3.1 Problem Formulation

Given a presentation slide and its accompanying speech, the goal is to establish fine-grained correspondence between the spoken content and specific regions within the slide. This requires grounding segments of speech in visually relevant areas so that the audience can easily follow the alignment between narration and visual elements. As shown in Figure 3.1, the proposed alignment module leverages outputs from OCR to extract textual content from the slide and ASR to transcribe the speech signal. These modalities are jointly processed to identify semantically consistent mappings between spoken phrases and the slide regions. Once the relevant regions are detected, they are forwarded to a highlighting module that emphasizes the corresponding areas on the slides. This module supports multiple visualization strategies, allowing the system to present aligned content in an intuitive and interpretable manner, thereby improving comprehension and reducing the cognitive effort required to track multimodal information during the presentation.

3.2 Dataset

To evaluate the proposed framework, we construct a multimodal dataset consisting of presentation slides, corresponding audio, and transcribed speech. The dataset is designed to capture real-world variability in both visual and textual content, enabling robust analysis of speech–slide alignment. It provides a diverse set of examples, including structured layouts, equations, figures, and varying text densities, making it suitable for studying fine-grained multimodal correspondence and grounding tasks.

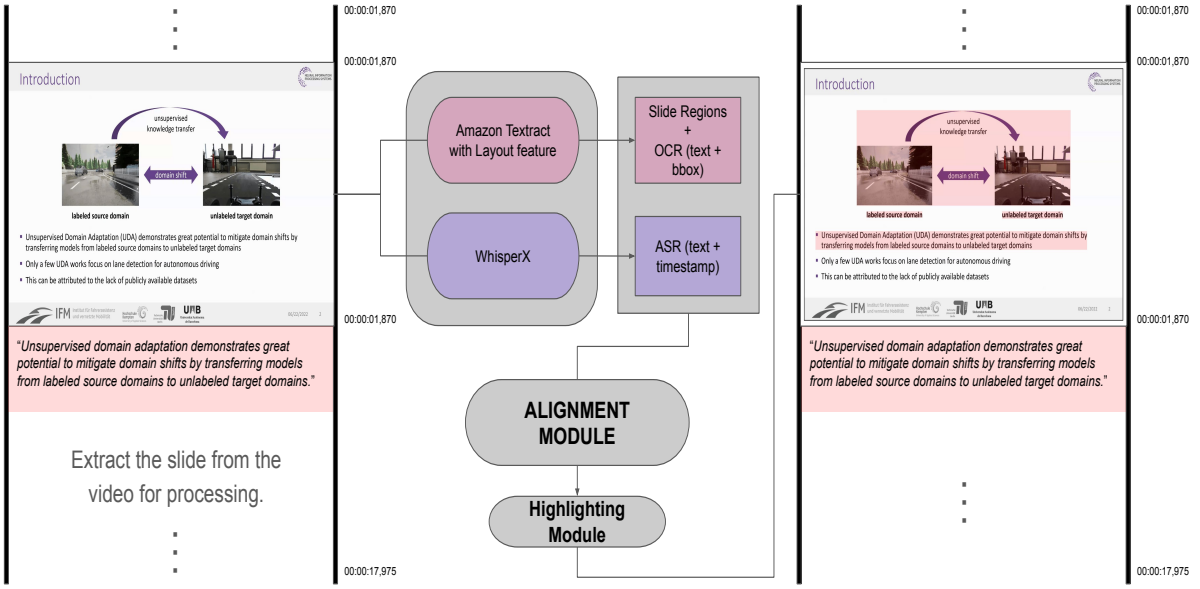


Figure 3.1: This figure illustrates the pipeline of our method. The Alignment Module takes OCR-extracted text and ASR-generated transcripts as inputs to establish a correspondence between spoken and visual content. These aligned results are then passed to the Highlighting Module, from which a suitable visualization can be chosen to highlight the relevant slide content.

3.2.1 Data Collection

We construct a dataset using 14 presentation videos from NeurIPS 2022 and ICML 2023, sourced from the SlidesLive platform. Each video is processed to extract individual slides, which are temporally aligned with their corresponding audio segments and speech transcripts. The dataset comprised 150 slides that exhibit substantial diversity in layout and content, including figures, equations, tables, and text rendered in varying fonts, sizes, and spatial arrangements. This diversity reflects realistic presentation scenarios and ensures a challenging benchmark for multimodal alignment. Representative examples from the dataset are illustrated in Figure 3.2, which demonstrates the range of visual and textual structures captured.

3.2.2 Dataset Statistics

The dataset spans over one hour of presentation content in total. Analysis of the temporal distribution shows that most slides are associated with audio segments ranging from 10 to 20 seconds, as depicted in Figure 3.4. The textual content extracted from slides includes 7,466 unique alphabetical words, while the corresponding speech transcripts contain 6,708 unique alphabetical words, indicating a rich yet partially overlapping vocabulary across modalities. Furthermore, Figure 3.3 presents detailed statistics, including the minimum, maximum, average, and median values for audio segment durations, as well as

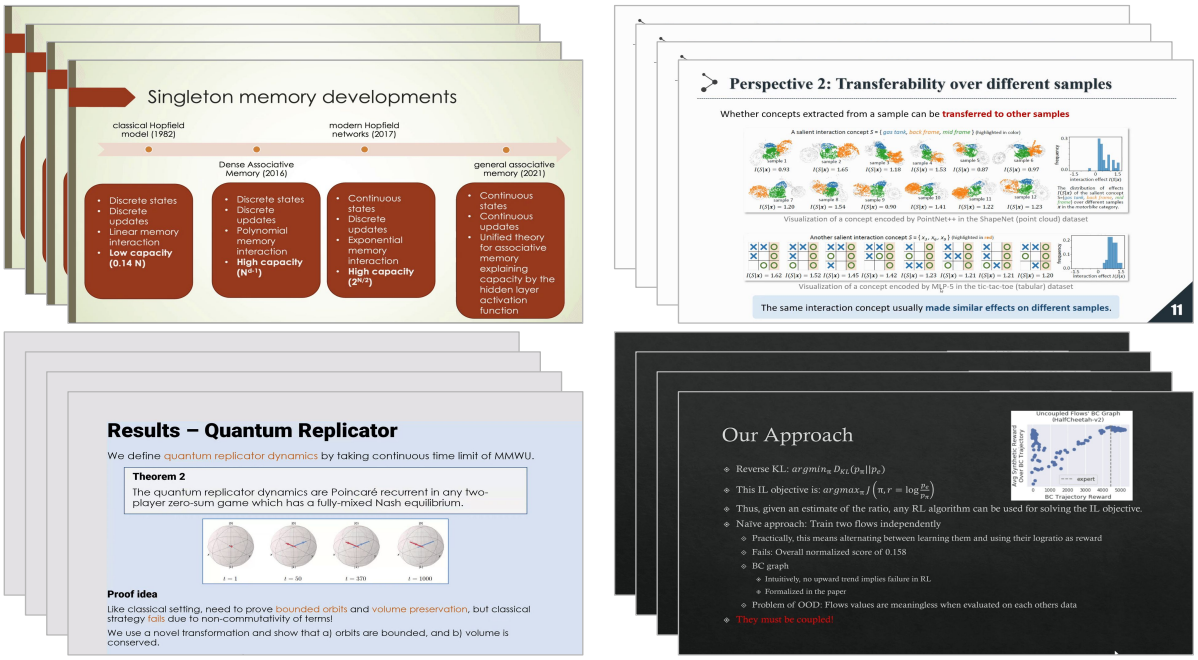


Figure 3.2: This figure showcases a set of presentation slides from our dataset. The dataset consists of slides with varying layouts, color schemes, and content structures, reflecting the diversity of real-world academic and professional presentations.

word counts derived from OCR and ASR. Overall, the dataset provides a comprehensive and realistic benchmark for evaluating multimodal alignment(within speech and text) and grounding techniques.

3.2.3 Data Preprocessing and Quality Assessment

Presentation slides often contain structured layouts, including tables, multi-column texts, and dense visual elements, making accurate segmentation essential for aligning visual regions with the corre-

	Min	Max	Avg	Median
Duration (in sec)	2	56	21.10	18
Word Count in OCR	1	799	99.93	83
Word Count in ASR	5	307	72.71	59

Figure 3.3: Statistics for Audio Segment Durations, Word Count in ASR and OCR

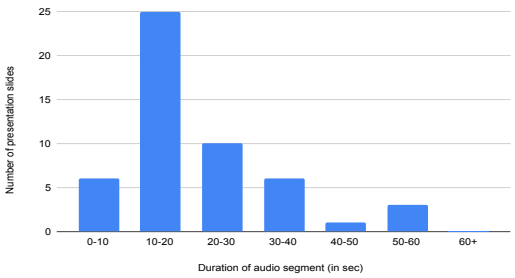


Figure 3.4: The distribution of presentation slides based on the duration of their corresponding audio segments.

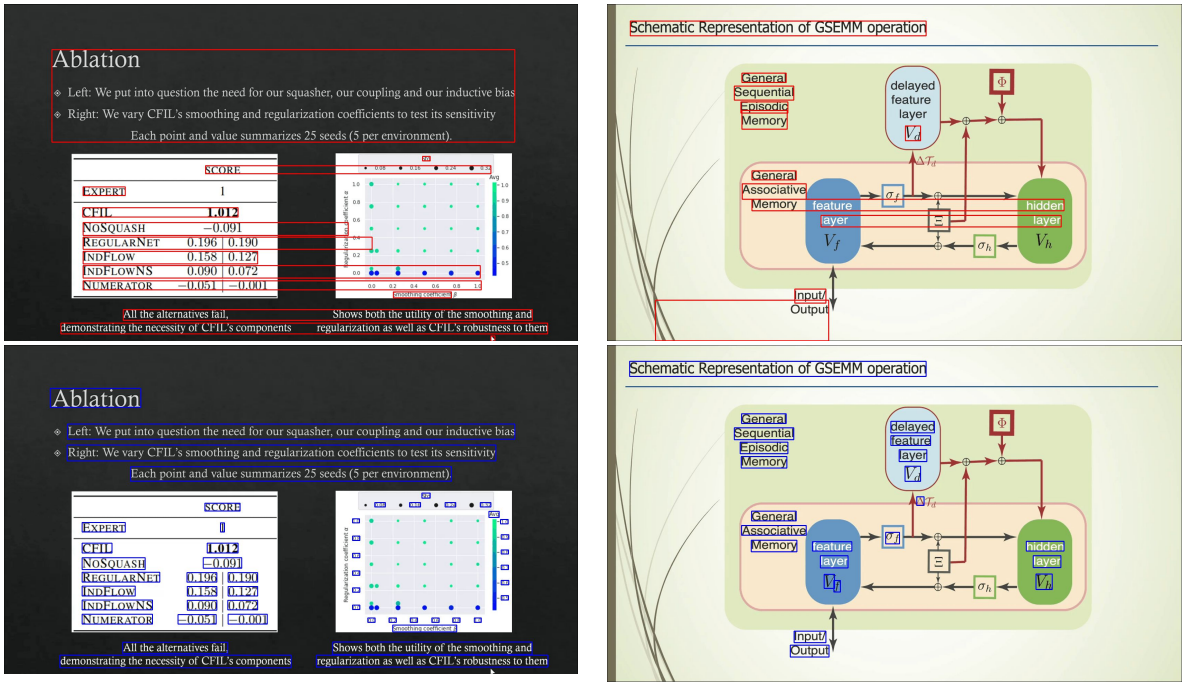


Figure 3.5: Comparison of OCR results from Tesseract (red) and AWS Textract (blue) on slides with complex layouts.

sponding spoken content. To address this, we employ Amazon Textract[1], which provides robust optical character recognition (OCR) along with layout analysis. This allows us to accurately extract text while preserving structural information and grouping content into meaningful regions, even in the presence of complex formatting. In contrast, traditional OCR tools such as Tesseract, often struggle with multi-line text, diverse font styles, and embedded elements, such as figures and equations (Figure 3.5), limiting their effectiveness for structured slide content.

For the speech modality, we utilize WhisperX [29], a state-of-the-art automatic speech recognition (ASR) system that achieves strong performance with low error rates (CER = 0.0129, WER = 0.0297; see Table 3.1). Despite its effectiveness, WhisperX may still produce errors when handling technical terminology, speaker names, and accent variations, as illustrated in Figure 3.13. Together, these OCR and ASR components provide reliable inputs for downstream alignment between spoken content and slide regions.

3.3 Speech–Text Alignment

Our approach aligns slide content with spoken narration by matching OCR-extracted text from slides with ASR-generated transcripts, as shown in Fig. 3.6. This alignment helps identify the specific regions

[1] https://aws.amazon.com/textract/

	CER	WER	Edit Distance	Precision	Recall	F1-score
WhisperX	0.0129	0.0297	1.0327	0.9745	0.9718	0.9731
Corrected	0.0366	0.0832	3.3224	0.9287	0.9266	0.9271

Table 3.1: Comparison of ASR performance before and after post-correction using OCR. The metrics include Character Error Rate (CER), Word Error Rate (WER), Edit Distance, Precision, Recall, and F1-SCORE. WhisperX represents the original ASR output, while "Corrected" refers to the ASR output refined using OCR-based post-correction.

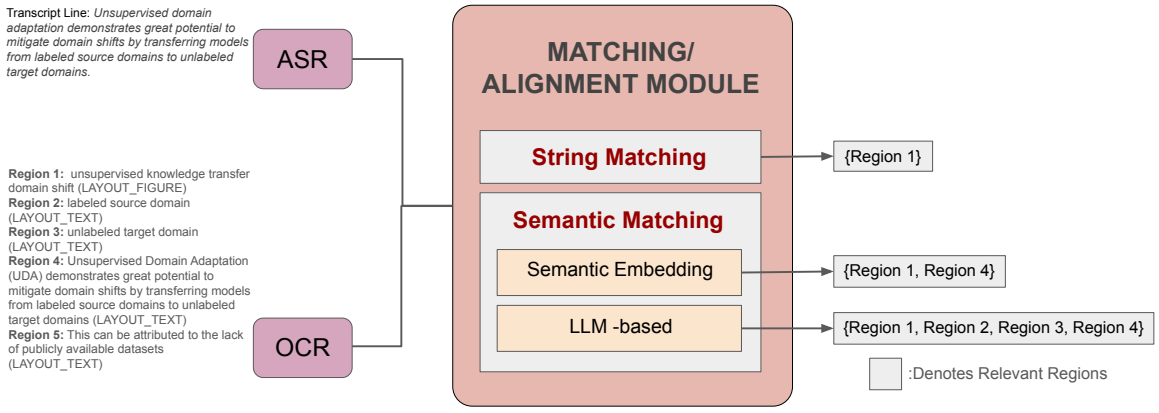


Figure 3.6: The figure illustrates the different categories in the matching module and how they function for an given example.

of a slide that correspond to what the speaker is discussing at any given moment. We explore two complementary matching strategies. The first is string matching, which focuses on detecting exact or near-exact overlaps between textual elements in transcripts and slide content. The second is semantic matching, which goes beyond surface-level similarity and captures contextual relationships between spoken and visual information.

3.3.1 Semantic Alignment

Semantic Embeddings: To enable a robust alignment between the spoken content and slide regions, we adopt a semantic similarity-based approach that goes beyond surface-level text matching. In this framework, transcript lines are compared with text extracted from slide regions using OCR via their semantic embeddings. Unlike fuzzy matching techniques, which depend on lexical overlap, semantic embeddings capture the underlying meaning of the text, making the alignment process more resilient to paraphrasing, synonym usage, and lexical variations. This is particularly important in presentation scenarios where speakers often rephrase or summarize slide content.

To generate these embeddings, we leverage a range of pre-trained models that are effective in capturing semantic relationships across different domains. These include:
MINILM [80]: A compact and efficient model that produces high-quality sentence embeddings while

maintaining low computational overhead, making it suitable for scalable semantic matching.
SciBERT [81]: A variant of BERT pretrained on scientific corpora, enabling it to better represent technical terminology and domain-specific language commonly found in academic presentations.
SPECTER [82]: A model trained on scholarly documents that generates embeddings capturing both semantic meaning and relational structure, making it particularly effective for scientific and research-oriented content.

LLM: In addition to embedding-based similarity, we incorporate large language models (LLMs) to further refine and validate the alignment process. Models such as Flan-T5 and Qwen-Instruct 2.5 are used to introduce reasoning capabilities and improve alignment decisions beyond purely vector-based comparisons.

Flan-T5 [83], an instruction-tuned version of the Text-to-Text Transfer Transformer, is particularly effective for binary relevance classification tasks. In our approach, each OCR region is evaluated against a given transcript line, and the model is prompted to determine whether the region is relevant through a yes/no response. This step helps filter out noisy or less relevant regions and improves the alignment precision.

Qwen-Instruct 2.5 [84] is a more powerful instruction-following model that operates in a global-selection setting. Given a transcript line along with all OCR-extracted regions from a slide, the model identifies and selects the most relevant regions that correspond to the spoken content. This allows for a more holistic and context-aware alignment, leveraging the reasoning capabilities of the model to capture complex relationships between speech and visual elements.

Together, these embedding and LLM-based strategies complement each other, enabling a more accurate and robust semantic alignment between the transcript content and slide regions.

3.3.2 Region-of-Interest Highlighting

With the correspondence between speech and slide regions established, the next step is to effectively visualize these aligned regions in synchrony with the spoken narrative. A range of visualization strategies can be employed for this purpose, including drawing bounding boxes around relevant regions, applying color-based shading, magnifying areas of interest, and selectively removing non-essential content from the slide. Each of these approaches offers a different trade-off between emphasis and the preservation of context. For instance, while removing background content can strongly focus attention on specific regions, it may also limit the user's ability to perceive the broader context of a slide. Conversely, preserving the full slide while highlighting the key regions allows users to follow the narration while exploring other parts of the slide as needed. These design considerations are illustrated in Figure 3.7, which demonstrates the different visualization strategies.

3.3.2.1 User Study Design and Analysis

To evaluate the effectiveness of the proposed visualization strategies, we conducted a user study involving 33 participants, including graduate students (MS, PhD, and those pursuing MS/PhD degrees).

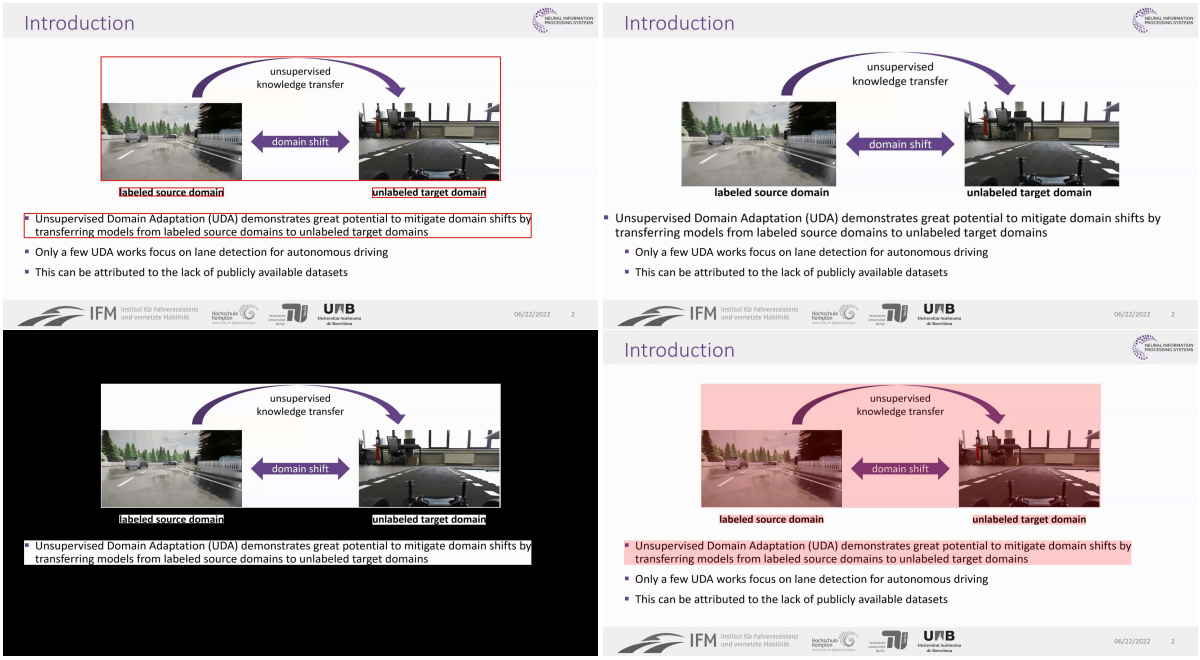


Figure 3.7: Illustration of different techniques for highlighting relevant regions in presentation slides, including transparent bounding boxes, color overlays, content removal, and magnification of key areas.

The participants were shown six video clips, each with an average duration of 35.03 seconds, selected from the middle portions of presentations to simulate a realistic scenario where users join a presentation midway. This setting allows us to assess how well the highlighting methods help users quickly catch up with the ongoing content.

After each clip, participants were asked to rate their experience using the following options: A) The highlighting was useful, and I preferred having it. B) I did not mind whether the highlighting was present. C) I did not prefer the highlighting and would rather watch the video without it.

At the end of the study, the participants were asked to identify their most preferred highlighting technique and provide qualitative feedback on their choices. The results, summarized in Figure 3.9, indicate that bounding box-based highlighting was the most preferred method. Participants found it effective for emphasizing important content while preserving full slide visibility and context. Shading techniques have also attracted attention, but have been reported to reduce readability due to dimming of the surrounding content. Magnification was the least preferred approach, as it often obscured surrounding context despite being useful for small or dense text. The hiding strategy received mixed feedback; while some users appreciated its ability to reduce distractions, many preferred retaining access to the full slide for improved comprehension. As shown in Table 3.8, approximately 50% of participants favored highlighting overall, particularly in scenarios where they joined presentations midway. These findings suggest that effective visualization should balance emphasis and contextual visibility to support better user understanding.

Category	A	B	C
Magnifying/emphasizing	14.70	61.76	22.05
Bounding box	83.82	8.82	4.41
Shading	57.35	10.29	32.35
Hiding the background	38.23	14.07	47.05
Total	49.26	24.26	26.47

Figure 3.8: User preference rating across different highlighting methods (all values in percentage).

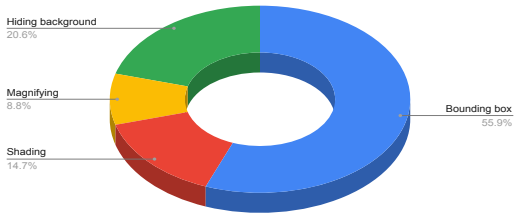


Figure 3.9: Distribution of user preferences across four different highlighting methods.

3.4 Metric

Correctness Score: The Correctness Score (S_c) measures the accuracy of the predicted slide regions concerning the corresponding transcript content. A higher score indicates that the majority of the identified regions correctly align with the relevant transcript segments, whereas a lower score reflects alignment errors, such as incorrectly selected regions or an excessive number of predictions for a given transcript line. In cases where no regions are predicted for a transcript line, the score is set to 1 by default. The Incorrectness Score, defined as $1 - S_c$, captures the proportion of misaligned or incorrect predictions.

Missing Score: The Missing Score (S_m) quantifies the extent to which relevant slide regions are not captured by the predictions. A higher value indicates that a significant number of relevant regions have been missed, resulting in incomplete alignment. It is computed as the ratio of missed alignments to the total number of expected alignments for a given transcript line. If no alignments are required for the transcript line, the score is assigned a value of 0.

F1-Score: The F1-Score provides a balanced evaluation by combining precision and recall. Recall is defined as $1 - S_m$, which reflects the proportion of relevant regions successfully retrieved. Precision measures the fraction of correctly predicted regions among all predicted regions. It is important to note that precision differs from the Correctness Score, as it directly evaluates the quality of the predicted set. In scenarios where no predictions are made for a transcript line, the precision is defined as 0 to reflect the absence of the retrieved regions.

3.5 Results and Analysis

3.5.1 Qualitative Analysis

Semantic alignment proves effective in establishing correspondences between spoken content and visual slide regions. As illustrated in Figure 3.10, T5 is able to accurately associate transcript lines with their corresponding regions on the slide. To further analyze model behavior, Figure 3.11 presents four

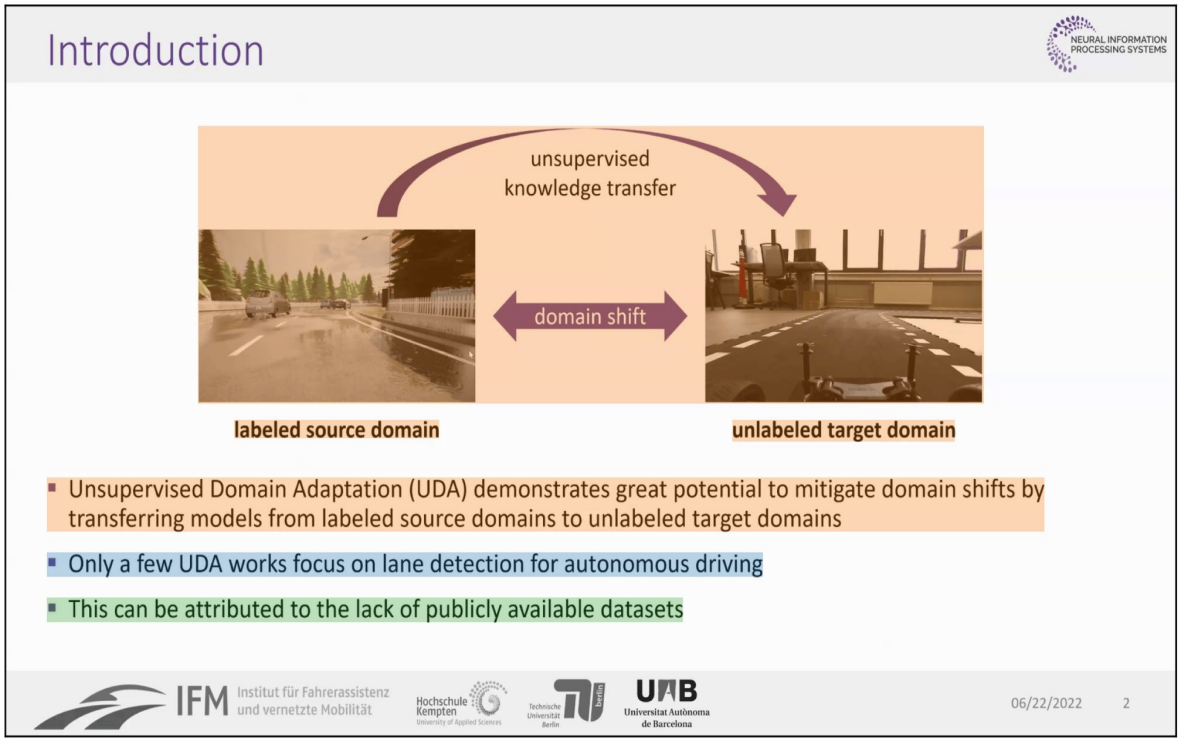
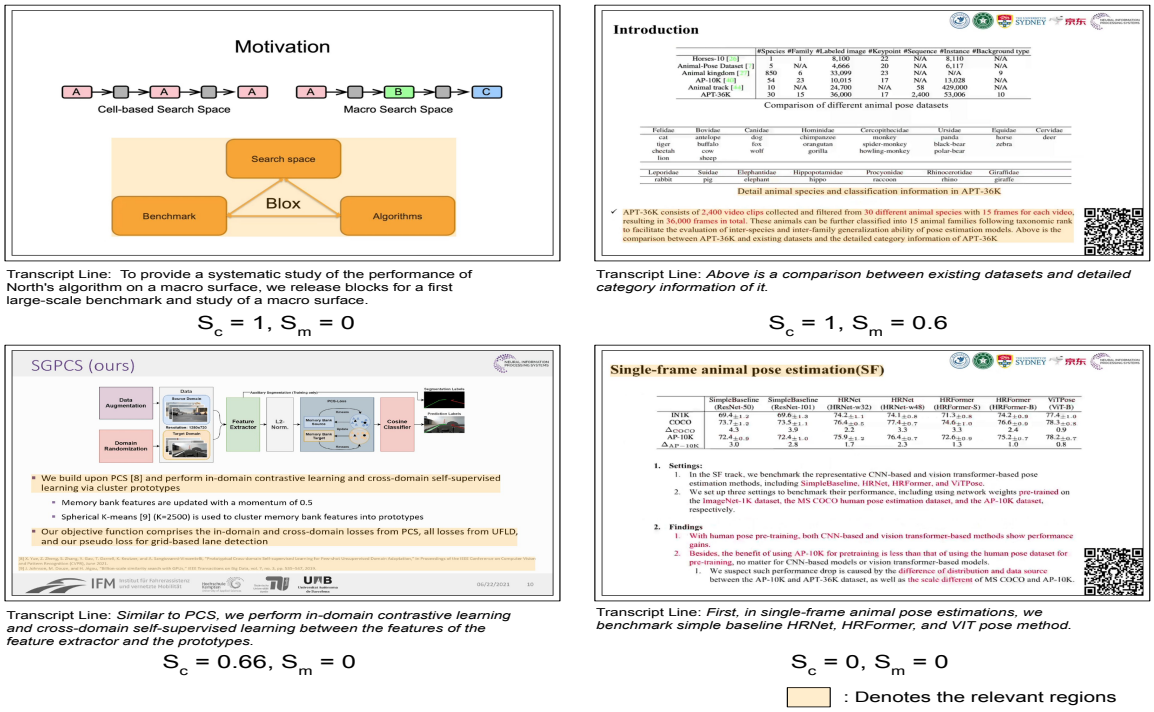


Figure 3.10: The figure illustrates the method's output, with each transcript line and its corresponding slide region highlighted in the same color.

representative cases for Flan-T5: (1) perfect alignment with $S_c = 1$, $S_m = 0$; (2) perfect correctness with additional mismatches ($S_c = 1$, $S_m = 0.6$); (3) perfect missing score with partial correctness ($S_c = 0.66$, $S_m = 0$); and (4) a failure case where predictions are made despite no true match ($S_c = 1$, $S_m = 0$).

Among the evaluated methods, S-BERT T-3 achieves the best overall F1 score. This can be attributed to the fact that while LLM-based models such as T5 and Qwen are capable of capturing complex semantic relationships, they may occasionally misinterpret generic or ambiguous phrases, leading to incorrect region predictions. For instance, Qwen tends to associate phrases like "That's all" or "Thank you for your attention" with conclusion sections, whereas T5 aligns them with the "Summary" label, as shown in Figure 3.12. In contrast, simpler baselines such as the Fuzzy method, rely on near-exact string matching and therefore generate fewer predictions. This conservative behavior can result in inflated correctness scores (e.g., $S_c = 1$) when no matches are predicted, even when relevant alignments are missed. The Missing Score component helps mitigate this issue by accounting for such omissions, thereby enabling a more balanced evaluation of the alignment quality.



Transcript Line: To provide a systematic study of the performance of North's algorithm on a macro surface, we release blocks for a first large-scale benchmark and study of a macro surface.

$S_c = 1, S_m = 0$

Transcript Line: *Above is a comparison between existing datasets and detailed category information of it.*

$S_c = 1, S_m = 0.6$

Transcript Line: *Similar to PCS, we perform in-domain contrastive learning and cross-domain self-supervised learning between the features of the feature extractor and the prototypes.*

$S_c = 0.66, S_m = 0$

Transcript Line: *First, in single-frame animal pose estimations, we benchmark simple baseline HRNet, HRFormer, and VIT pose method.*

$S_c = 0, S_m = 0$

Figure 3.11: t5-based alignment results across three examples: the first shows perfect correctness and missing scores ($S_c = 1, S_m = 0$); the second has perfect correctness ($S_c = 1$); and the third has perfect missing score ($S_m = 0$).

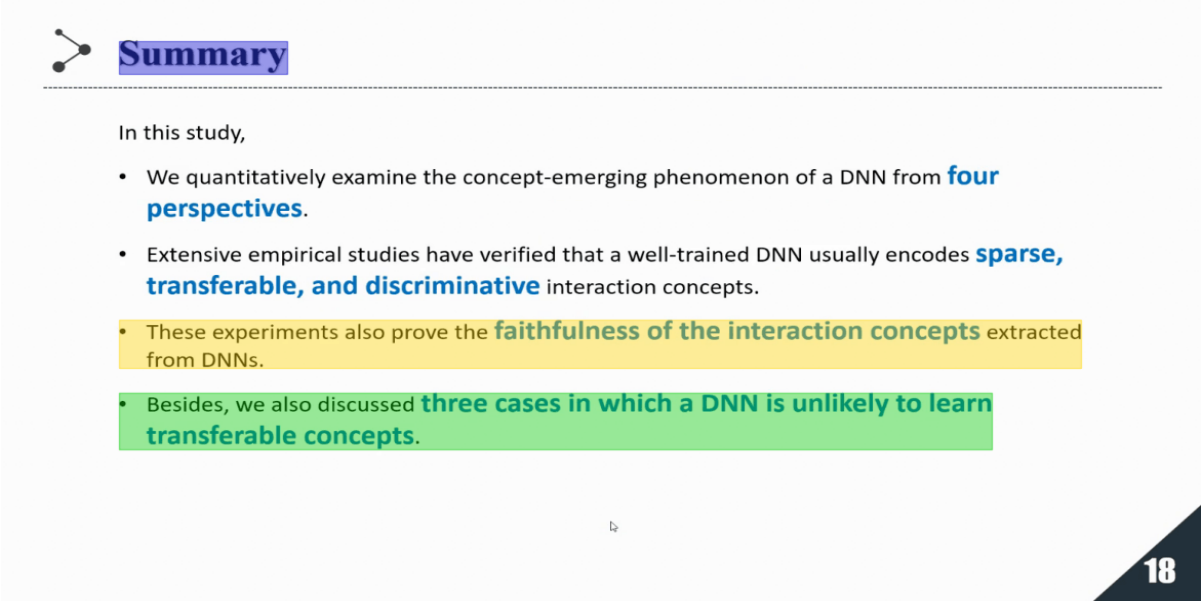


Figure 3.12: The figure shows that for "That's all," T5 highlights blue and Qwen highlights yellow; for "Thank you for your attention," T5 highlights blue and Qwen highlights green.

GT: To bridge the gap, we released METS-CoV data set.
Original: To bridge the gap, we released MassCov data set.
Corrected: To bridge the gap, we released METS-CoV data set.

GT: Hey, everyone, I'm Peilin Zhou.
Original: Hey, everyone, I'm Pei Leng Zhou.
Corrected: Hey, everyone, I'm Peilin Zhou.

GT: Tulane incorporates balanced and domain randomized images from simulation as the source domain and the well-known TuSimple dataset as the target domain.
Original: Tulane incorporates balanced and domain randomized images from simulation as the source domain and the well-known to simple data set as the target domain.
Corrected: Tulane incorporates balanced and domain randomized images from simulation as the source domain and the well-known TuSimple dataset as the target domain.

GT: In total, there are 45 unique blocks leading to a search base size of more than 91,000 models.
Original: In total, there are 45 release blocks leading to a search base size of more than 91,000 models.
Corrected: In total, there are 45 unique blocks leading to a search base size of more than 91,125 models.

GT: This can be attributed to the lack of publicly available data
Original: This can be attributed to the lack of publicly available data
Corrected: This can be attributed to the lack of publicly available datasets.

Figure 3.13: Comparison of ASR-generated text (GT) with the original and OCR-guided corrected versions.

3.5.2 Quantitative Analysis

The effectiveness of the proposed alignment method is evaluated using the Correctness Score (S_c) and Missing Score (S_m), as reported in Table 3.2. Fuzzy matching consistently achieves the highest correctness ($S_c \approx 0.99$); however, it exhibits a significantly high missing score, particularly at T-1 ($S_m = 0.663$), due to its inability to capture more challenging or semantically varied matches. In contrast, embedding-based approaches such as Sci-BERT and SPECTER perform comparatively poorly, with their correctness scores declining sharply across different thresholds.

Among the evaluated methods, MiniLM T-1 achieves the best overall F1 score, closely followed by T5. LLM-based models provide a more balanced trade-off between correctness and missing scores. For instance, Flan-T5 attains $S_c = 0.687$ and $S_m = 0.272$, whereas Qwen-2.5 achieves a higher correctness of $S_c = 0.843$ but with a corresponding increase in missing score ($S_m = 0.490$).

Method	**Avg** S_c	**Avg** S_m	**Avg** F_1	**Avg** S_c (W)	**Avg** S_m (W)	**Avg** F_1 (W)
Fuzzy [T-1]	0.998	0.663	0.254	1.000	0.673	0.246
Fuzzy [T-2]	0.998	0.603	0.307	1.000	0.617	0.295
Fuzzy [T-3]	0.989	0.525	0.364	0.989	0.526	0.360
S-BERT [T-1]	0.983	0.578	0.335	0.993	0.624	0.299
S-BERT [T-2]	0.943	0.460	0.403	0.966	0.507	0.373
S-BERT [T-3]	0.868	0.372	0.429	0.902	0.391	0.430
Sci-BERT [T-1]	0.524	0.356	0.280	0.636	0.474	0.279
Sci-BERT [T-2]	0.313	0.169	0.254	0.307	0.169	0.252
Sci-BERT [T-3]	0.222	0.056	0.003	0.219	0.053	0.220
SPECTER [T-1]	0.318	0.139	0.277	0.322	0.159	0.276
SPECTER [T-2]	0.211	0.035	0.219	0.211	0.0345	0.220
SPECTER [T-3]	0.174	0.010	0.191	0.175	0.011	0.191
T5	0.687	0.272	0.420	0.680	0.314	0.400
Qwen2.5	0.843	0.490	0.342	0.845	0.499	0.323

Table 3.2: Comparison of the average correctness score S_c and the average missing score S_m for different alignment methods under varying threshold settings. W denotes results obtained using the original transcript before correction. **T-1**, **T-2**, and **T-3** represent different threshold values for textual and visual regions: **T-1:** Textual region threshold = 0.8, Visual region threshold = 0.6, **T-2:** Textual region threshold = 0.7, Visual region threshold = 0.6 **T-3:** Textual region threshold = 0.6, Visual region threshold = 0.6

When comparing alignments on original (ASR) and corrected transcripts, LLM-based models such as Qwen show improvements in the F1-score, with gains of up to 0.019. Additionally, most methods benefit from reduced missing scores when using corrected transcripts. Overall, while fuzzy matching excels in exact-match scenarios due to its high correctness, LLM-based approaches demonstrate stronger and more robust semantic alignment capabilities, particularly when operating on improved or corrected input transcripts.

3.5.3 Impact of ASR Post-Correction

To analyze the impact of the correction process, we evaluate the ASR performance before and after applying OCR-guided refinement on a subset of 96 samples from the dataset. Table 3.1 summarizes the results, showing that the corrected ASR exhibits higher Character Error Rate (CER) and Word Error Rate (WER) than the original output. This increase is a direct consequence of enforcing alignment with text extracted from slide images during the correction process.

A closer inspection of Figure 3.13, particularly the last two examples, provides further insights into these corrections. In the third example, the corrected ASR transcribes the number as 91,125 instead of 91,000. This occurs because the speaker mentioned an approximate value during the presentation,

whereas the OCR-based correction enforces the exact figure present in the slide. Similarly, in the fourth example, the word "data" is replaced with "datasets" to align the transcription with the corresponding textual content extracted via OCR.

Despite leading to higher error metrics, these corrections are beneficial for the task at hand because they improve alignment with the visual content of the slides and facilitate more accurate localization of relevant information. Additionally, as illustrated in the first three examples of Figure 3.13, the OCR-guided approach enhances the recognition of technical terms and key concepts, thereby improving the overall quality and consistency of the transcription in technical presentation settings.

3.6 Summary

This work introduces a framework for aligning spoken content with visual regions in presentation slides, addressing the challenge of connecting speech with visual attention in multimodal presentations. By leveraging OCR to extract structured textual content from slides and ASR to transcribe spoken audio, we construct a unified representation that enables a fine-grained correspondence between speech and slide regions. This alignment facilitates the identification and highlighting of relevant visual content in synchrony with the narration, thereby improving the interpretability of presentations and enhancing the overall user experience.

We systematically evaluate our approach using the Correctness Score (S_c), Missing Score (S_m), and F1-score to capture both the accuracy and completeness of the alignment. Through extensive experiments, we compare a spectrum of techniques, ranging from traditional string matching to semantic embedding methods and LLM-based reasoning. Our analysis reveals inherent trade-offs between precision and recall: while exact matching methods achieve high correctness, they often fail to capture semantically relevant regions, whereas LLM-based approaches provide more robust and context-aware alignment. Additionally, our user study highlights the importance of intuitive visualization strategies that balance clarity, context preservation, and focus, and demonstrate that users benefit from effective highlighting mechanisms. This work opens several promising directions, including the integration of VLMs to better capture spatial and non-verbal cues, as well as extending the framework to real-time alignment in live presentation scenarios, paving the way for more interactive and intelligent multimodal systems.

This chapter focused on highlighting relevant content in presentation slides to reduce cognitive load and improve information accessibility during presentation understanding. The study demonstrates how selective emphasis and multimodal cues can support more effective comprehension of educational and presentation-based content.

Chapter 4

Educational Videos for Physics Learning

In this chapter, we explore the application of text-to-video (T2V) models for generating intuitive physics-related videos. As a foundational step, we present a benchmark and corresponding evaluation method to enable a systematic assessment of these models.

4.1 PhyEduVideo

Recent advances in Text-to-Video (T2V) models have opened new possibilities for generating dynamic visual content from natural language, yet their ability to accurately represent scientific and educational concepts remains insufficiently explored. This limitation is particularly evident in domains such as physics, where precise visual reasoning and faithful representation of underlying principles are essential for effective learning. To address this gap, we introduce a structured benchmark designed to evaluate the effectiveness of T2V models in visualizing foundational physics concepts in a pedagogically meaningful manner. This motivates the introduction of *PhyEduVideo*, a dedicated benchmark specifically designed to systematically assess the capability of T2V models in generating scientifically accurate and educationally aligned video content for physics.

4.1.1 Benchmark Construction

The proposed benchmark was carefully designed to systematically assess the ability of text-to-video (T2V) models to faithfully and accurately visualize foundational physics concepts in an educational context. Rather than focusing on generic video generation quality, our goal is to evaluate how well these models can capture and represent structured scientific knowledge through dynamic visualizations. To this end, the benchmark comprises a curated set of 60 core physics concepts drawn from seven major domains of classical physics: *Mechanics* (38.33%), *Waves & Oscillations* (6.67%), *Thermodynamics* (16.67%), *Electricity and Magnetism (Electromagnetism)* (20.00%), *Fluids* (10%), and *Optics* (8.33%) (see Figure 4.1 ②). Among these, *Mechanics* constitutes the largest portion of the benchmark, reflecting its central and foundational role in standard physics education curriculum. To ensure consistency and usability, we also define a standardized representation format, as illustrated in Figure 4.1 ③. The

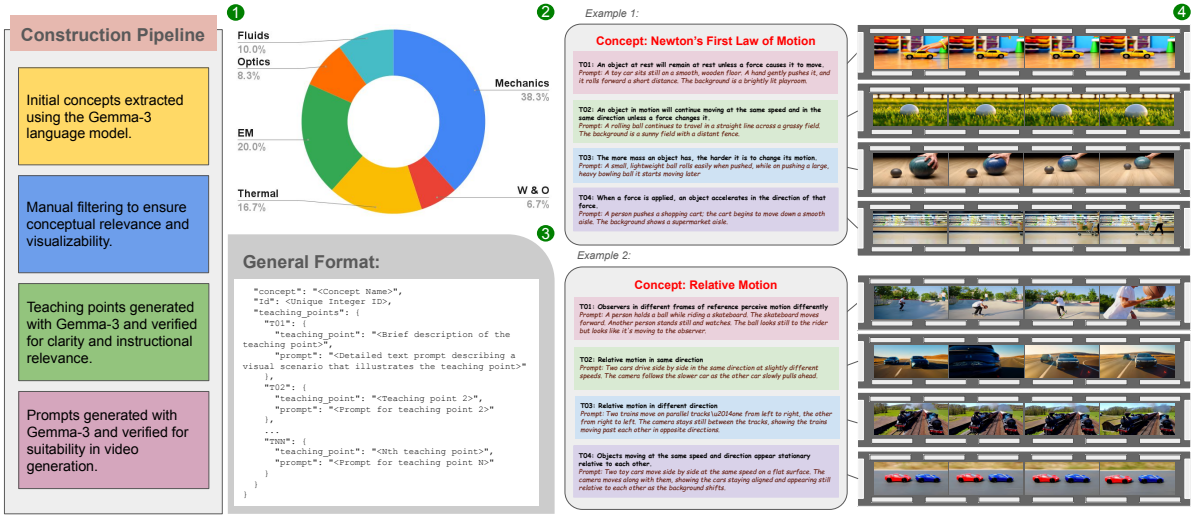
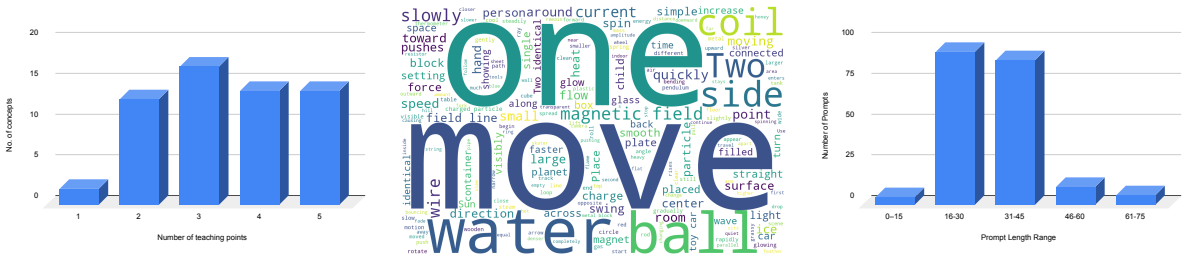


Figure 4.1: **Overview of the PhyEduVideo Benchmark.** 1 The construction pipeline, from concept extraction to prompt generation. 2 Concept distribution across five major physics domains: *Mechanics*, *Electromagnetism* (*EM*), *Optics*, *Thermodynamics* (Thermo), *Fluids*, and *Waves & Oscillations* (*W&O*). 3 Standardized representation of each concept, detailing key teaching points and corresponding video prompts. 4 Example concepts with teaching points, visual prompts, and representative generated video frames. As shown, current T2V models often fail to produce videos that are both semantically aligned and physically plausible—for example, in T04 (Relative Motion), the two toy cars were intended to move side by side at the same speed, but the generated video deviates from this.

construction of the benchmark followed a carefully structured multistage pipeline, as illustrated in Figure 4.1 1. Each stage was designed to progressively refine and validate the concepts, ensuring both the breadth of coverage and pedagogical relevance.

1. **Concept Identification:** We initiated the process by leveraging the Gemma-3 language model [85] to automatically extract a broad and diverse set of classical physics concepts from widely used K-12 and undergraduate physics curricula. This automated extraction step enabled efficient coverage across a wide conceptual spectrum, ensuring that the initial pool of concepts was comprehensive and representative of standard educational material.

2. **Manual Filtering:** Following automated extraction, the resulting list of concepts underwent rigorous manual review by domain experts in physics. This step was critical to ensure the pedagogical quality and visual feasibility of the selected concept. During this process, we retained only those concepts that are both essential for learning and suitable for visual representation. Highly abstract, redundant, or mathematically intensive topics, such as Lagrangian mechanics, tensor calculus, or complex integrals, were intentionally excluded. Instead, emphasis was placed on intuitive, observable, and visually demonstrable phenomena, including Newton's laws of motion, simple harmonic motion, and conservation of energy. This filtering process ensured that the benchmark remains grounded in concepts that can be effectively visualized and understood through, generated videos.



(a) (b) (c)

Figure 4.2: **Overview of benchmark statistics for the PhyEduVideo dataset.** (a) Distribution of teaching points across physics concepts, (b) Word cloud of frequent prompt terms, (c) Distribution of prompt lengths.

3. **Decomposition into Teaching Points:** To enable a more fine-grained evaluation of model capabilities, each validated concept was further decomposed into multiple *teaching points*. These teaching points represent smaller, pedagogically meaningful sub-concepts that capture specific physical behaviors, principles the relationships within the broader concept. For instance, as illustrated in Figure 4.1 4, the concept of "Newton's First Law" is broken down into four distinct teaching points: objects at rest, constant motion, inertia, and force-induced acceleration. This structured decomposition allows us to evaluate T2V models not only on their ability to represent broad conceptual ideas, but also on their capacity to accurately depict fine-grained physical interactions and behaviors.

4. **Prompt Generation and Refinement:** In the final stage, each teaching point was translated into candidate textual prompts designed to guide video generation. These prompts were initially generated automatically using the Gemma-3 model to ensure scalability and consistency in prompt formulation. Subsequently, human annotators refined the prompts to improve clarity, specificity, and pedagogical alignment. The refinement process ensured that each prompt is concise, unambiguous, and suitable for eliciting high-quality, conceptually accurate video outputs. Examples of the final refined prompts, along with their corresponding generated videos, are presented in Figure 4.1 4.

Through this multi-stage pipeline, the benchmark achieves a careful balance between automated scalability and expert-driven curation, resulting in a high-quality evaluation framework that is both comprehensive and pedagogically grounded.

4.1.2 Benchmark Statistics

The final benchmark consists of a total of 205 prompts, which are systematically derived from 60 carefully selected physics concepts. Each of these concepts is further decomposed into between one and five fine-grained teaching points, allowing for a more detailed and structured representation of the underlying physical ideas (see Figure 4.2(a)). This hierarchical design ensures that the benchmark captures both the breadth of classical physics and the depth required to evaluate specific conceptual

understandings. Each prompt is crafted as a self-contained, visually descriptive scenario that directly aligns with a clearly defined teaching objective, thereby enabling the precise evaluation of how well the models can translate conceptual knowledge into coherent visual narratives.

In terms of prompt complexity, the average prompt length ranged from 16 to 45 words. This variation in length is intentional, as shorter prompts typically correspond to simpler or more focused physical scenarios, whereas longer prompts provide richer contextual details necessary to describe more complex or multi-step phenomena. These longer prompts are important for testing whether models can maintain coherence and accuracy when presented with extended descriptions involving multiple interacting elements (Figure 4.2(c)).

Figure 4.2(b) illustrates the vocabulary distribution within the benchmark prompts, highlighting diverse and well-balanced terminology. The prompts incorporate a wide range of physical entities, such as "ball," "coil," and "current," alongside action-oriented verbs like "move," "push," and "show." This combination of entities and actions reflects both linguistic diversity and conceptual richness, ensuring that the benchmark spans a broad spectrum of physical interactions and phenomena. As a result, the dataset not only tests surface-level language understanding but also evaluates the model's ability to ground language in meaningful physical dynamics.

Collectively, these characteristics make PhyEduVideo a comprehensive and pedagogically grounded benchmark for evaluating text-to-video models. It is specifically designed to assess key aspects, such as scientific accuracy, temporal consistency, and visual fidelity, in the context of physics education. By focusing on structured teaching points and well-defined visual scenarios, the benchmark provides a rigorous testbed for analyzing how effectively models can bridge the gap between textual descriptions and scientifically meaningful video representations.

Compared to existing benchmarks, PhyEduVideo offers several notable distinctions. For instance, PhyGenBench comprises 160 prompts across 27 physical laws, whereas T2VPhysBench includes 84 prompts covering twelve laws. In contrast, VideoPhy emphasizes interaction-driven scenarios rather than structured conceptual learning. Unlike these benchmarks, PhyEduVideo is explicitly designed from an education-oriented perspective, grounded in a systematic decomposition of physics concepts into teaching points. This structured approach enables more fine-grained evaluation and aligns closely with real-world educational use cases, thereby distinguishing it as a more comprehensive and pedagogically aligned benchmark for evaluating physics-focused T2V models.

4.2 Metric

To comprehensively evaluate T2V models within the PhyEduVideo benchmark, we propose a structured evaluation framework that captures both the perceptual quality of the generated videos and their adherence to physics-based concepts. Because the primary goal of this benchmark is to assess how effectively models can support educational understanding, the evaluation is designed to examine not only

visual realism but also semantic and conceptual correctness. Accordingly, the framework is organized into two major components: *Video Quality Assessment* and *Conceptual Correctness*.

4.2.1 Video Quality Assessment

This component focuses on evaluating the visual fidelity and temporal consistency of the generated videos, ensuring that they are visually coherent and free of distracting artifacts.

- **Motion Smoothness (MS):** This metric evaluates whether the motion of objects in the generated video appears continuous and natural over time. A high-quality video should exhibit fluid transitions without abrupt jumps, unnatural accelerations, or irregular motion patterns. Smooth motion is particularly important for physics-based scenarios, where realistic depiction of movement is essential for conveying correct physical intuition.

- **Temporal Flickering (TF):** This metric assesses the presence of visual inconsistencies across consecutive frames, such as sudden changes in object appearance, lighting, or scene structure. Temporal flickering can disrupt the viewer's perception and reduce the overall realism of videos. Minimizing these inconsistencies is critical for maintaining temporal coherence and ensuring that the generated content is visually stable.

4.2.2 Conceptual Correctness

Beyond visual quality, it is equally important to evaluate whether the generated videos accurately reflect the underlying physics concepts described in input prompts. This component measures how well the models capture both the semantic meaning and the physical plausibility of the described scenarios, with reference to the structured prompts (Figures 4.8, 4.9, 4.10, 4.11, 4.12, 4.13)

- **Semantic Alignment (SA):** [57], [58], [60] This metric evaluates how closely the generated video aligns with the key elements of the input prompt, including the main objects, actions, and overall scenario. For instance, given a prompt such as "A rolling ball continues in a straight line," a correctly aligned video should depict a ball moving steadily along a straight trajectory without deviation. Semantic Alignment is scored on a scale from 0 to 3 using InternVL3.5 [86], which jointly considers object presence and action correctness. Higher scores indicate better alignment between textual descriptions and visual outputs.

- **Physics Commonsense (PC):** [57], [58], [60] This metric assesses whether the generated video adheres to fundamental physical principles and correctly reflects the intended teaching point. For example, in a scenario where ice is placed in water, the ice should gradually melt at $0°C$, with the water level increasing accordingly as the phase transition occurs.

  Following PhyGenBench [57], Physics Commonsense is further decomposed into three sub-components:

Figure 4.3: Comparison of SA (Semantic Adherence) and PC (Physics Commonsense) scores assigned by the VideoPhy, Automatic Evaluator (PhyEduVideo) and humans. Detailed videos are available on the GitHub page.

1. *Key Physical Phenomena Detection:* This evaluates whether the essential physical behavior described in the prompt is correctly represented in the video. For example, in projectile motion, an object should follow a parabolic trajectory rather than an unrealistic straight path.

2. *Physics Order Verification:* This component examines whether the sequence of events in the video follows a logically and physically valid order. For instance, in a pendulum system, the object must be released before it begins oscillating. This evaluation is performed using LLaVA-Interleave [87].

3. *Overall Naturalness Evaluation:* This evaluates the physical plausibility of the generated video by categorizing it into one of four levels: *fantastical*, *clearly unrealistic*, *slightly unrealistic*, and *realistic*. Fantastical descriptions correspond to highly imaginative or physically impossible scenarios, whereas clearly unrealistic cases violate fundamental physical laws (e.g., objects passing through solid surfaces). Slightly unrealistic descriptions generally

follow physical laws but contain minor inconsistencies, whereas realistic descriptions fully adhere to real-world physics. InternVideo2 [88] is used to compare the generated videos against these categories and assign the most appropriate label.

Figure 4.4: Guidelines and rules given to human annotators to ensure consistent and reliable evaluation.

4.3 Human Evaluation

To measure the extent to which automatic evaluation metrics reflect human perception, we conducted a human evaluation study using 500 generated videos. The annotators participating in this study had completed formal physics education up to the 12^{th} grade, ensuring sufficient familiarity with the underlying physical concepts and instructional content.

As illustrated in Figure 4.5, each video was assessed by human evaluators using two targeted questions specifically designed to measure the quality and correctness of the generated content. To maintain consistency across annotations, all evaluators followed a unified set of guidelines, presented in Figure 4.4. The evaluation primarily focused on two aspects: how accurately the generated video followed the provided prompt and whether it successfully conveyed the intended educational or physics-related concepts. These human annotations serve as reference points for assessing the extent to which automatic metrics align with human judgment.

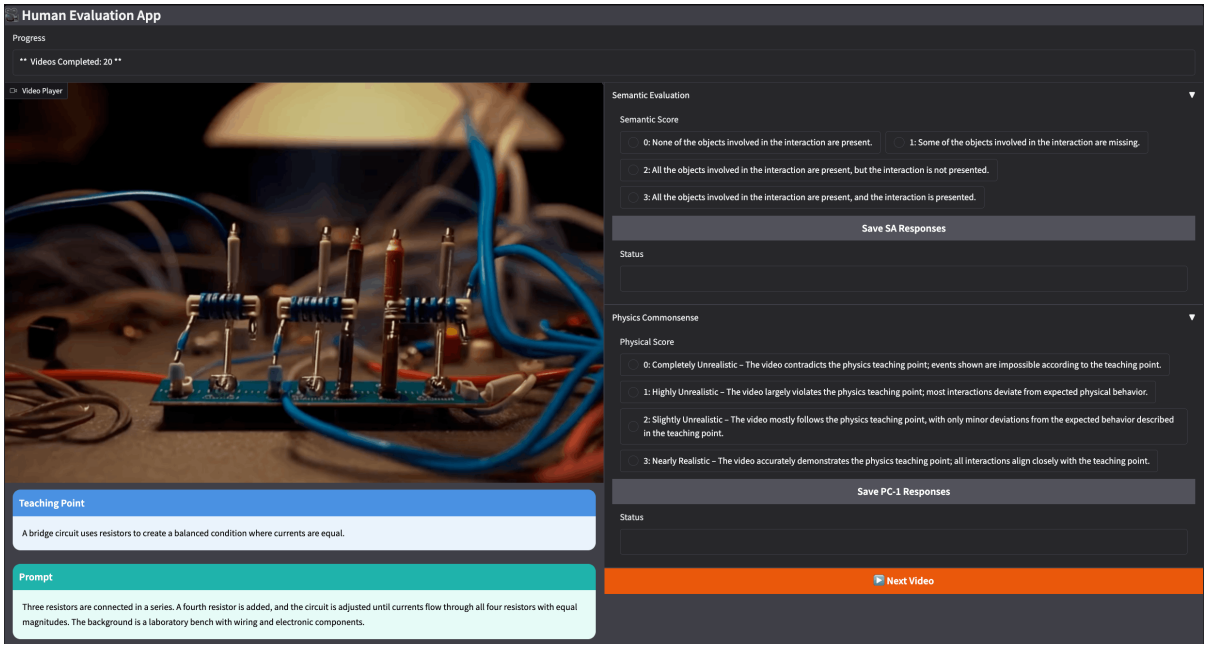


Figure 4.5: Questions provided for human evaluation and their respective scoring schemes are illustrated in the diagram above.

The quantitative results, summarized in Tables 4.1, show that **PhyEduVideo** achieves substantially stronger agreement with human evaluations than **VideoPhy** [57]. VideoPhy is primarily designed to evaluate whether text-to-video generation models satisfy general physical commonsense properties such as correct object interactions, material behavior, semantic consistency, and adherence to physical laws.

Across both Semantic Alignment (SA) and Physics Commonsense (PC), the strongest correlations are observed in the *Thermodynamics* category, while comparatively weaker correlations appear in *Electromagnetism* and *Optics*. Overall, for SA, PhyEduVideo achieves a Spearman correlation of 0.509 and a Kendall correlation of 0.462, outperforming VideoPhy by margins of 0.071 and 0.122, respectively. Similarly, for PC, PhyEduVideo obtains Spearman and Kendall correlations of 0.392 and 0.363, whereas VideoPhy achieves only 0.008 and 0.006, corresponding to absolute improvements of 0.384 and 0.357.

These consistent improvements indicate that our benchmark more effectively captures human judgment and aligns more closely with educationally meaningful evaluation criteria. In particular, the stronger correlations suggest that **PhyEduVideo** better reflects whether generated videos communicate the intended teaching concepts rather than merely exhibiting visually plausible motion. Figure 4.3 further provides qualitative comparisons between human annotations and the predictions produced by VideoPhy and PhyEduVideo, illustrating that our benchmark yields more accurate and interpretable evaluation behavior.

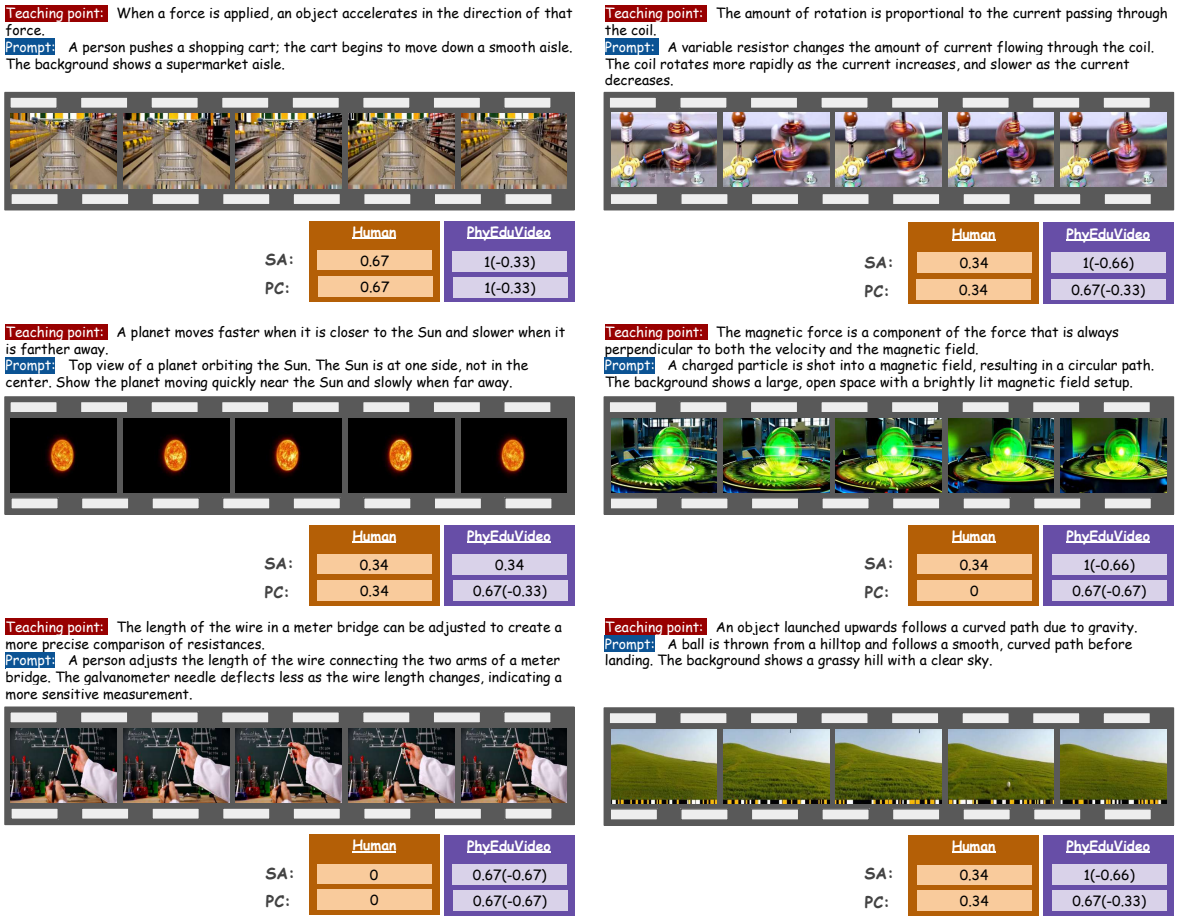


Figure 4.6: Comparison of SA (Semantic Adherence) and PC (Physics Commonsense) scores assigned by the Automatic Evaluator (PhyEduVideo) and humans.

4.3.1 Analysis of Discrepancies Between PhyEduVideo and Human Judgments

We analyze instances where the scores assigned by PhyEduVideo for Semantic Alignment (SA) and Physics Commonsense (PC) diverge from human judgments, aiming to understand the underlying reasons for such discrepancies (Figure 4.6). Overall, the model demonstrates strong agreement with human evaluations in simple and well-defined scenarios. For example, in a case involving the application of force to a shopping cart, both human annotators and the model assign consistent and high scores, indicating correct understanding and representation.

However, in more complex scenarios, PhyEduVideo tends to overestimate correctness, highlighting limitations in capturing nuanced physics reasoning and subtle semantic relationships. For instance, in the rotating coil example, human annotators assigned relatively low scores (SA = 0.34, PC = 0.34) due to only partial recognition of the relationship between the electric current and rotational motion. In contrast, PhyEduVideo assigned higher scores (SA = 1.0, PC = 0.67), reflecting an overestimation of correctness. Similarly, in the planetary orbit and charged particles in a magnetic field cases, the model

Metric	*EM*		*Mech*		*Fluids*		*Thermal*		*Optics*		*W&O*		Avg	
	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$
VideoPhy-SA	0.24	0.19	0.31	0.24	0.49	0.40	0.28	0.21	0.45	0.36	0.47	0.37	0.44	0.34
VideoPhy-PC	-0.01	-0.01	0.11	0.09	-0.11	-0.09	-0.05	-0.04	0.17	0.14	0.07	0.06	0.01	0.01
PhyEduVideo-SA	0.46	0.41	0.48	0.45	0.59	0.51	0.66	0.60	0.45	0.41	0.42	0.39	0.51	0.46
PhyEduVideo-PC	0.30	0.27	0.56	0.52	0.35	0.33	0.59	0.55	0.30	0.27	0.57	0.54	0.39	0.36

Table 4.1: Domain-wise correlations between human and model scores using Spearman's ρ and Kendall's τ. Models (VideoPhy, PhyEduVideo) are split into SA = Semantic Alignment and PC = Physics Commonsense. Domains are abbreviated as follows: EM: Electromagnetism, Mech: Mechanics, Thermal: Thermodynamics, W&O: Waves and Oscillations, and Avg: Average across all domains.

produced higher scores than humans, likely because it identified general motion patterns or the presence of fields without accurately capturing critical physical details such as orbital speed variation or circular motion trajectories.

A comparable trend is observed in the meter bridge wire adjustment and projectile motion on a hill scenarios, where PhyEduVideo again assigns higher SA and PC scores than human evaluators. In these cases, the model appears to misinterpret superficial visual cues as indicators of correct semantic and physical adherence, while human annotators identify subtle inconsistencies related to the intended purpose or motion dynamics.

In summary, while PhyEduVideo shows strong alignment with human judgments in straightforward cases, it tends to overestimate correctness in more complex scenarios that require fine-grained reasoning. This behavior likely arises from subtle physics nuances, partial contextual cues, and the model's reliance on detecting visible objects and motion patterns rather than fully understanding the underlying physical principles.

4.4 Experiments

4.4.1 Models Evaluated

We evaluated five state-of-the-art text-to-video (T2V) generation models on our benchmark: CogVideoX [14], Wan2.1 [16], VideoCrafter2 [12], Video-MSG [89], and PhyT2V [60]. Among these, CogVideoX-5B, which has demonstrated strong performance in physics-oriented evaluations, was considered a representative baseline for physics-grounded video generation due to its consistently high scores on physics-following benchmarks. Wan2.1, a robust general-purpose T2V model, was included for its strong performance across a wide range of benchmarks, thereby providing insights into the generalization capabilities of current large-scale video generation systems. VideoCrafter2, known for producing high-resolution and visually coherent outputs, was selected to evaluate the visual fidelity and fine-grained details in the generated videos.

In addition to these general-purpose models, we evaluated models with more specialized architectures and design strategies. Video-MSG adopts a training-free structured guidance pipeline designed

to closely adhere to the input prompt. The generation process is carried out in three stages: (1) *Background Planning*, where a multimodal large language model (MLLM, specifically GPT-4o) generates detailed background descriptions, which are then rendered using a text-to-image (T2I) model and subsequently animated using an image-to-video (I2V) model; (2) *Foreground Object Layout and Trajectory Planning*, where object placements and motion trajectories are determined under the guidance of the MLLM; and (3) *Video Generation*, where the planned layout is progressively denoised to produce the final video. This compositional and structured approach has demonstrated strong performance on T2V-CompBench [90], and was evaluated in our work for its ability to generate visually consistent and pedagogically meaningful physics-oriented content.

PhyT2V [52], on the other hand, was specifically designed for physics-aware video generation. It integrated large language models with physics-based simulation priors to iteratively refine the generated content, ensuring adherence to fundamental physical laws while maintaining both semantic and temporal consistency. Its strong performance on PhyGenBench [60] highlighted its effectiveness in improving physical plausibility and instructional clarity, making it particularly well-suited for evaluating T2V systems in educational and scientifically grounded contexts.

4.4.2 Quantitative Analysis

The quantitative results, summarized in Tables 4.2 reveal clear trends in both perceptual quality and correctness-oriented performance of the T2V models. Across all evaluated systems, Motion Smoothness (MS) and Temporal Flickering (TF) scores remain consistently high, typically exceeding 0.85, indicating that the current models are capable of producing visually coherent and temporally stable videos. However, this strength does not translate to correctness-based metrics such as Semantic Adherence (SA) and Physics Commonsense (PC), which are essential for generating educationally meaningful and conceptually accurate content.

Among the evaluated models, Wan2.1 [16] emerges as the strongest overall performer, achieving the highest SA and PC scores in most categories. It is closely followed by PhyT2V [52], which demonstrates competitive reasoning capabilities while maintaining strong visual consistency. In contrast, VideoCrafter2 [12] records the lowest performance in SA and PC, despite exhibiting strong results in terms of temporal smoothness and flickering. Similarly, Video-MSG [89] achieves high scores in video quality metrics but does not show significant improvements in physics commonsense, suggesting that compositional control alone is insufficient for capturing complex physical-reasoning.

A closer examination of the physics domains highlights notable variations in the model performance. *Mechanics* appears to be a relatively well-handled domain, with models achieving SA scores above 0.75 and PC scores exceeding 0.50, indicating that visually grounded phenomena such as motion and collisions are more effectively represented. *Fluids* and *Optics* perform particularly well across all four metrics, with the highest SA values reaching up to 0.90 and PC scores up to 0.71, suggesting that visually distinctive behaviors like fluid flow and light interactions are more easily captured by current models. In contrast, *Thermodynamics* and *Electromagnetism* present greater challenges, with lower correctness

scores observed across models. For example, in Thermodynamics, VideoCrafter2's performance drops to an SA of 0.51 and a PC of 0.38, whereas in electromagnetism, most models record PC values below 0.50. *Waves & Oscillations* exhibit intermediate performance, outperforming Thermodynamics and Electromagnetism but lagging behind Fluids and Optics.

These findings highlight a persistent gap between visual quality and conceptual accuracy. While current T2V models are effective in generating smooth and visually appealing videos, their ability to faithfully represent semantic meaning and adhere to physical principles remains limited. In this context, Wan2.1 and PhyT2V demonstrate comparatively better performance, showing stronger stability, coherence, and alignment with physical concepts, making them more suitable for physics-oriented educational applications.

A key challenge is particularly evident in domains such as *Electromagnetism*, where core concepts involve invisible entities such as electric and magnetic fields. For educational purposes, it is essential to represent abstract phenomena in a visually interpretable manner to support learning. Our benchmark addresses this limitation by going beyond traditional physics-based evaluation. Instead of solely assessing whether generated videos obey physical laws, it emphasizes the educational dimension, requiring models to effectively visualize both observable and abstract concepts in a way that facilitates understanding.

	Model	SA ↑	PC ↑	MS ↑	TF ↑
Mechanics	VideoCrafter2	0.75	0.52	0.94	0.92
	CogVideoX	0.85	0.57	0.98	0.97
	Wan2.1	0.86	0.66	0.99	0.98
	Video-MSG	0.75	0.53	0.99	0.99
	PhyT2V	0.80	0.59	0.98	0.97
W&O	VideoCrafter2	0.72	0.49	0.92	0.90
	CogVideoX	0.79	0.59	0.98	0.98
	Wan2.1	0.87	0.59	0.99	0.98
	Video-MSG	0.69	0.46	0.99	0.99
	PhyT2V	0.72	0.59	0.98	0.97
Fluids	VideoCrafter2	0.58	0.48	0.89	0.87
	CogVideoX	0.71	0.58	0.98	0.97
	Wan2.1	0.90	0.63	0.99	0.98
	Video-MSG	0.67	0.58	0.99	0.99
	PhyT2V	0.85	0.63	0.98	0.97
Thermal	VideoCrafter2	0.51	0.38	0.89	0.86
	CogVideoX	0.75	0.52	0.98	0.98
	Wan2.1	0.93	0.52	0.99	0.98
	Video-MSG	0.71	0.39	0.99	0.99
	PhyT2V	0.75	0.49	0.98	0.97
EM	VideoCrafter2	0.54	0.50	0.89	0.88
	CogVideoX	0.73	0.65	0.98	0.98
	Wan2.1	0.65	0.57	0.99	0.98
	Video-MSG	0.60	0.48	0.99	0.99
	PhyT2V	0.75	0.62	0.98	0.98
Optics	VideoCrafter2	0.64	0.62	0.88	0.83
	CogVideoX	0.69	0.60	0.99	0.98
	Wan2.1	0.78	0.64	0.99	0.98
	Video-MSG	0.69	0.64	0.99	0.99
	PhyT2V	0.80	0.71	0.99	0.98
Average	VideoCrafter2	0.62	0.50	0.90	0.88
	CogVideoX	0.75	0.59	0.98	0.98
	Wan2.1	0.83	0.60	0.99	0.98
	Video-MSG	0.68	0.52	0.99	0.99
	PhyT2V	0.78	0.60	0.98	0.97

Table 4.2: Comparison of five video generation models across six physics domains, along with their overall averages. Best scores are highlighted in cyan, and second-best in light cyan.

4.4.3 Qualitative Analysis

The qualitative results presented in Figure 4.7 provide a comparative analysis of model outputs across six key physics education domains—*Mechanics*, *Waves & Oscillations*, *Fluids*, *Thermodynamics*, *Electromagnetism*, and *Optics*. These examples highlight the strengths and limitations of current T2V models in visually representing scientific concepts. For each domain, the figure includes a representative teaching point,

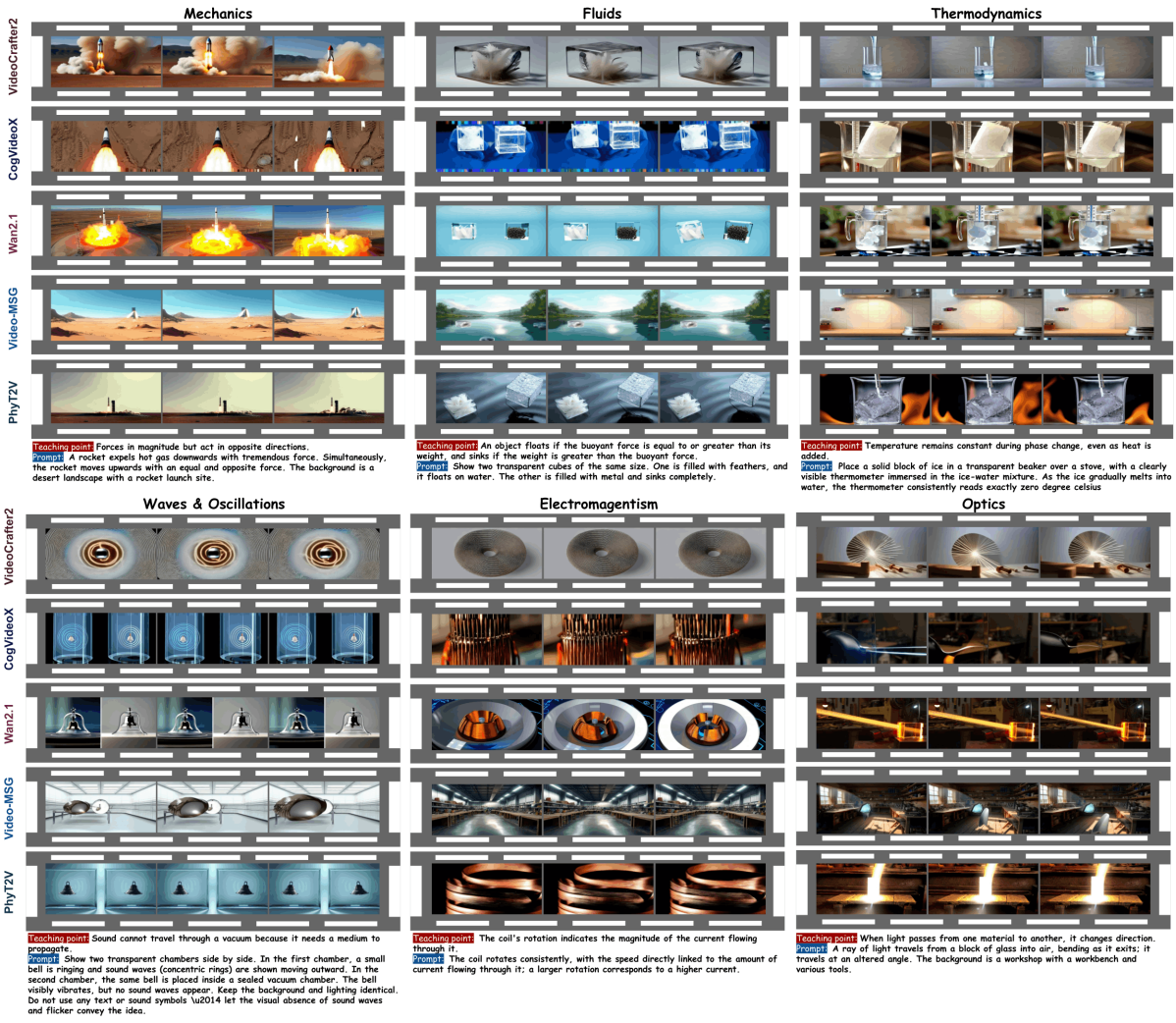


Figure 4.7: Qualitative comparisons of generated videos across six classical physics categories—*Mechanics*, *Waves & Oscillations*, *Fluids*, *Thermodynamics*, *Electromagnetism*, and *Optics*—for five T2V models: VideoCrafter2, CogVideoX, Wan2.1, Video-MSG, and PhyT2V. Detailed videos are available on the GitHub page.

the corresponding textual prompt used to query the T2V models, and representative frames extracted from the beginning, middle, and end of the generated video. The complete set of generated videos is available on the project's GitHub page.

Wan2.1 [16] demonstrates strong potential for educational applications by producing coherent and semantically aligned video sequences that adhere closely to physical principles, such as realistic representations of projectile motion in *Mechanics* and light refraction in *Optics*. CogVideoX [14] performs well in domains such as *Mechanics* and *Fluids*, where interactions between objects are relatively simple and strongly supported by visual cues; however, its outputs are sometimes affected by structural inconsistencies.

VideoCrafter2 [12] is capable of generating visually smooth and high-quality videos but often falls short in preserving semantic fidelity, which limits its effectiveness in instructional contexts, particularly in more abstract domains like *Electromagnetism* and *Waves*. Video-MSG [89] maintains strong temporal stability and shows promising results in more controlled settings such as *Mechanics* and *Thermodynamics*, but struggles to effectively capture deeper causal relationships and complex dynamic behaviors that are crucial for physics education.

In contrast, PhyT2V [52], which is explicitly designed with physics awareness, achieves a strong balance between visual consistency and conceptual accuracy. It performs particularly well in scenarios that require accurate physical reasoning, such as current-induced effects in *Electromagnetism*. Overall, these results highlight a clear gap between visual quality and conceptual fidelity in existing models, underscoring the need for physics-aware generation approaches to support accurate and meaningful science education.

4.5 Summary

This work presents a benchmark for evaluating text-to-video (T2V) generation in the context of physics education. Unlike prior efforts that primarily focus on physical plausibility, our benchmark emphasizes educational effectiveness. Each physics concept is decomposed into fine-grained teaching points, with prompts specifically designed to elicit visual explanations of these ideas. This enables a more meaningful assessment of whether generated videos go beyond visual realism to support learning by making abstract or invisible phenomena—such as electric fields, charge interactions, and wave propagation—intuitively understandable.

Using this benchmark, we evaluate a range of state-of-the-art models, including CogVideoX, Wan2.1, VideoCrafter2, Video-MSG, and PhyT2V. Although these models are generally capable of producing coherent motion with satisfactory temporal smoothness, they frequently fall short of capturing semantic adherence and incorporating physics commonsense. Among the evaluated models, Wan2.1 and PhyT2V demonstrate relatively stronger performance, yet still exhibit noticeable limitations in faithfully representing the underlying physical principles.

These findings highlight a critical gap between visual realism and conceptual accuracy, underscoring the need for T2V systems that are both physics-aware and education-oriented. Our benchmark provides a step toward bridging this gap, offering a structured framework to guide the development of models that can better align generative capabilities with educational objectives.

This chapter explored the use of text-to-video generation for creating intuitive physics educational videos that can support concept understanding through dynamic visualizations. Through our study, we highlight both the potential of current text-to-video models for educational content generation and the existing gaps in scientific accuracy, and pedagogical effectiveness that still need to be addressed.

Teaching point: If no external forces act on a system, the total linear momentum of the system remains the same before and after a collision.
Prompt: Two identical ice skaters glide toward each other on a frictionless ice rink. They collide gently and move together slowly after the collision. The background is a quiet, empty ice rink with no distractions.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: skater, ice rink
ACTION: A smaller skater collides with a larger stationary skater, and they slide together slowly across a frictionless ice rink.

KEY SEQ IDENTIFICATION

Q. After the collision, are both skaters moving together across the ice? Yes

Q. Is the two identical skaters gliding towards each other before the collision? Yes

ORDER VERIFICATION

Retr. prompt: At the moment the two skaters make contact

Description1: Two identical skaters are gliding towards each other.

Description2: both skaters slide together slowly after the collision.

OVERALL NATURALNESS EVALUATION

1) The skaters levitate, pass through each other without interaction, or explode in a flash of light after collision — completely ignoring momentum conservation.

2) The skaters bounce off each other and move away faster than before, or one suddenly speeds up while the other stops instantly, defying conservation laws.

3) The skaters' speeds or paths change slightly too early or too late relative to the moment of collision, or there's slight unnatural jittering, but overall momentum conservation is mostly maintained.

4) The skaters approach at equal speed, collide, and move together at the correct slower combined speed immediately after — matching the conservation of linear momentum almost perfectly.

Wan2.1
SA: 0.67
PC: 0.67

PhyT2V
SA: 0.67
PC: 0.67

Figure 4.8: Domain: Mechanics. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: Temperature describes the average kinetic energy of particles in a substance.
Prompt: Split the screen into two parts. On one side, show cold gas particles moving slowly and spaced far apart. On the other side, show hot gas particles moving rapidly and bouncing around quickly. Include a digital thermometer above each container showing low and high temperatures.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: gas particles, thermometer
ACTION: Cold gas particles move slowly; hot gas particles move rapidly.

KEY SEQ IDENTIFICATION

Q. Is the ice block completely melted after being in contact with the hot metal rod? Yes

Q. Is the metal rod no longer glowing after all the ice has melted? Yes

ORDER VERIFICATION

Retr. prompt: When the particles in both containers are visibly moving at different speeds

Description1: both containers show particles at rest or with minimal motion, and thermometers indicate similar low temperatures.

Description2: particles in the hot container move rapidly and collide frequently, while those in the cold container move slowly and less often, with thermometers showing a clear temperature difference.

OVERALL NATURALNESS EVALUATION

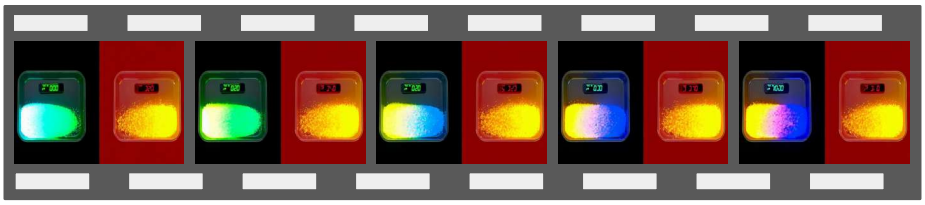
1) The animation shows gas particles on the cold side moving backward in time, changing shape, or merging and splitting at random. Thermometers flicker with nonsensical symbols. Particles teleport or transform into non-physical objects like animals or geometric shapes. The visual sequence is magical and completely ignores physical laws.

2) The cold gas particles are moving faster than the hot gas particles, or both sides have particles moving at the same speed regardless of thermometer reading. Particles may stop abruptly or pass through container walls without bouncing. The thermometer readings do not correlate with the observed particle speeds, clearly breaking the connection to kinetic energy.

3) The vast majority of particle motion is correct, but there are minor issues: perhaps a few collisions look awkward or a couple of particles move slightly faster or slower than they should for their side. The thermometer may have a slight delay in updating when the particle speeds change, but these are minor deviations that do not seriously undermine the teaching point.

4) The animation accurately shows cold gas particles moving slowly and spaced apart, and hot gas particles moving rapidly and bouncing energetically. Thermometers above each container display low and high temperatures that match the observed motion. All visual details closely align with the expected physical behavior and teaching point, with no noticeable errors.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 1
PC: 0.67

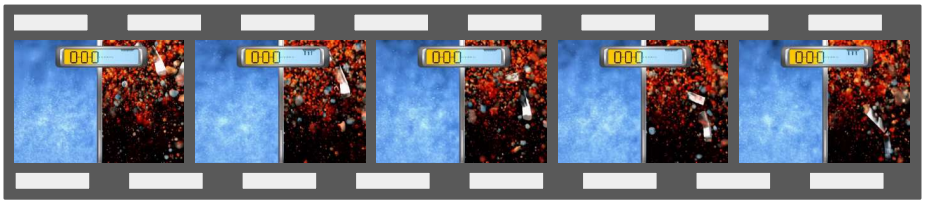


Figure 4.9: Domain: Thermodynamics. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: Temperature describes the average kinetic energy of particles in a substance.
Prompt: Split the screen into two parts. On one side, show cold gas particles moving slowly and spaced far apart. On the other side, show hot gas particles moving rapidly and bouncing around quickly. Include a digital thermometer above each container showing low and high temperatures.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: hair dryer, ping pong ball
ACTION: A ping pong ball floats in an upward stream of air and falls when the air stops.

KEY SEQ IDENTIFICATION

Q. Is the ping pong ball floating stably above the upward air stream from the hair dryer? Yes

Q. Does the ping pong ball start to fall as soon as the air stream stops? Yes

ORDER VERIFICATION

Retr. prompt: When the hair dryer turns off and the ball starts falling.

Description1: the ball floats steadily above the hair dryer in the fast-moving air stream.

Description2: the hair dryer is off and the ball falls straight down due to gravity.

OVERALL NATURALNESS EVALUATION

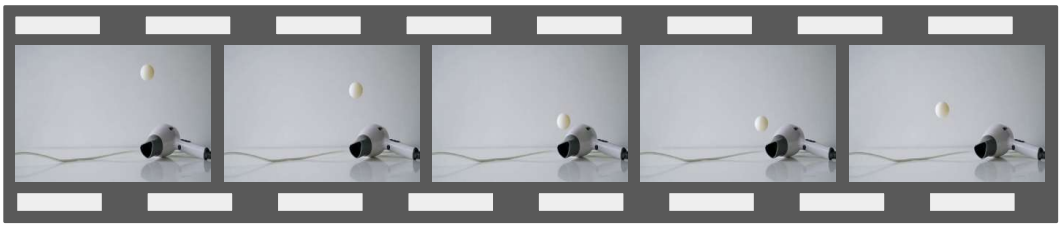
1) The ping pong ball levitates above the hair dryer, but it glows, spins in place with no air movement, and occasionally floats side to side or hovers even after the hair dryer is off. The ball might even rise higher when the dryer turns off or move in impossible ways, completely ignoring gravity and airflow.

2) The ball hovers, but its motion is inconsistent with airflow: it may drift far outside the airstream and still stay aloft, or it falls very slowly after the dryer is turned off, appearing to ignore gravity for several seconds. The ball might also bounce up and down repeatedly without any plausible reason.

3) The ball mostly stays in the air stream, levitating as expected, but there may be a slight lag between the dryer turning off and the ball beginning to fall, or the ball's motion is a bit jerky when it stabilizes in the air. The fall looks mostly natural but might be a bit too smooth or too abrupt.

4) The ping pong ball remains directly above the hair dryer, stably floating in the upward air stream; when the hair dryer is turned off, the ball immediately and naturally falls straight down under gravity. The timing and motion match real-world expectations of the Bernoulli effect and airflow.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 1
PC: 0.67

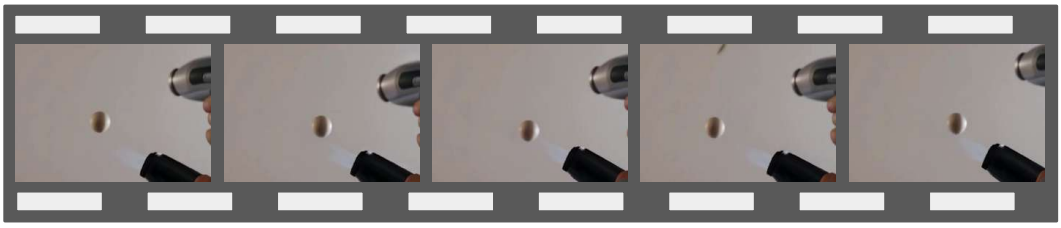


Figure 4.10: Domain: Fluids. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: When light passes from one material to another, it changes direction.
Prompt: A ray of light travels from a block of glass into air, bending as it exits; it travels at an altered angle. The background is a workshop with a workbench and various tools.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: light ray, glass block
ACTION: Light ray exits glass into air and bends away from normal.

KEY SEQ IDENTIFICATION

Q. Does the light ray bend at the boundary between glass and air? Yes

Q. Is the angle of the light ray in air different from its angle in glass? Yes

ORDER VERIFICATION

Retr. prompt: Ray going from glass to air - the point of ray enters air.

Description1: the ray of light moves through the glass block.

Description2: the ray of light exits the glass block into air, bending away from the normal; the angle of the ray changes, showing refraction, while the background and other objects remain the same.

OVERALL NATURALNESS EVALUATION

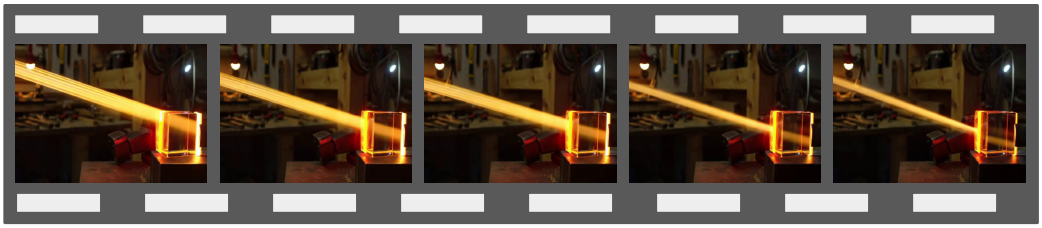
1) The light ray passes through the glass and into the air without changing direction at all, or bends in an impossible way (e.g., it loops, zigzags, or splits into multiple rays of different colors spontaneously). Alternatively, the ray transforms into a physical object or displays magical effects like sparking or levitating tools in the workshop.

2) The light ray noticeably ignores the interface between glass and air: it continues in a straight line, or bends in the wrong direction (toward the normal instead of away), or moves erratically for much of its path. The timing or sequence is inconsistent with normal behavior, such as the ray pausing mid-air or reflecting off surfaces that should be transparent.

3) The ray exits the glass and bends at the interface, but the angle is slightly off (e.g., a small deviation from what Snell's Law would predict), or the bending animation seems abrupt or a little delayed. There might be a tiny visual glitch, like the ray edge blurring, but the overall sequence matches the teaching point.

4) The ray clearly changes direction as it exits the glass block into air, following the correct angle relative to the normal—bending away as expected. The transition is smooth and matches the physical principle, with no distracting artifacts or unrealistic motion. The background elements (workbench, tools) remain neutral and do not interfere with the physics depiction.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 0.67
PC: 0.67

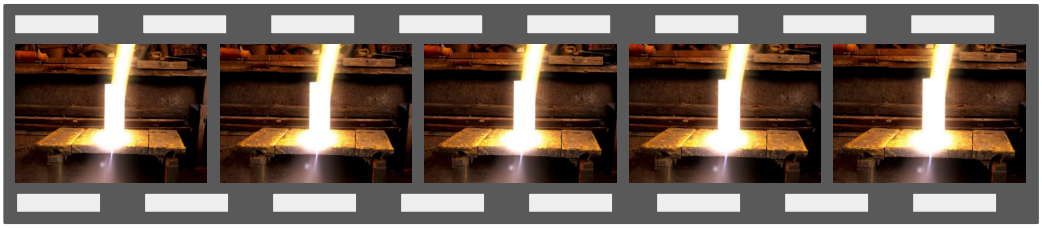


Figure 4.11: Domain: Optics. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: A capacitor stores electrical energy in an electric field created between its plates.
Prompt: Two metal plates are positioned close to each other. An arrow visually indicates the flow of electrons from one plate to the other, creating a visible electric field between the plates. A faint glow emanates from the region between the plates.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: metal plate, electrons, electric field
ACTION: Electrons flow between plates, generating electric field glow.

KEY SEQ IDENTIFICATION

Q. Is there a visible electric field (such as lines or a glow) shown between the two plates? Yes

Q. Is there an arrow showing electrons moving from one plate to the other?Yes

ORDER VERIFICATION

Retr. prompt: Middle Frame

Description1: From the first to the middle frame, the plates start out neutral, and then one plate becomes more negatively charged while the other becomes more positively charged. The electric field between the plates begins to form, and the faint glow starts to appear.

Description2: From the middle frame to the last frame, the electric field between the plates becomes stronger and the faint glow intensifies, indicating increased energy storage. The charge separation on the plates is now at its maximum, with the field fully established.

OVERALL NATURALNESS EVALUATION

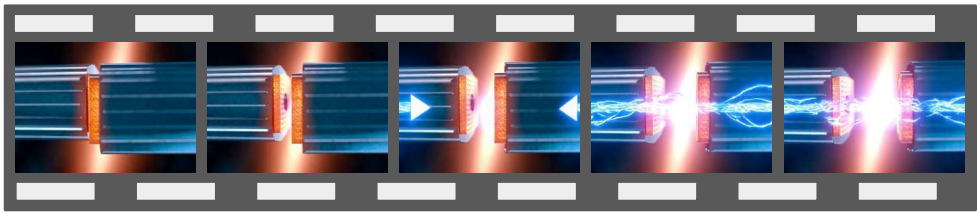
1) The plates float in midair emitting swirling, multicolored lightning bolts. Electrons visibly teleport between plates, and the plates levitate or morph shape. The 'electric field' manifests as an animated, pulsating wave that lifts objects or produces magical effects. The glow between plates pulses to the beat of music. None of these effects correspond to real physical behavior.

2) Electrons are shown moving in continuous loops between the plates even after the power source is removed, or the electric field causes the plates to attract or move towards each other dramatically. The glow becomes intensely bright, illuminating the whole scene. Arrows reverse direction randomly, and the plates spark or vibrate violently. These effects clearly contradict basic physical expectations for a capacitor.

3) The electron flow and field formation are correct, but there is a slight delay between electron motion and the appearance of the electric field. The faint glow between the plates may fade in or out a bit too slowly, or the electron arrow wiggles awkwardly. The sequence is almost correct, with only minor, brief timing or motion oddities.

4) Electrons are shown moving from one plate to the other in a brief, clear burst, with the electric field appearing steadily and symmetrically between the plates. The faint glow grows smoothly as the field builds, with all elements behaving as expected. The sequence accurately reflects the physical process of energy storage in a capacitor, with no noticeable deviations.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 0.34
PC: 0.34

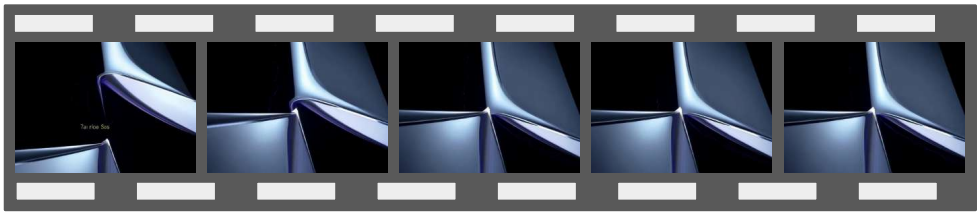


Figure 4.12: Domain: Electromagnetism. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: Resonance occurs when a system is driven by an external force at its natural frequency, leading to large amplitude oscillations.
Prompt: Show a child sitting on a swing. Initially, the pushes are irregular, and the swing barely moves. Then, demonstrate the child being pushed at regular intervals matching the swing\u2019s natural back-and-forth motion. With each well-timed push, the swing\u2019s amplitude increases noticeably. Clearly highlight that the energy transfer is most efficient when the pushing frequency matches the swing\u2019s natural frequency.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: child, swing
ACTION: Regular pushes matching swing's frequency increase its amplitude efficiently.

KEY SEQ IDENTIFICATION

Q. Is the swing reaching a much higher amplitude when the pushes are given at regular intervals matching its natural frequency? Yes

Q. IDoes the swing remain at a low amplitude when the pushes are irregular? Yes

ORDER VERIFICATION

Retr. prompt: Show the frame where the swing first begins to noticeably increase its arc due to well-timed pushes (after the irregular pushes).

Description1: Between the first frame (where the swing barely moves with irregular pushes) and the retrieval frame (the first frame showing well-timed pushes), the swing\u2019s arc starts to noticeably increase, and the child swings higher than before.

Description2: Between the retrieval frame (first noticeable increase in arc) and the last frame (after several well-timed pushes), the swing\u2019s arc grows even larger, and the child reaches a much greater height, clearly showing the effect of resonance.

OVERALL NATURALNESS EVALUATION

1) The swing begins to levitate, spin, or move in impossible ways regardless of how or when pushes are applied. The child might fly off at random, or the swing reaches infinite amplitude instantly. There are magical effects such as glowing energy waves, or the swing responds to pushes even when no one is pushing, completely disregarding the laws of motion.

2) The swing's motion does not correspond at all to the timing or strength of pushes: for example, the swing slows down or stops entirely when pushed at its natural frequency, or gains maximum height from random, weak, or mistimed pushes. The amplitude might decrease or stay constant no matter how well-timed the pushes are, contradicting resonance. The sequence shows persistent impossible behaviors (e.g., the swing passes through the support structure, or pushes act with a visible delay of many seconds).

3) Most of the animation matches expected behavior, but there are small flaws: the swing might respond a bit too quickly or slowly to changes in push timing, or the amplitude increases are slightly exaggerated. There could be a brief moment where a mistimed push has a larger effect than expected, or the swing's motion looks a little awkward or jerky, but overall the resonance effect is clear and mostly accurate.

4) The swing only gains significant amplitude when pushed at regular intervals matching its natural frequency; irregular pushes have little effect as expected. The amplitude builds up gradually over several well-timed pushes, and the swing's motion is smooth and physically plausible. Energy transfer is clearly most efficient at resonance, and all details (timing, amplitude growth, damping if included) faithfully reflect the real physics of resonance in a playground swing.

Wan2.1
SA: 1
PC: 0.67

PhyT2V
SA: 1
PC: 0.67

Figure 4.13: Domain: Waves & Oscillations. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Chapter 5

Fine-grain correspondence between various scientific formats

In this chapter, we address the problem of fine-grained correspondence in scientific content. We present a dataset and a structured evaluation framework, followed by a detailed analysis of the results.

5.1 Introduction

Scientific knowledge is increasingly communicated through multiple complementary formats, including research papers, presentation slides, recorded talks, and explanatory videos. While each modality conveys the same underlying ideas, they differ significantly in structure, level of detail, and style of presentation. Effectively connecting these heterogeneous sources requires understanding fine-grained correspondences across modalities—linking specific parts of a paper to the exact moments in a talk, regions in slides, or segments in explanatory videos where the same concepts are discussed.

In this work, we address the problem of fine-grained cross-modal correspondence across scientific formats. Unlike coarse alignment tasks that operate at the document or section level, our focus is on establishing precise mappings between smaller semantic units, such as paragraphs, figures, equations, and algorithmic components in papers, and their counterparts in slides and videos. This level of granularity is essential for applications such as multimodal retrieval, content navigation, and assistive learning systems, where users often seek specific pieces of information rather than entire documents.

To facilitate this, we introduce the Multimodal Conference Dataset (MCD), a curated dataset that captures aligned representations of research content across papers, presentation materials, and explanatory videos. The dataset is designed to reflect both formal and informal modes of scientific communication, enabling comprehensive analysis of how information is presented, transformed, and interpreted across modalities.

5.2 Multimodal Conference Dataset

The *Multimodal Conference Dataset (MCD)* is a curated collection of research papers, presentation slides, presentation recordings, and explanatory videos, each representing different perspectives of the same scholarly work. It aims to facilitate the study of how information is structured and linked across these complementary modalities. Table 5.1 compares MCD with existing multimodal academic datasets.

5.2.1 Dataset Construction

To build the Multimodal Conference Dataset (MCD), we systematically gathered materials that refer to the same underlying research work across multiple modalities. The dataset was organized into two complementary subsets: a *presentation set* and an *explanation set*, each designed to capture different forms of scientific communication.

The presentation set consisted of paper–slide–presentation video triplets curated from the NeurIPS 2023 conference. We first collected the research papers from *arXiv*, ensuring access to standardized and publicly available versions. The corresponding presentation slides and recorded talks were then obtained from the *SlidesLive* platform, which hosts official conference presentations along with the slide decks used by the authors. This alignment ensured that each triplet represented a consistent and coherent view of the same research contribution across textual, visual, and spoken modalities. In addition to formal conference presentations, we constructed an explanation set to capture more interpretive and pedagogical forms of content. This subset comprised paper–explanatory video pairs, where the explanatory videos were sourced from the *Yannic Kilcher* YouTube channel[1]. For each video, the corresponding research paper was retrieved using the links provided in the video descriptions, ensuring that the exact versions discussed in the explanations were included. This careful pairing enabled the dataset to reflect not only how research is formally presented but also how it is explained and interpreted for broader audiences.

Dataset	**Contains**					**Annot.**	**Task**
	Sl.	Doc	PV	EV	Pos		
DOC2PPT [72]	✓	✓	✗	✗	✗	A	Slide Gen
Automatic Slide Gen [77]	✓	✓	✗	✗	✗	M	Slide Gen
SciDuet [91]	✓	✓	✗	✗	✗	A	Slide Gen
Persona-D2S [79]	✓	✓	✗	✗	✗	Auto	Slide Gen
SlideAVSR [92]	✗	✗	✓	✗	✗	Auto+Man	AVSR
DocVideoQA [93]	✗	✗	✓	✗	✗	A+M	VideoQA
CS-PaperSum [94]	✗	✗	✓	✗	✗	A	Summ.
Paper2Poster [73]	✗	✓	✗	✗	✓	A	Poster Gen
Doc2Present [76]	✓	✓	✓	✗	✗	A	Video Gen
Paper2Video [74]	✓	✓	✓	✗	✗	A	Video Gen
MCD	✓	✓	✓	✓	✗	A+M	Cross Trav

Table 5.1: Comparison of multimodal academic datasets based on included formats (Sl.: slides, Doc: documents, PV: presentation videos, EV: explanatory videos, Pos: posters), annotation type (A: automatic, M: manual), and supported task.

5.2.2 Data Processing

To ensure consistency and high-quality alignment across modalities, we applied a sequence of preprocessing steps throughout the dataset. For the presentation set, particular care was taken to reduce redundancy while preserving meaningful contextual information.

Slides containing animations, where content was progressively revealed over multiple frames, were first identified. To avoid redundancy introduced by these incremental builds, only the final version of each slide—containing the complete set of visual elements—was retained. To preserve the full narrative context, transcript segments associated with the intermediate slides were concatenated with the transcript corresponding to the retained final slide.

[1] `https://www.youtube.com/@YannicKilcher/videos`

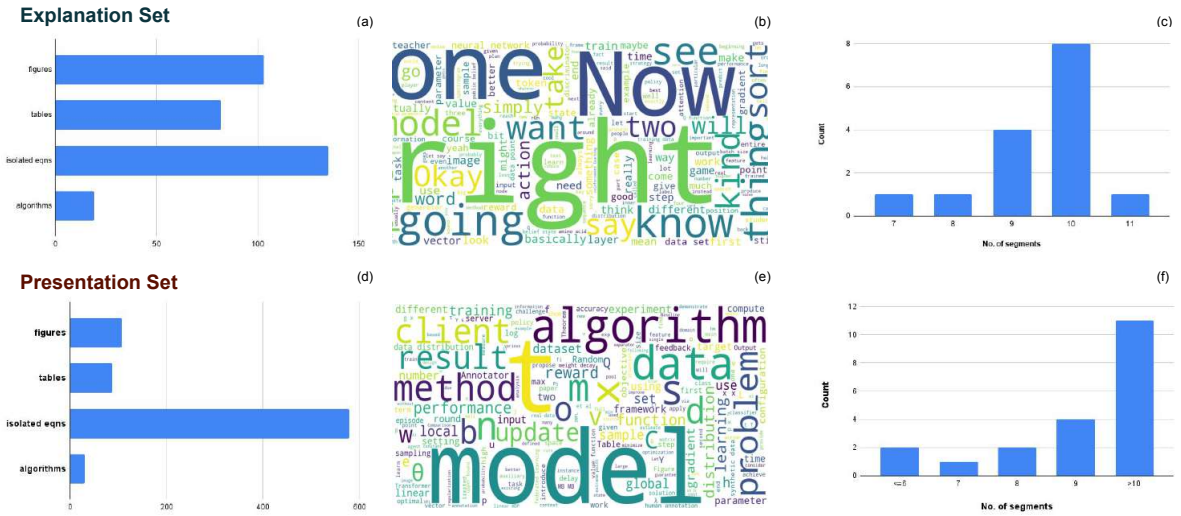


Figure 5.1: **Statistics of the Explanation and Presentation sets:** (a) and (d) show the distribution of algorithms, equations, tables, and figures in the papers; (b) presents the word cloud generated from ASR transcripts; (e) presents the word cloud generated from slide text and ASR transcripts; (c) and (f) depict the number of videos per segment category.

We additionally removed slides that do not contribute meaningful semantic information, including title slides, outline slides, acknowledgement slides, and closing slides such as "Thank You" pages. The associated papers were then segmented into paragraph-level units. Segments containing only a single character were discarded because they lacked informative content (Figure 5.2). Furthermore, segments marked as "abandon" by `PDFExtractor` were filtered out to improve overall data quality.

To handle overlapping visual regions, figure and equation areas intersecting with textual content were carefully processed so that only meaningful and non-redundant components were preserved. These preprocessing steps collectively produced a clean and well-structured dataset suitable for downstream retrieval and alignment tasks.

All video transcripts were generated using *WhisperX*, ensuring a consistent and high-quality automatic speech recognition pipeline across the dataset. To verify transcription quality, we manually inspected a sampled subset of videos and observed an accuracy greater than 90%.

We do not apply explicit OCR to slide images. Instead, slide-based modalities directly rely on the original visual inputs. In the *S→PP* setting, models are provided with complete slide images, requiring joint interpretation of visual structure and embedded textual content. In the *PV→PP* setting, slide images are paired with ASR transcripts, allowing complementary use of visual and spoken information.

Temporal alignment between videos and their corresponding content is obtained using reliable source-specific signals rather than post hoc synchronization methods. For *EV→PP*, we use topic timestamps explicitly provided in YouTube videos, which divide the content into semantically meaningful segments. For *PV→PP*, we utilize the native synchronization between slides and speech available in SlidesLive recordings.

	Count	Min	Max	Avg
EV (min.)	15	25.55	72.36	39.77
Slides	20	4	19	10.60
PV (min.)	20	2.25	5.33	4.58
Papers (all)	35	9	43	18.2

(a)

Modality	**Avg.**	**Max.**	**Min.**
Slides (S)	108.17	476	6
Slide + Transcript (PV)	156.80	582	39
Transcript (EV)	615.84	2759	28

(b)

Category	**Pres.**	**Exp**
Algorithm	30	19
Equation	579	135
Image + Caption	107	103
Table + Caption	88	82

(c)

Table 5.2: Dataset statistics: (a) Overview of dataset composition, including the number of papers and slides, along with video durations (in minutes) for Explanation Videos (EV) and Presentation Videos (PV); (b) Word count statistics for different modalities (EV, PV, and S), reporting average, maximum, and minimum values; (c) Distribution of paper segments, including algorithms, equations, images, and tables, across the Presentation and Explanation sets.

5.2.3 Dataset Statistics

The dataset is organized into two primary subsets: the *Presentation Set* and the *Explanation Set*, each capturing distinct forms of scientific communication. The Presentation Set comprises 20 paper–slide–presentation video triplets, where each presentation video has an average duration of approximately 5 minutes. For this subset, segmentation is performed at the slide level, such that each segment corresponds to an individual slide along with its associated ASR transcript, enabling fine-grained alignment between visual and spoken content.

In contrast, the Explanation Set consists of 15 paper–explanatory video pairs, with explanatory videos having a significantly longer average duration of approximately 40 minutes. Rather than relying on uniform segmentation, this subset leverages predefined topic boundaries, and the provided segments are used directly to preserve coherent thematic units and maintain the natural structure of the explanations. Transcripts for all videos across both subsets are generated using WhisperX, ensuring consistent and high-quality speech recognition.

A summary of the dataset is provided in Table 5.2. Specifically, Table 5.2a reports overall dataset statistics, including the number of papers, slides, and the durations of Explanation Videos (EV) and Presentation Videos (PV). Table 5.2b presents word count statistics across different modalities, while Table 5.2c summarizes the distribution of paper segments, including algorithms, equations, images, and tables, across the Presentation and Explanation sets.

Figure 5.1 further illustrates the dataset characteristics. Subfigures (a) and (d) show the distribution of content types within the papers, including algorithms, equations, tables, and figures. Subfigures (b) and (e) present the most frequently occurring words from the respective sources—ASR transcripts for the Explanation Set, and a combination of slide OCR and ASR transcripts for the Presentation Set. The Explanation Set contains relatively more conversational terms (e.g., “see,” “right”), whereas the Presentation Set is dominated by technical vocabulary such as “model,” “data,” and “algorithm.” Finally, subfigures (c) and (f) illustrate the distribution of segment counts per video, indicating relatively uniform segmentation in the Explanation Set, while the Presentation Set contains more videos with over ten segments.

5.2.4 Annotation Protocol

To support fine-grained cross-modal alignment, we carried out a comprehensive manual annotation procedure that explicitly links content across modalities with their corresponding research papers. In particular, each source segment—originating from either a slide, a presentation video segment (consisting of a slide and its aligned transcript), or an explanatory video segment (derived from transcript content)—was carefully matched with the relevant portions of the associated paper, including paragraphs, figures, equations, and algorithms. This process ensured that all semantically connected paper segments corresponding to a given source segment were systematically identified, enabling accurate alignment across textual, visual, and spoken modalities.

An initial pilot annotation study was conducted on a sampled subset of the MCD dataset using four annotators, all of whom achieved complete agreement. This result indicates that the annotation task is largely clear and unambiguous within the scientific literature domain. Such consistency is likely due to the structured and precise nature of academic content. Based on this observation, the remaining dataset was annotated by a single expert annotator to maintain consistency while improving annotation efficiency.

Although the segment categories are clearly defined, particular attention was given to establishing standardized criteria for identifying valid correspondences, especially in cases involving partial or seemingly ambiguous matches. The annotation process emphasizes semantic and conceptual alignment rather than strict surface-level similarity. For example, figures appearing in slides are considered aligned with figures in papers even when they differ in formatting, cropping, layout, or visual styling, provided they communicate the same underlying concept; both the visual elements and associated captions are considered during annotation. Similarly, equations are treated as corresponding when they represent the same concept or serve an equivalent explanatory purpose, even if their symbolic representations are not exactly identical. These guidelines help ensure coherent, systematic, and reliable annotations across the dataset.

The final annotated dataset consists of 15 explanatory videos and 20 presentation videos. The query set comprises 460 queries, each containing at least one paragraph. Among these, 147 queries include at least one relevant figure, 121 include at least one relevant equation, and 56 include at least one relevant algorithm. This diverse annotation framework supports comprehensive evaluation of multimodal retrieval and alignment tasks by covering a broad range of content types and varying levels of semantic complexity.

5.2.5 Benchmark Utility

The primary goal of this benchmark is to enable effective alignment and navigation across diverse forms of scientific communication, including research papers, presentation slides, conference talks, and explanatory videos.

Although these modalities communicate related scientific concepts, establishing meaningful connections between them remains challenging for both researchers and learners. Content presented in slides or presentation videos often corresponds not to a single section of a paper, but to multiple related components, such as explanatory paragraphs, mathematical expressions, and supporting figures. Similarly, explanatory videos may align with paper parts that provide theoretical intuition, derivations, or additional context. Accurately capturing these one-to-many relationships requires fine-grained multimodal alignment together with strong semantic understanding.

Furthermore, it provides a useful testbed for studying the grounding and reasoning abilities of multimodal large language models in scientific domains. As such, the benchmark is primarily intended as an evaluation resource for future research systems rather than a direct end-user application.

5.2.6 Paper Segments

Research papers inherently consist of multiple complementary elements—including textual descriptions, figures, algorithms, and equations—that together convey the full scope of the work. To systematically capture and represent this multimodal structure, we decompose each paper into four distinct segment types: *paragraphs*, *equations*, *figures* (along with their associated captions), and *algorithms*.

To extract these components, we employ a combination of specialized tools. Figures are identified and extracted using PDFFIGURES [95]. Algorithmic regions are detected using PP-DOCLAYOUT-L [96], which provides bounding boxes for structured layout elements; the corresponding textual content within these regions is then extracted using PADDLEOCR [97]. Equations are detected and recognized using the PDFEXTRACTOR toolkit[2], specifically leveraging its formula detection and recognition modules.

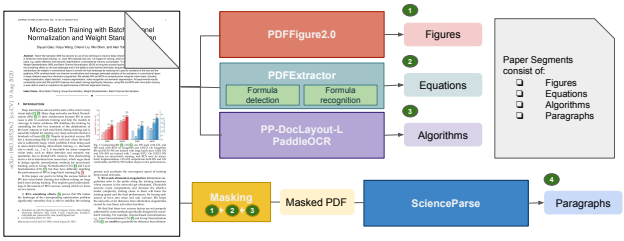


Figure 5.2: Pipeline for extracting paper segments—figures, equations, algorithms, and paragraphs—from the paper PDF.

To ensure clean separation between modalities, a preprocessing step is applied to resolve overlaps between extracted figures and equations, guaranteeing that each component is uniquely represented. These identified regions are subsequently removed from the original PDF, and the remaining textual content is processed using SCIENCEPARSE[3]. This step segments the document into structured components such as *abstract*, *sections*, and *authors*, from which the final *paragraph* segments are derived.

Overall, this pipeline enables a structured and non-overlapping decomposition of research papers into meaningful units, facilitating precise multimodal alignment and downstream retrieval tasks. Figure 5.2 illustrates the complete extraction process.

[2] https://github.com/opendatalab/PDF-Extract-Kit?tab=readme-ov-file
[3] https://reurl.cc/e62LXL

5.3 Cross-Format Retrieval

We consider the problem of linking multimodal research materials—explanatory videos, presentation videos, and slides—to their corresponding content in scientific papers. Although these modalities describe the same underlying work, they differ substantially in how information is presented: explanatory videos typically emphasize conceptual narratives, presentation videos combine visual cues with spoken explanations, and slides provide concise, visually structured summaries. Bridging these heterogeneous representations requires establishing precise correspondences across formats.

A research paper itself is composed of multiple complementary elements, including paragraphs, figures, equations, and algorithms. In this context, a figure is defined as the visual content along with its associated caption, with tables and their captions also treated as part of the figure category. Given a source segment from any modality—whether an explanatory video, a presentation video, or a slide—the objective is to identify and rank the most relevant components of the paper, thereby establishing a *fine-grained, one-to-many correspondence*. Formally, the task can be defined as:

$$f : x \to \{p_i\}_{i=1}^{N},$$

where x denotes an input segment from a source modality and $\{p_i\}$ represents the set of semantically related paper elements. This formulation reflects the fact that a single segment may correspond to multiple parts of the paper that collectively convey the same concept.

We evaluate this problem under three distinct retrieval settings, each defined by the nature of the query. For explanatory videos, the query consists of the transcript of a video segment (EV→PP). For slides, the query is the slide image itself (S→PP). For presentation videos, the query combines both the slide image and its corresponding transcript (PV→PP). Performance is measured using *NDCG@K* for paragraphs, figures, and equations, capturing how well the retrieved rankings align with human-annotated relevance. For algorithms, we adopt a threshold of 0.6 and report recall to evaluate retrieval effectiveness.

5.4 Evaluated Models

We evaluated six representative models spanning both embedding-based approaches and vision–language models (VLMs), enabling a comprehensive assessment across different paradigms of multimodal understanding. The embedding-based models included E5-V [68], GME (2.2B and 8.2B) [69], and ColQwen2-VL [98]. E5-V produces universal embeddings by adapting multimodal large language models (MLLMs) through text-pair training and prompt-based bridging, effectively aligning representations across modalities and demonstrating strong performance in image–text and document retrieval tasks. GME supports single-modal, cross-modal, and fused-modal retrieval, with both 2.2B and 8.2B variants (with and without instruction tuning). Its large-scale, instruction-driven training on fused-modal datasets enables the learning of unified representations across text, images, and visual documents, mak-

ing it particularly well-suited for capturing fine-grained correspondences between source segments and structured paper elements. ColQwen further extends this paradigm by integrating multimodal signals into a unified embedding space, facilitating efficient and coherent cross-modal retrieval.

Complementing these approaches, we included three VLMs—InternVL-4B, InternVL3.5-38B, and Qwen2.5-32B—selected for their strong capabilities in OCR, document understanding, and multimodal reasoning. InternVL-4B offers a lightweight and computationally efficient solution, while InternVL3.5-38B provides high-quality multimodal comprehension at a larger scale. Qwen2.5-32B demonstrates strong visual–linguistic alignment, making it effective for tasks requiring joint reasoning over textual and visual inputs. Although VLMs have shown strong performance on tasks such as summarization, rephrasing, and question answering, their ability to establish precise cross-modal correspondences remains less explored. We therefore probe this capability through the task of cross-linking and cross-grounding across modalities. Both embedding-based and VLM approaches are systematically evaluated across three traversal settings—EV→PP, S→PP, and PV→PP—spanning single-modal, cross-modal, and fused-modal retrieval.

5.4.1 Baseline Details

This section describes the embedding-based and vision–language model baselines used in our experiments. All models are evaluated in their original off-the-shelf configurations, without any task-specific fine-tuning, parameter updates, or architectural modifications. Each model is used with the default inference setup provided in its official implementation, including its native image resolution handling, context length, and decoding configuration. Closed-source systems, such as Gemini, are similarly evaluated using their standard API configurations.

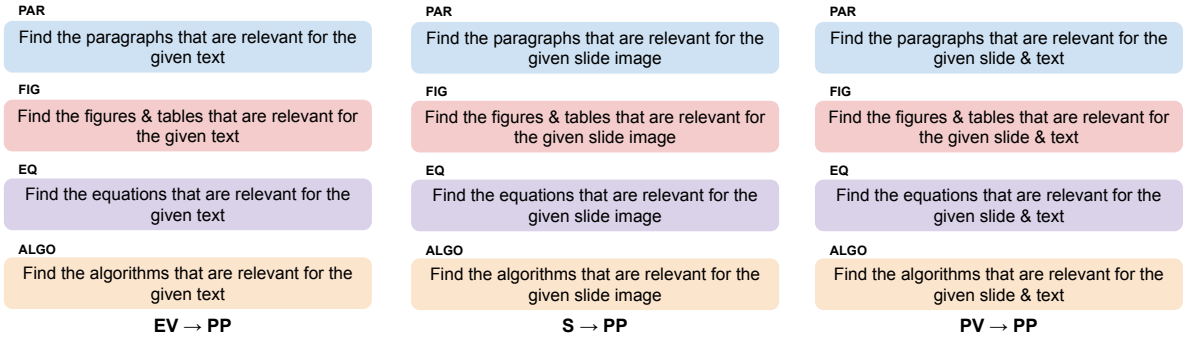


Figure 5.3: The figure illustrates the instruction prompts used for the GME models across the three traversal settings: EV → PP, S → PP, and PV → PP.

5.4.1.1 Embedding-based Models

The embedding-based baselines consist of ColQwen, E5-V, and four variants of GME. The GME models include two parameter scales (2.2B and 8.2B), each evaluated in both instruction-tuned and non-instruction settings. The instruction templates used for the GME variants are illustrated in Figure 5.3.

For ColQwen and E5-V, multimodal inputs containing combined information (e.g., an image together with its caption) are processed separately for each modality component. Similarity is computed independently, and the highest score is selected as the final similarity measure. In contrast, GME represents fused multimodal content using a unified embedding space and therefore produces a single joint embedding for the entire input.

All image inputs are supplied in their original resolutions without applying external resizing or handcrafted preprocessing steps. Image scaling and normalization are handled internally by the corresponding model processors according to their default pipelines. Since image resolution can affect representation quality and retrieval behavior, relying on the native preprocessing strategy of each model ensures a fairer and more consistent evaluation setting while avoiding biases introduced by manual preprocessing choices.

5.4.1.2 Vision–Language Models

For the vision–language model baselines, prompts explicitly include terms such as "direct" and "fine-grained" to encourage detailed cross-modal correspondence reasoning. Similar to the embedding models, all VLMs are evaluated using their default settings without additional adaptation or fine-tuning. The prompt template used in our experiments is shown in Figure 5.7.

Providing an entire paper and all candidate segments simultaneously often exceeds the context limitations of current VLMs, while evaluating each segment individually is computationally inefficient. To address this issue, candidate segments are divided into manageable subsets and processed incrementally. For each subset, the model evaluates every segment independently and assigns a relevance score with respect to the query. The resulting scores are then ranked to compute retrieval-based evaluation metrics.

To ensure that the benchmark primarily measures multimodal alignment rather than raw text recognition capability, we performed preliminary verification experiments confirming that the evaluated VLMs can reliably read and understand the images.

5.5 Evaluation Details

For paragraph, figure, and equation retrieval tasks, we use NDCG@K as the primary evaluation metric to measure the quality of ranked retrieval results. Algorithm retrieval, however, is evaluated differently because most papers contain at most a single relevant algorithm, making metrics such as NDCG@1 less meaningful, as successful retrieval directly yields a perfect score.

To evaluate algorithm retrieval more effectively, we adopt a recall-based evaluation using similarity thresholds. For the $EV \rightarrow PP$, $S \rightarrow PP$, and $PV \rightarrow PP$ settings, similarity thresholds are systematically

varied from 0.3 to 0.95 with a step size of 0.05. Based on empirical observations, a threshold value of 0.6 was chosen as a stable operating point. Smaller threshold values tend to increase false-positive matches, whereas larger thresholds become overly restrictive and significantly reduce recall.

We do not report precision for algorithm retrieval because many queries contain only one relevant algorithm candidate, although this is not always the case. Under such conditions, recall provides a more informative measure of retrieval performance. In several instances, the retrieval problem effectively reduces to a binary decision involving a single correct algorithm, which explains the occurrence of extreme metric values such as 0 or 100.

5.6 Performance Evaluation

In this section, we present a comprehensive analysis of retrieval performance across all evaluated models, considering both modality-specific traversal settings and diverse paper segment types. Specifically, we examine performance under three traversal scenarios—explanatory video to paper (EV $\rightarrow$ PP), slides to paper (S $\rightarrow$ PP), and presentation video to paper (PV $\rightarrow$ PP)—each representing a distinct form of cross-modal alignment between source content and structured scientific documents. These settings differ in the nature of the input signal, ranging from purely textual (transcripts) to purely visual (slides) and combined multimodal inputs (slide + transcript), thereby allowing us to probe model behavior under varying levels of modality complexity.

In addition to traversal settings, we analyze retrieval performance across different categories of paper segments, including paragraphs (par), figures (fig), equations (eq), and algorithms (algo) (Table 5.3). This fine-grained breakdown enables us to assess not only overall retrieval quality but also how effectively models capture different forms of scientific content, ranging from descriptive text to structured visual and mathematical elements.

To provide a structured and in-depth evaluation, our analysis is organized along four key dimensions: (i) traversal-specific trends, which highlight how performance varies across different source modalities; (ii) embedding-based models, focusing on their ability to learn unified representations for cross-modal retrieval; (iii) vision–language models, examining their strengths and limitations in grounding and reasoning across modalities; and (iv) the effect of model size, which investigates how scaling influences retrieval performance and alignment quality. Together, these perspectives offer a holistic understanding of the challenges and capabilities involved in fine-grained multimodal correspondence.

5.7 Performance Analysis

We now present a detailed analysis of retrieval performance across all evaluated models, examining both modality-specific traversal settings and diverse paper segment types. Specifically, we consider three traversal scenarios—explanatory video to paper (EV $\rightarrow$ PP), slides to paper (S $\rightarrow$ PP), and presentation video to paper (PV $\rightarrow$ PP)—each capturing a different form of cross-modal correspondence. These set-

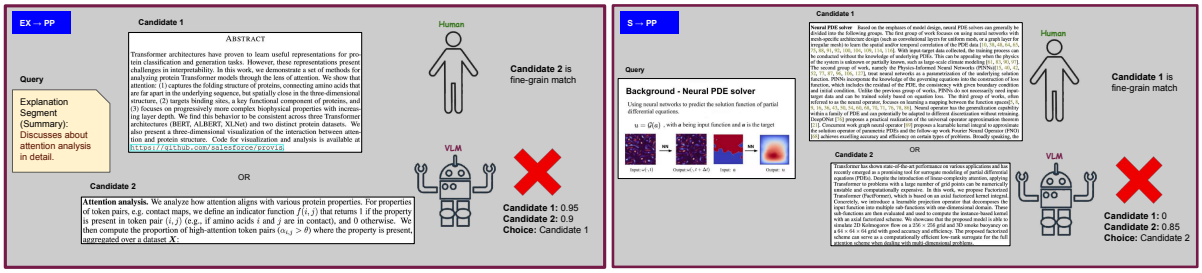


Figure 5.4: The figure presents two qualitative examples illustrating typical VLM failure cases. In the left example, the explanation segment discusses attention analysis in detail, as described in Candidate 2. However, the VLM assigns a higher similarity score to the abstract, which offers only a broad overview rather than the fine-grained correspondence expected. Similarly, in the right example, for the slide image, Candidate 1 provides a fine-grained match that closely aligns with the slide content, whereas Candidate 2 conveys only a general overview, yet receives a higher score from the VLM.

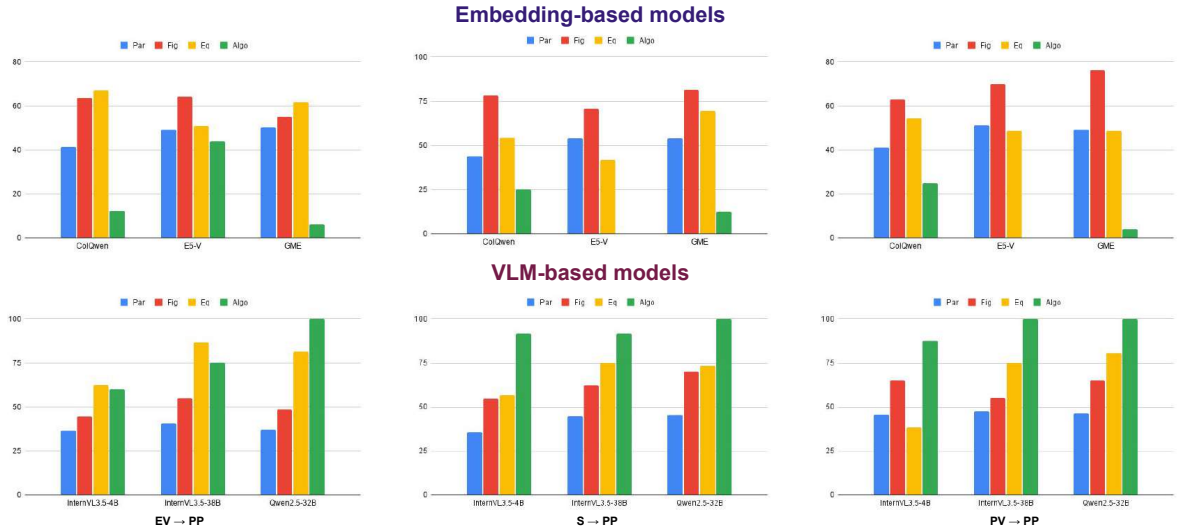


Figure 5.5: Performance comparison of embedding-based and VLM-based models across all three traversal settings (EV → PP, S → PP, PV → PP). NDCG@2 is reported for paragraphs, figures, and equations, while recall (threshold = 0.6) is used for algorithms.

tings vary in the nature of the input signal, ranging from narrative transcripts to visual representations and their combination, allowing us to systematically study how modality influences alignment. Performance is further analyzed across four segment types: paragraphs (par), figures (fig), equations (eq), and algorithms (algo) (Table 5.3). To provide a structured understanding, we organize our analysis along four dimensions: traversal-specific trends, embedding-based models, vision–language models, and the effect of model scale. A qualitative example is shown in Figure 5.4.

5.7.1 Across Traversals

Performance varies noticeably across the three traversal settings—EV→PP, S→PP, and PV→PP highlighting how differences in multimodal correspondence directly influence retrieval quality (Table 5.3). Among these, S→PP generally achieves the strongest performance across most metrics. This

can be attributed to the relatively tight alignment between slide content and the corresponding paper material. Slides typically act as structured condensations of the paper, where figures, equations, and key textual descriptions are either directly reused or lightly rephrased. This structural and semantic proximity makes it easier for models to establish correct mappings, particularly for visually grounded components such as figures and equations. However, even in this setting, algorithm-related retrieval remains comparatively difficult for most embedding-based models, likely due to its procedural nature and weaker visual grounding.

In contrast, EV→PP represents the most challenging traversal setting. Explanatory videos often rely on spoken narration that expands, reorders, or reinterprets paper content in a more conversational and pedagogical style. This introduces a substantial semantic gap between the narrative form of the video transcripts and the concise, formal phrasing used in research papers. Despite this difficulty, models such as GME-Qwen2VL-2B (with instruction tuning) still maintain relatively stable performance for figure-level retrieval, suggesting that visual or conceptual anchors embedded in the narration help preserve partial alignment with the underlying paper structure. Notably, paragraph-level retrieval remains particularly challenging in this setting, with even the best-performing embedding model achieving only up to ≤ 53 NDCG@1.

The PV→PP traversal lies between these two extremes and effectively combines characteristics of both settings. Presentation videos typically integrate slide-based visual structure with spoken explanation, thereby providing both explicit visual grounding and additional interpretive context. This dual modality helps reduce ambiguity and improves alignment compared to EV→PP, while still being slightly less structured than pure S→PP. Consequently, its performance is often comparable to S→PP in several cases, while consistently outperforming EV→PP across most metrics.

Overall, the observed trend S→PP $>$ PV→PP $>$ EV→PP holds in most cases, indicating a clear relationship between retrieval performance and the degree of semantic and structural continuity between modalities. Settings with stronger alignment and more explicit structural correspondence facilitate more reliable grounding, whereas settings characterized by stylistic and linguistic divergence pose greater challenges for fine-grained cross-modal retrieval.

5.7.2 Embedding-Based Models

Figure 5.6 visualizes query embeddings—for explanatory videos (transcripts), slides (images), and presentation videos (slide + transcript)—alongside candidate embeddings of paragraphs, equations, figures, and algorithms across EV $\rightarrow$ PP, S $\rightarrow$ PP, and PV $\rightarrow$ PP traversals. A clear structural pattern emerges in the embedding space, where equation embeddings form compact and well-separated clusters with minimal mixing with textual and visual embeddings. In the case of ColQwen, this separation is particularly pronounced, with equation clusters appearing almost entirely isolated from other modalities. This suggests strong modality-specific encoding, but at the same time indicates weaker cross-modal integration. In contrast, E5-V and GME exhibit partial mixing behavior: while equation clusters remain

clearly identifiable, they show mild overlap with paragraph and figure embeddings, indicating a more shared semantic representation space while still preserving modality distinctions.

Table 5.3 provides the corresponding quantitative evaluation, reinforcing several of these observed trends. Across all traversal settings, GME consistently demonstrates strong overall performance. In particular, the 2.2B variants outperform other embedding-based baselines for paragraph, figure, and equation retrieval in most cases. However, a notable limitation is observed in algorithm retrieval, where these models are unable to identify any relevant matches. Furthermore, across both the 2.2B and 7B variants, incorporating instruction prompts does not lead to consistent improvements in performance, suggesting limited sensitivity to prompt-based refinement in this setting. The GME-7B variant also fails to retrieve relevant algorithms in the S $\rightarrow$ PP traversal.

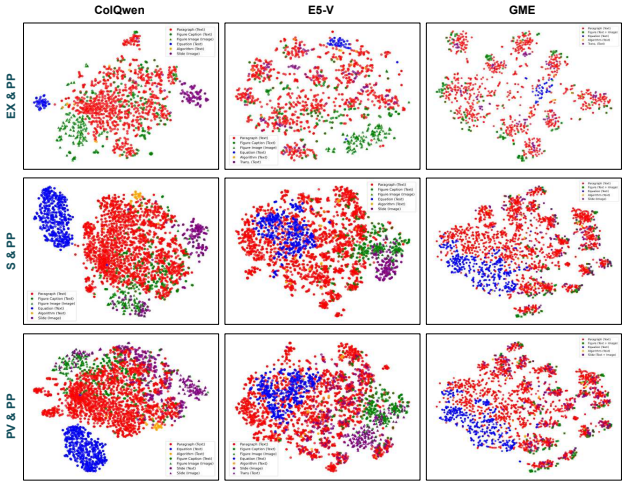


Figure 5.6: Visualization of query and candidate embeddings for representative instances across all three settings: EX $\rightarrow$ PP, S $\rightarrow$ PP, and PV $\rightarrow$ PP. Zoom in for a better view.

E5-V shows reasonably competitive performance overall, but remains below GME in most retrieval categories and traversal settings. ColQwen2-VL, on the other hand, exhibits a different behavior pattern: it achieves comparatively lower scores for paragraph and figure retrieval, but performs better on equation retrieval in EV $\rightarrow$ PP and S $\rightarrow$ PP, while showing moderate performance in PV $\rightarrow$ PP. Notably, it achieves perfect scores for all algorithms across all traversals, distinguishing it from the other models in this specific category. Finally, Figure 5.5 (top row) presents NDCG@2 scores for all three models across different paper segments, further illustrating these comparative trends.

5.7.3 Vision Language Models

Vision–language models (VLMs) exhibit a fundamentally different behavior compared to embedding-based approaches. While they possess strong multimodal reasoning capabilities, their performance on fine-grained retrieval tasks is generally lower, particularly for paragraphs and figures. This can be attributed to their broad semantic generalization: VLMs tend to identify multiple partially relevant candidates, assigning high similarity scores to a wider set of segments. Although this reflects strong conceptual understanding, it reduces precision when exact alignment is required.

Interestingly, VLMs perform relatively better on equations, likely due to the structured and symbolic nature of mathematical expressions, which provide clearer alignment cues. Algorithm retrieval performance is moderate, indicating partial success in grounding structured procedural content. Despite

the use of few-shot prompting designed to emphasize fine-grained matching, VLMs continue to favor coarse semantic associations over precise correspondence.

A notable contrast is observed with closed-source Gemini models, which consistently outperform open-source VLMs across all segment types and traversal settings. Gemini-2.5 Pro achieves the strongest overall performance, particularly for paragraphs and figures, while Gemini Flash offers a competitive trade-off between performance and efficiency. These models demonstrate a better balance between semantic generalization and fine-grained discrimination, resulting in more accurate retrieval. Figure 5.5 (bottom row) highlights these trends.

5.7.4 Effect of Model Size

Within the GME variants (2.2B and 8.2B), increasing model size does not lead to a consistent improvement in performance. In several cases, the 2.2B variant matches or even surpasses the 8.2B model across different traversal settings, indicating that parameter scaling alone is not sufficient to ensure stronger multimodal alignment. Overall, no clear or stable correlation is observed between model size and retrieval quality within embedding-based methods. A similar pattern is seen more broadly among larger embedding-based models as well: even models such as E5-V (8.4B) and GME-8.2B do not demonstrate clear or substantial gains over smaller counterparts like GME-2.2B or ColQwen2-VL (2.2B), reinforcing the lack of a consistent scaling benefit in this class of approaches.

In contrast, a different trend emerges among VLM-based models, where increasing model size leads to more consistent and noticeable improvements in performance. This improvement is particularly evident within the InternVL family, where larger variants perform better across the evaluated settings. The gain is especially prominent for equation retrieval, where larger models show clear advantages across all three traversal settings (EV$\rightarrow$PP, S$\rightarrow$PP, and PV$\rightarrow$PP). This suggests that scaling in vision–language architectures more effectively strengthens alignment for structurally grounded components such as equations.

Overall, these observations indicate that retrieval performance is jointly influenced by traversal type, segment modality, and model architecture. Embedding-based models generally struggle more with algorithm-related retrieval, whereas VLM-based models show comparatively weaker performance on paragraph-level retrieval but benefit more clearly from scaling, particularly for equations. This highlights that effective fine-grained cross-modal retrieval depends not only on model size, but also on how well the architecture supports modality alignment across different forms of scientific content.

5.8 Ablation

5.8.1 0-Shot and k-Shot Performance Analysis

Table 5.4 shows that the k-shot setting generally performs better than the zero-shot setting across most tasks. Providing a small number of input–output examples helps vision–language models learn

finer-grained cross-modal correspondences, leading to consistent improvements in paragraph, figure, and equation retrieval.

These results suggest that example-based prompting improves the model's ability to align semantic and structural information across modalities in scientific content. However, the gains for algorithm retrieval remain marginal, indicating that k-shot examples provide limited additional benefit for this modality.

5.9 Summary

This work introduced the *Multimodal Conference Dataset (MCD)*, a benchmark designed to study fine-grained correspondence across multiple scientific formats, including research papers, slides, presentation videos, and explanatory videos. Modern scientific communication is inherently multimodal, yet existing approaches largely operate at a coarse level, overlooking the nuanced relationships between specific elements across formats. MCD addresses this gap by enabling precise segment-level alignment between heterogeneous modalities, capturing how the same concept is expressed, condensed, or elaborated in different forms.

We formulated the problem as fine-grained cross-modal retrieval and evaluated six representative models—spanning embedding-based and vision–language paradigms—across three traversal settings: EV $\rightarrow$ PP, S $\rightarrow$ PP, and PV $\rightarrow$ PP. The evaluation further considered multiple paper segment types, including paragraphs, figures, equations, and algorithms, allowing for a detailed analysis of model behavior across diverse forms of scientific content. Our results reveal a clear dependency between retrieval performance and the degree of structural and semantic alignment between modalities. Slides, which closely resemble papers, provide strong grounding signals, whereas explanatory videos, characterized by narrative abstraction, pose significant challenges.

Beyond traversal-specific insights, our analysis highlights fundamental differences between model classes. Embedding-based methods exhibit strong performance in structured retrieval, particularly for textual and visual elements, but struggle with more complex or procedural components such as algorithms. In contrast, vision–language models demonstrate broader semantic understanding and cross-modal reasoning, but often lack the precision required for exact, fine-grained alignment. We further observe that model scaling impacts these approaches differently, while larger VLMs benefit from increased capacity, embedding-based models do not consistently improve with size, emphasizing the role of architecture and representation learning over scale alone.

Taken together, these findings underscore a central challenge in multimodal scientific understanding: bridging the gap between high-level semantic comprehension and precise cross-format grounding. MCD provides a unified and rigorous framework for studying this problem, offering both a benchmark and a diagnostic tool for analyzing model behavior. We hope this work will encourage the development of next-generation models that can accurately align information across modalities at a granular level, ultimately enabling more effective systems for scientific search, navigation, and learning.

This chapter investigated fine-grained correspondences across diverse scientific formats through the analysis of embedding-based and vision-language models for multimodal alignment and retrieval. The study highlights the strengths and limitations of current approaches in capturing detailed semantic relationships across heterogeneous scientific content, contributing toward more effective multimodal understanding and knowledge integration. Overall, our research focused on exploring the potential and limitations of current multimodal models in educational settings. Through studies spanning document understanding, educational video generation, and fine-grained scientific correspondence, we identify key challenges in alignment, evaluation, and contextual understanding, highlighting important directions for developing more reliable and effective AI systems for learning-oriented applications.

Model	Size	Par			Fig		Eq		Algo
		K=1	K=2	K=3	K=1	K=2	K=1	K=2	
Traversal: EV→PP									
ColQwen2-VL [98]	2.21B	47.27	41.52	40.37	61.40	63.80	63.16	67.01	12.50
E5-V [68]	8.4B	50.29	49.05	47.50	62	64.40	39.71	50.84	44
GME-Qwen2VL-2B w/o instr. [69]	2.2B	53.64	48.10	46.28	54.39	56.96	57.89	61.75	6.25
GME-Qwen2VL-2B with instr. [69]	2.2B	53.64	50.21	49.70	50.88	55.23	52.63	61.84	6.25
GME-Qwen2VL-7B w/o instr. [69]	8.2B	47.27	44.90	45.18	38.60	45.67	47.37	47.90	62.5
GME-Qwen2VL-7B with instr. [69]	8.2B	50.91	48.19	46.47	47.37	51.30	47.37	53.26	62.5
InternVL3.5-4B [86]	4B	36.29	36.45	34.79	39.34	44.68	61.90	62.39	60
InternVL3.5-38B [86]	38B	39.09	40.72	40.71	47.37	55.12	84.21	86.78	75
Qwen2.5-VL-32B [99]	32B	34.55	37.19	38.98	43.86	48.72	73.68	81.61	100
Gemini-2.5 Flash	-	53.64	53.60	53.72	52.63	56.38	73.68	82.36	75
Gemini-2.5 Pro	-	62.73	60.84	60.28	71.93	77.89	94.74	88.63	75
Traversal: S→PP									
ColQwen2-VL [98]	2.21B	47.88	44.02	45.03	68.18	78.22	50.98	54.21	25
E5-V [68]	8.4B	56.97	54.02	54.80	63.64	70.81	43.14	41.82	0
GME-Qwen2VL-2B w/o instr. [69]	2.2B	58.79	53.58	54.48	75	83.60	68.63	70.54	8.33
GME-Qwen2VL-2B with instr. [69]	2.2B	58.18	54.08	54.76	72.73	81.33	68.63	69.78	12.50
GME-Qwen2VL-7B w/o instr. [69]	8.2B	52.12	50	51.45	77.27	83.01	66.67	66.39	0
GME-Qwen2VL-7B with instr. [69]	8.2B	55.15	52.35	53.89	77.27	85.88	68.63	68.35	0
InternVL3.5-4B [86]	4B	33.33	35.92	37.71	40.91	54.69	50.98	56.97	91.67
InternVL3.5-38B [86]	38B	42.42	44.58	47.55	50.00	62.03	70.59	75.06	91.67
Qwen2.5-VL-32B [99]	32B	46.06	45.60	50.60	59.09	70.01	64.71	73.17	100
Gemini-2.5 Flash	-	59.39	57.91	60.68	77.27	86.43	80.39	87.81	100
Gemini-2.5 Pro	-	66.06	62.85	64.87	79.55	85.84	90.20	88.40	100
Traversal: PV→PP									
ColQwen2-VL [98]	2.21B	46.49	41.16	41.90	47.83	62.91	50.98	54.21	25
E5-V [68]	8.4B	54.05	51.29	53.91	60.87	69.94	47.06	48.49	0
GME-Qwen2VL-2B w/o instr. [69]	2.2B	52.97	48.06	48.57	65.22	76.72	49.02	49.70	4.17
GME-Qwen2VL-2B with instr. [69]	2.2B	55.14	49.24	49.85	67.39	76.15	45.10	48.53	4.17
GME-Qwen2VL-7B w/o instr. [69]	8.2B	55.14	49.39	49.63	69.57	79.70	60.78	61.26	0
GME-Qwen2VL-7B with instr. [69]	8.2B	55.68	49.30	50.02	65.22	78.40	62.75	62.47	4.17
InternVL3.5-4B [86]	4B	46.56	45.88	47.32	59.18	65.12	37.25	38.21	87.50
InternVL3.5-38B [86]	38B	49.73	47.70	48.98	47.83	55.21	75.00	75.00	100
Qwen2.5-VL-32B [99]	32B	44.86	46.39	49.41	52.17	65.05	74.51	80.70	100
Gemini-2.5 Pro	-	59.46	58.66	61.20	78.26	81.53	86.27	89.99	100
Gemini-2.5 Flash	-	59.46	56.22	57.22	76.09	81.57	82.35	86.06	100

Table 5.3: Retrieval results for traversals from explanatory video (EV→PP), slides (S→PP), and presentation video (PV→PP) to paper. Evaluation is performed over paper candidates—including paragraphs (Par), figures (Fig), and equations (Eq)—using NDCG@K, while for algorithms (Algo), only candidates with a threshold greater than 0.6 are considered, using recall as the metric. All values are reported in percentage (%).

Model	Size	Par			Fig		Eq		Algo
		K=1	K=2	K=3	K=1	K=2	K=1	K=2	
Traversal: EV→PP									
InternVL3.5-4B (0-shot) [86]	4B	29.87	32.51	32.30	25.58	33.15	50.00	47.06	0
InternVL3.5-4B (k-shot) [86]	4B	36.29	36.45	34.79	39.34	44.68	61.90	62.39	60
InternVL3.5-38B (0-shot) [86]	38B	35.06	32.37	35.93	34.88	41.89	61.11	70.27	100
InternVL3.5-38B (k-shot) [86]	38B	39.09	40.72	40.71	47.37	55.12	84.21	86.78	75
Qwen2.5-VL-32B (0-shot) [99]	32B	25.00	32.51	34.66	18.60	35.08	66.67	75.03	100
Qwen2.5-VL-32B (k-shot) [99]	32B	34.55	37.19	38.98	43.86	48.72	73.68	81.61	100
Traversal: S→PP									
InternVL3.5-4B (0-shot) [86]	4B	30.30	33.83	35.50	40.91	54.69	50.98	53.45	90.48
InternVL3.5-4B (k-shot) [86]	4B	33.33	35.92	37.71	40.91	54.69	50.98	56.97	91.67
InternVL3.5-38B (0-shot) [86]	38B	42.31	43.44	44.75	42.86	51.58	65.52	72.29	91.67
InternVL3.5-38B (k-shot) [86]	38B	42.42	44.58	47.55	50.00	62.03	70.59	75.06	91.67
Qwen2.5-VL-32B (0-shot) [99]	32B	43.03	45.19	50.82	50.00	66.10	62.75	73.20	91.67
Qwen2.5-VL-32B (k-shot) [99]	32B	46.06	45.60	50.60	59.09	70.01	64.71	73.17	100
Traversal: PV→PP									
InternVL3.5-4B (0-shot) [86]	4B	38.92	38.11	40.09	47.83	61.01	31.37	36.80	87.50
InternVL3.5-4B (k-shot) [86]	4B	46.56	45.88	47.32	59.18	65.12	37.25	38.21	87.50
InternVL3.5-38B (0-shot) [86]	38B	48.11	47.04	49.21	43.48	54.98	75.51	75.51	100
InternVL3.5-38B (k-shot) [86]	38B	49.73	47.70	48.98	47.83	55.21	75.00	75.00	100
Qwen2.5-VL-32B (0-shot) [99]	32B	40.54	45.32	47.84	52.17	65.00	74.50	80.00	95.65
Qwen2.5-VL-32B (k-shot) [99]	32B	44.86	46.39	49.41	52.17	65.05	74.51	80.70	100

Table 5.4: Zero-shot and k-shot retrieval performance for traversals from explanatory video (EV→PP), slides (S→PP), and presentation video (PV→PP) to paper content. Models are evaluated over paragraph (Par), figure (Fig), and equation (Eq) candidates using NDCG@K, while algorithm (Algo) retrieval is evaluated using recall over candidates exceeding a similarity threshold of 0.6. Zero-shot rows contain no in-context examples, whereas one-shot rows incorporate a single in-context demonstration. All values are reported in percentage (%).

PROMPT

Your task is to identify how relevant each paper segment is to a given video segment.

Inputs:

- Query: Transcript (EV) | Slide image (S) | Slide Image + Transcript (PV) (any one)
- Candidate:
  - Paper Paragraphs: {par_group}

    OR

  - Paper Segment: consists of a figure & caption: <image>
    Caption: {caption_figure}
    Fig_id: {fig_id}

    OR

  - Paper Segment: consists of an algorithm:
    {cand_algo_text}
    Algo_id: {algo_id}

    OR

  - Paper Segment: Equations:
    {all_eqns}

Instructions:

1. Determine how relevant each paper segment is to the video segment.
2. If relevant, briefly explain why.
3. If not relevant, state that it is not relevant and explain why.
4. There must be a direct, fine-grained overlap between the video content and the equation.
   Only assign a high score if most of the concepts mentioned in the video appear explicitly in the equation.

When giving the relevance score consider the paper segment in isolation (do not compare it with others).

Output Format (strict JSON):

{{
"results": {{"<paper_segment_id_1>": <float_between_0_and_1>, "<paper_segment_id_2>": <float_between_0_and_1>, ...}},
"explanation": "<few lines summarizing reasoning for the assigned scores>"
}}

When giving the relevance score consider the paper segment in isolation (do not compare with others).
Provide a score for all input paper segments; do not skip any.

Figure 5.7: Prompt used for VLM evaluation. Queries are provided as: transcripts (explanatory videos), slides (slides), or slides+transcripts (presentation videos). Each query is evaluated against four paper segment types: paragraph, figure, equation, and algorithm.

Chapter 6

Conclusions and Future Work

This thesis explored multimodal understanding from three complementary perspectives: (i) aligning spoken content with visual regions in presentations, (ii) evaluating generative models for educational video synthesis through *PhyEduVideo*, and (iii) establishing fine-grained correspondence across scientific formats using the *Multimodal Conference Dataset (MCD)*. Together, these works address a unified challenge—how models perceive, generate, and align information across diverse modalities such as speech, text, images, and videos.

In the first line of work, we introduced a framework for aligning spoken narration with visual regions in presentation slides. By leveraging OCR for extracting structured slide content and ASR for transcribing speech, we constructed a shared representation that enables fine-grained correspondence between spoken language and visual elements. Our evaluation, based on Correctness Score (S_c), Missing Score (S_m), and F1-score, demonstrated clear trade-offs between exact matching, semantic embedding, and LLM-based approaches. While traditional methods achieved high precision, LLM-based techniques provided more context-aware and semantically meaningful alignments. The accompanying user study further emphasized the importance of intuitive visualization for effectively guiding user attention.

Building on alignment, the second contribution—*PhyEduVideo*—shifted focus to generation and evaluation in the context of physics education. By decomposing concepts into fine-grained teaching points and introducing metrics that capture both visual quality and conceptual correctness, this benchmark revealed a critical gap in current text-to-video (T2V) models. Although these models produce visually smooth and coherent outputs, they often struggle to faithfully represent underlying physical principles or convey them in a pedagogically meaningful manner. This finding highlights an important limitation of existing generative systems: visual realism alone is insufficient for educational applications where conceptual correctness and clarity are equally critical.

The evaluation further demonstrated that models incorporating stronger semantic grounding or domain-specific priors, such as Wan2.1 and PhyT2V, show relatively improved alignment with physics concepts. Nevertheless, even these approaches face difficulties in representing abstract or non-visible phenomena such as fields, forces, and wave interactions. These observations emphasize the need for future generative models that integrate scientific reasoning and educational awareness alongside visual synthesis capabilities.

Extending beyond alignment and generation, the third contribution—*MCD*—addressed fine-grained correspondence across scientific formats, including papers, slides, and videos. Through systematic evaluation across multiple traversal settings and model families, we demonstrated that retrieval performance is strongly influenced by structural and semantic alignment between modalities. Embedding-based models excel in structured retrieval tasks, whereas vision–language models capture broader semantics but lack precision in exact grounding. This highlights a persistent gap between high-level understanding and fine-grained alignment.

Taken together, these works present a cohesive view of multimodal intelligence: effective systems must not only generate or understand content, but also align it accurately across modalities and representations. Bridging these aspects is essential for applications in education, scientific communication, and knowledge access. The thesis further demonstrates that multimodal learning should move beyond isolated tasks toward integrated systems capable of jointly reasoning over speech, text, images, documents, and videos in a coherent and interpretable manner.

Future Scope: Several promising directions emerge from this research. First, integrating domain knowledge—such as physics priors, symbolic reasoning, or structured document representations—into both generative and retrieval models could improve conceptual fidelity and alignment accuracy. In the context of educational video generation, future systems should move beyond optimizing visual realism alone and instead focus on pedagogical effectiveness, interpretability, and conceptual consistency. Developing evaluation metrics that better capture educational usefulness, beyond surface-level visual quality, remains an important open challenge.

Second, incorporating multimodal supervision through diagrams, annotations, textual explanations, and instructional cues could guide models toward more semantically grounded and educationally meaningful outputs. Extending video generation systems to support regional languages and adaptive content generation based on learner grade levels may further improve accessibility and inclusiveness in educational settings.

Third, future work can move beyond OCR-centric alignment approaches by integrating VLMs that jointly reason over textual, visual, and spatial information. Such systems would enable alignment not only through textual similarity but also through visual semantics, layout structures, diagrams, and contextual cues. Incorporating temporal modeling across sequences of slides and speech can further improve coherence and enable more accurate alignment in dynamic presentation scenarios.

Another important direction is extending these systems to real-time and interactive environments, including live presentations, adaptive educational assistants, and intelligent tutoring systems. Incorporating non-verbal cues such as gestures, pointer movements, gaze, and emphasis signals could provide a more holistic understanding of presentation dynamics and human communication. Additionally, unified frameworks that jointly model alignment, retrieval, and generation may enable tighter coupling between understanding and synthesis, leading to more robust multimodal systems.

Finally, advancing fine-grained multimodal grounding—particularly for abstract, scientific, or non-visible concepts—remains a challenging research problem requiring stronger reasoning capabilities and

richer representations. Human-in-the-loop evaluation and feedback mechanisms will also play a critical role in ensuring that multimodal systems remain interpretable, reliable, and aligned with real-world user needs.

Overall, this thesis contributes toward building more robust, interpretable, and educationally meaningful multimodal systems, while highlighting key challenges that must be addressed to fully realize their potential across education, scientific communication, and knowledge-driven applications.

Bibliography

[1] Ø. Anmarkrud, A. Andresen, and I. Bråten, "Cognitive load and working memory in multimedia learning: Conceptual and measurement issues," *Educational psychologist*, vol. 54, no. 2, pp. 61–83, 2019.

[2] J. L. Plass and B. D. Homer, "Cognitive load in multimedia learning: The role of learner preferences and abilities," in *Proceedings of the International Conference on Computers in Education*, ser. ICCE '02, USA: IEEE Computer Society, 2002, p. 564.

[3] Y. Zhu et al., "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *ICCV*, 2015, pp. 19–27.

[4] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman, "Aligning subtitles in sign language videos," in *ICCV*, Oct. 2021, pp. 11 552–11 561.

[5] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial intelligence in education: A systematic literature review," *Expert systems with applications*, vol. 252, p. 124 167, 2024.

[6] V. Heilala, R. Araya, and R. Hämäläinen, "Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling," in *SIGAPP*, 2025.

[7] A. Bewersdorff et al., "Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education," *Learning and Individual Differences*, 2025.

[8] A. Yaacoub, S. Tarnpradab, P. Khumprom, Z. Assaghir, L. Prevost, and J. Da-Rugna, "Enhancing ai-driven education: Integrating cognitive frameworks, linguistic feedback analysis, and ethical considerations for improved content generation," in *2025 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2025, pp. 1–7.

[9] Khan Academy, *Harnessing ai so that all students benefit: A nonprofit approach for equal access*, `https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/`, 2024.

[10] Google, *Socratic by google help center*, `https://support.google.com/socratic/?hl=en`.

[11] H. Chen et al., "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[12] H. Chen et al., "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *CVPR*, 2024.

[13] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," in *ICLR*, 2023.

[14] Z. Yang et al., "Cogvideox: Text-to-video diffusion models with an expert transformer," in *ICLR*, 2025.

[15] X. Sun et al., "Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent," *arXiv preprint arXiv:2411.02265*, 2024.

[16] T. Wan et al., "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[17] J. Wang et al., "Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation," *arXiv preprint arXiv:2506.19774*, 2025.

[18] Y. Liu et al., "Sora: A review on background, technology, limitations, and opportunities of large vision models," *arXiv preprint arXiv:2402.17177*, 2024.

[19] L. Wang, N. Vincent, J. Rukanskaitė, and A. X. Zhang, "Pika: Empowering non-programmers to author executable governance policies in online communities," in *CHI*, 2024.

[20] O. Bar-Tal et al., "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia*, 2024.

[21] D. W. Lee, C. Ahuja, P. P. Liang, S. Natu, and L.-P. Morency, "Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos," in *ICCV*, 2023, pp. 20 030–20 041.

[22] B. Bahrani and M.-Y. Kan, "Multimodal alignment of scholarly documents and their presentations," in *ACM MM*, 2013, pp. 281–284.

[23] T. Hayama, H. Nanba, and S. Kunifuji, "Alignment between a technical paper and presentation sheets using a hidden markov model," in *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia: IEEE, Sep. 2005, pp. 102–106, ISBN: 0-7803-9035-0. DOI: 10.1109/AMT.2005.1505278.

[24] Q. Fan, K. Barnard, A. Amir, and A. Efrat, "Robust spatiotemporal matching of electronic slides to presentation videos," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2315–2328, 2011.

[25] X. Wang and M. Kankanhalli, "Robust alignment of presentation videos with slides," in *Pacific-Rim Conference on Multimedia*, Springer, 2009, pp. 311–322.

[26] K. Anderer, A. Reich, and M. Wölfel, "Mavils, a benchmark dataset for video-to-slide alignment, assessing baseline accuracy with a multimodal alignment algorithm leveraging speech, ocr, and visual features," in *Proceedings of Interspeech 2024*, Kos, Greece: ISCA, Sep. 2024, pp. 1375–1379.

[27] M. Tani, M. Manuguerra, and S. Khan, "Can videos affect learning outcomes? evidence from an actual learning environment," *Educational technology research and development*, vol. 70, Aug. 2022.

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, PMLR, 2023, pp. 28 492–28 518.

[29] M. Bain, J. Huh, T. Han, and A. Zisserman, *Whisperx: Time-accurate speech transcription of long-form audio*, 2023. arXiv: 2303.00747 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2303.00747.

[30] A. Shahabaz and S. Sarkar, "Increasing importance of joint analysis of audio and video in computer vision: A survey," *IEEE Access*, 2024.

[31] D. S. S, A. Gupta, C. V. Jawahar, and M. Tapaswi, "Unsupervised audio-visual lecture segmentation," in *WACV*, IEEE, 2023, pp. 5221–5230.

[32] M. Naphade and T. Huang, "Extracting semantics from audio-visual content: The final frontier in multimedia retrieval," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 793–810, 2002.

[33] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020.

[34] L. Kumari, S. Singh, V. Rathore, and A. Sharma, *Lexicon and attention based handwritten text recognition system*, 2022. arXiv: 2209.04817 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2209.04817.

[35] L. Kumari, S. Singh, V. V. S. Rathore, and A. Sharma, *A comprehensive handwritten paragraph text recognition system: Lexiconnet*, 2023. arXiv: 2205.11018 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2205.11018.

[36] E. Lakomkin, C. Wu, Y. Fathullah, O. Kalinli, M. L. Seltzer, and C. Fuegen, "End-to-end speech recognition contextualization with large language models," in *ICASSP*, IEEE, 2024, pp. 12 406–12 410.

[37] A. Antonacopoulos, B. Gatos, and D. Bridson, "Page segmentation competition," in *ICDAR*, vol. 2, 2007, pp. 1279–1283.

[38] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Icdar2017 competition on recognition of documents with complex layouts - rdcl2017," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 1404–1410. DOI: 10.1109/ICDAR.2017.229.

[39] W. Yu et al., "Icdar 2023 competition on structured text extraction from visually-rich document images," in *ICDAR*, Springer, 2023, pp. 536–552.

[40] H. Yang and W. H. Hsu, "Vision-based layout detection from scientific literature using recurrent convolutional neural networks," in *ICPR*, 2021, pp. 6455–6462.

[41] H. Yang and W. Hsu, "Transformer-based approach for document layout understanding," in *ICIP*, Oct. 2022, pp. 4043–4047.

[42] M. Haurilet, A. Roitberg, M. Martinez, and R. Stiefelhagen, "Wise — slide segmentation in the wild," in *ICDAR*, 2019, pp. 343–348.

[43] N. Daddaoua, J. Odobez, and A. Vinciarelli, "Ocr based slide retrieval," in *ICDAR*, 2005, 945–949 Vol. 2.

[44] R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito, "Slidevqa: A dataset for document visual question answering on multiple images," in *AAAI*, vol. 37, 2023, pp. 13 636–13 645.

[45] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *CVPR*, 2018.

[46] Z. Ding, X.-Y. Liu, M. Yin, and L. Kong, "Tgan: Deep tensor generative adversarial nets for large image generation," *arXiv preprint arXiv:1901.09953*, 2019.

[47] J. Ho et al., "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[48] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[49] Y. Guo et al., "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," in *ICLR*, 2024.

[50] L. Khachatryan et al., "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *ICCV*, 2023.

[51] U. Singer et al., "Make-a-video: Text-to-video generation without text-video data," in *ICLR*, 2023.

[52] Q. Xue, X. Yin, B. Yang, and W. Gao, "Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation," in *CVPR*, 2025.

[53] Z. Huang et al., "Vbench: Comprehensive benchmark suite for video generative models," in *CVPR*, 2024.

[54] Z. Huang et al., "Vbench++: Comprehensive and versatile benchmark suite for video generative models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–18, Nov. 2025. DOI: 10.1109/TPAMI.2025.3633890.

[55] D. Zheng et al., "Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness," *arXiv preprint arXiv:2503.21755*, 2025.

[56] M. Liao et al., "Evaluation of text-to-video generation models: A dynamics perspective," *NeurIPS*, 2024.

[57] H. Bansal et al., "Videophy: Evaluating physical commonsense for video generation," in *ICLR*, 2025.

[58] H. Bansal, C. Peng, Y. Bitton, R. Goldenberg, A. Grover, and K.-W. Chang, "Videophy-2: A challenging action-centric physical commonsense evaluation in video generation," *arXiv preprint arXiv:2503.06800*, 2025.

[59] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos, "Do generative video models understand physical principles?" In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 948–958.

[60] F. Meng et al., "Towards world simulator: Crafting physical commonsense-based benchmark for video generation," in *Proceedings of the 42nd International Conference on Machine Learning*, A. Singh et al., Eds., ser. Proceedings of Machine Learning Research, vol. 267, PMLR, Jul. 2025, pp. 43 781–43 806. [Online]. Available: https://proceedings.mlr.press/v267/meng25c.html.

[61] A. Frome et al., "Devise: A deep visual-semantic embedding model," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2013, pp. 2121–2129.

[62] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[63] J.-B. Alayrac et al., "Self-supervised multimodal versatile networks," in *Advances in Neural Information Processing Systems*, 2020.

[64] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019.

[65] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[66] Y. Li et al., "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems*, 2021.

[67] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[68] T. Jiang et al., "E5-V: universal embeddings with multimodal large language models," *CoRR*, vol. abs/2407.12580, 2024. DOI: 10.48550/ARXIV.2407.12580. arXiv: 2407.12580. [Online]. Available: https://doi.org/10.48550/arXiv.2407.12580.

[69] X. Zhang et al., "Bridging modalities: Improving universal multimodal retrieval by multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 9274–9285.

[70] N.-V. Nguyen, M. Coustaty, and J.-M. Ogier, "Multi-modal and cross-modal for lecture videos retrieval," in *ICPR*, 2014, pp. 2667–2672. DOI: 10.1109/ICPR.2014.461.

[71] H. Chen, M. Cooper, D. Joshi, and B. Girod, "Multi-modal language models for lecture video retrieval," in *ACM MM*, 2014, pp. 1081–1084. DOI: 10.1145/2647868.2654964.

[72] T.-J. Fu, W. Wang, D. McDuff, and Y. Song, "Doc2ppt: Automatic presentation slides generation from scientific documents," in *AAAI*, 2022, pp. 634–642. DOI: 10.1609/aaai.v36i1.19943.

[73] T. Sun et al., "P2p: Automated paper-to-poster generation and fine-grained benchmark," in *NeurIPS 2025*, 2025. [Online]. Available: https://arxiv.org/abs/2505.17104.

[74] Z. Zhu, K. Q. Lin, and M. Z. Shou, *Paper2video: Automatic video generation from scientific papers*, 2025. arXiv: 2510.05096 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2510.05096.

[75] J. Miao, J. R. Davis, Y. Zhang, J. K. Pritchard, and J. Zou, *Paper2agent: Reimagining research papers as interactive and reliable ai agents*, 2025. arXiv: 2509.06917 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2509.06917.

[76] J. Shi et al., "Presentagent: Multimodal agent for presentation video generation," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2025, pp. 760–773.

[77] L. Cagliero and M. La Quatra, "Automatic slides generation in the absence of training data," in *IEEE Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, 2021, pp. 103–108. DOI: 10.1109/COMPSAC51774.2021.00025.

[78] M. Sravanthi, C. R. Chowdary, and P. Kumar, "Slidesgen: Automatic generation of presentation slides for a technical paper using summarization," in *Proceedings of the International Conference on Intelligent Agent & Multi-Agent Systems (IAMA)*, IEEE, 2009. DOI: 10.1109/IAMA.2009.5228050. [Online]. Available: https://doi.org/10.1109/IAMA.2009.5228050.

[79] I. Mondal, S. S, A. Natarajan, A. Garimella, S. Bandyopadhyay, and J. Boyd-Graber, "Presentations by the humans and for the humans: Harnessing llms for generating persona-aware slides from documents," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2664–2684. DOI: 10.18653/v1/2024.eacl-long.163. [Online]. Available: https://aclanthology.org/2024.eacl-long.163/.

[80] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in neural information processing systems*, vol. 33, pp. 5776–5788, 2020.

[81] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3615–3620.

[82] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," in *ACL*, 2020.

[83] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3615–3620.

[84] B. Hui et al., *Qwen2.5-coder technical report*, 2024. arXiv: 2409.12186 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2409.12186.

[85] G. Team et al., "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[86] W. Wang et al., *Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency*, 2025. arXiv: 2508.18265 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2508.18265.

[87] F. Li et al., "LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models," in *ICLR*, 2025.

[88] Y. Wang et al., "Internvideo2: Scaling foundation models for multimodal video understanding," in *ECCV*, 2024.

[89] J. Li, S. Yu, H. Lin, J. Cho, J. Yoon, and M. Bansal, "Training-free guidance in text-to-video generation via multimodal planning and structured noise initialization," *ArXiv2504.08641*, 2025.

[90] K. Sun et al., "T2v-compbench: A comprehensive benchmark for compositional text-to-video generation," in *CVPR*, 2025.

[91] E. Sun, Y. Hou, D. Wang, Y. Zhang, and N. X. R. Wang, "D2S: Document-to-slide generation via query-based text summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova et al., Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 1405–1418. DOI: 10.18653/v1/2021.naacl-main.111. [Online]. Available: https://aclanthology.org/2021.naacl-main.111/.

[92] H. Wang, S. Kurita, S. Shimizu, and D. Kawahara, "Slideavsr: A dataset of paper explanation videos for audio-visual speech recognition," in *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, 2024, pp. 129–137.

[93] H. Wang, K. Hu, and L. Gao, "Docvideoqa: Towards comprehensive understanding of document-centric videos through question answering," in *ICASSP*, IEEE, 2025, pp. 1–5.

[94] J. Liu, A. Vats, and Z. He, "Cs-papersum: A large-scale dataset of ai-generated summaries for scientific papers," *CoRR*, vol. abs/2502.20582, 2025. DOI: 10.48550/ARXIV.2502.20582. arXiv: 2502.20582. [Online]. Available: https://doi.org/10.48550/arXiv.2502.20582.

[95] C. Clark and S. Divvala, "Pdffigures 2.0: Mining figures from research papers," in *ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL)*, 2016, pp. 143–152. DOI: 10.1145/2910896.2910904. [Online]. Available: https://doi.org/10.1145/2910896.2910904.

[96] T. Sun, C. Cui, Y. Du, and Y. Liu, *Pp-doclayout: A unified document layout detection model to accelerate large-scale data construction*, 2025. arXiv: 2503.17213 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2503.17213.

[97] Y. Du et al., *Pp-ocr: A practical ultra lightweight ocr system*, 2020. arXiv: 2009.09941 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2009.09941.

[98] M. Faysse et al., "Colpali: Efficient document retrieval with vision language models," in *International Conference on Learning Representations (ICLR)*, 2025.

[99] Qwen et al., *Qwen2.5 technical report*, 2025. arXiv: 2412.15115 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.15115.