Unifying Scientific Communication: Fine-Grained Correspondence Across Scientific Media

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

Vineeth N. Balasubramanian
Microsoft Research India & IIT Hyderabad
vineeth.nb@microsoft.com

C.V. Jawahar
IIIT Hyderabad
jawahar@iiit.ac.in

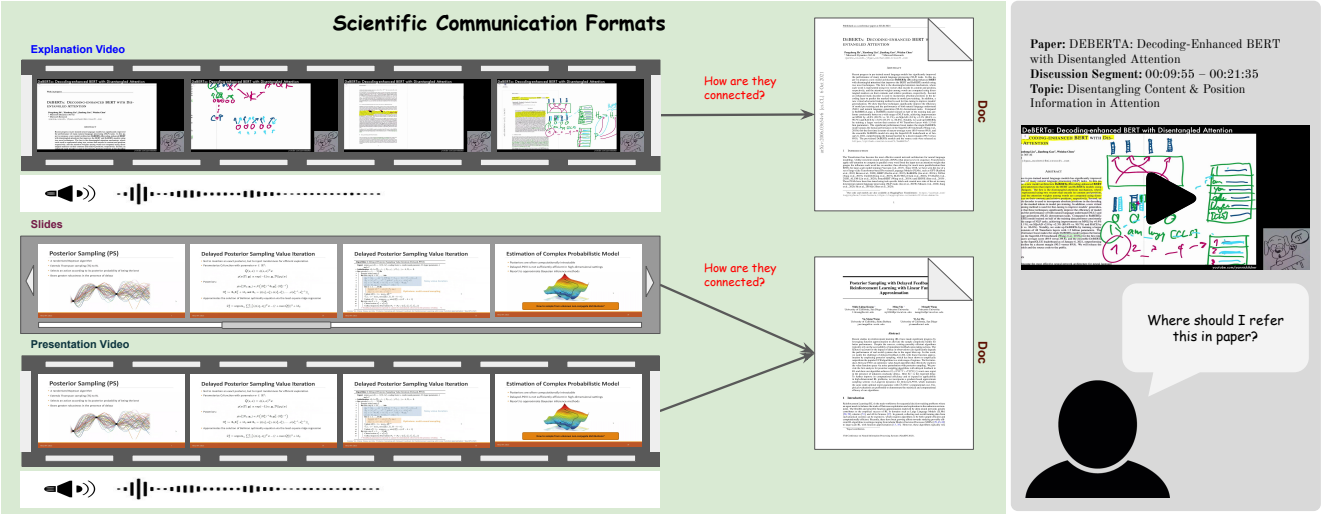

Figure 1. **Scientific communication formats and their interconnections:** The figure illustrates how research knowledge is represented and shared across multiple formats, including research papers (Docs), presentation slides, conference videos, and explanation videos. These formats are not isolated; rather, strong semantic and structural connections exist between 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.

Abstract

The communication of scientific knowledge has become increasingly multimodal, spanning text, visuals, and speech through materials such as research papers, slides, and recorded presentations. These different representations collectively convey a study's reasoning, results, and insights, offering complementary perspectives that enrich understanding. However, despite their shared purpose, such materials are rarely connected in a structured way. The absence of explicit links across formats makes it difficult to trace how concepts, visuals, and explanations correspond, limiting unified exploration and analysis of research content. To address this gap, we introduce the Multimodal Conference Dataset (MCD), the first benchmark that integrates research papers, presentation videos, explanatory videos, and slides from the same works. We evaluate a range of embedding-based and vision–language models to assess their ability to discover fine-grained cross-format correspondences, establishing the first systematic benchmark for this task. Our results show that vision–language models are robust but struggle with fine-grained alignment, while embedding-based models capture text–visual correspondences well but equations and symbolic content form distinct clusters in the embedding space. These findings highlight both the strengths and limitations of current ap-

proaches and point to key directions for future research in multimodal scientific understanding. To ensure reproducibility, we release the resources for MCD at link.

1. Introduction

In contemporary research, scientific communication extends far beyond traditional written papers. Discoveries and ideas are now shared through a rich ecosystem of materials—formal manuscripts, visual summaries, recorded presentations, and explanatory videos—each offering a distinct perspective on the same body of work. A written paper captures the complete technical depth, articulating methods, experiments, and analyses with precision. Visual summaries, such as slides highlight essential insights and illustrate complex ideas through concise design. Recorded presentations add the researcher's voice, tone, and emphasis, revealing intent and interpretation that are often missing from written text. Explanatory videos, on the other hand, take a step toward accessibility, recontextualizing dense scientific content into forms that are easier to grasp and communicate to wider audiences. Figure 1 shows these scientific communication formats—papers, slides, presentation videos, and explanation videos—and indicates their potential relationships using arrows.

Although these different forms describe the same research, they tend to exist in isolation. The links between them—how a figure in a paper connects to a slide image [16], or how a spoken explanation corresponds to an equation—are rarely documented [2, 3, 8, 13, 16, 34]. This lack of correspondence creates information silos, where valuable insights conveyed through one medium remain disconnected from others. Consequently, searching for, comparing, or comprehensively understanding research across its various representations has become challenging. For students, this limits opportunities to learn from complementary materials; for researchers, it restricts automated analysis and knowledge discovery; and for systems that aim to organize or recommend research content, it reduces interpretability and coherence.

Bridging these gaps requires structured alignment across different research representations.. Establishing meaningful connections between them can unlock new ways of exploring scientific knowledge, where a concept introduced in a paper can be directly linked to its visual explanation or a spoken commentary can guide a reader to the relevant figure or section. Such integration not only enhances accessibility and understanding but also supports the development of intelligent systems capable of reasoning across multiple forms of research communications. Motivated by this vision, we introduce a unified collection that brings together these diverse materials and offers a comprehensive view of how scientific ideas are conveyed, interpreted, and understood across formats.

We introduce the *Multimodal Conference Dataset (MCD)*, the first benchmark for evaluating fine-grained correspondences across research papers, slides, presentation videos, and explanatory videos. Using MCD, we evaluate six models—embedding-based and vision–language—across three traversal settings: EV $\rightarrow$ PP(explanatory video to paper), S $\rightarrow$ PP(slide to paper), and PV $\rightarrow$ PP(presentation video to paper), covering paper segments: paragraphs, figures, equations, and algorithms. The results are analyzed across traversals, model types, and sizes. Our study provides several key insights. Vision–language models are robust across modalities, but their broad generality can make fine-grained alignment challenging. In embedding-based models, equations/symbolic content form distinct clusters in embedding space, showing limited mixing with text and visuals. Despite this, models achieve effective retrieval. GME-2.2B and InternVL3.5-38B are demonstrating solid performance in these segments. Through this comprehensive evaluation, we uncover trends and limitations across modalities, highlighting where models succeed, where they fail, and what challenges remain for robust cross-format understanding in scientific communication.

2. Related Work

2.1. Cross-Modal Retrieval

Cross-modal retrieval aims to identify semantically corresponding information across different content types(text, images, audio, and video) enabling queries in one modality to retrieve relevant information in another. While widely studied in general domains such as image–text retrieval [1, 10, 25], video–text alignment [12, 21], and vision–language understanding [15, 17], these approaches typically focus on broad visual–linguistic content. Recent methods like E5-V [14] adapt multimodal large language models (MLLMs) to produce universal embeddings across modalities via prompt-based representations. At the same time, GME [36] fine-tunes MLLM retrievers on large fused-modal datasets to support single-modal, cross-modal, and fused-modal retrieval. In educational and research contexts, content is highly structured and semantically dense, combining precise text, abstract visuals, and spoken explanations. Early work aligned slides with papers [3, 13], later enhanced with visual cues and sequential modeling [3]. Research expanded to link presentation videos with slides [2, 8, 34], integrating speech transcripts, visual frames, and temporal synchronization [5, 23], as in the Google I/O dataset [5]. Fine-grained figure–text associations were explored [16], but prior work mostly considers pairwise correspondence. Our work addresses this gap by enabling comprehensive cross-modal retrieval across papers, slides, and videos through a unified benchmark.

2.2. AI for Research

The integration of Artificial Intelligence (AI) into the research lifecycle is transforming the creation, communication, and understanding of scientific knowledge. Traditionally, researchers manually prepared papers, slides, and presentations, but large language and multimodal models now enable automation and augmentation across these stages [11, 20, 26, 30, 37]. AI-driven systems assist in summarizing complex papers, generating multimodal representations such as slides [4, 11, 22, 27], videos [26, 37], and posters [30], and linking concepts across textual, visual, and spoken formats, enhancing efficiency and accessibility. Recent work has focused on automatic content generation from research papers. Slide generation models [4, 11, 27] select key sections, figures, and equations to produce coherent decks, while video generation systems [26, 37] synthesize narrated, visual presentations. Poster frameworks [30] provide concise, visually engaging summaries. Beyond static outputs, systems such as Paper2Agent [20] transform papers into interactive AI agents that explain methods, reason about results, and engage in dialogue with users. Emerging methods also align spoken content in talks with corresponding regions in papers or slides, enabling highlighting of relevant sections during presentations [19]. This supports audiences in following explanations while navigating materials and maintaining context. Such alignment reduces cognitive load [24, 38] and improves comprehension by directing attention to the most pertinent content. Together, these advances illustrate the role of AI in bridging modalities, reducing the effort required to prepare research artifacts, and enabling more interactive, multimodal, and accessible engagement with scientific knowledge. This convergence of AI and scholarly communication points to a dynamic, interpretable, and interconnected future for research dissemination.

3. Multimodal Conference Dataset

The *Multimodal Conference Dataset (MCD)* is a curated collection of research papers, presentation slides, presentation videos, and explanatory videos that capture multiple perspectives of the same research work. It is designed to support the study of how related information is connected and expressed across these formats. Table 1 presents a comparison of MCD with existing multimodal academic datasets.

Data Collection: To construct MCD, we compiled materials that co-refer to the same research work. The dataset comprises two sets: the presentation set and an explanation set. The presentation set consists of paper–slide–presentation video triplets collected from the NeurIPS 2023 conference. Research papers were retrieved from *arXiv*, while presentation slides and recorded talks were obtained from the *SlidesLive* platform, which archives official conference presentations along with slide decks used by speakers. The explanation set contains paper–explanatory video pairs, where explanatory videos were sourced from the *Yannic Kilcher* YouTube channel[1]. The corresponding papers were downloaded from the links provided in the video descriptions to ensure that the exact versions referenced in the explanations were used.

Dataset	Contains					Annot.	Task
	Sl.	Doc	PV	EV	Pos		
DOC2PPT [11]	✓	✓	✗	✗	✗	A	Slide Gen
Automatic slides generation [4]	✓	✓	✗	✗	✗	M	Slide Gen
SciDuet [28]	✓	✓	✗	✗	✗	A	Slide Gen
Persona-Aware D2S [22]	✓	✓	✗	✗	✗	A	Slide Gen
SlideAVSR [31]	✗	✗	✓	✗	✗	A+M	AVSR
DocVideoQA [32]	✗	✗	✓	✗	✗	A+M	VideoQA
CS-PaperSum [18]	✗	✗	✓	✗	✗	A	Summ.
Paper2Poster [30]	✗	✓	✗	✗	✓	A	Poster Gen
Doc2Present [26]	✓	✓	✓	✗	✗	A	Video Gen
Paper2Video [37]	✓	✓	✓	✗	✗	A	Video Gen
MCD	✓	✓	✓	✓	✗	A+M	Cross Trav

Table 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.

Data Preprocessing: For the presentation set, animated slides—where elements appeared incrementally across multiple slides-were identified, and only the final version containing all elements was retained to avoid redundancy. In such cases, the corresponding transcript segments up to the retained slides were concatenated to preserve the complete verbal context. Non-content slides, such as title, outline, acknowledgement, and closing (e.g., "Thank You") slides were removed. After forming paragraph-level paper segments, segments containing only a single character were discarded 3. Additionally, segments categorized as "abandon" by `PDFExtractor` were filtered out from the paragraph set, and overlapping figure and equation elements were handled to ensure that only meaningful textual and visual components were preserved for subsequent alignment/retrieval tasks.

Data Statistics: The dataset comprises two primary subsets: the Presentation Set and the Explanation Set. The Presentation Set consists of 20 paper–slide–presentation video triplets, where each presentation video has an average duration of approximately 5 minutes. Segmentation is performed at the slide level, with each segment corresponding to an individual slide and its associated ASR transcript. In contrast, the Explanation Set includes 15 paper–explanatory video pairs, with explanatory videos averaging approximately 40 minutes in duration. For this set, segmentation follows predefined topic boundaries, and the provided segments are used directly to preserve coherent thematic units. Transcripts for all videos across both sets are generated us-

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

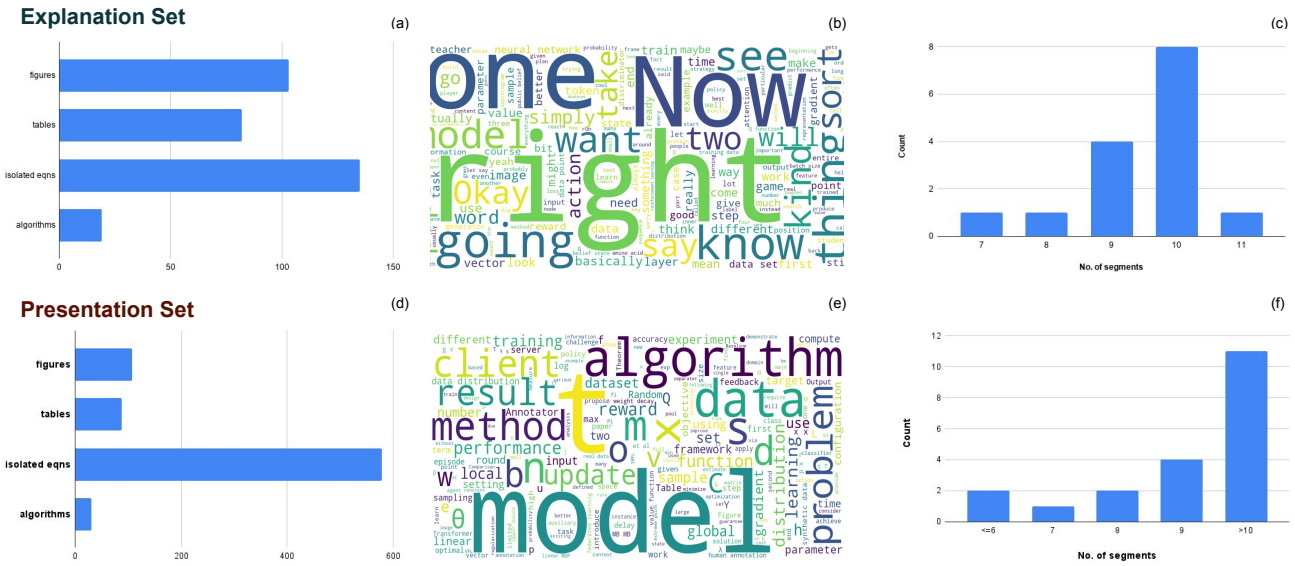

Figure 2. **Statistics of the Explanation and Presentation sets:** (a) and (d) show the distribution of algorithms, equations, tables, and figures in the papers; (b) presents the word cloud generated from ASR transcripts; (e) presents the word cloud generated from slide text and ASR transcripts; (c) and (f) depict the number of videos per segment category.

ing WhisperX. Table 2 summarizes the dataset, including the number of papers and slides, as well as the durations of Explanation Videos (EV) and Presentation Videos (PV).

Figure 2 summarizes the dataset statistics for both subsets. Subfigures (a) and (d) show the distribution of content types (algorithms, equations, tables, and figures) in the papers. Subfigures (b) and (e) present frequent words from the sources (Explanation Set: ASR; Presentation Set: slide OCR + ASR): the Explanation Set features conversational terms (e.g., "see," "right"), while the Presentation Set highlights technical terms (e.g., "model," "data," "algorithm"). Subfigures (c) and (f) depict segment counts per video, with the Explanation Set showing a more uniform distribution and the Presentation Set containing more videos with over ten segments.

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

Table 2. Dataset content overview: counts of papers and slides, and durations (in minutes) for videos (EV = Explanation Video, PV = Presentation Video).

Paper Segments: Each research paper comprises diverse elements such as text, figures, algorithms, and equations that collectively convey the study's content. To systematically capture these components, we categorized the paper into four segment types: *paragraphs*, *equations*, *figures* (including associated captions), and *algorithms*. Figures are extracted using PDFFIGURES [6]. Algorithmic regions are detected with PP-DOCLAYOUT-L [29], from which bounding boxes are obtained and subsequently processed with PADDLEOCR [7] to extract the text. Equations are identified and recognized using the PDFEXTRACTOR toolkit[2], specifically employing its formula detection and recognition modules. A preprocessing stage ensures that extracted equations and figure regions are non-overlapping. These detected components are then removed from the PDF, and the remaining text is processed with SCIENCEPARSE[3], which segments the document into structured components such as *abstract*, *sections*, and *authors*. The *paragraph* segments were subsequently obtained from these textual components. Figure 3 illustrates the overall extraction pipeline.

Annotation: To establish fine-grained cross-modal correspondence, each source segment—whether a slide, a presentation video segment (slide + transcript), or an explanatory video segment (transcript)—is manually aligned with its corresponding paper segments. This ensures that all semantically relevant parts of the paper are accurately identified for every source segment. The dataset contains 15 explanatory videos and 20 presentation videos. The query set comprises 460 queries containing at least one paragraph, including 147 with at least one relevant figure, 121 with at least one relevant equation, and 56 with at least one relevant algorithm.

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

[3] https://reurl.cc/e62LXL

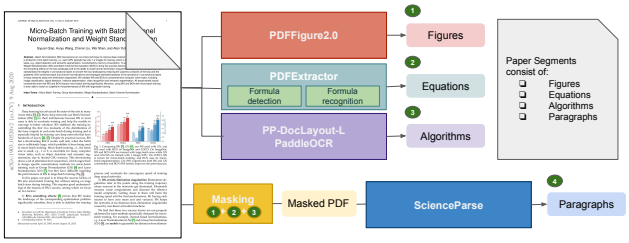

Figure 3. Pipeline for extracting paper segments—figures, equations, algorithms, and paragraphs—from the paper PDF.

4. Cross Linking Across Modalities

4.1. Cross Modal Grounding

We address the task of linking multimodal research materials—explanatory videos, presentation videos, and slides—to their corresponding paper content. Each source provides a distinct perspective on the same work: the explanatory video emphasizes conceptual flow, the presentation video blends visuals and narration, and the slides convey condensed visual cues.

A research paper comprises multiple elements, including paragraphs, figures, equations, and algorithms. Here, a figure is defined as the visual image along with its associated captions (tables and their associated captions are also considered part of the figure). Given a source segment from any format—explanatory video, presentation video, or slide—the goal is to establish *fine-grained, one-to-many correspondence* with the paper by identifying and ranking the most relevant elements. Formally, the task is:

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

where x denotes an input segment from a source modality and $\{p_i\}$ represents the set of paper elements semantically related to x. This formulation captures that a single segment may correspond to multiple paper components conveying the same underlying concept.

We consider three retrieval settings, each with a distinct query type. For explanatory videos, the query consists of the transcript of the video segment (EV→PP). For slides, the query is the slide image (S→PP). For presentation videos, the query combines both the slide image and the corresponding transcript (PV→PP). Performance is assessed using *NDCG@K* for paragraphs, figures, and equations, which measures how well the ranked paper elements align with human-annotated relevance judgments. For algorithms, we set a threshold of 0.6 and compute recall to evaluate retrieval.

4.2. Are MLLMs ready?

We evaluate six models covering embedding-based and vision–language models (VLMs). The embedding-based models include E5-V [14], GME (2.2B, 8.2B) [36], and ColQwen [9]. E5-V produces universal embeddings by adapting multimodal large language models (MLLMs) with text-pair training and prompt-based bridging to align modalities, showing strong performance in image–text and document retrieval. GME supports single-modal, cross-modal, and fused-modal retrieval, with 2.2B and 8.2B variants (with and without instruction). Its large-scale instruction-based training on fused-modal datasets enables unified embeddings for text, images, and visual documents, making it effective for fine-grained correspondence between source segments and paper elements. ColQwen integrates multimodal information to produce unified representations for cross-modal retrieval.

The VLMs—InternVL-4B, InternVL3.5-38B, and Qwen2.5-32B—are selected for their strong performance in OCR and document understanding. InternVL-4B is lightweight and efficient, InternVL3.5-38B provides high-quality multimodal comprehension, and Qwen2.5-32B offers strong visual–linguistic alignment. VLMs appear to understand individual media in tasks such as summarization, rephrasing, and question answering. We probe the depth of this understanding through the task of cross-linking and cross-grounding. Together, these embedding-based and VLM approaches enable evaluation across the three traversal tasks—EV→PP, S→PP, and PV→PP—covering single-, cross-, and fused-modal retrieval. This setup ensures that each model's strengths are leveraged to assess fine-grained multimodal correspondence in structured research content.

5. Performance Analysis

In this section, we analyze the retrieval performance of all models across the three traversal settings—explanatory video to paper (EV → PP), slides to paper (S → PP), and presentation video to paper (PV → PP)—and across different paper segment types, including paragraphs (par), figures (fig), equations (eq), and algorithms (algo) (Table 3). Our analysis is structured along four dimensions: (i) traversal-specific trends, (ii) embedding-based models, (iii) vision–language models, and (iv) the effect of model size, providing a comprehensive assessment of fine-grained multimodal alignment.

5.1. Across Traversals

Performance varies across the three traversals—EV→PP, S→PP, and PV→PP—reflecting differences in multimodal correspondence (Table 3). The S→PP traversal achieves the highest scores in majority cases due to strong visual–textual alignment between slides and paper content. Figures and equations on slides often directly reuse or paraphrase material from papers; however, algorithm retrieval remains challenging for most embedding-based models. In contrast, EV→PP is the most difficult to traverse. Explanatory video

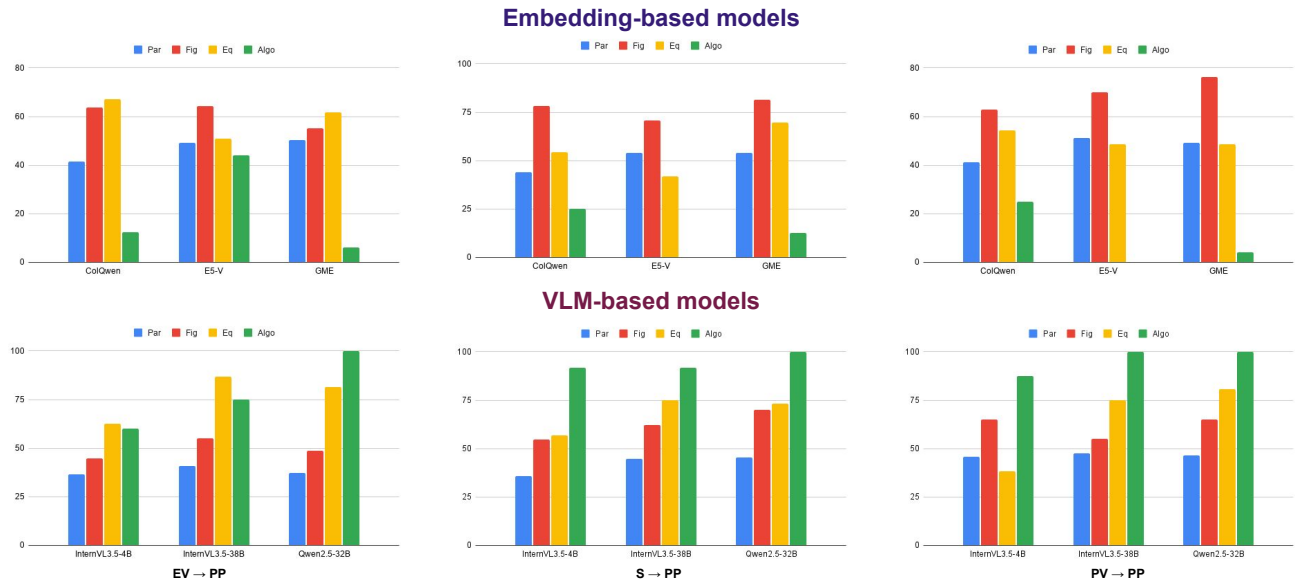

Figure 4. Illustrates the performance of embedding-based and VLM-based models across all three traversal settings (EV → PP, S → PP, PV → PP). NDCG@2 scores are used for paragraphs, figures, and equations. For algorithms, recall is calculated using a threshold of 0.6.

transcripts contain rich narrative language that diverges from concise paper phrasing, reducing paragraph-level retrieval accuracy ($\leq$ 53 NDCG@1 for the best embedding model). However, models such as GME-Qwen2VL-2B (with instructions) maintain stable figure-level performance, suggesting that conceptual anchors in narration aid alignment. PV→PP combines the strengths of the other traversals. Slides provide visual grounding, and narration adds interpretive context. The performance is comparable to S→PP and exceeds EV→PP in the majority of cases. Overall, the trend S→PP > PV→PP > EV→PP shows that retrieval correlates with semantic and structural continuity: clearer alignment supports stronger grounding, whereas stylistic divergence challenges fine-grained retrieval.

5.2. Embedding-Based Models

Figure 5 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 → PP, S → PP, and PV → PP traversals. Equation embeddings form compact, isolated clusters with limited mixing with textual and visual embeddings. In ColQwen, this separation is most pronounced, with equation clusters almost entirely distinct, indicating strong modality-specific encoding but weaker cross-modal integration. E5-V and GME show partial mixing: equation clusters remain identifiable but slightly overlap with paragraphs and figures, implying moderately shared semantic space.

Table 3 reports quantitative performance. GME consistently performs strongly across traversals, with the 2.2B variants outperforming other embedding-based baselines

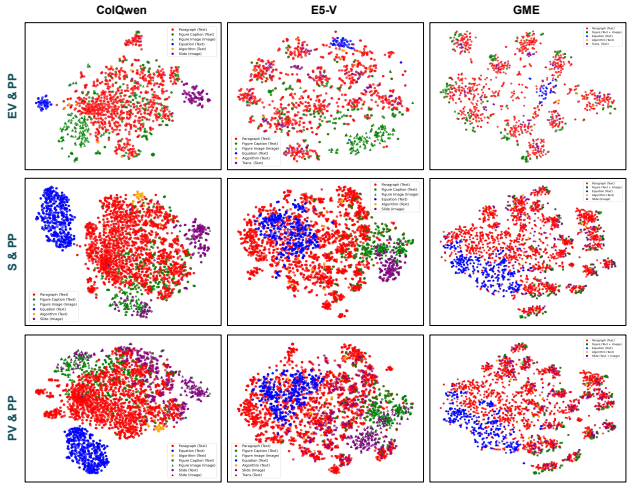

Figure 5. Visualization of query and candidate embeddings for representative instances across all three settings: EV → PP, S → PP, and PV → PP. Zoom in for a better view.

for paragraphs, figures, and equations in most cases (but struggles with algorithms—unable to find any relevant matches). Across both the 2.2B and 7B variants, the use of instruction prompts does not lead to consistent improvements in performance. The GME-7B variant failed to retrieve relevant algorithms for S → PP. E5-V shows overall decent performance, but remains lower than GME in most cases. ColQwen2-VL has lower paragraph and figure scores but comparatively better equation retrieval in EV → PP and S → PP, with moderate performance in PV → PP. Notably, it achieves perfect scores for all algorithms across all traversals. Figure 4 (top row) presents NDCG@2 scores for all three models across paper segments.

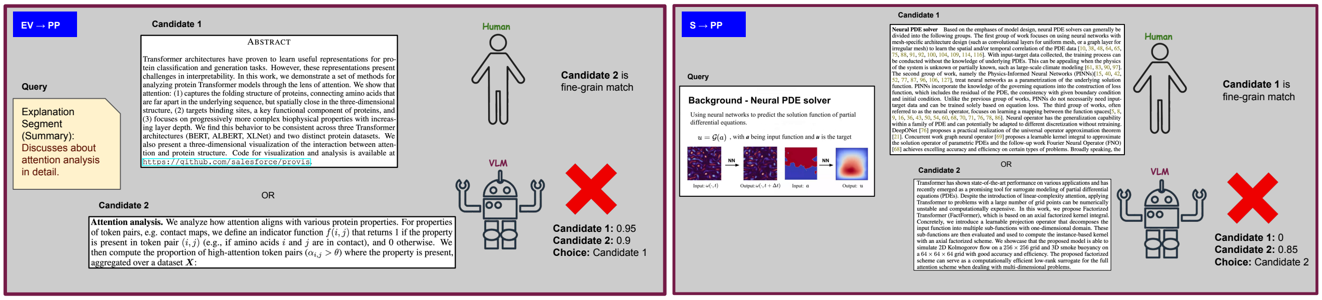

Figure 6. The figure presents two qualitative examples illustrating typical VLM failure cases. In the left example, the explanation segment discusses attention analysis in detail, as described for Candidate 2. However, the VLM assigns a higher similarity score to the abstract, which offers only a broad 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.

5.3. Vision Language Models

We evaluate three vision–language models (VLMs)—two from the InternVL3.5 family (4B and 38B) and one from Qwen2.5-VL (32B)—along with proprietary Gemini models. Across most traversal settings, open-source VLMs exhibit lower performance all paragraph and figure retrieval in to embedding-based methods. This can be attributed to their inherently broad semantic sensitivity: they are robust at identifying subtle cross-modal relationships, often assigning high similarity scores to multiple partially relevant paper segments. While this reflects strong semantic understanding, it reduces precision of fine-grained segment-level correspondence, which is central to our task. In other words, although many candidate segments may be topically relevant, the task requires exact alignment, and VLMs' generalization tendency leads to weaker discrimination among closely related candidates (see Figure 5).

Interestingly, VLMs perform comparatively better in equation retrieval, likely due to the structured and symbolic nature of equations, which provide clearer alignment cues. The performance of the algorithms is consistently better than that of embedding-based models. All reported results for VLMs are based on few-shot inferences. Despite task-specific prompting and examples emphasizing fine-grained matching, these models tend to favor coarse semantic retrieval, excelling at capturing general relationships but lacking precision in detailed alignment.

In contrast, the closed-source Gemini models (Flash and Pro) consistently outperform open-source VLMs across all segment types and traversals. Gemini-2.5 Pro achieves the strongest overall results, with notable gains in paragraph and figure retrieval, while Flash offers competitive performance at a lower computational cost. Unlike open-source VLMs, Gemini models strike a better balance between semantic generalization and fine-grained matching, enabling more accurate retrieval.

Overall, the performance gap highlights a clear distinction: open-source VLMs are effective at capturing broad multimodal semantics but struggle with precise grounding, whereas closed-source models demonstrate stronger fine-grained alignments. Figure 4 (bottom row) shows the NDCG@2 scores for InternVL-38B, InternVL-4B, and Qwen-32B(Open-source VLMs) across different paper segments.

5.4. Effect of Model Size

Within the GME variants 2.2B and 8.2B, an increase in model size does not consistently translate to improved performance. The 2.2B variant often matches or even surpasses the 8.2B model across multiple traversals, indicating that scaling alone does not guarantee better multimodal alignments. No clear trend correlating the model size with the retrieval quality is observed among the embedding-based methods. Similarly, larger embedding-based models such as E5-V (8.4B) and GME-8.2B do not exhibit substantial gains over smaller counterparts like GME-2.2B or ColQwen2-VL (2.2B). In contrast, among the VLM-based models, increasing model size leads to significant and consistent improvements, particularly within the InternVL family. The performance rise is especially pronounced for equation retrieval, where the larger models demonstrate a substantial advantage across all three traversal settings (EV→PP, S→PP, and PV→PP), highlighting the benefits of scaling in vision–language architectures.
Overall, these observations indicate that model performance depends on a combination of traversal type, segment modality, and model architecture. Embedding-based models tend to struggle with algorithms, while VLMs struggle with paragraphs; however, VLMs show notable gains from scaling, particularly for equations. This underscores that effective fine-grained retrieval relies on both modality alignment and model design, shaping performance across papers, slides, and videos.

6. Conclusion

We introduce the *Multimodal Conference Dataset (MCD)*, the first benchmark for fine-grained correspondences across

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 [9]	2.21B	47.27	41.52	40.37	61.40	63.80	63.16	67.01	12.50
E5-V [14]	8.4B	50.29	49.05	47.50	62	64.40	39.71	50.84	44
GME-Qwen2VL-2B w/o instr. [36]	2.2B	53.64	48.10	46.28	54.39	56.96	57.89	61.75	6.25
GME-Qwen2VL-2B with instr. [36]	2.2B	53.64	50.21	49.70	50.88	55.23	52.63	61.84	6.25
GME-Qwen2VL-7B w/o instr. [36]	8.2B	47.27	44.90	45.18	38.60	45.67	47.37	47.90	62.5
GME-Qwen2VL-7B with instr. [36]	8.2B	50.91	48.19	46.47	47.37	51.30	47.37	53.26	62.5
InternVL3.5-4B [33]	4B	36.29	36.45	34.79	39.34	44.68	61.90	62.39	60
InternVL3.5-38B [33]	38B	39.09	40.72	40.71	47.37	55.12	84.21	86.78	75
Qwen2.5-VL-32B [35]	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 [9]	2.21B	47.88	44.02	45.03	68.18	78.22	50.98	54.21	25
E5-V [14]	8.4B	56.97	54.02	54.80	63.64	70.81	43.14	41.82	0
GME-Qwen2VL-2B w/o instr. [36]	2.2B	58.79	53.58	54.48	75	83.60	68.63	70.54	8.33
GME-Qwen2VL-2B with instr. [36]	2.2B	58.18	54.08	54.76	72.73	81.33	68.63	69.78	12.50
GME-Qwen2VL-7B w/o instr. [36]	8.2B	52.12	50	51.45	77.27	83.01	66.67	66.39	0
GME-Qwen2VL-7B with instr. [36]	8.2B	55.15	52.35	53.89	77.27	85.88	68.63	68.35	0
InternVL3.5-4B [33]	4B	33.33	35.92	37.71	40.91	54.69	50.98	56.97	91.67
InternVL3.5-38B [33]	38B	42.42	44.58	47.55	50.00	62.03	70.59	75.06	91.67
Qwen2.5-VL-32B [35]	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 [9]	2.21B	46.49	41.16	41.90	47.83	62.91	50.98	54.21	25
E5-V [14]	8.4B	54.05	51.29	53.91	60.87	69.94	47.06	48.49	0
GME-Qwen2VL-2B w/o instr. [36]	2.2B	52.97	48.06	48.57	65.22	76.72	49.02	49.70	4.17
GME-Qwen2VL-2B with instr. [36]	2.2B	55.14	49.24	49.85	67.39	76.15	45.10	48.53	4.17
GME-Qwen2VL-7B w/o instr. [36]	8.2B	55.14	49.39	49.63	69.57	79.70	60.78	61.26	0
GME-Qwen2VL-7B with instr. [36]	8.2B	55.68	49.30	50.02	65.22	78.40	62.75	62.47	4.17
InternVL3.5-4B [33]	4B	46.56	45.88	47.32	59.18	65.12	37.25	38.21	87.50
InternVL3.5-38B [33]	38B	49.73	47.70	48.98	47.83	55.21	75.00	75.00	100
Qwen2.5-VL-32B [35]	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 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 (%).

research papers, slides, presentation videos, and explanatory videos. MCD evaluates six embedding-based and vision–language models across three traversals—EV → PP, S → PP, and PV → PP. Vision–language models are robust but less precise for fine-grained alignment, while embedding-based models capture text–visual correspondences; equations and symbolic content form distinct clusters, showing limited mixing yet remaining retrievable. GME-2.2B and InternVL3.5-38B perform well. Overall, MCD provides a unified framework for benchmarking cross-format scientific retrieval, highlighting model strengths and limitations and guiding future research in fine-grained correspondence discovery.

Acknowledgments

This work is supported by the MeitY, Government of India, through the NLTM Bhashini project (https://bhashini.gov.in). We sincerely thank the anonymous reviewers for their valuable feedback, which helped improve the quality of this paper.

References

[1] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, Andrew Zisserman, and Joao Carreira. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems*, 2020. 2

[2] Katharina Anderer, Andreas Reich, and Matthias 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*, pages 1375–1379, Kos, Greece, 2024. ISCA. 2

[3] Bamdad Bahrani and Min-Yen Kan. Multimodal alignment of scholarly documents and their presentations. In *ACM MM*, page 281–284, 2013. 2

[4] Luca Cagliero and Moreno La Quatra. Automatic slides generation in the absence of training data. In *IEEE Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, pages 103–108, 2021. 3

[5] Huizhong Chen, Matthew Cooper, Dhiraj Joshi, and Bernd Girod. Multi-modal language models for lecture video retrieval. In *ACM MM*, pages 1081–1084, 2014. 2

[6] Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL)*, pages 143–152, 2016. 4

[7] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. *CoRR*, abs/2009.09941, 2020. 4

[8] Quanfu Fan, Kobus Barnard, Arnon Amir, and Alon Efrat. Robust spatiotemporal matching of electronic slides to presentation videos. *IEEE Transactions on Image Processing*, 20(8):2315–2328, 2011. 2

[9] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations (ICLR)*, 2025. 5, 8

[10] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 2121–2129, 2013. 2

[11] Tsu-Jui Fu, William Wang, Daniel McDuff, and Yale Song. Doc2ppt: Automatic presentation slides generation from scientific documents. In *AAAI*, pages 634–642, 2022. 3

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[13] Tessai Hayama, Hidetsugu Nanba, and Susumu 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)*, pages 102–106, Melbourne, Australia, 2005. IEEE. 2

[14] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: universal embeddings with multimodal large language models. *CoRR*, abs/2407.12580, 2024. 2, 5, 8

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 2

[16] Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *ICCV*, pages 20030–20041, 2023. 2

[17] Yuan Li, Haoxuan Lin, Deyao Zhou, Bin Zhao, Zhiqiang Guan, Jinqiao Wang, and Shiliang Pu. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, 2021. 2

[18] Javin Liu, Aryan Vats, and Zihao He. Cs-papersum: A large-scale dataset of ai-generated summaries for scientific papers. *CoRR*, abs/2502.20582, 2025. 3

[19] Megha Mariam K M and C. V. Jawahar. Attend to what i say: Highlighting relevant content on slides. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2025. 3

[20] Jiacheng Miao, Joe R. Davis, Yaohui Zhang, Jonathan K. Pritchard, and James Zou. Paper2agent: Reimagining research papers as interactive and reliable ai agents, 2025. 3

[21] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[22] Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan 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)*, pages 2664–2684, St. Julian's, Malta, 2024. Association for Computational Linguistics. 3

[23] Nhu-Van Nguyen, Mickaël Coustaty, and Jean-Marc Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *ICPR*, pages 2667–2672, 2014. 2

[24] Jan L. Plass and Bruce D. Homer. Cognitive load in multimedia learning: The role of learner preferences and abilities. In *Proceedings of the International Conference on Computers in Education*, page 564, USA, 2002. IEEE Computer Society. 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 2

[26] Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. PresentAgent: Multimodal agent for presentation video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 760–773, Suzhou, China, 2025. Association for Computational Linguistics. 3

[27] 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. 3

[28] Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy 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*, pages 1405–1418, Online, 2021. Association for Computational Linguistics. 3

[29] Ting Sun, Cheng Cui, Yuning Du, and Yi Liu. Pp-doclayout: A unified document layout detection model to accelerate large-scale data construction, 2025. 4

[30] Tao Sun, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongliang Li, Wenhao Huang, Ge Zhang, Jian Yang, and Zhoujun Li. P2p: Automated paper-to-poster generation and fine-grained benchmark. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. 3

[31] Hao Wang, Shuhei Kurita, Shuichiro Shimizu, and Daisuke 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)*, pages 129–137, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[32] Haochen Wang, Kai Hu, and Liangcai Gao. Docvideoqa: Towards comprehensive understanding of document-centric videos through question answering. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE, 2025. 3

[33] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, JingJing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025. 8

[34] Xiangyu Wang and Mohan Kankanhalli. Robust alignment of presentation videos with slides. In *Advances in Multimedia Modeling: 16th International Conference, MMM 2010*, pages 311–322, Chongqing, China, 2010. Springer. 2

[35] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. 8

[36] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 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)*, pages 9274–9285, 2025. 2, 5, 8

[37] Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. Paper2video: Automatic video generation from scientific papers. *CoRR*, abs/2510.05096, 2025. 3

[38] Anette Andresen Øistein Anmarkrud and Ivar Bråten. Cognitive load and working memory in multimedia learning: Conceptual and measurement issues. *Educational Psychologist*, 54(2):61–83, 2019. 3