Unifying Scientific Communication: Fine-Grained Correspondence Across Scientific Media

https://meghamariamkm2002.github.io/mcd_2026/

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

A. Benchmark Utility

The core contribution of this benchmark is to enable seamless navigation and alignment across different forms of scientific communication—such as research papers, presentation slides, recorded talks, and explanatory videos—each of which presents information with a distinct structure, level of detail, and perspective. While all these media aim to convey the same underlying concepts, they do so in complementary ways: papers provide formal rigor and completeness, slides emphasize key ideas and visual summaries, presentation videos capture the narrative flow and intent of the speaker, and explanation videos often simplify and reinterpret concepts for clarity.

This diversity, while valuable, creates a practical challenge for learners and researchers: connecting corresponding pieces of information across modalities is non-trivial. A concept introduced in slides or a presentation video may not map to a single element, but instead correspond to multiple components within a paper—for example, a detailed paragraph, a formal equation, and an associated figure. Similarly, content from explanatory videos may align with complementary parts of the paper that provide intuition, derivation, or formal grounding. Identifying these one-to-many correspondences requires both semantic understanding and fine-grained alignment across modalities.

Additionally, by enabling such fine-grained alignment, the benchmark can also support the evaluation of paper-to-slide and paper-to-video generation systems, as well as the probing of multimodal LLM grounding in scientific domains. It is not intended for end users today, but rather for evaluating systems that will power such tools.

B. Dataset Details

The dataset consists of 15 explanatory videos and 20 presentation videos. It includes 142 explanatory video segments, of which 128 have at least one corresponding relevant paper segment. Additionally, there are 204 slides and 204 presentation video segments, of which 200 have at least one relevant paper segment. 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.

B.1. Preprocessing and Data Quality

ASR Transcripts. All video transcripts are generated using *WhisperX*, ensuring consistent and high-quality automatic speech recognition across the dataset. To validate transcription reliability, we performed manual spot-checks on a sampled subset of videos, observing an accuracy of greater than 90%.

Slide Content We do not apply explicit OCR to slide images. Instead, slide-based modalities rely directly on raw visual inputs. In the *S→PP* setting, models are provided with full slide images, requiring them to jointly interpret visual structure and embedded text. In the *PV→PP* setting, slide images are paired with ASR transcripts, enabling complementary use of visual and spoken information.

Timestamp Alignment. Temporal alignment between video and corresponding content is derived from reliable, source-specific signals rather than post hoc synchronization. For *EV→PP*, we use explicit topic timestamps provided in YouTube videos, which segment content into semantically meaningful units. For *PV→PP*, we leverage

the native alignment between slides and speech available in SlidesLive recordings.

B.2. Annotation Details

A pilot study with four annotators on a sampled MCD subset showed complete agreement, indicating that the task is largely unambiguous in the scientific literature domain. This high level of agreement is likely due to the structured and precise nature of scientific content. Based on this observation, the remainder of the dataset was annotated by a single high-quality expert to ensure consistency and efficiency.

While the segment types are clearly defined, special care was taken to standardize the criteria for determining correct correspondences, particularly in cases that may appear ambiguous or partial. In general, annotations are guided by conceptual relevance rather than strict surface-level matching. For instance, figures in slides are considered aligned with those in papers even when they are cropped, reformatted, or stylistically modified, provided they convey the same underlying concept; in such cases, both visual content and accompanying captions are taken into account. Similarly, equation alignment does not require exact symbolic equivalence, but instead focuses on whether the expressions represent the same concept or play an equivalent role in the explanation. These guidelines facilitate systematic and consistent annotation.

C. Baseline Details

In this section, we discuss the details regarding the baseline models, both vision language models and embedding-based models.

All models are evaluated in their default, off-the-shelf settings without any fine-tuning or parameter modification. This includes using each model with its standard resolution, context length, and inference configurations as provided by their respective implementations. Closed-source models such as Gemini are also used under their default API settings

C.1. Embedding-based models

For the embedding-based baselines, ColQwen, E5-V, and four configurations of GME are evaluated, corresponding to two model sizes (2.2B and 8.2B), each used with and without instructions. Figure 1 shows the instruction prompts used for the GME variants.

For ColQwen and E5-V, when a candidate or query contains fused content (e.g., an image and its caption), similarity scores are computed independently for each component and the maximum is taken. In contrast, GME uses a single fused embedding for such multimodal inputs.

All images are provided to the respective model processors in their original resolution, without any manual resizing or preprocessing, allowing each model to internally handle image scaling according to its default pipeline. Image resolution can influence embedding quality and retrieval performance; using the native preprocessing of each model ensures consistency in evaluation and avoids introducing model-specific biases through external adjustments.

C.2. Vision language models

To encourage fine-grained correspondence, we include keywords such as "direct" and "fine-grained" in the VLM prompts. We use all vision-language models in their default settings, without any additional fine-tuning or parameter modification. Because supplying all paper segments (e.g., paragraphs) often exceeds the model's maximum context length—and evaluating segments one by one is computationally expensive—we provide the model with a subset of segments at a time. The VLM is instructed to evaluate each segment in isolation and assign a relevance score. These scores are then sorted in descending order to compute rank-based retrieval metrics. Figure 2 shows the prompt used.

Since slide images are provided without explicit OCR, the retrieval performance of VLMs inherently depends on their implicit ability to read and interpret textual content within images. To ensure that the task remains well-posed and does not disproportionately disadvantage certain models, we conducted preliminary checks verifying that the evaluated VLMs can reliably extract and reproduce the textual content from slides. This ensures that the task primarily evaluates cross-modal alignment rather than raw text recognition ability.

D. Evaluation Details

For paragraph, figure, and equation retrieval, we report NDCG@K to evaluate ranking quality. Algorithm retrieval is treated separately, as most papers contain at most one algorithm, making NDCG@1 uninformative (i.e., trivially 100 when the correct item is retrieved).

We instead evaluate algorithm retrieval using recall at a similarity threshold. Specifically, we sweep thresholds from 0.3 to 0.95 in increments of 0.05 for *EV*→*PP*, *S*→*PP*, and *PV*→*PP*, and select 0.6 as a stable operating point. Lower thresholds tend to introduce false positives, while higher thresholds are overly strict and reduce recall. Precision is not reported, as most queries have only one relevant algorithm(not all), making recall the more informative metric. There are cases where queries have a single relevant algorithm, making retrieval effectively binary, which is one of the reasons we observe extreme values (0 or 100) for algorithms.

D.1. Annotation Granularity

Paper segments are annotated at four levels: paragraphs (text), figures (image + caption, including tables), equations (text), and algorithms (text). Queries are defined as ASR

transcripts for explanatory videos, slide images for slides, and both for presentation videos. .

E. Ablation

E.1. 0-Shot and k-Shot Performance Analysis

From Table 3, we observe that, in most cases, the k-shot setting outperforms the zero-shot setting. Providing a small number of input–output examples helps vision–language models establish finer-grained correspondences. Notably, improvements are consistent across paragraph, figure, and equation retrieval. In contrast, the performance gains for algorithm retrieval are negligible, suggesting that example-based guidance offers limited additional benefit for this modality.

F. Additional Dataset Details

Modality	Avg (#)	Max (#)	Min (#)
Slides (S)	108.17	476	6
Slide + Transcript (PV)	156.8	582	39
Transcript (EV)	615.84	2759	28

Table 1. Word count statistics across the different content types(EV, PV, S), showing average, maximum, and minimum counts.

Category	Presentation Set	Explanation Set
Algorithm	30	19
Equation	579	135
Image + Caption	107	103
Table + Caption	88	82

Table 2. Distribution of paper segments—algorithms, equations, images, and tables—across the Presentation and Explanation Sets.

References

[1] 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. 4

[2] 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. 4

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

Table 3. 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 (%).

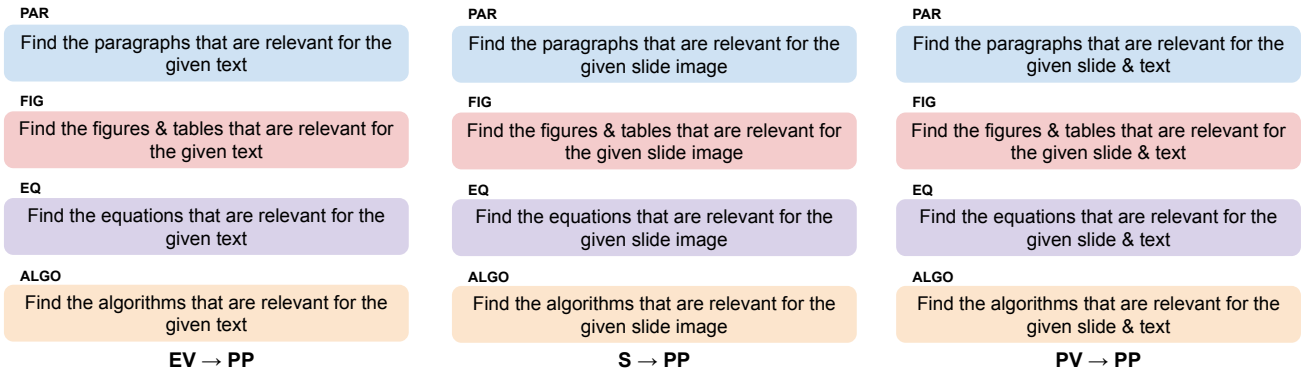

Figure 1. The figure illustrates the instruction prompts used for the GME models across the three traversal settings: EV → PP, S → PP, and PV → PP.

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