POSTERSUM: A Multimodal Benchmark for Scientific Poster Summarization

Rohit Saxena **Pasquale Minervini** **Frank Keller**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
rohit.saxena@ed.ac.uk p.minervini@ed.ac.uk keller@inf.ed.ac.uk

Abstract

Generating accurate and concise textual summaries from multimodal documents is challenging, especially when dealing with visually complex content like scientific posters. We introduce POSTERSUM[1][2], a novel benchmark to advance the development of vision-language models that can understand and summarize scientific posters into research paper abstracts. Our dataset contains 16,305 conference posters paired with their corresponding abstracts as summaries. Each poster is provided in image format and presents diverse visual understanding challenges, such as complex layouts, dense text regions, tables, and figures. We benchmark state-of-the-art Multimodal Large Language Models (MLLMs) on POSTERSUM and demonstrate that they struggle to accurately interpret and summarize scientific posters. We propose SEGMENT & SUMMARIZE, a hierarchical method that outperforms current MLLMs on automated metrics, achieving a 3.14% gain in ROUGE-L. This will serve as a starting point for future research on poster summarization.

1 Introduction

Scientific posters play a critical role in academic communication, offering a visually rich medium that combines text, images, charts, and other graphical elements to present research findings. Summarizing these visually complex posters into concise and accurate textual abstracts presents a unique challenge, requiring models to integrate multimodal information effectively. Multimodal Large Language Models (MLLMs; OpenAI et al., 2024; Grattafiori et al., 2024) demonstrated remarkable capabilities in vision-and-language tasks, including image captioning (Fu et al., 2024; Koh et al., 2023; Yu et al., 2024; Garg et al., 2024) and visual question answering (Liu et al., 2024e; Yue et al., 2024). While these models exhibit strong generalization across various domains, their performance often declines when applied to scientific text (Li et al., 2024; Lu et al., 2024; Pramanick et al., 2024). Additionally, the complexity of poster layouts, the use of technical terminology, and the intricate interplay between text, tables, and figures make summarizing scientific posters a particularly challenging task, which has remained under-explored due to the lack of specialized datasets.

To address this gap, we introduce POSTERSUM, a novel multimodal benchmark for summarizing scientific posters into research paper abstracts. Our dataset consists of 16,305 scientific posters and corresponding abstracts as summaries collected from the main Machine Learning conferences, namely ICLR, ICML, and NeurIPS. These posters cover a broad range of scientific disciplines and present unique challenges, including complex layouts and intricate combinations of text, tables, and figures as shown in Fig. 1. Information is often distributed across the poster, requiring careful navigation and integration of diverse elements to identify and summarize the key points effectively.

We benchmark state-of-the-art MLLMs on POSTERSUM and demonstrate that, despite their impressive performance on a range of other multimodal tasks, these models face significant limitations when summarizing scientific posters. For instance, the best-performing closed-source model in our experiments, GPT-4o (OpenAI et al., 2024), achieves a ROUGE-L score of 22.30, underscoring the difficulty of this task specifically with the posters with figures and tables.

To address this challenge, we propose SEGMENT & SUMMARIZE, a hierarchical approach inspired by the divide-and-conquer principle (Chen and Zhao, 2023). The method involves three key steps: (1) Segmentation: we segment each poster into coherent regions; (2) Localized Summarization: a multimodal large language model extracts

[1]The dataset is available at rohitsaxena/PosterSum.
[2]The code is available at this link.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1828–1844

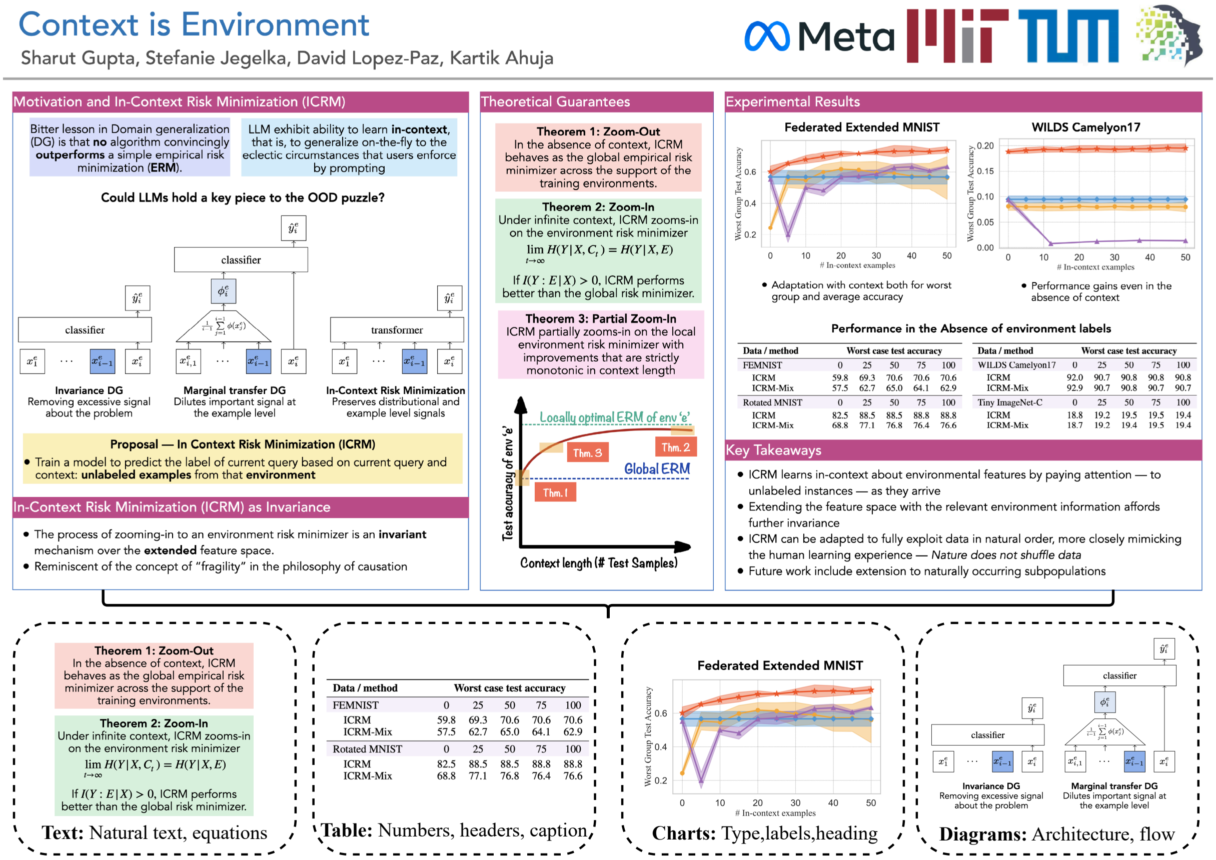


Figure 1: An example of scientific poster from the PosterSum dataset. The poster, describing the work in Gupta et al. (2024), contains visual elements such as structured tables with numerical results, charts, diagrams, and textual sections, demonstrating the multimodal complexity present in the dataset.

and interprets the text within each segment, generating localized summaries for each region; and (3) Global Summarization: these localized summaries are combined using text-based large language model to produce a cohesive abstract that spans the entire poster. Notably, this approach does not require additional training or fine-tuning. Local summaries allow the model to focus on fine details within that specific area, which are useful for tables and figures. Also, it aligns with the inherent structure of the poster, which has sections with a specific focus. This approach achieves a ROUGE-L score of 24.18, outperforming both closed-source and open-source models, setting a new benchmark for scientific poster summarization. Our contributions can be summarized as follows:

- We introduce PosterSum, a large-scale multimodal dataset of 16,305 scientific posters paired with their abstracts, tailored for research poster summarization.
- We benchmark state-of-the-art MLLMs on PosterSum, showing their limitations in processing and summarizing scientific posters.
- We propose Segment & Summarize, a hierarchical approach that segments each poster into coherent regions, extracts the textual content from those regions, and then composes a final summary; we also demonstrate PosterSum's utility for fine-tuning MLLMs, showing promising improvements over zero-shot results.

2 Related Work

Multimodal Large Language Models. After the emergence of LLMs, recent work (Liu et al., 2023; Wang et al., 2024b; Alayrac et al., 2022) investigated their use in processing multimodal inputs, giving rise to Multimodal Large Language Models (MLLMs). The core idea in this line of research is to align visual and textual features by using shared representations. This framework typically involves using a pre-trained visual encoder to extract visual features, a projection layer to map visual representations into corresponding text representations, and a pre-trained LLM to generate textual responses, allowing the model to condition the output on visual and textual inputs. MLLM architectures such as LLaVA (Liu et al., 2023) and MiniCPM (Yao et al., 2024) demonstrated impressive zero-shot generalization across diverse visual and language tasks. However, most existing MLLMs focus on general domain tasks and relatively simple visual inputs; the challenge of understanding complex and information-dense visual documents like scientific posters remains under-explored.

Summarization in Scientific Domains. *Scientific summarization* consists of generating concise summaries for scientific content (Yasunaga et al., 2019; Cachola et al., 2020; Ju et al., 2021; Sotudeh and Goharian, 2022). Several scientific summarization benchmarks have been proposed, designed to process modalities such as videos (Lev et al., 2019; Chen et al., 2024), slides (Tanaka et al., 2023), surveys (Liu et al., 2024d), and research papers (Takeshita et al., 2024; Liu et al., 2024a). While scientific posters are widespread in scientific communication, no poster summarization benchmark has been proposed in the literature. Our proposed POSTERSUM aims to address this gap.

Document Layout Analysis and Segmentation. Understanding document layouts plays a significant role in processing complex visual documents like scientific posters. Recent work in document layout analysis (Peng et al., 2022; Wang et al., 2024a; Luo et al., 2024; Appalaraju et al., 2024) aims at identifying and classifying different regions within a document considering spatial relationships and content type. Previous work has also focused on understanding individual elements in documents, such as charts (Masry et al., 2022) and tables (Zheng et al., 2024). However, most existing approaches are designed for either standard documents or individual elements like charts and tables and do not capture the complex layouts and the rich multimodal structure of scientific posters, which typically consist of text, charts, equations, and tables.

3 The POSTERSUM Dataset

We introduce POSTERSUM, a novel dataset and benchmark for multimodal abstractive summarization of scientific posters. POSTERSUM consists of 16,305 pairs of academic posters as images (PNG format) and their corresponding research paper abstracts. These posters were collected from major machine learning and artificial intelligence conferences, which accept papers from various subfields of machine learning, including computer vision, natural language processing, optimization, and computational biology.

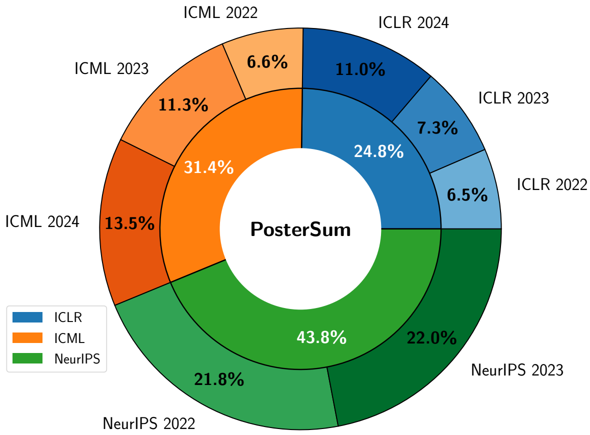


Figure 2: Distribution of the POSTERSUM dataset.

POSTERSUM captures the diverse and heterogeneous nature of academic posters, which are commonly used at conferences to present research findings. These posters vary in layout, content, and visual complexity–some are text-heavy, while others emphasize visual elements such as charts, graphs, and figures, as shown in Fig. 1. This variability presents a significant challenge for MLLMs, requiring them to interpret and summarize multimodal information effectively.

Each poster in the dataset is paired with its corresponding abstract, which serves as the ground-truth summary. The abstract highlights the key contributions and findings of the research, making it an ideal summary for the poster. Unlike image captioning, poster summarization requires a deeper understanding of multiple elements in the poster to generate a comprehensive and meaningful abstract-based summary; and unlike other scientific summarization works, it focuses on extracting information from a poster rather than a scientific paper.

3.1 Dataset Creation

The POSTERSUM dataset was collected from the websites of top-tier machine learning and artificial intelligence conferences: ICLR, ICML, and NeurIPS. We selected these conferences based on the availability of research posters. We first collected research paper links and paper identifiers from the conference websites. We filtered out any entries where the poster of the paper was not available, ensuring that only papers with accessible posters were included in the dataset. We exclusively collected posters from the years 2022 to 2024, as shown in Fig. 2. Additionally, we man-

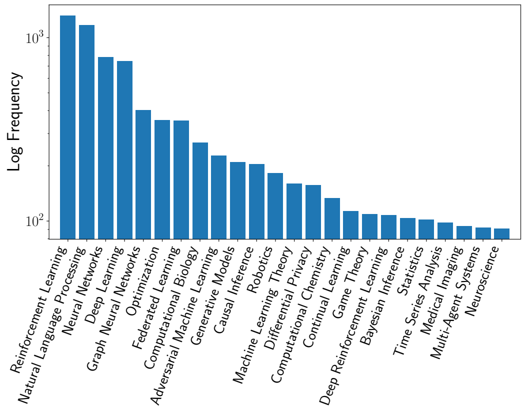


Figure 3: Distribution of the most frequent 25 topics for the posters in POSTERSUM.

POSTERSUM Statistics	
Total number of posters-summary	16,305
Total number of unique categories	137
Mean token length of the summary	224
Mean summary sentences	7.21
Train/Val/Test size	10305/3000/3000
Mean CLIP score	29.08
Year range	2022–2024

Table 1: Statistics of the POSTERSUM dataset.

% Novel n-grams in Summary			
1-grams	**2-grams**	**3-grams**	**4-grams**
54.54	81.13	88.67	91.41

Table 2: Statistics for percentage of novel n-grams in the POSTERSUM summaries.

ually reviewed the dataset to remove any posters with placeholder images. We assume that the research reported in the posters is of a high standard, and the posters are of high quality, as the corresponding papers appeared at top machine learning conferences.

To build a robust summarization dataset, it was essential to pair each poster with a human-written summary. We collected the research paper abstracts from the corresponding paper pages using the paper identifiers. These abstracts serve as the summaries for the posters, as they highlight the core findings and contributions of the research. For papers where the abstract was missing from the webpage, we manually extracted the abstract from the research paper's PDF to ensure completeness.

3.2 Dataset Statistics and Analysis

This process resulted in the 16,305 poster-summary pairs, providing a comprehensive multimodal resource for evaluating abstractive summarization of academic research posters.

Table 1 provides an overview of key statistics for the dataset. The average length of the poster summaries is 224 word-piece tokens, with an average of seven sentences per summary. The poster images are of high-resolution, with a mean size of 3547×2454. We randomly split the dataset into training, validation, and test sets using a 10305/3000/3000 split, which can be utilized for training and fine-tuning models.

To better understand the diversity within the dataset, we categorized posters into topics. Since topics were not available on the conference websites, we employed the GPT-4o vision model to generate topic labels by prompting the model in a zero-shot setting using the images of the posters. As a result, we identified 137 distinct topics within machine learning and artificial intelligence, spanning areas such as reinforcement learning, natural language processing (NLP), computational biology, and healthcare applications. Fig. 3 illustrates the distribution of the most frequent 25 topics.

To assess the abstractiveness of the poster summaries, we report the percentage of novel n-grams in the summaries compared to the Optical Character Recognition (OCR) extracted text from the posters. We used MMOCR (Kuang et al., 2021) to extract the text. While most posters do not explicitly include abstracts, we found that approximately 8% of the total posters may contain an abstract in poster, based on the occurrence of the word "abstract" in the OCR text. As shown in Table 2, a significant portion of the summaries contains novel content, particularly in the 3-gram and 4-gram categories. This demonstrates that the summaries are not simple restatements of poster text but instead provide a more comprehensive abstraction.

We also find a mean CLIPScore (Hessel et al., 2021) of 29.08 when we evaluate the alignment between the images of the posters and their summaries. This score was computed at the sentence level and averaged across the dataset. The relatively low CLIPScore highlights the challenge that POSTERSUM poses for existing MLLMs. Unlike image-captioning tasks, where captions directly describe visual features, academic posters are composed of diverse and complex visual elements, such as charts, graphs, equations, and dense textual explanations. This complexity makes it more difficult for models to capture the semantic relationships between these elements and the corresponding abstract summaries.

4 Multimodal Poster Summarization

4.1 Task Formulation

Given a scientific poster I in image format as input, the objective is to generate a textual summary $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$ that encapsulates the key points and essential content of the poster. Formally, a model M_θ, parameterized by θ, takes the poster I as input, optionally accompanied by a prompt P, and generates a summary $\hat{Y}$. The key challenge in this task is that model M_θ must effectively abstract from the diverse visual and textual elements present in the poster, including text, charts, diagrams, and equations, to produce a coherent and informative summary.

4.2 Baselines

We evaluate various multimodal models, both open-source and closed-source, to assess their performance on the abstractive summarization task for scientific posters. As the posters include textual elements, we also evaluate OCR-based methods as baselines. For MLLMs, evaluation is conducted in a zero-shot and Chain-of-Thought (CoT) setting to assess the capability of models to generate accurate summaries. Additionally, we explore parameter-efficient fine-tuning techniques on selected open-source models. Below are the categories of models used in our experiments.

Optical Character Recognition (OCR). For OCR-based baselines, we used two OCR methods (MMOCR (Kuang et al., 2021) and Pytesseract[3]) to extract text from the poster images and concatenated the results to generate a summary. Additionally, we combined the best OCR output with a text-based large language model (LLM). In this approach, we first extract text from the posters and then use the Llama-3.1-8B-Instruct (Grattafiori et al., 2024) model for summarization. This allows us to evaluate the performance of text-only LLMs when provided with OCR-extracted text.

Closed-source MLLMs. We evaluated GPT-4o (OpenAI et al., 2024), Claude 3.5 Sonnet (Anthropic, 2024), and Gemini 2.0 (Anil et al., 2024) as closed-source MLLMs. All the models were prompted with the image of the poster in a zero-shot setting to generate abstractive summaries based on the input. The prompt template can be found in Appendix A.

[3]https://github.com/h/pytesseract

Open-source MLLMs. As open-source/open-weights models, we evaluated Llama-3.2-11B-Vision-Instruct (Meta, 2024), Qwen2-VL-7B-Instruct (Yang et al., 2024), LLaVA-NeXT (Liu et al., 2024c,b), mPLUG-DocOwl2 (Hu et al., 2024), and MiniCPM-Llama3-V-2.5 (Yao et al., 2024). Each model was evaluated in both zero-shot and CoT settings. The CoT prompt was used to steer the models to extract relevant information, such as the title, research problem, methods, results, and conclusion, from the poster. We report the full prompt template in Appendix A.

Fine-tuned Models (LoRA). We also evaluated the fine-tuned Llama-3.2-11B-Vision-Instruct and LLaVA-NeXT models. We used parameter-efficient fine-tuning using the Low-rank Adaptation (LoRA; Hu et al., 2022) method to fine-tune both of these models using the training and validation set of the POSTERSUM dataset.

4.3 SEGMENT & SUMMARIZE

We now introduce SEGMENT & SUMMARIZE, a hierarchical approach inspired by the divide-and-conquer principle. Rather than processing the entire poster I as a single input, SEGMENT & SUMMARIZE decomposes the task into three key steps: (1) Segmentation and Clustering (2) Localized Summarization, and (3) Global Summarization. The SEGMENT & SUMMARIZE pipeline is outlined in Fig. 4.

1. Segmentation and Clustering. Given the image of a poster I, the first step is to segment it into n coherent regions $M = \{M_1, M_2, \ldots, M_n\}$. This is achieved using a segmentation model S_ϕ, parameterized by ϕ, Since the number of regions n can be large and can contain redundant and small segments, the regions are further clustered into groups R with the number of the clustered regions as k using a clustering algorithm C such that $k \ll n$. The clustering step groups similar regions together, reducing redundancy and ensuring complete coverage of the poster. Formally, $M = S_\phi(I)$ and $R = C(M)$.

By segmenting the poster and summarizing each region independently, the method ensures a detailed and accurate understanding of the content.

2. Localized Summarization. For each clustered region R_i, a localized summary $\hat{Y}_i = \{\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{ik}\}$ is generated using an MLLM V_ϕ. The model is used to extract and inter-

	R-1	R-2	R-L	RLSum	SBLEU	Met	BS_p	BS_r	BS_{f1}	CLIPS
Closed-Source Models										
Gemini	39.89	12.38	20.89	36.21	6.57	22.34	59.46	59.6	59.53	24.41
Claude-3.5 Sonnet	43.45	11.42	19.51	39.08	7.72	28.43	59.3	60.3	59.8	25.02
GPT-4o	44.98	13.12	22.30	40.55	10.05	30.29	60.31	60.22	60.77	25.06
OCR										
Pytesseract	26.27	1.03	9.26	17.07	0.06	21.18	34.89	41.15	37.71	18.21
MMOCR	24.35	8.96	12.73	23.4	4.03	27.62	34.32	49.39	40.40	18.49
MMOCR + Llama	28.37	5.37	15.49	24.94	2.42	25.0	52.51	56.88	54.58	19.78
Zero-Shot										
Llama-3.2-11B-V	20.7	4.29	11.01	18.88	1.75	18.07	43.51	44.46	43.75	18.91
Qwen2-VL-7B	20.63	1.93	12.08	18.97	0.63	16.13	46.81	48.35	47.53	17.34
LLaVA-NeXT	29.89	6.61	16.0	27.02	3.41	19.57	53.02	51.10	51.89	21.67
mPLUG-DocOwl2	35.62	8.79	19.06	32.07	3.36	18.35	58.35	55.69	56.99	23.65
MiniCPM	39.88	11.11	20.14	35.45	7.18	23.76	59.54	58.91	59.22	25.50
Chain of Thought										
Llama 3.2-11B-V	20.05	3.4	10.77	18.14	1.7	8.57	42.43	45.89	43.86	19.57
Qwen2-VL-7B	25.58	2.92	13.75	23.24	1.52	15.65	54.48	51.97	53.16	19.68
LLaVA-NeXT	30.25	6.16	16.25	27.48	2.95	24.53	48.79	50.89	49.78	21.56
mPLUG-DocOwl2	37.04	9.15	19.71	33.45	3.98	19.6	58.59	56.26	57.40	23.78
MiniCPM	41.50	11.68	21.04	37.08	8.60	26.34	59.32	58.29	58.80	25.76
Fine-tuning MLLMs										
Qwen2-VL-7B	28.77	6.11	15.18	26.32	2.66	19.09	53.78	51.99	52.83	18.96
LLaVA-NeXT	31.77	9.94	18.25	28.7	6.21	27.18	51.29	58.89	54.67	22.77
Llama-3.2-11B-V	35.16	13.33	20.64	31.75	8.91	28.32	50.65	58.82	54.19	23.77
SEGMENT & SUMMARIZE										
Ours	**46.68**	**15.73**	**24.18**	**42.5**	**12.63**	**30.87**	**61.21**	**61.62**	**61.37**	**27.63**

Table 3: Summarization results on the POSTERSUM dataset. The results show ROUGE scores (R-1, R-2, R-L, R-LSum), BERTScores (BS_p, BS_r, BS_{f1}), SacreBLEU, CLIPScore, and METEOR scores for all baselines and models. All the scores are percentages.

pret the content within R_i, including text, figures, and tables, to generate a localized summary for that specific region. This also helps in processing the high-resolution image.

3. Global Summarization. The localized summaries $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$ are combined into a cohesive global summary $\hat{Y}$ using a text-based large language model L_ω, parameterized by ω. The model L_ω takes as input the individual summaries and generates a single, well-structured output that represents the overall content of the poster. This step ensures that the final abstract is not only comprehensive but also maintains logical flow and coherence. Formally, $\hat{Y} = L_\omega(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k)$.

This processing pipeline helps summary generation through a structured, localized, and hierarchical approach. By segmenting the poster and summarizing each region independently, the method captures fine-grained details that might be overlooked in a global approach. This also aligns with the structure of these posters, which are mostly divided into sections. This approach does not require additional training or fine-tuning, and both the models (V_ϕ, L_ω) are frozen.

5 Experimental Details

All models in each category were evaluated using the same hyperparameter settings for a fair evaluation. We generate at most 768 new tokens for all the experiments. For closed-source models, we used the default platform settings. Open-source models were evaluated with a beam size of 4 with greedy decoding to ensure reproducibility. The fine-tuning experiments were conducted for 10 epochs with a batch size of 4. More details on the hyperparameters and prompt templates can be found in Appendices A and D.

For SEGMENT & SUMMARIZE, we used the Segment Anything Model (Kirillov et al., 2023) for segmentation with k-Means for clustering. The num-

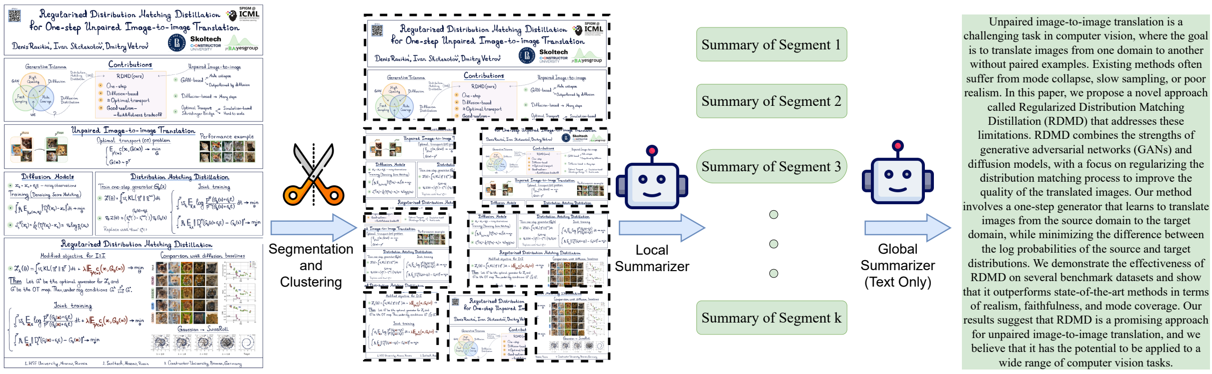


Figure 4: Illustration of our SEGMENT & SUMMARIZE pipeline. The poster, describing the work in Rakitin et al. (2024), is first divided into segments, each of which is summarized by a MLLM. These localized summaries are subsequently merged by a text-based large language model to generate a single, coherent summary.

ber of clusters (k) was set to 8 based on the analysis in Appendix C. We used MiniCPM-Llama3-V-2.5 as the local summarizer (V_ϕ) and Llama 3.1-8B-Instruct as the global summarizer (L_ω). We used the training set for fine-tuning and the validation set for hyperparameter tuning. All the final results are evaluated on the test set.

Evaluation Metrics. We use ROUGE F1 (R-1/2/L/LSum) scores (Lin, 2004), SacreBLEU (SBLEU; Post, 2018), METEOR (MET; Banerjee and Lavie, 2005), CLIPScore (CLIPS; Hessel et al., 2021), and BERTScore (Zhang et al., 2020) to evaluate the accuracy of all models.

6 Results

Table 3 presents the poster summarization performance of all baselines alongside our proposed SEGMENT & SUMMARIZE method, evaluated on the POSTERSUM test set. Our method outperforms both open-source and closed-source models, achieving the best results across all metrics.

Closed-source Models. GPT-4o achieves relatively high performance among the closed-source models across all metrics, with ROUGE-1/2/L scores of 44.98, 13.12, and 22.30, respectively. Claude-3.5 Sonnet also performs well, attaining a ROUGE-L score of 19.51.

OCR Baselines. The two OCR-based methods, MMOCR and Pytesseract, achieve relatively low scores across all metrics. This is likely due to the limitation of concatenating raw OCR text without leveraging other visual elements. Combining OCR with the text-only Llama-3.1 model results in a substantial improvement, with ROUGE-L increasing from 12.73 to 15.49. Interestingly, these OCR methods still outperform certain multimodal models, indicating that text extraction remains a challenge for some MLLMs.

Open-source Models. Among the open-source MLLMs evaluated in zero-shot settings, MiniCPM-Llama3-V-2.5 obtains the highest ROUGE-1/L score (39.88/20.14) and a strong BERTScore-F1 of 59.22. Meanwhile, mPLUG-DocOwl2 achieves a competitive ROUGE-L of 19.06 and a BERTScore-F1 of 56.99.

Chain of Thought (CoT). Adding an explicit CoT prompt improves the performance of most models. For instance, MiniCPM-Llama3-V-2.5 improves its ROUGE-1/L/METEOR/CLIPScore scores to 41.50/21.04/26.34/25.76, while mPLUG-DocOwl2's performance also increases (ROUGE-1/L of 37.04/19.71). Additionally, LLaVA-NeXT and Qwen2-VL-7B exhibit similar gains. Although the performance boosts are not large, these results suggest that guiding models via CoT prompt can help in extracting relevant poster content.

Fine-tuned Models. Using LoRA substantially boosts performance for both MLLMs. In particular, Llama-3.2-11B-Instruct demonstrates notable improvements in ROUGE, ScareBLEU, and METEOR scores, though it does not surpass the best CoT variants of MPLUG-DOCOWL2 and MINICPM-LLAMA3-V-2.5, which likely benefit from pre-training on multimodal scientific data.

SEGMENT & SUMMARIZE. Our proposed method outperforms all other models, including closed-source models, on all metrics, achieving ROUGE-1/2/L scores of 46.68, 15.73, and 24.18, respectively, with a 3.14% gain on ROUGE-L

Methods	R1	R-2	R-L	Met
Without clustering	42.25	14.30	22.76	23.97
With clustering	46.68	15.73	24.18	30.87

Table 4: Comparison of SEGMENT & SUMMARIZE with and without clustering — clustering the segments yields more accurate results.

Methods	R1	R-2	R-L	Met
mPLUG-DocOwl2	37.04	9.15	19.71	19.6
Ours with DocOwl2	42.48	11.18	20.61	26.72
Ours with MiniCPM	46.68	15.73	24.18	30.87

Table 5: Comparison of using mPLUG-DocOwl2 as local summarize. Applying SEGMENT & SUMMARIZE shows improvement compared to using the model itself.

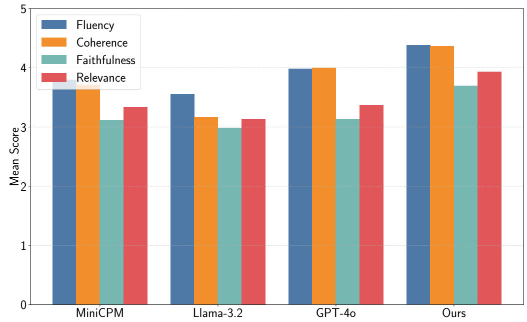


Figure 5: Mean 5-point Likert ratings for Fluency, Coherence, Faithfulness, and Relevance across four methods. SEGMENT & SUMMARIZE (ours) achieves the highest scores across all the dimensions.

compared to open-source models. It also attains a substantially higher ScareBLEU score (12.63), BERTScore-F1 of 61.37, and a CLIPScore of 27.63. These results indicate that local-region summaries effectively preserve small details and handle posters of varying complexity by processing each region independently rather than attempting to analyze the entire poster as a single input.

7 Ablation Studies and Analysis

Effect of Clustering on Summarization. To quantify the impact of clustering in our SEGMENT & SUMMARIZE approach, we conduct an ablation study that removes the clustering step. Specifically, we select the top-k segments (with $k = 8$) based on their region size to generate local and global summaries. Table 4 shows that clustering improves the ROUGE-1 score by +4.43, ROUGE-2 by +1.43, and ROUGE-L by +1.42 over the non-clustered baseline. We hypothesize that clustering helps reduce redundant segments and improves context aggregation.

Effect of Local Vision Summarization. To assess the role of the local summarization model in SEGMENT & SUMMARIZE, we replaced MiniCPM-Llama3-V-2.5 with mPLUG-DocOwl2, which previously ranked second among open-source models under the CoT setting. Table 5 shows that using mPLUG-DocOwl2 with our hierarchical approach boosts ROUGE-1 to 42.48 and METEOR to 26.72 compared to using the model in the CoT setting. However, it does not outperform our method using MiniCPM. These findings highlight that the segmentation and summarization approach substantially improves performance compared to using the poster as a single input.

Human Evaluation We conducted a human evaluation to compare the quality of summaries generated by our method against the best models in each category (MiniCPM CoT, Llama-3.2-11B-V LoRA, GPT-4o ZS). Forty crowdworkers were recruited via Prolific (all L1 English speakers, master's/doctoral degree holders, and at least 100 previously approved submissions) and compensated at $17/hr. We randomly sampled 40 posters, and participants viewed the poster image, the reference abstract, and one candidate summary, resulting in 160 (4x40) poster–summary evaluations. They rated each summary on 5-point Likert scales for each of four dimensions: **Fluency**, **Coherence**, **Faithfulness**, and **Relevance**. Across all dimensions, SEGMENT & SUMMARIZE received the highest mean ratings (see Figure 5). A one-way ANOVA followed by Tukey's HSD confirmed that SEGMENT & SUMMARIZE significantly outperformed MiniCPM and Llama-3.2-11B-V on every dimension ($p < .01$ for all) and surpassed GPT-4o on Faithfulness and Relevance ($p < .05$). However, differences with GPT-4o in Fluency and Coherence did not reach significance. More statistical details and instructions are available in Appendix E and F.

8 Conclusions

We presented POSTERSUM, a multimodal benchmark for scientific poster summarization comprising 16,305 poster-abstract pairs. Our experiments show that even state-of-the-art MLLMs struggle with key aspects of scientific poster summarization. Furthermore, we propose SEGMENT & SUMMARIZE, a hierarchical approach that outperforms existing models by breaking down the summarization task into localized segments before generating a cohesive abstract. We find that our method outperforms MLLMs in both zero-shot and fine-tuned settings and that there remains significant room for improvement in multimodal understanding of complex scientific documents such as posters. We believe POSTERSUM will be a valuable resource for developing and evaluating MLLMs capable of processing information-dense scientific content.

Acknowledgments

This work was supported in part by the School of Informatics at the University of Edinburgh. Pasquale Minervini was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), and a donation from Accenture LLP. This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

Limitations

While our work advances scientific poster summarization, we should highlight a few limitations. First, our dataset is restricted to machine learning conference posters from 2022 to 2024, which may limit the generalization to other scientific domains. Second, while practical, automated topic labeling using GPT-4o may introduce biases or inaccuracies in the topic distribution. The proposed SEGMENT & SUMMARIZE method relies heavily on the quality of the initial segmentation: a suboptimal segmentation can lead to fragmented or redundant local summaries. Our method also assumes that the content can be meaningfully decomposed into spatial regions, which may not hold for posters with complex cross-referencing or interdependent visual elements. We considered the abstract as a ground-truth summary of the poster, but the poster may sometimes differ from the paper.

Ethics Statement

Dataset. All the scientific posters and abstracts in our dataset are sourced from publicly accessible conference resources. Additionally, we sought permission from the conference website contacts to use the publicly available data for research purposes.

Multimodal Large Language Models. This paper utilizes pre-trained multimodal large language models, which have been shown to exhibit various biases, occasionally hallucinate, and generate non-faithful text. Therefore, summaries generated using our dataset should not be released without automatic filtering or manual verification to ensure accuracy and reliability.

Bias. Despite efforts to include a wide range of posters, the dataset may not fully represent the diversity of research poster styles, languages, or scientific disciplines. As a result, models trained on POSTERSUM may exhibit biases towards the types of posters included in the dataset. Future work should consider expanding the dataset to encompass a broader spectrum of academic fields and visual formats to mitigate potential biases.

References

Natalie Abreu, Nathan Vaska, and Victoria Helus. 2022. Addressing mistake severity in neural networks with semantic knowledge. *CoRR*, abs/2211.11880.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, and 1330 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Anthropic. 2024. Claude 3.5 - sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2024-12-06.

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. 2024. Docformerv2: Local features for document understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 709–718. AAAI Press.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Ex-*

trinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. 2022. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. In *Advances in Neural Information Processing Systems*.

Shi Chen and Qi Zhao. 2023. Divide and conquer: Answering questions with object factorization and compositional reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6736–6745. IEEE.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M^{3}av: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9041–9060. Association for Computational Linguistics.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. 2024. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 93–127, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sharut Gupta, Stefanie Jegelka, David Lopez-Paz, and Kartik Ahuja. 2024. Context is environment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *Preprint*, arXiv:2409.03420.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4091–4098, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *ICCV*, pages 3992–4003. IEEE.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17283–17300. PMLR.

Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, Kai Chen, Wayne Zhang, and Dahua Lin. 2021. Mmocr: A comprehensive toolbox for text detection, recognition and understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 3791–3794, New York, NY, USA. Association for Computing Machinery.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. TalkSumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2125–2131, Florence, Italy. Association for Computational Linguistics.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024a. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024d. SumSurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9632–9651, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024e. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *CVPR*, pages 15630–15640. IEEE.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Yuhui Cao, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3744–3756. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multi-modal question answering on scientific papers. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Denis Rakitin, Ivan Shchekotov, and Dmitry Vetrov. 2024. Regularized distribution matching distillation for one-step unpaired image-to-image translation. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Sajad Sotudeh and Nazli Goharian. 2022. TSTR: Too short to represent, summarize with details! intro-guided extended summary generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–335, Seattle, United States. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13636–13645.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024b. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499. Curran Associates, Inc.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chang Zhou et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. 2024. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14022–14032.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt Templates

Prompt Template for Zero-Shot
Write an abstract for an AI conference paper for the given research poster image.

Prompt Template for CoT
Analyze the research poster image step by step. First, identify the title and main research problem. Then, briefly describe the methodology used. Next, summarize the key findings or results. Finally, note the conclusion or implications. Using this information, write an abstract for the given research poster image.

Prompt Template for Local Summary
Describe all the text, tables, figures, and equations in the image.

B Effect of Poster Text Content on Summarization Performance

To investigate whether posters with a high amount of text result in better summarization performance, we analyze the relationship between OCR-extracted text length and ROUGE-L scores using our SEGMENT & SUMMARIZE method. Specifically, we use MMOCR to extract text from each poster and compute its total length in characters (not in tokens).

Fig. 6 presents the mean ROUGE-L scores across different OCR text-length bins. The dotted line represents the number of posters in each text-length bin. We observe that summarization performance tends to improve as the amount of text in poster increases. However, the correlation remains weak (*Pearson* $r = 0.213$, *Spearman* $r = 0.210$), suggesting that text in poster alone is not a strong predictor of summarization quality. Low performance in posters with minimal text also highlights the need for more robust multimodal understanding of figures, charts, equations, and tables.

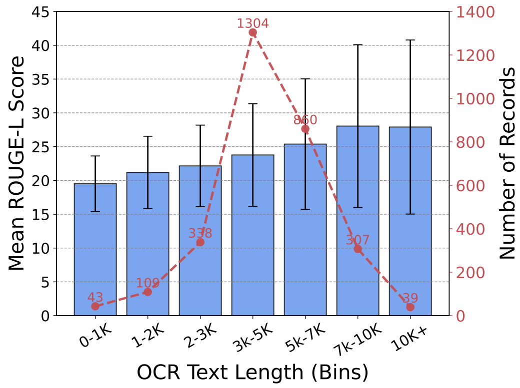


Figure 6: Effect of text present in the poster on summarization. We report mean ROUGE-L scores for different OCR-extracted character-length bins. The red dashed line represents the number of posters in each bin.

Figure 7: Effect of varying the number of clusters on ROUGE-L performance on SEGMENT & SUMMARIZE

C Selecting the Number of Clusters

To select the number of clusters (k) for our SEGMENT & SUMMARIZE, we conducted an empirical analysis on a subset of 100 posters from the validation set, varying the number of clusters from 2 to 10. Fig. 7 presents the mean ROUGE-L score for each cluster configuration. In these experiments, the local and global summarization components remained fixed.

We observe that the best performance is achieved at $k = 8$ which was used in our final experiments. Additionally, we limit the maximum number of clusters to 10 in the analysis to keep the inference time of our local summarization manageable.

D Additional Experiment Details

Table 6 summarizes the versions of the closed-source models used in our experiments. For fine-tuning, we use a learning rate of 1×10^{-4} with the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$) and a cosine learning rate schedule. We

Model	Version
GPT-4o	gpt-4o-2024-08-06
Gemini 2.0	gemini-2.0-flash-exp
Claude 3.5 Sonnet	claude-3-5-sonnet-20241022

Table 6: Details of the closed-sourced models.

employ LoRA with rank $r = 8$, $\alpha = 8$, and a dropout rate of 0.1.

All images are processed and scaled by the respective model's image processor for model specific sizes. In the case of closed-source models, we scale each image to a maximum width of 2048 while preserving the original aspect ratio due to size limitations. All the models were trained using 2 A100 GPU with 80GB memory. We used the Huggingface *evaluate* library for the implementation of the metrics. Our method's additional wall-clock time per batch is approximately 1.75 seconds for the segmentation and clustering stage and 6.02 seconds for the two stages.

E Human Evaluation Statistical Analysis

Model	Fl	C	Fa	R
MiniCPM (CoT)	3.80	3.72	3.12	3.33
Llama-3.2-11B-V (LoRA)	3.55	3.17	2.98	3.13
GPT-4o (ZS)	3.98	4.00	3.13	3.37
SEGMENT & SUMMARIZE	4.38	4.37	3.70	3.93

Table 7: Mean Likert ratings (1–5) for each model across the four dimensions. Fl: Fluency, C: Coherence, Fa: Faithfulness, R: Relevance

Mean Likert ratings for each model are provided in Table 7. We conducted one-way ANOVAs to assess whether there were statistically significant differences among the models across the four dimensions. The results showed a significant difference across all models:

- **Fluency**: $F = 9.20, p < 0.001$
- **Coherence**: $F = 20.33, p < 0.001$
- **Faithfulness**: $F = 6.27, p = 0.0004$
- **Relevance**: $F = 6.64, p = 0.0003$

To identify the specific differences among the models, Tukey's HSD post-hoc tests were performed for all the dimensions. SEGMENT & SUMMARIZE method significantly outperformed all the models on Faithfulness and Relevance.

- **Faithfulness**: +0.58 vs. MiniCPM (p = 0.007), +0.72 vs. Llama (p=0.0005), +0.57 vs. GPT-4o (p = 0.0098)
- **Relevance**: +0.60 vs. MiniCPM (p=0.009), +0.80 vs. Llama (p=0.0002), +0.57 vs. GPT-4o (p=0.0155)

Against GPT-4o, SEGMENT & SUMMARIZE 's advantages in Fluency (+0.40, p=0.0717) and Coherence (+0.37, p=0.0987) did not reach significance, although it remained significantly higher than MiniCPM and Llama on those dimensions:

- **Fluency**: +0.58 vs. MiniCPM (p=0.0025), +0.83 vs. Llama (p<0.001)
- **Coherence**: +0.65 vs. MiniCPM (p=0.0003), +1.20 vs. Llama (p<0.001)

F Instructions for Human Evaluation

In this task, you will assess the quality of computer-generated summaries of scientific posters by comparing each against the poster and its reference summary. For each trial, you will be shown:

1. Poster Image.
2. Reference Summary.
3. Generated Summary.

Your task is to rate the Generated Summary on four dimensions using a 5-point Likert scale (1 = Poor, 5 = Excellent).

Dimensions of Evaluation

Fluency This dimension evaluates whether the generated summary is grammatically correct, easy to read, and well-structured.

Coherence This dimension assesses whether the sentences in the generated summary flow logically and maintain a consistent narrative.

Faithfulness This dimension checks if all the facts presented in the generated summary are accurate and can be directly inferred from the poster image and reference summary.

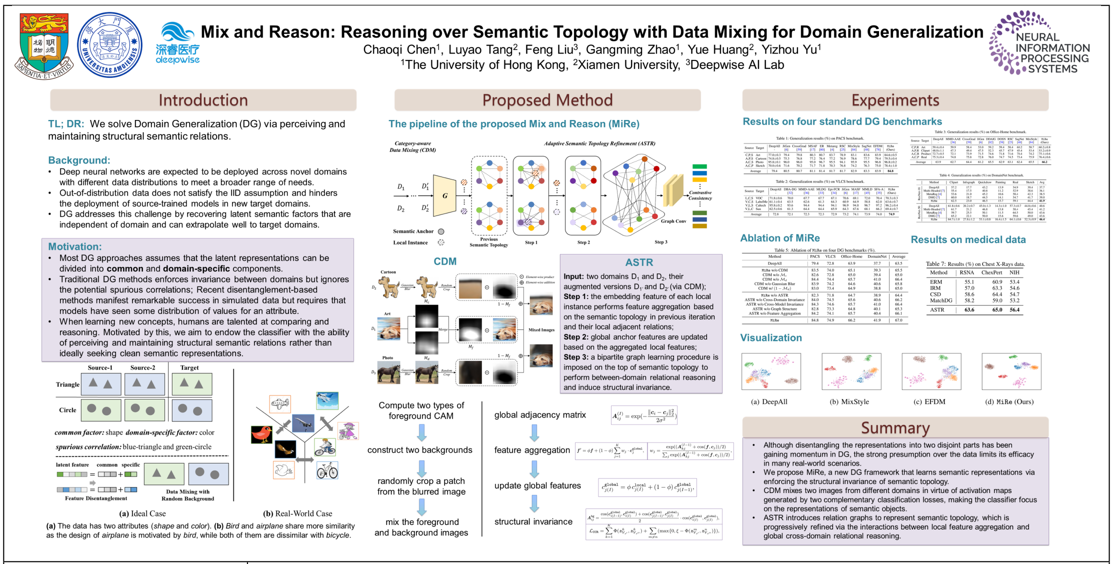
Relevance This dimension evaluates whether the generated summary includes the key findings and contributions shown in the poster and reference summary, without omitting important information.

Rating Procedure

For each poster–summary pair:

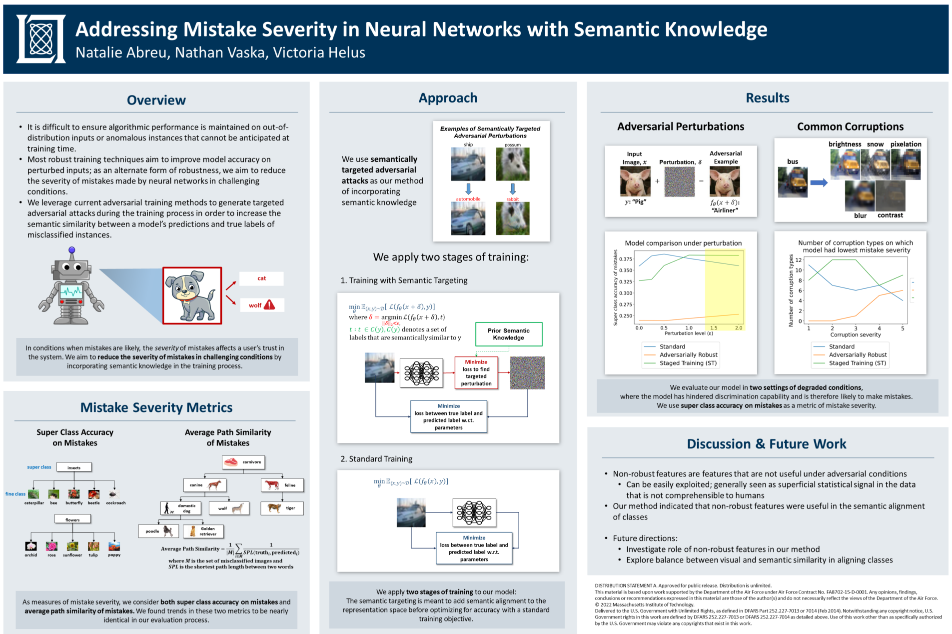
1. View the poster image and reference summary carefully.
2. Read the generated summary in its entirety.
3. Assign a score (1–5) for each of the four dimensions, based only on the definitions above.
4. Minor typos or formatting issues should not lower your score unless they impede understanding.

G Dataset Examples with Model Summaries



Model	Output
Reference	Domain generalization (DG) enables generalizing a learning machine from multiple seen source domains to an unseen target one. The general objective of DG methods is to learn semantic representations that are independent of domain labels, which is theoretically sound but empirically challenged due to the complex mixture of common and domain-specific factors. Although disentangling the representations into two disjoint parts has been gaining momentum in DG, the strong presumption over the data limits its efficacy in many real-world scenarios. In this paper, we propose Mix and Reason (MiRe), a new DG framework that learns semantic representations via enforcing the structural invariance of semantic topology. MiRe consists of two key components, namely, Category-aware Data Mixing (CDM) and Adaptive Semantic Topology Refinement (ASTR). CDM mixes two images from different domains in virtue of activation maps generated by two complementary classification losses, making the classifier focus on the representations of semantic objects. ASTR introduces relation graphs to represent semantic topology, which is progressively refined via the interactions between local feature aggregation and global cross-domain relational reasoning. Experiments on multiple DG benchmarks validate the effectiveness and robustness of the proposed MiRe.
MiniCPM CoT	The paper is about Mix and Reason, a methodology for domain generalization that reasons over semantic topology to mix data from different domains and construct new target domains. It addresses the problem of limited training data by leveraging the IID assumption and the representations learned from the source domain to perform well in the target domain. The proposed methodology involves a pipeline of the proposed Mix and Reason (MiRe) component, which takes the adaptive semantic topology refinement (ASTR) component as its backbone. The key results show that the proposed methodology outperforms state-of-the-art methods on four benchmark datasets, including CelebA, CUB, Stanford 2007, and COCO.
Llama-3.2-11B-V FT	Deep neural networks (DNNs) are expected to be deployed across novel domains. However, out-of-distribution (OOD) data does not satisfy the assumption and hinders DNNs from generalizing well to target domains. In this paper, we propose Mix and Reason (MIRa) for domain generalization (DG) via perceiving and reasoning over semantic topology. Specifically, MIRa divides input common and domain-specific components and captures their potential spurious correlations. Reasoning over semantic topology enables MIRa to disentangle attribute and domain-specific representations and learn domain-invariant representations. Extensive experiments on four standard DG benchmarks show that MIRa achieves state-of-the-art performance. Besides, MIRa can be applied to many DG methods and significantly improve their performance.
GPT-4o	Domain Generalization (DG) for deep neural networks poses a significant challenge in adapting models to varied out-of-distribution data. This paper introduces "Mix and Reason" (MiRe), a novel framework leveraging semantic topology reasoning and data mixing to enhance domain generalization. MiRe employs a two-part approach, utilizing Anchored Semantic Topology Reasoning (ASTR) to retain relational features across domains and Contrastive Data Mixing (CDM) to generate robust feature representations. Experiments were conducted on four standard DG benchmarks, demonstrating substantial improvements in model adaptability and performance. Additionally, results on medical datasets underscore the frameworkś efficacy in maintaining domain-invariant characteristics across diverse conditions. This research highlights the strength of integrating semantic relationships with innovative data augmentation techniques, offering a scalable solution for realistic multi-domain scenarios.
SEGMENT & SUMMARIZE	Domain Generalization (DG) is a critical challenge in machine learning, where models are expected to generalize well across novel domains with different data distributions. Existing DG methods often rely on strong assumptions about the data or ignore potential spurious correlations. In this work, we propose a novel DG framework, Mix and Reason (MiRe), which learns semantic representations by enforcing the structural invariance of semantic topology. MiRe consists of two key components: Category-Aware Data Mixing (CDM) and Adaptive Logic Semantic Topology Refinement (ASTR). CDM mixes two images from different domains using activation maps generated by complementary classification losses, focusing the classifier on semantic object representations. ASTR introduces relation graphs to represent semantic topology, which is progressively refined via interactions between local feature aggregation and global cross-domain relational reasoning. We evaluate MiRe on four standard DG benchmarks and medical data, achieving state-of-the-art performance and outperforming existing methods. Our results demonstrate the effectiveness of MiRe in handling domain shifts and provide insights into how to better perceive and maintain structural semantic relations in machine learning.

Table 8: Sample of poster image from Chen et al. (2022) with gold reference and model generated summaries



Model	Output
Reference	Robustness in deep neural networks and machine learning algorithms in general is an open research challenge. In particular, it is difficult to ensure algorithmic performance is maintained on out-of-distribution inputs or anomalous instances that cannot be anticipated at training time. Embodied agents will be deployed in these conditions, and are likely to make incorrect predictions. An agent will be viewed as untrustworthy unless it can maintain its performance in dynamic environments. Most robust training techniques aim to improve model accuracy on perturbed inputs; as an alternate form of robustness, we aim to reduce the severity of mistakes made by neural networks in challenging conditions. We leverage current adversarial training methods to generate targeted adversarial attacks during the training process in order to increase the semantic similarity between a model's predictions and true labels of misclassified instances. Results demonstrate that our approach performs better with respect to mistake severity compared to standard and adversarially trained models. We also find an intriguing role that non-robust features play with regards to semantic similarity.
MiniCPM CoT	The paper is about the problem of mistake severity in neural networks and the use of semantic knowledge to address it. The research problem is the inability of neural networks to accurately predict the severity of mistakes, leading to poor model accuracy. The methodology involves incorporating semantic knowledge into the training process to generate targeted adversarial attacks. The key results show that using semantic knowledge to guide the attacks leads to better model accuracy and improved performance on benchmarks. The conclusion suggests that the use of semantic knowledge is a promising approach to addressing the problem of mistake severity in neural networks.
Llama-3.2-11B-V FT	It is difficult to ensure algorithmic performance is maintained at out-of-distribution (OOD) training time for neural networks due to the presence of perturbed inputs. An alternate form of robustness, we aim to reduce the conditions in which mistakes are made by neural networks. We leverage semantic similarity between a model's predictions and labels of OOD data as a measure of mistake severity. We propose a two-stage training method that generates targeted perturbations of semantic similarity between a model's predictions and true labels of OOD data. Our experiments show that our method improves upon state-of-the-art methods in aligning class similarity between a model's predictions and true labels of OOD data. Our code is available at https://github.com/natale-abreu/semi-targeted-adversarial-training.git.
GPT-4o	This paper addresses the challenge of reducing mistake severity in neural networks under adversarial and anomalous conditions by incorporating semantic knowledge into the training process. Traditional robust training techniques typically focus on improving model accuracy on perturbed inputs, but often overlook the semantic alignment between model predictions and true labels. We propose a novel method utilizing semantically targeted adversarial attacks to train neural networks, aiming to enhance semantic similarity and reduce the impact of severe errors.Our approach encompasses a two-stage training process: firstly integrating semantic knowledge through targeted adversarial perturbations, followed by standard training protocols. The methodology was tested using metrics such as super class accuracy on mistakes and average path similarity of mistakes, which measure semantic closeness between mistaken and true labels. This research highlights the value of non-robust features in capturing semantic alignments and suggests future exploration into balancing visual and semantic similarities for enhanced model reliability.
SEGMENT & SUMMARIZE	Ensuring algorithmic performance on out-of-distribution inputs or anomalous instances is a significant challenge in neural networks. Traditional robust training techniques focus on improving model accuracy on perturbed inputs, whereas we propose an alternate form of robustness by reducing the severity of mistakes made by neural networks in challenging conditions. We leverage current adversarial training methods to generate targeted adversarial attacks during the training process, increasing the semantic similarity between a model's predictions and true labels of misclassified instances. Our approach involves two stages of training: Semantic Targeting and Prior Semantic Knowledge. We demonstrate that by incorporating semantic knowledge in the training process, we can reduce the severity of mistakes in challenging conditions, thereby improving user trust in the system. Our results show that the proposed method outperforms traditional robust training techniques in terms of reducing mistake severity, making it a promising approach for addressing mistake severity in neural networks.

Table 9: Sample of poster image from the work Abreu et al. (2022) with gold reference and model generated summaries