	Flowchart			Results			Architecture			Summary		
Models (IV)	R	BS	CS	R	BS	CS	R	BS	CS	R	BS	CS
DALLE-3	0.12	0.45	0.23	0.11	0.34	0.21	0.25	0.45	0.25	0.04	0.10	0.18
Automatikz	0.20	0.48	0.25	0.20	0.50	0.29	0.24	0.45	0.28	0.27	0.56	0.45
SciDoc2Diagram												
w/ GPT4-o (ZS)	0.22	0.58	0.43	0.32	0.49	0.30	0.24	0.50	0.34	0.33	0.61	0.62
w/ GPT4-o (FS)	0.28	0.67	0.43	0.40	0.56	0.38	0.32	0.57	0.45	0.45	0.67	0.62
w/ GPT4-o-SR	0.30	0.70	0.45	0.44	0.56	0.38	0.38	0.65	0.54	0.47	0.68	0.69
w/ GPT4-o-SumMAF	0.35	0.74	0.48	0.50	0.57	0.39	0.34	0.64	0.49	0.50	0.74	0.74
w/ GPT4-o-SeqMAF	0.39	0.79	0.53	0.49	0.49	0.36	0.37	0.58	0.49	0.45	0.67	0.62

Table 2: Automatic evaluation of models on various diagrams on **SciDoc2DiagramBench-Gold** using ROUGE (R), BERTScore (BS), CLIPScore (CS) with the highest values shown in green, highlighting that flowcharts, architectures are of the best quality after sequential refinement, whereas the other ones after summarization-based refinement.

Model	Flowchart			Summary			Architecture			Results		
	C	F	L	C	F	L	C	F	L	C	F	L
GPT4-o w/o refinement	3.3	3.5	4.0	3.9	4.0	4.2	2.8	4.0	4.0	3.7	4.0	4.0
GPT4-o w/ refinement	3.9	4.0	4.0	4.5	4.1	4.2	3.2	4.0	4.0	3.7	4.5	3.8
Phi-3 w/o refinement	3.2	4.0	4.0	3.8	4.0	4.0	2.8	4.0	4.0	3.5	3.6	3.8
Phi-3 w/ refinement	3.8	4.0	4.3	4.4	4.1	4.2	3.0	4.3	4.0	3.8	4.1	3.8

Table 3: Comparison of Completeness (C), Faithfulness (F), and Layout (L) assessed by humans for different model variants on **SciDoc2DiagramBench-Extended**, indicating that on the complex set, Flowcharts and Summary tables improve significantly after refinement, but Layout-satisfaction is still comparatively lower.

Prior Work and Baselines. We have compared our approaches with existing text-to-image generation methods like DALLE3 (Rombach et al., 2021) and Automatikz (Belouadi et al., 2023). For experimenting with Diagram Planning in SciDoc2Diagrammer, we evaluate using zero-shot and few-shot (with 3 in-context exemplars) versions of LMs as mentioned in Table 8 (Prompts in 15, 16, 18, 19). We use Self-Refine by Madaan et al. (2023) as another baseline (Prompt can be found in 25).

Evaluation Metrics. SciDoc2DiagramBench-Gold and SciMultiDoc2DiagramBench-Gold include diagrams explicitly crafted by humans for each document or multiple documents, which can be served as a gold standard, allowing us to benchmark the quality of our generated diagrams. To evaluate our diagrams, we use three automated metrics: **BERTScore** (Zhang et al., 2019), **ROUGE-1** (Lin, 2004), and **CLIPScore** (Hessel et al., 2021).

However, simple token evaluations and neural evaluations like BERTScore and CLIPScore are only measuring string similarity and lacking a more finegrained evaluation framework that correlates more with human preferences (Li et al., 2024c). Thus, we go beyond string similarity scores and include a comprehensive analysis of more finegrained aspects through human evaluation and the GPT4-V Evaluation. We assess **completeness** (measures the degree to which all relevant and necessary information is in the generated figure), **faithfulness** (assesses how well the figure adheres to the facts, data, or specific instructions provided, ensuring that the content is correct and not misleading or misrepresented), and **layout** (measures visual clarity, focusing on how well the elements are structured and arranged) on a scale from 1 to 5.

6 Results and Findings

Our main objective is to explore the creation of 'Extrapolated' scientific visuals from documents using zero-shot/few-shot settings of advanced VLMs, and also analyze whether the quality of these visuals without dataset-specific fine-tuning, can be improved using our proposed SciDoc2Diagrammer-MAF. To answer this broad question, we focus our experiments on answering the following research questions as follows:

6.1 Main Research Questions and Results

RQ1: What is the best base LM for generating diagrams before refinement? GPT4-o is clearly the best-performing base/refiner model on human and automatic judgement. In Table 13, GPT-4o consistently outperforms other models across all categories of diagrams on the SciDoc2DiagramBench.