FlowGen: Synthesizing Diverse Flowcharts to Enhance and Benchmark MLLM Reasoning

*Kaiwen Shi, Sichen Liu, Ziyue Lin, Hangrui Guo, Gong Cheng**

State Key Laboratory for Novel Software Technology, Nanjing University

Mar 12, 2026

: kaiwenshi@smail.nju.edu.cn

NANJING UNIVERSITY

Background

- Flowcharts are widely used to represent processes, workflows, and decision logic in many real-world scenarios, such as:
  - Software engineering documentation
  - Business process modeling (BPMN)
  - Scientific communication
  - Education and tutorials
- By combining text, graphical symbols, and directed connections, flowcharts allow humans to quickly understand complex procedures.

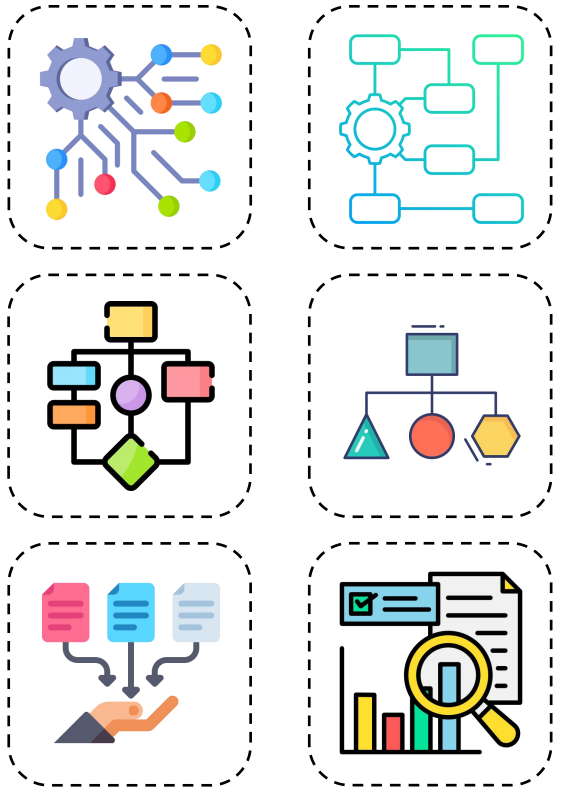
Background

Challenges for MLLM Flowchart Understanding

Structural Complexity

Visual Diversity

Data Scarcity



Background

- To advance flowchart understanding, we need datasets that:
  - Provide large-scale training data
  - Cover diverse visual rendering styles
  - Allow fine-grained control over structural complexity
  - Support structual evaluation of MLLM reasoning

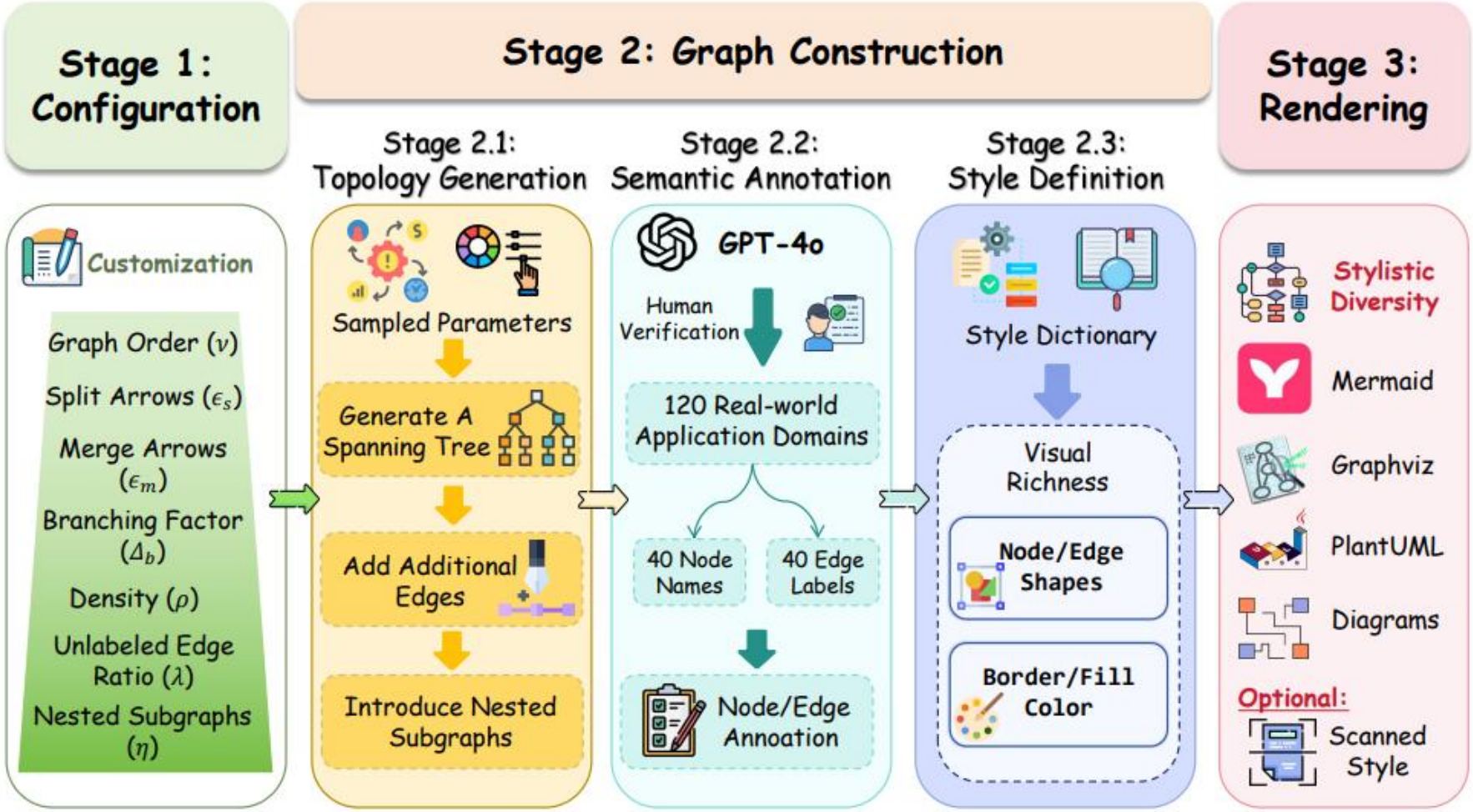
FlowGen is proposed to address these challenges by synthesizing controllable and diverse flowcharts!

Methodology

- FlowGen adopts a three-stage synthesis pipeline for controllable flowchart generation:

  - **Configuration:**
    - ✓ Define generation parameters that control the structural complexity of flowcharts.
    - ✓ node scale, split/merge arrows, branching factor, edge density, ratio of unlabeled edges, nested subgraphs
  - **Graph Generation:**
    - ✓ Construct flowchart structures as directed graphs based on the configured parameters.
    - ✓ covering 120 application domains, each with 40 node names and 40 edge labels
  - **Rendering:**
    - ✓ Support Mermaid, Graphviz, PlantUML, and Diagrams, with extensibility to scanned-style flowcharts.

Methodology



Experiments

1. Flowchart parsing evaluation on six open-source flowchart datasets.

   - **Performance improvement after SFT on the FlowGen-generated training set.**
   - **Some open-source models surpass proprietary models after training.**

Model	FlowVQA		CBD		FC_A		FC_B		hdBPMN		FlowLearn	
	F1	rF1	F1	rF1	F1	rF1	F1	rF1	F1	rF1	F1	rF1
Proprietary Multimodal Large Language Models												
GPT-4o	88.2	89.1	64.5	67.1	41.9	44.3	52.6	55.3	26.5	35.8	53.9	58.6
Gemini-2.5-Flash	62.3	63.3	63.2	66.2	38.5	40.9	52.9	55.8	8.1	10.5	82.1	86.2
GLM-4V-Plus	74.0	74.5	55.6	58.7	33.0	34.0	46.8	49.3	8.6	11.3	31.6	40.8
Open-Source Multimodal Large Language Models												
Qwen2.5-VL-3B	12.7	13.0	17.9	18.3	2.8	3.0	8.6	9.1	2.3	3.3	20.6	25.4
+*SFT*	**51.3**	**52.9**	**49.8**	**52.0**	**22.4**	**23.0**	**34.6**	**37.2**	**8.4**	**11.3**	**41.6**	**55.5**
Qwen2.5-VL-7B	59.4	59.7	49.1	51.2	25.8	27.2	32.0	33.2	16.6	20.6	43.2	48.3
+*SFT*	**70.9**	**71.9**	**55.4**	**57.7**	**29.5**	**31.4**	**40.8**	**43.7**	**18.3**	**24.2**	**60.1**	**71.2**
InternVL3-2B	15.7	17.0	18.4	19.5	4.6	5.7	8.3	9.1	**3.2**	**4.3**	21.8	32.6
+*SFT*	**34.6**	**39.4**	**37.8**	**40.7**	**13.7**	**14.7**	**21.6**	**24.3**	2.1	2.8	**13.2**	**25.2**
MiniCPM-V2.6-8B	17.6	18.5	20.4	20.9	7.5	8.7	8.8	10.2	3.4	5.3	9.1	12.7
+*SFT*	**44.0**	**47.2**	**37.9**	**39.5**	**13.4**	**14.3**	**22.9**	**24.9**	**3.6**	**5.7**	**20.0**	**31.0**
Llava-V1.6-Mistral-7B-HF	2.6	3.3	12.2	13.1	**2.8**	**3.7**	4.5	5.3	1.0	1.3	3.6	8.9
+*SFT*	**4.9**	**8.7**	**14.0**	**17.4**	2.3	2.7	**4.8**	**5.4**	**1.1**	**1.5**	**4.4**	**13.5**
Gemma3-4B-IT	8.4	9.9	17.3	17.8	3.6	5.4	4.2	6.3	**2.2**	**2.9**	17.0	24.8
+*SFT*	**19.6**	**29.8**	**28.6**	**35.7**	**15.3**	**18.8**	**7.5**	**11.2**	1.2	1.7	**11.4**	**21.0**
Gemma3-12B-IT	36.4	40.9	36.2	39.0	11.6	14.8	12.7	19.0	**10.1**	**13.7**	29.7	37.3
+*SFT*	**38.9**	**46.3**	**39.3**	**45.5**	**13.1**	**15.8**	**17.9**	**23.1**	3.7	4.6	**24.4**	**39.0**

Experiments

2. Flowchart QA evaluation on four open-source QA datasets.

   - **Parsing results provided by FlowGen further improve QA accuracy.**
   - **MiniCPM-V2.6-8B surpasses GPT-4o on FlowLearn.**

Model	FlowVQA	FlowLearn	AI2D	MISS-QA
GPT-4o	90.2	83.2	71.7	63.0
Gemini-2.5-Flash	88.0	83.4	63.2	67.3
GLM-4V-Plus	86.4	89.8	74.6	57.0
Qwen2.5-VL-3B	64.6	72.3	50.7	35.6
+ Gold-Standard Triplets	**75.6**	**89.1**	**53.5**	**47.3**
+ Self-Extracted Triplets	66.2	77.1	51.2	37.2
+ FlowGen-Extracted Triplets	73.4	88.2	51.4	38.7
Qwen2.5-VL-7B	74.6	71.1	60.9	42.1
+ Gold-Standard Triplets	**83.2**	**73.9**	**62.5**	**55.3**
+ Self-Extracted Triplets	77.1	70.7	60.5	44.6
+ FlowGen-Extracted Triplets	81.1	68.6	61.1	49.2
MiniCPM-V2.6-8B	61.6	80.9	32.0	31.1
+ Gold-Standard Triplets	**71.4**	**85.9**	**45.0**	**37.8**
+ Self-Extracted Triplets	61.0	81.1	32.6	31.5
+ FlowGen-Extracted Triplets	69.1	85.5	42.8	33.4

Experiments

3. FlowGen Test Set Experiments

- **Structural Complexity:** difficulty is controlled by adjusting node scale, split/merge arrows, and the number of nested subgraphs.

- **Scanned-style Rendering:** introduces image-level scanning artifacts such as blur, perspective distortion, and lossy compression.

Model	Test Subsets					
	Graph-Easy	Graph-Medium	Graph-Hard	Scanned-Easy	Scanned-Medium	Scanned-Hard
GPT-4o	38.4	16.2	12.0	24.0	23.2	20.5
Gemini-2.5-Flash	43.2	13.7	9.7	22.9	23.1	21.6
GLM-4V-Plus	30.0	9.4	5.3	15.3	19.3	18.3
Qwen2.5-VL-3B	20.4	6.2	2.9	10.2	10.0	9.0
+ SFT on Graph Easy	**50.2**	24.0	16.0	33.0	30.5	29.0
+ SFT on Graph Medium	36.3	**32.9**	28.2	33.1	33.1	31.3
+ SFT on Graph Hard	26.9	30.5	**30.3**	29.4	29.8	28.4
+ SFT on Scanned Easy	45.1	31.8	29.0	34.8	35.3	34.1
+ SFT on Scanned Medium	43.6	32.2	28.5	34.8	35.5	33.8
+ SFT on Scanned Hard	43.7	32.6	29.3	35.6	35.5	**35.0**
+ SFT on Combined	49.9	31.0	24.6	**36.3**	**35.7**	33.1
Qwen2.5-VL-7B	35.7	10.9	7.7	18.9	18.5	17.2
+ SFT on Graph Easy	**74.9**	35.0	21.2	45.3	44.3	41.9
+ SFT on Graph Medium	67.3	**52.4**	**44.0**	**56.2**	**55.3**	**52.3**
+ SFT on Graph Hard	39.1	43.6	43.3	42.6	42.9	40.8
+ SFT on Scanned Easy	66.8	47.8	41.6	53.9	53.1	49.7
+ SFT on Scanned Medium	68.8	48.1	42.0	54.4	53.6	51.0
+ SFT on Scanned Hard	66.0	47.1	42.0	53.0	52.6	50.2
+ SFT on Combined	66.5	47.2	42.6	52.8	51.7	50.2
MiniCPM-V2.6-8B	7.3	2.1	2.1	4.0	4.3	3.5
+ SFT on Graph Easy	33.0	13.1	9.0	18.8	18.1	17.7
+ SFT on Graph Medium	23.9	19.6	15.2	20.0	19.9	18.1
+ SFT on Graph Hard	17.5	17.4	17.4	17.8	17.9	16.8
+ SFT on Scanned Easy	29.6	19.1	16.5	22.2	22.1	20.8
+ SFT on Scanned Medium	29.6	19.0	16.1	22.3	22.3	20.4
+ SFT on Scanned Hard	27.1	18.0	16.3	21.3	20.4	19.4
+ SFT on Combined	**39.5**	**24.2**	**20.6**	**29.0**	**28.8**	**27.1**

Conclusions

In this paper, Our contribution can be summarized as follows:

- We propose **FlowGen**, a controllable flowchart synthesizer with explicit structural configurations and diverse rendering styles.
- We provide **scalable training data** and **a challenging test set** for flowchart parsing, showing that current MLLMs struggle with high structural complexity and diverse rendering styles.
- Our experiments demonstrate that fine-tuning on FlowGen **improves** the robustness and generalization of MLLMs.
- In the future, we plan to incorporate more real-world flowchart features and extend FlowGen to more complex reasoning tasks, such as IT root cause analysis.

Q & A

Thank you!

Kaiwen Shi

State Key Laboratory for Novel Software Technology, Nanjing University

Mar 12, 2026

南京大学
NANJING UNIVERSITY