From EduVisBench to EduVisAgent: A Benchmark and Multi-Agent Framework for Reasoning-Driven Pedagogical Visualization

Haonian Ji* UNC-Chapel Hill
Shi Qiu* UNC-Chapel Hill
Siyang Xin* UNC-Chapel Hill
Siwei Han* UNC-Chapel Hill

Zhaorun Chen University of Chicago
Dake Zhang Rutgers University
Hongyi Wang Rutgers University
Huaxiu Yao UNC-Chapel Hill

Abstract

Foundation models (FMs) have shown promise in educational contexts, yet struggle to generate pedagogically effective visual explanations for complex reasoning tasks. Current approaches focus primarily on textual outputs, overlooking the critical role of structured visualizations in supporting conceptual understanding. To address this gap, we introduce EduVisBench, a multi-domain benchmark featuring 1154 STEM problems requiring visually grounded solutions, evaluated using a fine-grained rubric informed by pedagogical theory. Our analysis reveals that existing models frequently fail to decompose complex reasoning into interpretable visual representations aligned with human cognitive processes. To overcome these limitations, we propose EduVisAgent, a multi-agent collaborative framework coordinating specialized agents for instructional planning, reasoning decomposition, metacognitive prompting, and visualization design. Experimental results demonstrate that EduVisAgent substantially outperforms all baselines, achieving a 40.2% improvement in generating educationally effective visualizations.

1 Introduction

While foundation models (FMs) have been extensively adopted in educational domains [Chu et al., 2025, Wang et al., 2024], their applications have predominantly focused on text-based interactions. However, creating compelling visualizations is crucial for cognitive comprehension and learning effectiveness in K-12 education [Presmeg, 2006]. Current FMs show limited ability to generate visually grounded pedagogical explanations.

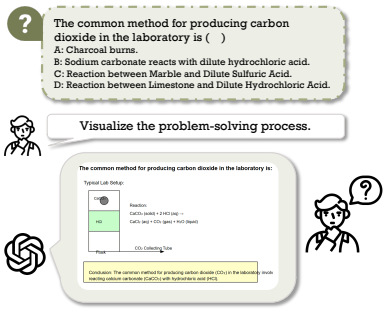

Figure 1: GPT-4o fails to illustrate its problem-solving with high-quality, logical, and explanatory visualization.

Generating pedagogically effective visualizations poses three core challenges: (1) decomposing complex reasoning into steps that align with human cognitive processes, (2) producing visual aids that optimally support learners for each sub-step, and (3) adapting visualization styles across different educational domains. These obstacles stem from the complex task of translating abstract pedagogical concepts into intuitive visual narratives. More detailed discussions are provided in Appendix A.

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

To address these challenges, we introduce **EduVisBench**, a multi-domain benchmark designed to evaluate FMs' capacity to generate step-by-step visual reasoning for education. EduVisBench comprises 1,154 structured problems across diverse STEM domains, each requiring visualizations that prioritize interpretability, cognitive alignment, and instructional clarity. We develop a fine-grained evaluation rubric focusing on pedagogical criteria including contextual relevance, visual clarity, multimodal coherence, reasoning support, and interactive engagement. Our systematic evaluation reveals that current models achieve correct textual analyses but frequently fail to generate useful visualizations, as shown in Figure 1. Specific challenges include semantic misalignments between text and visuals, omission of critical reasoning steps, and structural inconsistencies in visual outputs.

To overcome these limitations, we propose **EduVisAgent**, a multi-agent collaborative framework that simulates expert instructional design through specialized agents for visualization design, cognitive scaffolding, and metacognitive regulation. A central planning agent orchestrates six expert agents, with outputs synthesized into interactive learning experiences. Experimental results show EduVisAgent achieves an average improvement of 40.2% over current state-of-the-art methods, demonstrating the effectiveness of modular specialization and collaborative integration for educational visualization.

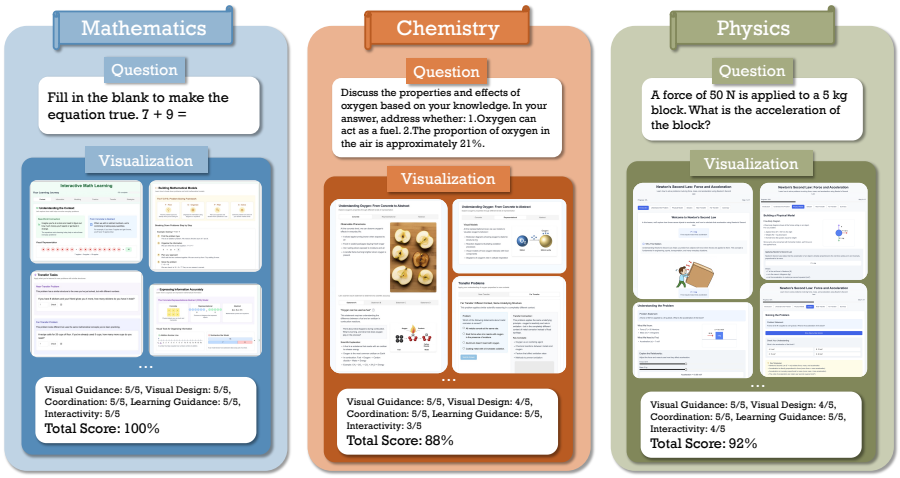

Figure 2: Representative examples from EduVisBench, featuring questions from Maths, Chemistry, and Physics alongside their corresponding high-scoring visual explanations.

2 EduVisBench Benchmark

We introduce EduVisBench, a benchmark designed to evaluate models' capability to generate pedagogically effective visualizations. EduVisBench comprises 1,154 STEM questions across three subjects: Mathematics, Physics, Chemistry(Figure 2) and 15 domains, organized into three difficulty levels (Figure 3). Unlike traditional benchmarks focusing solely on answer accuracy, EduVisBench emphasizes visual reasoning quality—evaluating how well models communicate problem-solving processes through structured, interpretable visualizations.

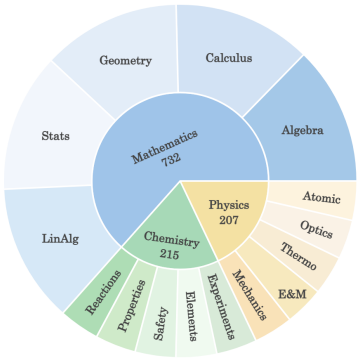

Figure 3: Dataset distribution of EduVisBench, covering 15 comprehensive pedagogical scenarios.

Models receive multimodal inputs and generate diverse outputs including interactive webpages and diagrams. We evaluate outputs using a five-dimensional rubric: **(1) Context Visualization**—situating problems in relevant contexts; **(2) Diagram Design**—clarity and accuracy of visual representations; **(3) Text-Graphic Integration**—coherence between explanations and visuals; **(4) Thought Guidance**—support for reasoning processes; and **(5) Interactivity**—engagement and manipulation features.

Our evaluation protocol standardizes all outputs to raster images for fair comparison. Interactive webpages are rendered via headless browser with systematic state traversal. GPT-4o evaluates each

dimension on a 0-5 scale, yielding normalized scores (0-100). Detailed dataset curation, rubrics, and evaluation procedures are provided in Appendix B.

3 EduVisAgent

Systematic evaluation on EduVisBench reveals that existing models perform poorly, with average scores below 50 on a 0-100 scale (detailed results in Section 4). This underperformance stems from the inherent challenge of decomposing complex reasoning and translating it into visual representations aligned with human cognitive processes, a task that remains highly non-trivial for monolithic architectures.

To address these challenges, we propose EduVisAgent, a multi-agent system inspired by pedagogical theories that emulates expert instructional design through division of labor and collaborative reasoning. EduVisAgent consists of five specialized agents:

Task Planning Agent structures the instructional objective by: (1) breaking problems into coherent subgoals, (2) clarifying expected reasoning at each step, (3) aligning steps with domain-specific principles, and (4) anticipating student misconceptions. This provides a pedagogically grounded foundation for downstream agents. **Conceptual Mapping Agent** extracts and organizes core problem components using the Concrete–Representational–Abstract (CRA) instructional model [Nugroho and Jailani, 2019]. It classifies information into concrete entities, representational elements, and abstract constructs, bridging the gap between concrete problem elements and abstract solution principles. **Reasoning Decomposition Agent** applies the memory-oriented FOPS strategy [Miller and Cohen, 2020]—*Find* the problem type, *Organize* structure via equations/diagrams, *Plan* the solution path, and *Solve* the task. It identifies critical instructional points requiring visual scaffolding or interactive guidance. **Metacognitive Reviewer** generates reflective prompts grounded in metacognitive theory [Schraw and Moshman, 1995], fostering self-questioning and self-correction to help learners evaluate their problem-solving approaches. **Visualization Agent** constructs pedagogically effective visual representations including number lines, charts, schematic illustrations, and interactive diagrams. Rather than decorative visuals, it emphasizes abstract yet educationally meaningful representations tightly aligned with underlying concepts, rendered via `v0` [Vercel, 2025].

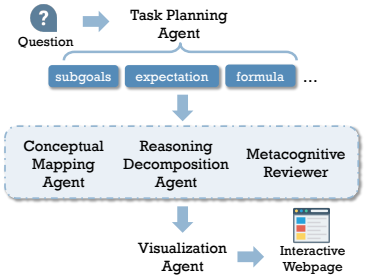

Figure 4: The structure of EduVisAgent.

The agents operate sequentially yet interdependently. As shown in Figure 4, the Task Planning Agent first structures the instructional task, followed by Conceptual Mapping and Reasoning Decomposition agents analyzing and organizing content. The Metacognitive Reviewer adds reflective elements, while the Visualization Agent synthesizes outputs into interactive learning experiences. This modular design introduces pedagogical interpretability by embedding distinct instructional roles directly into the workflow, enabling targeted improvements and transparent reasoning processes.

4 Experiments

We evaluate various foundation models on EduVisBench to address: (1) How do existing models perform on educational visualization? (2) Can EduVisAgent outperform current approaches? (3) What performance patterns emerge across model types and domains?

4.1 Experimental Setup

We evaluate three categories: **(1) Diffusion Models**: Flux.1-dev [Labs, 2024], Stable Diffusion 3.5 Large (SD3.5) [IT Admin, 2024], and Stable Diffusion XL Base 1.0 (SDXL) [Podell et al., 2023]. (2) Large Vision-Language Models (LVLMs): Deepseek-VL2 [Wu et al., 2024], GLM-4V-9B [GLM et al., 2024], MiniCPM-V2.6 [Yao et al., 2024], Mistral-Small-3.1-24B-Instruct-2503 [Mistral AI, 2025], Phi-3.5-Vision-Instruct [Abdin et al., 2024], Phi-4-Multimodal-Instruct [Abouelenin et al., 2025], Qwen2.5-VL-72B [Team, 2025], GPT-4o [Hurst et al., 2024], Claude 3.7 Sonnet [Anthropic,

2025], and Gemini 2.0 Flash [Mallick and Kilpatrick, 2025]. (3) Specialized Visualization Agent: v0 [Vercel, 2025]. (4) Our EduVisAgent. All outputs are standardized to images. Interactive pages are captured via automated screenshot traversal. GPT-4o scores each visualization on our five-dimensional rubric (0-5 per dimension, normalized to 0-100).

4.2 Results and Analysis

Table 1: Performance of Diffusion Models, Large Vision Language Models, v0 and EduVisAgent on EduVisBench.

Method	Vis. Type	Maths			Physics			Chemistry			Avg
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	
Diffusion Model											
Flux.1-dev	Image	13.8	13.4	13.2	11.7	8.5	10.0	20.0	16.6	16.0	13.8
SD3.5	Image	17.3	20.3	18.8	16.8	13.0	12.0	22.8	21.7	34.0	18.4
SDXL	Image	17.3	23.3	25.5	18.9	15.4	24.0	33.6	30.2	24.0	21.8
Large Vision Language Model											
Deepseek VL2	Webpage	20.3	17.1	15.7	17.9	17.0	20.0	16.4	13.8	14.0	17.5
GLM-4V-9B	Webpage	22.3	21.1	19.4	24.5	21.5	24.0	22.3	21.5	16.0	21.9
MiniCPM-V-2.6	Webpage	24.1	17.3	15.5	19.1	17.4	20.0	14.5	15.2	12.0	19.3
Mistral-Small-3.1	Webpage	29.1	31.6	32.2	32.3	33.5	20.0	30.6	27.5	24.0	30.2
Phi-3.5	Webpage	25.3	20.7	19.1	21.2	19.5	12.0	20.0	18.6	20.0	21.8
Phi-4	Webpage	26.1	25.1	22.9	27.8	25.5	24.0	31.2	27.5	12.0	26.4
Qwen2.5-VL-72B	Webpage	24.3	18.1	15.8	19.7	17.1	24.0	18.2	16.4	12.0	20.0
Claude 3.7 Sonnet	SVG	61.2	26.7	23.6	18.5	16.9	14.0	47.5	47.2	18.0	42.0
Claude 3.7 Sonnet	Webpage	56.2	57.5	55.6	44.8	42.6	24.0	61.1	60.6	64.0	54.6
GPT-4o	Webpage	47.6	39.3	37.9	25.7	24.2	24.0	34.3	32.6	36.0	38.1
GPT-4o	SVG	36.1	19.7	19.5	13.0	12.8	4.0	30.0	27.5	22.0	26.3
Gemini 2.0 Flash	Webpage	46.9	9.5	15.7	31.7	26.5	24.0	32.0	25.8	30.0	43.6
Visualization Agent											
v0	Webpage	63.0	37.6	47.2	53.3	58.5	52.0	**74.7**	52.8	68.0	58.2
Multi-Agent Visualization Framework											
EduVisAgent(ours)	Webpage	**90.2**	**64.5**	**65.0**	**85.3**	**81.7**	**84.0**	69.0	**76.3**	**76.0**	**81.6**

Table 1 shows that existing models struggle with educational visualization. Diffusion models perform poorly (13.8-21.8%), as static image generation cannot capture the nuanced requirements of explanatory visualizations. Most LVLMs score 17.5-30.2%, with top performers like Claude 3.7 Sonnet reaching 54.6% for webpage generation. The specialized v0 agent achieves 58.2%, highlighting the advantage of task-specific design.

EduVisAgent Performance. Our multi-agent system achieves 81.6% average score, representing a 40.2% relative improvement over the best baseline (v0 at 58.2%). EduVisAgent demonstrates consistent performance across domains: Mathematics (73.2%), Physics (83.7%), and Chemistry (73.8%). This substantial improvement validates our hypothesis that collaborative specialization and pedagogical theory integration effectively address the limitations of monolithic approaches.

Key Insights. The results reveal three critical findings: (1) *Specialization matters*: Task-specific agents (v0, EduVisAgent) significantly outperform general-purpose models; (2) *Multi-agent collaboration is effective*: EduVisAgent's collaborative approach surpasses even specialized single agents; (3) *Pedagogical theory integration works*: Our theory-grounded design principles translate into measurable performance gains across all evaluation dimensions. We present additional experiments and analysis in Appendix F.

5 Conclusion

We addressed the challenge of generating pedagogically effective visual explanations through two key contributions: **EduVisBench**, a comprehensive benchmark revealing significant limitations in existing

models' educational visualization capabilities, and **EduVisAgent**, a multi-agent framework that leverages pedagogical theories to achieve 40.2% performance improvement. Our work demonstrates that collaborative specialization and educational theory integration are crucial for advancing AI-driven educational visualization, opening new directions for intelligent tutoring systems and educational technology.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Young Jin Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Praneetha Vaddamanu, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Andre Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, Xiren Zhou, and Yifan Yang. Phi-3 technical report: A highly capable language model locally on your phone. ArXiv, abs/2404.14219, 2024. URL https://api.semanticscholar.org/CorpusID:269293048.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv preprint arXiv:2503.01743, 2025.

Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, February 2025. Accessed: 2025-05-16.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. arXiv preprint arXiv:2410.07095, 2024.

Jiawen Chen, Jianghao Zhang, Huaxiu Yao, and Yun Li. Celltypeagent: Trustworthy cell type annotation with large language models. arXiv preprint arXiv:2505.08844, 2025a.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? arXiv preprint arXiv:2407.04842, 2024a.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. Advances in Neural Information Processing Systems, 37:130185–130213, 2024b.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. arXiv preprint arXiv:2403.00425, 2024c.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. arXiv preprint arXiv:2402.11452, 2024d.

Zhaorun Chen, Mintong Kang, and Bo Li. Shieldagent: Shielding agents via verifiable safety policy reasoning. arXiv preprint arXiv:2503.22738, 2025b.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. Llm agents for education: Advances and applications. arXiv preprint arXiv:2503.11733, 2025.

Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. arXiv preprint arXiv:2410.14148, 2024.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. Plotgen: Multi-agent llm-based scientific data visualization via multimodal feedback. arXiv preprint arXiv:2502.00988, 2025.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. arXiv preprint arXiv:2503.13964, 2025.

Jiayi Hong, Christian Seto, Arlen Fan, and Ross Maciejewski. Do llms have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation. IEEE Transactions on Visualization and Computer Graphics, 2025.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

IT Admin. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, October 2024. Updated October 29, 2024; Accessed: 2025-05-20.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pages 1–28, 2025.

Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. Theoremexplainagent: Towards multimodal explanations for llm theorem understanding. arXiv preprint arXiv:2502.19400, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. arXiv preprint arXiv:2305.20050, 2023.

Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, flash-lite and pro. https://developers.googleblog.com/en/gemini-2-family-expands/, February 2025. Accessed: 2025-05-20.

Chad M. Miller and Jonathan D. Cohen. Metacognitive prompts in multimedia learning: A meta-analysis. Educational Psychology Review, 32(3):979–1003, 2020. doi: 10.1007/s10648-020-09525-3. URL https://files.eric.ed.gov/fulltext/EJ1262620.pdf.

Mistral AI. Mistral small 3.1: Sota. multimodal. multilingual. apache 2.0. https://mistral.ai/news/mistral-small-3-1, March 2025. Accessed: 2025-05-20.

Inderjeet Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. arXiv preprint arXiv:2410.08058, 2024.

Fan Nie, Lan Feng, Haotian Ye, Weixin Liang, Pan Lu, Huaxiu Yao, Alexandre Alahi, and James Zou. Weak-for-strong: Training weak meta-agent to harness strong executors. arXiv preprint arXiv:2504.04785, 2025.

Satria Nugroho and Jailani Jailani. The effectiveness of concrete representational abstract approach (cra) approach and problem solving approach on mathematical representation ability at elementary school. KnE Social Sciences, 06 2019. doi: 10.18502/kss.v3i17.4620.

Saugat Pandey and Alvitta Ottley. Benchmarking visual language models on standardized visualization literacy tests. arXiv preprint arXiv:2503.16632, 2025.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. ArXiv, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

Luca Podo, Muhammad Ishmal, and Marco Angelini. Vi (e) va llm! a conceptual stack for evaluating and interpreting generative ai-based visualizations. arXiv preprint arXiv:2402.02167, 2024.

Norma Presmeg. Research on visualization in learning and teaching mathematics. Handbook of research on the psychology of mathematics education, pages 205–235, 2006.

M. Rohith. High school physics. https://huggingface.co/datasets/mrohith29/high-school-physics, 2023. Accessed: 2025-05-15.

Gregory Schraw and David Moshman. Metacognitive theories. Educational Psychology Review, 7(4): 351–371, 1995. ISSN 1040726X, 1573336X. URL http://www.jstor.org/stable/23359367.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Pere-Pau Vázquez. Are llms ready for visualization? In 2024 IEEE 17th Pacific Visualization Conference (PacificVis), pages 343–352. IEEE, 2024.

Vercel. v0: Ai chat interface for web automation. https://v0.dev, 2025. Accessed: 2025-05-20.

Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. Llm4vis: Explainable visualization recommendation using chatgpt. arXiv preprint arXiv:2310.07652, 2023a.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. arXiv preprint arXiv:2403.18105, 2024.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605, 2025.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning? arXiv preprint arXiv:2310.07018, 2023b.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. Medco: Medical education copilots based on a multi-agent framework. arXiv preprint arXiv:2408.12496, 2024.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. Mathchat: Converse to tackle challenging math problems with llm agents. arXiv preprint arXiv:2306.01337, 2023.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. Eduagent: Generative student agents in learning. arXiv preprint arXiv:2404.07963, 2024.

Yiying Yang, Wei Cheng, Sijin Chen, Xianfang Zeng, Jiaxu Zhang, Liao Wang, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. Omnisvg: A unified scalable vector graphics generation model. arXiv preprint arXiv:2504.06263, 2025.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, et al. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. arXiv preprint arXiv:2402.11453, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In International Conference on Learning Representations (ICLR), 2023.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, Shufei Zhang, Mao Su, Hansen Zhong, Yuqiang Li, and Wanli Ouyang. Chemllm: A chemical large language model, 2024.

Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D Lindsay, and Johannes Bjerva. Sefl: Harnessing large language model agents to improve educational feedback systems. arXiv preprint arXiv:2502.12927, 2025.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, et al. Anyprefer: An agentic framework for preference data synthesis. arXiv preprint arXiv:2504.19276, 2025.

A Related Work

LLM for Pedagogical Assistance. Foundation models (FMs), including diffusion models and large vision-language models (LVLMs), have been increasingly adopted in educational contexts [Chu et al., 2025, Wang et al., 2024] to support teaching and classroom interactions. EduAgent [Xu et al., 2024] and Teachtune [Jin et al., 2025] enhance the problem-solving process through automated simulations of student-teacher dialogues, collaborative learning, and task-oriented reasoning. Agents such as SEFL [Zhang et al., 2025] and PROF [Nair et al., 2024] synthesize immediate, on-demand feedback to support large-scale instructional scenarios. Furthermore, domain-specific agents such as MathChat [Wu et al., 2023], NEWTON [Wang et al., 2023b], and MEDCO [Wei et al., 2024] further provide textual explanations tailored to scientific and medical education. While these systems address diverse pedagogical needs, their focus remains largely on text-based interactions [Wu et al., 2023, Xu et al., 2024, Cui et al., 2024], overlooking the critical role of visualization in fostering conceptual understanding and improving learning outcomes [Presmeg, 2006]. While valuable, these text-centric systems do not address the large body of educational research highlighting the unique cognitive benefits of visual learning. Despite its pedagogical importance, the capacity of FMs and agents to generate logical, explanatory visual illustrations remains underexplored. EduVisBench is the first comprehensive benchmark designed to systematically evaluate FMs' ability to produce pedagogically effective, step-by-step visual reasoning, covering 15 diverse visually grounded educational scenarios with multi-level problem sets and multimodal-centric solutions.

LLM for Scientific Visualization. While some existing works have preliminarily explored the potential of FMs in supporting visual scaffolding [Podo et al., 2024, Chen et al., 2024c, Pandey and Ottley, 2025, Hong et al., 2025], they are typically fragmented, lack pedagogical grounding, and fail to generalize across diverse educational tasks [Wang et al., 2023a, Ku et al., 2025]. For instance, Visual Sketchpad [Hu et al., 2024] attempts to illustrate problem-solving processes with sketches generated from code. However, these visuals are often low in quality, lack logical coherence, and fall short in explanatory depth [Wang et al., 2025]. Other approaches like MatplotAgent [Yang et al., 2024], PlotGen [Goswami et al., 2025], and OmniSVG [Yang et al., 2025] leverage plotting and SVG tools to produce more accurate, data-grounded visualizations. Still, these methods are limited in scope, often addressing only isolated steps rather than providing systematic, end-to-end visual explanations of multi-step problem-solving tasks [Vázquez, 2024, Chen et al., 2024a, 2025b]. Our framework, in contrast, is designed to manage the entire pedagogical workflow, from problem deconstruction to the final interactive explanation. To overcome these limitations, we propose a multi-agent collaborative framework, EduVisAgent, that simulates the full learning journey—from initial problem exposure to deep conceptual understanding—by coordinating specialized agents to generate coherent, pedagogically aligned visualizations throughout the reasoning process.

LLM-based Education Agents. Recent advancements in LLM-based agents have led to the development of specialized architectures capable of long-horizon planning, tool use, and memory management across a range of real-world domains [Yao et al., 2023, Chan et al., 2024, Chen et al., 2024b, 2025a, Nie et al., 2025, Han et al., 2025, Zhou et al., 2025]. In the educational domain, AI agents such as EduAgent [Xu et al., 2024] and Teachtune [Jin et al., 2025] simulate student-teacher dialogues, collaborative learning activities, and task-oriented reasoning to enhance problem-solving instruction. Agents like SEFL [Zhang et al., 2025] and PROF [Nair et al., 2024] generate on-demand feedback for large-scale educational settings, while domain-specific tools such as MathChat [Wu et al., 2023], NEWTON [Wang et al., 2023b], and MEDCO [Wei et al., 2024] provide textual explanations for scientific and medical learning. Despite these advances, limited research has investigated collaborative, multi-agent approaches tailored to educational reasoning and visualization. EduVisAgent is the first systematic multi-agent framework that coordinates specialized agents and provides a comprehensive approach to supporting step-by-step pedagogical problem-solving.

Current Difficulties in Pedagogical Intepretation. Currently, generating visually grounded elements for pedagogical reasoning poses several challenges: (1) decomposing complex reasoning into representable steps that align closely with human cognitive processes is non-trivial [Yang et al., 2024, Chen et al., 2024d]; (2) precisely producing visual aids for each sub-step to optimally support learners is challenging [Hong et al., 2025]; and (3) different educational domains require distinct visualization styles and formats, which makes consistent and adequate visual aid delivery difficult [Pandey and Ottley, 2025]. This difficulty stems not just from technical rendering challenges, but from the complex task of translating abstract pedagogical concepts into intuitive visual narratives. Addressing

these obstacles first requires a picture of how current FMs perform, so that future models can be purpose-built to close the gaps. Consequently, a comprehensive evaluation platform is critical for systematically assessing FMs on visual pedagogical reasoning.

B Curation and Evaluation of EduVisBench

B.1 Dataset Curation

EduVisBench is built from several high-quality public educational resources that we carefully curated, translated, and adapted to support multimodal visualization learning tasks. Specifically, the chemistry questions are sourced from the *C-MHChem-Benchmark* [Zhang et al., 2024], originally presented in Chinese and meticulously translated into English with careful attention to scientific accuracy and terminology. The physics questions are drawn from the *high-school-physics* [Rohith, 2023] dataset, which includes a range of conceptual and quantitative exercises suitable for secondary-level learners. The mathematics component combines easy-level problems from the Illustrative Mathematics curriculum with medium- to hard-level questions selected from the *MATH-500* [Lightman et al., 2023] dataset.

Furthermore, each domain encompasses diverse sub-domains, collectively covering 15 comprehensive scenarios, as illustrated in Figure 3. All data sources were standardized into a unified format and consolidated to enable consistent and comprehensive evaluation across subjects.

B.2 Evaluation Rubrics

To comprehensively evaluate the quality of generated visualizations in supporting student understanding and learning, we introduce a fine-grained scoring metric grounded in five pedagogically motivated dimensions: **(1) Context Visualization**—situating problems in relevant contexts; **(2) Diagram Design**—clarity and accuracy of visual representations; **(3) Text-Graphic Integration**—coherence between explanations and visuals; **(4) Thought Guidance**—support for reasoning processes; and **(5) Interactivity**—engagement and manipulation features. We detail these

B.3 Evaluation Protocol

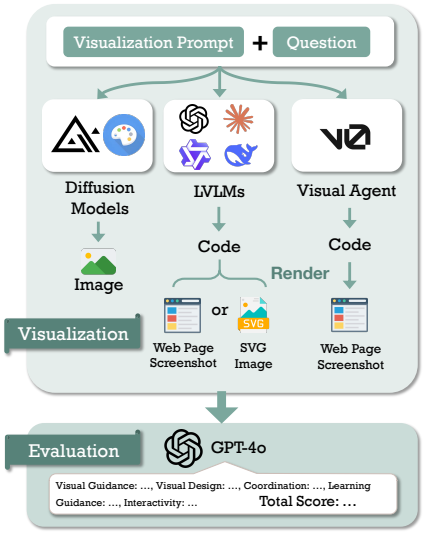

Figure 5: Detailed workflow for evaluation process.

As shown in Figure 5, models are provided with a visualization prompt together with a question and are asked to generate visual outputs. To enable fair comparison across heterogeneous outputs, we first canonicalize every model result to a raster image prior to scoring.

B.3.1 Output Standardization Process

This standardization is a crucial step that ensures all systems are evaluated on a level playing field, independent of their native modality or file format, and prevents format-specific rendering artifacts from biasing the assessment. Visuals produced directly as SVG or PNG are used as-is. Web pages (HTML or Next.js) are rendered in a headless browser and captured as screenshots of the primary view; when lightweight interactivity is present (e.g., buttons, tabs, or toggles), we systematically traverse the reachable states and retain one representative screenshot per state.

B.3.2 Scoring Methodology

All resulting images are then evaluated by GPT-4o along five dimensions to compute an overall performance score. Each dimension is rated on a 0-5 scale; the ratings are summed (0-25) and, when appropriate, normalized to a percentage to yield the final overall score.

C Visualization disciplines

Table 2 illustrates the disciplines and types in our EduVisBench.

Table 2: Representative Visualization Types Across Academic Disciplines

Discipline	Common Visualization Types
Mathematics	Number lines, function graphs, and other formalized visual tools.
Physics	Diagrams involving levers, rigid body motion, forces, and fields.
Chemistry	Molecular structures and schematic representations of standard laboratory apparatus.

D Evaluation Metric

Context Visualization The category of "Context Visualization" outlines different levels of visualizing mathematical concepts, progressing from basic text-only representations to highly integrated visual-text formats. Through five defined levels, the framework demonstrates how visual elements can enhance students' understanding and engagement with abstract ideas, guiding instructional designers to gradually enrich scenarios, add annotations, and strengthen contextual connections—ultimately achieving the goal of visually presenting the full flow and conceptual structure of the content.The five levels of Context Visualization are as follows:

Diagram Design The category of "Diagram Design" describes progressive levels of visual elements used to support students' systematic understanding of quantities and relationships. It ranges from no visual aids to complex integrated dashboards that deeply connect data and model structures. Through five levels, the framework guides designers to improve clarity, coherence, and contextual richness of visual illustrations, enhancing students' analytic and comparative abilities.

Text–Graphic Integration The category of "Text–Graphic Integration" describes levels of alignment and integration between textual content and visual elements within images. This progression ranges from complete disconnection to seamless fusion, enabling students to effectively map and synthesize text, formulas, and graphics. The framework guides designers in strengthening links between verbal and visual information to enhance comprehension and structural understanding.

Thought Guidance The category of "Thought Guidance" describes the progressive inclusion of visualized problem-solving strategies and reflective cues in images. From presenting only problem

Table 3: Five Levels of Context Visualization

Level	Description
Level 1	The image contains no scenes or illustrations, presenting only text and formulas. It lacks contextual visual cues, failing to spark interest or connect the concepts to real-life situations.
Level 2	The image includes a single static illustration or low-fidelity mockup with minimal labeling that does not highlight variables or key objects, offering limited context and poor immersion.
Level 3	Multiple static schematic diagrams or sketch-style illustrations appear in the image, labeling core objects, variables, and simple steps, providing basic visual guidance but lacking layered coherence.
Level 4	The image integrates scenario illustrations, storyboard panels, and infographics to present the process in multiple views and steps, with annotations and captions guiding students through mapping abstract concepts to context.
Level 5	Storyboard-style illustrations and infographics are fused into a single image, including overview, detailed close-ups, and key pathway diagrams with comprehensive annotations, allowing students to grasp the entire flow and conceptual network at a glance.

Table 4: Five Levels of Diagram Design

Level	Description
Level 1	The image contains no charts, axes, or flow diagrams—only text. Without embedded visual tools, students cannot systematically organize or analyze quantities and relationships.
Level 2	The image presents a static number line and colored bar chart with complete scales and a legend, helping students gain a basic understanding of numerical changes. However, it lacks comparison and contextual layering.
Level 3	The image presents a static number line and colored bar chart with complete scales and legends, helping students grasp basic numerical changes visually, though comparison and context layering are absent.
Level 4	The image combines number lines, flowcharts, infographics, and arrow annotations; multiple visuals are juxtaposed or overlaid to show processes and variable changes for a coherent modeling view.
Level 5	The image presents a dashboard-style visualization integrating axes, bar charts, flow diagrams, heatmaps, etc., with linked elements that deeply visualize data relationships and model structure.

statements to complex integrated dashboards, this framework guides designers to scaffold students' strategic thinking and metacognitive reflection through visual tools, enabling deeper reasoning and transfer of learning.

Interactivity The category of "Interactivity" outlines levels of incorporating feedback, hints, and tailored assistance into images, evolving from static presentations to dynamic, student-responsive visual supports. This framework encourages designers to embed interactive elements that adapt to learner needs, promoting engagement and personalized problem-solving.

Table 5: Five Levels of Text–Graphic Integration

Level	Description
Level 1	Text and illustrations in the image are completely disconnected, with no labels, legends, or connectors—students cannot use visuals to understand text or formulas.
Level 2	Text occasionally prompts "see diagram" or "refer to the illustration," but the image lacks legends or clear labels, so mapping between text and graphics remains ambiguous.
Level 3	Text descriptions and image elements share consistent numbering, color blocks, or arrows linked to a simple legend, explaining core symbols and variables to support initial mapping.
Level 4	Text paragraphs are laid out alongside corresponding visuals within the same image, with detailed legends and color-coded annotations enabling simultaneous reading and mapping.
Level 5	Text, formulas, and legends are fully integrated in one image, using consistent colors, numbering, and layered layout to achieve seamless text–graphic fusion for complete structural understanding.

Table 6: Five Levels of Thought Guidance

Level	Description
Level 1	The image offers no visualized problem-solving guidance, showing only the problem statement and formulas, leaving students without strategic cues or reflection prompts.
Level 2	The image embeds a simple flowchart or two title-style hints (e.g., "Identify problem type," "Check result"), but the flowchart is overly simplistic and hints lack hierarchical detail.
Level 3	The image displays a step-by-step flowchart template with key thinking nodes and self-check checkpoints, leaving annotation space for students to visually record their reasoning.
Level 4	The image combines a near-transfer exercise with a comparative thought diagram, visually highlighting strategy differences so students can apply existing reasoning to a new context.
Level 5	The image fuses near- and far-transfer exercises, concept mind maps, and a reflection panel into a dashboard-style layout, allowing students to review and extend their problem-solving network visually.

E Evaluation Prompt

The instructional web page evaluation prompt is structured as follows:

Table 7: Five Levels of Interactivity

Level	Description
Level 1	The image includes no feedback or support components—only a static problem statement and answer field—offering no hints, examples, or error cues and resulting in a nonresponsive visual.
Level 2	The image shows fixed hint boxes (e.g., "Hint: draw a number line," "Hint: check rounding"), but hints are not tailored to student responses, limiting personalized guidance.
Level 3	The image integrates multiple static correction tips and example solution modules (common mistakes and standard approaches), which students can reference visually but without intelligent recommendations.
Level 4	The image presents example solution workflows, text hints, and a common-errors analysis section highlighted with color blocks and arrows, providing diverse visual support in a single layout.
Level 5	The image displays a comprehensive visual support panel with difficulty suggestions, personalized hints, worked examples, and extension resource links, enabling students to select tailored guidance directly from the visual layout.

Evaluation Prompt

As a **professional evaluator of instructional web pages**, your task is to determine whether the generated web page meets expectations across five specific categories.

Instructions:

- Assign an **integer score from 0 to 5** for each of the five categories (1–5).
- **0** = completely missing or extremely poor
  5 = fully meets the highest standard
- Evaluation should be based solely on the specified aspect: *{category}*.
  The definition of *{category}* is: *{description}*.
- **Do not include any explanation, justification, or additional commentary. Refusing to provide a score is not allowed.**

Evaluation Output Format

```
{{RATING: {"1":score, "2":score, "3":score, "4":score, "5":score}}}
```

F Additional Analysis for Experiments

F.1 Reliability of GPT-based Scoring.

To validate the reliability of our automatic judge, we compared GPT-based evaluations with human evaluations. Specifically, we selected 50 samples from each subject category (Chemistry, Math, and Physics), and had both GPT and human evaluators independently rate them. Human evaluators were undergraduate students from top

Table 8: Cosine similarity and mean squared error across subjects. Math is the average of Math500 and IllustrativeMath, each with 50 samples.

Metric	Chemistry	Math	Physics	Average
Cosine Similarity ↑	0.9742	0.9557	0.9666	**0.9655**
MSE ↓	0.3895	0.7093	0.6118	**0.5702**

universities. We measured agreement using Cosine Similarity and Mean Squared Error (MSE). As shown in Table 8, high agreement—average cosine similarity 0.9655 and MSE 0.5702 across subjects indicates negligible practical discrepancy .

F.2 Case Analysis

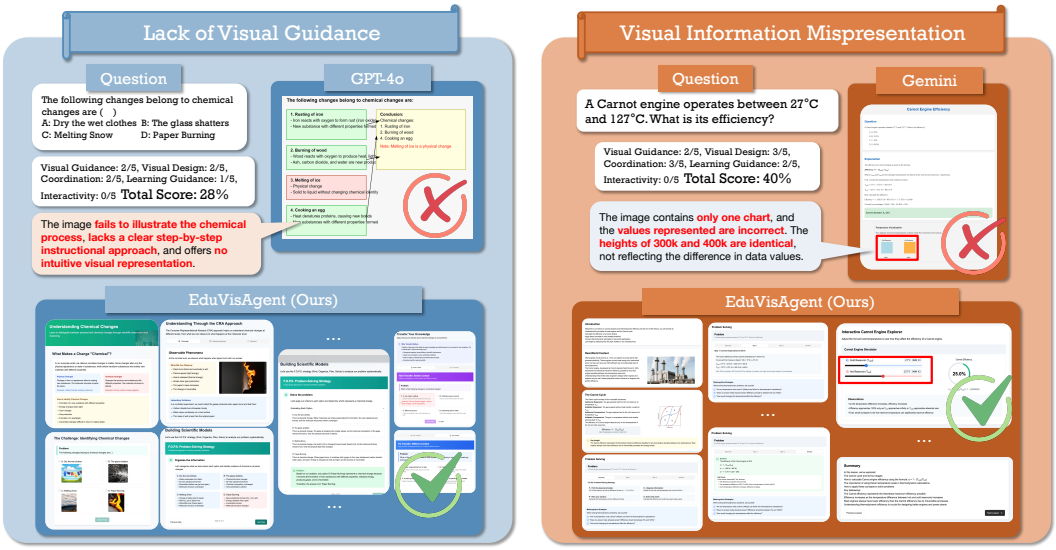

Figure 6: Baseline models versus our EduVisAgent. These examples clearly demonstrate the often poor output quality of baseline models, contrasting sharply with the high-quality, effective visualizations produced by EduVisAgent.

To further illustrate the limitations of existing baselines and how our approach addresses these challenges, we present two case studies in Figure 6. On the left, for a chemistry question, the GPT-4o-generated solution lacks intuitive visualization of the chemical processes, resulting in fragmented information without visual guidance—reflected in a low score of just 28%. In contrast, EduVisAgent begins by displaying background images of the relevant chemical elements, activating students' prior knowledge. This strategy effectively connects abstract chemical concepts to tangible, everyday experiences, a well-established method for enhancing comprehension and retention. It then contextualizes each of the four answer options with real-world scenarios, thereby enhancing students' understanding of the underlying chemical transformations.

Conversely, for the Carnot cycle efficiency physics problem (right side of Figure 6), the Gemini solution presents a single, flawed chart. Its depiction of 300K and 400K temperatures with identical heights introduces visual misinformation, failing to accurately represent data differences and thereby diminishing its pedagogical value. In stark contrast, EduVisAgent employs a multi-agent collaborative approach: it first generates a concrete factory scene to activate students' working memory of the "heat engine" concept. Subsequently, it constructs an accurate Carnot cycle diagram and offers a step-by-step problem breakdown, fostering clear conceptual understanding. Crucially, EduVisAgent provides interactive visualization components, enabling users to dynamically adjust temperatures via sliders and observe real-time changes in heat engine efficiency. This interactive element transforms the learner from a passive observer into an active participant, which is known to deepen engagement and learning. This interactive engagement significantly facilitates higher-order thinking skills.

Overall, through coordinated multi-agent optimization of image design, instructional structure, and learning pathways, EduVisAgent significantly outperforms traditional single-model approaches in accuracy, guidance, and interactivity.

F.3 Fine-Grained Analysis on Five Evaluation Dimensions

Figure 7 reveals distinct performance profiles for eight high-performing evaluated models. In Context Visualization and Diagram Design, most baselines, including SDXL, Claude 3.7, and `v0`, exhibit

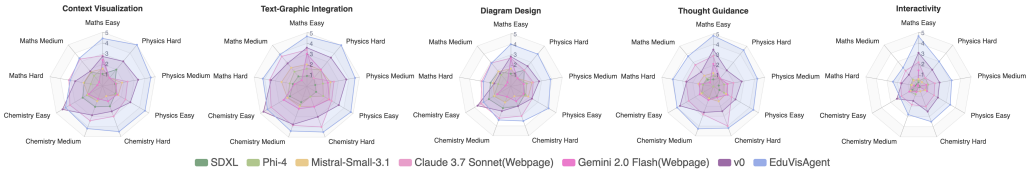

Figure 7: Fine-grained performance comparison across our five key evaluation dimensions.

moderate to low scores, often struggling with providing rich situational cues or pedagogically sound visual structures, especially for complex problems. `v0` and Claude show relatively better capabilities in Text-Graphic Integration and Thought Guidance compared to other FMs, which generally offer minimal support in these areas. However, all baseline models, including `v0`, are significantly limited in the Interactivity dimension, primarily due to their output format (static images/SVG or less dynamic webpages). In contrast, our EduVisAgent demonstrates consistently strong performance across all five dimensions. It particularly excels in creating rich context visualizations, well-structured diagram designs, and ensuring seamless text-graphic integration. Furthermore, EduVisAgent provides superior thought guidance and achieves notably high scores in Interactivity, areas where baseline models significantly lag. This comprehensive strength highlights EduVisAgent's advanced ability to generate not just visualizations, but truly effective and interactive pedagogical tools.