THINKING WITH COMICS: EFFICIENT MULTIMODAL REASONING THROUGH SEQUENTIAL VISUAL NARRATIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought reasoning has driven large language models to extend from thinking with text to thinking with images and videos. However, different modalities still have clear limitations: static images struggle to represent temporal structure and dynamic processes, while videos introduce substantial redundancy and computational cost. In this work, we propose Thinking with Comics (TwC), a visual reasoning paradigm that uses comics as a high information-density medium positioned between images and videos. Comics preserve temporal structure, embedded text, and narrative coherence while requiring significantly lower reasoning cost. To systematically evaluate reasoning over sequential visual narratives, we introduce GutterBench, a benchmark that measures whether models can represent, update, and apply intermediate states across comic panels rather than merely predict final answers. We further evaluate TwC across diverse reasoning tasks and visual understanding scenarios. Experimental results show that Thinking with Comics outperforms Thinking with Images on multi-step temporal and causal reasoning tasks, while remaining substantially more efficient than Thinking with Video. Further analysis reveals that comics provide an effective intermediate representation for modeling cross-panel state transitions, offering a promising direction for multimodal reasoning[1].

1 INTRODUCTION

Large language models (LLMs) have significantly improved their reasoning ability on complex tasks by adopting explicit Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022; Besta et al., 2024; Yao et al., 2023), making step-by-step textual reasoning (Think with text) a common paradigm. With the development of multimodal large language models (MLLM), this idea of explicit reasoning has extended from pure text to the visual domain.

Under the Thinking with Images (TWI) paradigm (Hurst et al., 2024; OpenAI, 2025a; Zhang et al., 2023; Wang et al., 2025; Chen et al., 2025), models not only use images as input signals but also generate intermediate visual representations during reasoning to supplement critical visual information (Li et al., 2025; Hu et al., 2024), thereby improving the reasoning performance of vision–language models (VLMs). Building on this, Thinking with Video further introduces temporal structure by generating short video sequences, enabling more complex forms of dynamic reasoning (Tong et al., 2025).

Despite the extension of reasoning paradigms from text to images and videos, each modality still exhibits clear limitations. Static images struggle to represent temporal structure and dynamic processes, while the absence of explicit textual cues complicates cross-modal alignment. Videos provide temporal information but introduce substantial redundancy and significantly higher computational overhead, which limits their efficiency for reasoning.

To address these limitations, we turn to a more natural reasoning medium from daily life—*comics*—and introduce the **Thinking with Comics** (TwC) paradigm. Comics are a distinctive narrative form. Compared with static images, they retain most key properties of video, including temporal logic,

[1] Our code and data will be available once the paper is accepted..

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

embedded text, and dynamic reasoning (Augereau et al., 2017). Yet compared with video, each panel is more information-dense and requires far lower reasoning cost. Recent generative models such as Gemini-3 Pro Image (Google DeepMind, 2025) can convert long text into coherent sequential panels while embedding text naturally within images. This allows comics to combine the high-density reasoning benefits of images with the dynamic logic of video. Thus, Thinking with Comics has strong potential to expand visual reasoning into a new research direction.

To support this exploration, we develop GutterBench, a benchmark for evaluating reasoning over sequential visual narratives. GutterBench examines whether models can represent, update, and apply intermediate states across panels, rather than merely predict final answers. It covers five reasoning families: paper folding, die tipping, egocentric navigation, tile transformation, and gossip propagation. These tasks form a progression from geometric transformation and spatial reasoning to relational propagation and long-horizon state tracking. GutterBench contains both text-centric tasks with explicit state descriptions and vision-centric tasks requiring visual state initialization, enabling systematic evaluation of cross-panel reasoning through controlled trajectories and reasoning depths.

Further analysis demonstrates the unique advantages of comics as a reasoning medium. First, different tasks benefit from different narrative structures. For example, detective-style comics are better suited for logical reasoning, while culture-oriented comics are more effective for knowledge understanding. Second, TwC exhibits scaling behavior similar to Chain-of-Thought, where more challenging reasoning tasks typically require a larger number of panels to support the evolution of intermediate states. Moreover, comic panels contain clear temporal and logical dependencies, and disrupting their order leads to substantial performance degradation, indicating that effective reasoning relies on modeling cross-panel state transitions rather than simply exploiting local visual cues. Meanwhile, compared with Thinking with Video, TwC preserves essential temporal structures while achieving significantly lower inference costs.

These findings indicate that visual expression still offers substantial room for exploration, and that comics provide a new reasoning medium positioned between static images and videos. We hope this work will inspire further exploration of Thinking with paradigms and help establish comics as an important component of a unified visual reasoning framework.

2 Related Works

Reasoning Paradigm Transfer. CoT enhances the interpretability of reasoning in LLMs by incorporating explicit intermediate reasoning steps, and significantly improves their reasoning capabilities (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022; Huang & Chang, 2023). Inspired by this paradigm, some works have further introduced it into MLLMs, developing the Thinking with Images paradigm (Hurst et al., 2024; Zhang et al., 2023; Zheng et al., 2023; Mitra et al., 2024; Gao et al., 2024), where MLLMs process original images or generate new ones and perform reasoning within an interleaved flow of textual and visual information. For both aforementioned paradigms, models typically employ large-scale reinforcement learning (Shao et al., 2024; Guo et al., 2025; Liu et al., 2025) or some training-free inference-time scaling methods (Kojima et al., 2022; Xu et al., 2025; Dhuliawala et al., 2024) to enhance their CoT reasoning abilities. Recently, addressing issues in the Thinking with Images paradigm, such as the lack of temporal information in single image and the relative independence between textual and visual modalities, Tong et al. (2025) proposed the Thinking with Video paradigm. This approach leverages video generation models like Sora 2 to integrate visual and textual reasoning within a unified temporal framework, where the video generation process itself constitutes the reasoning process.

Vision Generation Model. The development of visual generation has been driven by diffusion models, now the mainstream paradigm for image and video synthesis (Ho et al., 2020). A key milestone is Stable Diffusion (Rombach et al., 2022), a latent diffusion model enabling efficient high-resolution image generation. Building on this foundation, recent advances focus on improving text-to-image consistency, controllability, and fidelity. For instance, DALL·E 3 (Betker et al., 2023) leverages advanced captioning and multimodal training to produce detailed, contextually accurate images, addressing earlier compositional limitations. Similarly, Google's Nano Banana and Nano Banana Pro enhance controllability and prompt accuracy, supporting precise editing and high-quality image generation via natural language (Google DeepMind, 2025). Extending to video, OpenAI's Sora and Sora 2 apply spatiotemporal diffusion to generate coherent sequences with

realistic motion and long-range consistency, incorporating world simulation capabilities (OpenAI, 2025b). Meanwhile, Google DeepMind's Veo 3 advances audiovisual generation by integrating sound effects and dialogue with high-fidelity video via 3D latent diffusion, improving temporal coherence and multimodal expressiveness (Google, 2025). Collectively, these models mark a shift toward more versatile, integrated visual generative systems, enabling broader applications in creative domains and beyond.

3 THINKING WITH COMICS AND GUTTERBENCH

We introduce Thinking with Comics, a reasoning paradigm that externalizes intermediate reasoning into a sequence of comic panels ordered by temporal and causal structure. As shown in Figure 1, an image generation model first turns the question into a comic, and a multimodal large language model then answers the question with that comic as context.

Let $q \in \mathcal{Q}$ denote the input question and θ the parameters of the image generation model. The question is text-only on purely textual benchmarks, and $q = (t, v)$ whenever it is grounded in a given image, where t is its textual part and v the visual premise; the generator conditions on both. Given q and one fixed instruction that is identical for every task, the model produces a sequence of panels $\mathcal{C}$:

$$\mathcal{C} = G_\theta(q). \tag{1}$$

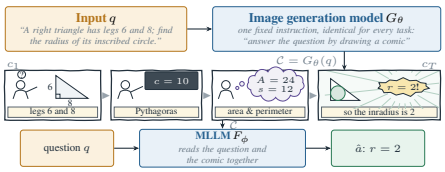


Figure 1: Overview of Thinking with Comics. The input is just the question q plus one fixed instruction, from which the generator produces a comic $\mathcal{C}$ whose panels carry the reasoning steps. The question and the comic are then passed to an MLLM, which outputs the answer.

The question and the comic are then given to a multimodal large language model F_ϕ with parameters ϕ:

$$\hat{a} = F_\phi(q, \mathcal{C}), \tag{2}$$

so that the predicted answer is

$$\hat{a} = \arg\max_a p(a \mid q, \mathcal{C}). \tag{3}$$

Compared with the textual intermediate variables of conventional chain-of-thought reasoning and with the single image used in Thinking with Images, the comic $\mathcal{C}$ encodes spatial structure, object relationships, and temporal evolution at the same time, giving the model a structured multimodal reasoning context.

We introduce **GutterBench**, a benchmark for evaluating comic-based reasoning over sequential visual narratives. Unlike existing benchmarks that mainly assess final answers, as shown in Figure 2, GutterBench evaluates whether models can represent, update, and use intermediate states across panels, distinguishing genuine cross-panel reasoning from loosely related image generation.

GutterBench contains 300 items from five spatial and relational reasoning families: paper folding, die tipping, egocentric navigation, tile transformation, and gossip propagation. It includes a text-centric track (200 items) with textual states and operations, and a vision-centric track (100 items) with an initial image state and textual operations. Each item is formulated as a state trajectory $s_0 \xrightarrow{\tau_1} s_1 \cdots \xrightarrow{\tau_D} s_D$, where $D \in \{2, 4, 6, 8\}$ and all transitions affect the final state. We provide gold storyboards $\mathcal{S} = (s_1, \ldots, s_D)$ to support process-level evaluation across text chains, image sequences, videos, and comics.

Table 1: GutterBench statistics. **Txt/Vis** are the text-centric and vision-centric tracks; **Fin./CF/Tal.** are final-state, counterfactual, and process-tally queries over both tracks; **Words** is the mean prompt length.

Family	Track		Query			Words
	Txt	Vis	Fin.	CF	Tal.	
Paper folding	41	–	16	14	11	153
Die tipping	46	30	27	27	22	136
Egocentric nav.	42	25	23	23	21	125
Tile transform.	32	20	26	26	–	157
Gossip propagation	39	25	22	21	21	139
Total	**200**	**100**	114	111	75	140

The benchmark includes final-state, counterfactual, and process-tally queries, and reports Acc@D, Trace-F1, Order-τ, and Anchor to evaluate final accuracy and intermediate-state tracking. All samples are generated from state-transition templates with controlled reasoning depth and manually verified after filtering invalid or degenerate cases. Table 1 summarizes the benchmark statistics. Full details on human verification, template diversity, and leakage control are provided in Technical Appendix E.3.

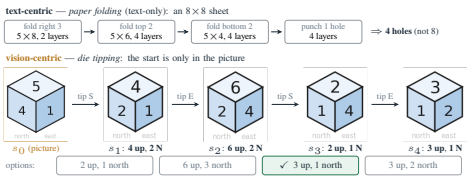


Figure 2: One released item per track, both at D=4. Top (`txt-folding-fina-d4-000`): paper folding has no visual premise, and because folds are off-centre the naive "every fold doubles" count (2^3=8) is offered as a distractor. Bottom (`img-cube-fina-d4-002`): the start is given only as a picture, checked by a depth-zero control; every wrong option is what dropping one step yields. Both rows are the gold storyboard $\mathcal{S}$, rendered as the benchmark ships it.

4 Experiments

4.1 Evaluation Datasets

We first evaluate TwC on **GutterBench**, our benchmark for process-level comic reasoning. To assess generalization beyond GutterBench, we further evaluate on downstream benchmarks organized according to its Text-Centric and Vision-Centric taxonomy.

Text-Centric benchmarks. We adopt MATH-500 (Lightman et al., 2024), GSM8K (Cobbe et al., 2021), and MathVista (Lu et al., 2023) from VideoThinkBench for direct comparison with TwV, together with DocVQA (Mathew et al., 2021) and CulturalBench (Chiu et al., 2024) (Easy/Hard).

Vision-Centric benchmarks. We evaluate ARC-AGI-2 (Chollet et al., 2025), Mazes, Visual Puzzles, and Eyeballing Puzzles following the VideoThinkBench protocol.

Category	Model / Method	Notes	Text-Centric (Acc %)					Vision-Centric (Acc %, VideoThinkBench protocol)									GutterBench				
			MATH	GSM	M-V	Doc	Cultu. (E/H)	ARC	Mazes			V-Puzzles		Eyeballing			Acc %		Process		
									Sq	Hex	Cir	Col	Shp	Pt	Ln	Shp	Txt	Vis	T-F1	Ord-τ	Anc
MLLM	GPT-5.2	direct	99.0	100.0	67.5	72.8	88.3/84.4	0.5	0.0	0.0	0.0	79.6	97.5	33.6	24.0	32.5	52.5	38.0	— *no trace* —		
	Gemini-3-Pro	direct	**100.0**	99.0	71.5	94.5	**90.4**/90.0	2.0	0.0	0.0	0.0	75.6	93.4	40.0	34.0	**60.0**	54.0	41.0	— *no trace* —		
	Claude-Sonnet 4.5	direct	99.0	100.0	72.5	92.6	87.2/76.5	**5.3**	0.0	0.0	0.0	85.6	68.6	36.2	26.3	50.5	49.5	36.0	— *no trace* —		
Text-COT†	DeepSeek-R1	CoT	90.4	96.1	— *text-only, not applicable* —		87.2/85.1	— *text-only, not applicable* —									61.0	*n/a*	38.5‡	**0.91**‡	**96.5**‡
	Qwen3-235B-A22B	CoT	92.4	94.3	— *text-only, not applicable* —		83.1/82.5	— *text-only, not applicable* —									57.5	*n/a*	36.0‡	0.88‡	95.0‡
Think w/ Image	G-IMG	G-t-R	70.2	69.4	63.6	67.5	69.7/71.4	0.3	0.0	0.0	0.0	71.5	74.0	35.0	27.5	29.0	40.5	34.0	22.0	–	44.0
Think w/ Video	Sora 2*	V-o-T	67.0	75.7	67.6	50.5	60.0/70.0	1.3	35.3	0.0	0.0	67.0	64.9	44.7	**38.0**	34.5	44.0	46.0	44.5	0.76	38.0
Think w/ Comic	**TwC (Ours)**	G-t-R	92.3	95.4	**85.8**	**99.4**	88.3/**90.2**	2.3	**41.0**	0.0	0.0	**86.0**	**84.0**	**47.0**	35.5	41.0	**68.5**	**59.0**	**61.5**	0.89	81.0

Table 2: Unified main results on text-centric and vision-centric benchmarks. **M-V** and **Cultu.** denote MathVista and CulturalBench, respectively. G-t-R is Generate-then-Reason, and V-o-T is Video-of-Thought. Vision-Centric tasks follow the VideoThinkBench evaluation protocol. **Sq/Hex/Cir**: square/hexagonal/circular mazes; **Col/Shp**: color/shape puzzles; **Pt/Ln/Shp**: point/line/shape estimation. For GutterBench, **Txt** and **Vis** are the text-centric (200 items) and vision-centric (100 items) tracks, reported as accuracy averaged over the four reasoning depths, and **Process** scores the emitted trace against the gold storyboard: **T-F1** is Trace-F1 over matched gold states, **Ord-τ** is Kendall's τ between gold order and trace order, and **Anc** is the share of matched states whose anchor facts are legible in the textual channel. Process metrics need an intermediate trace, so direct prompting cannot receive them; **Ord-τ** additionally needs two matched states, which a single generated image cannot supply (**–**). * results are reported by the original paper, † denotes text-only models that cannot process visual inputs, and ‡ marks process metrics computed on the text-centric track alone.

4.2 Models and Experimental Setup

Implementation. Unless otherwise specified, all models are evaluated in a zero-shot setting. For Thinking with Comics (TwC), we first generate a comic using Gemini-3 Pro Image, and then feed the comic together with the original question into Gemini-3 Pro for reasoning. All methods use unified prompts for fair comparison. The complete prompt templates, inference code and detailed case studies of the evaluated tasks are provided in the appendix of the supplementary material.

Baselines. We compare TwC with four groups of strong baselines: (i) *MLLMs*: GPT-5.2 (Singh et al., 2025), Gemini 3 Pro (Google DeepMind, 2025), and Claude Sonnet 4.5 (Anthropic, 2025)[2]; (ii) *Text-CoT LLMs*: DeepSeek-R1 (Guo et al., 2025) and Qwen3-235B-A22B (Yang et al., 2025)[3]; (iii) *Thinking with Images*: G-IMG (Cheng et al., 2025) (generating 1 image) built with the same Gemini-3 Pro and Gemini-3 Pro Image as TwC; and (iv) *Thinking with Video*: Sora 2.

[2]Versions: gpt-5.2-2025-12-11, gemini-3-pro-preview, and claude-sonnet-4-5-20250929.

[3]Versions: deepseek-r1-0528 and qwen3-235b-a22b-thinking-2507.

Metrics. For all benchmarks, we adopt accuracy as the primary evaluation metric, measuring the correctness of the final predicted answers. This applies uniformly to both text-centric tasks and vision-centric tasks. Since every method is evaluated on identical items, we treat all comparisons as paired and test them with a paired bootstrap over test items, using a Wilcoxon signed-rank test for the 1–5 human ratings and Holm correction whenever one condition is compared against several others; unless stated otherwise, the improvements of TwC reported below are significant at $p < 0.05$, while results quoted from prior work are excluded because they cannot be tested item-wise.

4.3 Main Results

Table 2 presents the main results of TwC on text-centric, vision-centric, and GutterBench evaluations.

On text-centric tasks, TwC achieves 85.8% accuracy on MathVista, outperforming TwV (67.6%) and the single-image TwI baseline (63.6%). For long-document understanding, TwC reaches 99.4% on DocVQA, surpassing all baselines including Gemini-3 Pro (94.5%). On text reasoning tasks where current frontier models already approach saturation (e.g., 99–100% on MATH-500), TwC remains competitive with 92.3% accuracy. These results demonstrate that comic-based reasoning is not limited to visual reasoning tasks, but can also provide effective intermediate representations for complex context understanding.

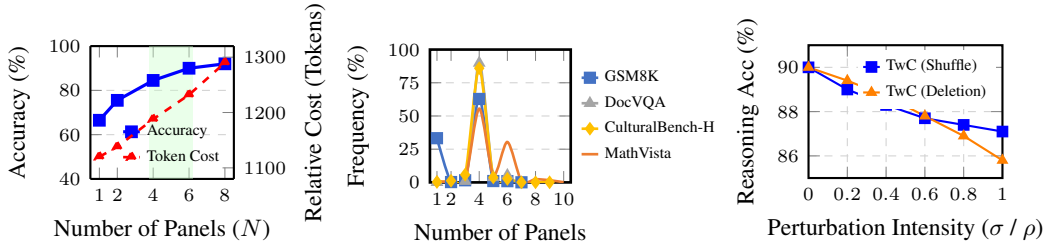
On vision-centric tasks, TwC demonstrates advantages over both TwI and Thinking with Video on spatial perception tasks. For Mazes (Square), TwC achieves 41.0% accuracy, outperforming Sora 2 (35.3%) and TwI (0.0%), indicating that the structured multi-panel layout helps models decompose spatial navigation processes. On Visual Puzzles, TwC achieves 86.0% and 84.0% accuracy on the Color and Shape subsets, consistently existing baselines, including GPT-5.2 (79.6%/97.5%) and Sora 2 (67.0%/64.9%). On Eyeballing tasks, TwC achieves the highest accuracy on Point estimation (47.0%) and remains competitive on Line estimation (35.5%). However, abstract pattern induction tasks (e.g., ARC-AGI-2) and non-standard maze configurations (Hex/Circle) remain challenging for current visual reasoning paradigms.

Finally, GutterBench evaluates whether models can preserve and utilize intermediate states throughout sequential visual narratives. TwC achieves the best performance on both the text-centric track (68.5%) and vision-centric track (59.0%), outperforming existing image- and video-based reasoning methods. Beyond final-answer accuracy, TwC also achieves the process-level performance, with 61.5 Trace-F1, 0.89 Order-τ, and 81.0 Anchor. These results demonstrate that comic-based representations can better preserve state transitions and temporal dependencies during the reasoning process.

5 Analysis Experiment

5.1 Role-playing Narrative Alignment

We investigate how specific Role-playing narrative frameworks—such as documentary-style, detective-style, and slice-of-life comic pictures—serve as "Role-playing Narratives" to induce specific reasoning paths in TwC. We compare three comic-mediated styles (documentary, detective,



(a) Performance-cost curve on MATH-500.

(b) Panel distribution across tasks.

(c) Effect of temporal perturbations.

Figure 3: Panel-level analysis experiments. (a) Marginal accuracy gains diminish beyond $N \in [4, 6]$ (shaded) while token cost keeps growing. (b) Harder tasks shift the panel distribution toward more panels. (c) Shuffle and deletion perturbations degrade accuracy, confirming that panel order carries information.

and slice-of-life) on 100 questions randomly sampled from each of the MathVista and GSM8K benchmarks, and observe performance variance when the model handles complex spatial and logical deduction tasks.

Experimental results, as shown in Table 3, reveal that the detective-style significantly outperforms the standard documentary-style comic in logical reasoning tasks. Averaged across the two benchmarks, accuracy increases from (60.0 + 68.0)/2 = 64.0 to (85.0 + 100.0)/2 = 92.5, yielding a 28.5-point absolute gain. This corresponds to a relative improvement of 28.5 / 64.0 = 44.5% over the documentary baseline. This suggests that role-playing narrative style is not merely a visual decoration but a potent Visual System Prompt. The results confirm that specific role-playing narrative structures established via comic panels can effectively activate the potential of MLLM for causal reasoning, leading to a more focused inference path.

Table 3: Narrative Style Ablation on MathVista and GSM8K. **Detective** style acts as the most effective visual prompt for tasks.

Style (Visual Prompt)	**M-Vista**	**GSM8K**	**Avg. Δ**
Documentary (Base)	60.0	68.0	—
Slice-of-Life	80.0	86.0	+19.1
Detective Style	**85.0**	**100.0**	**+28.5**

5.2 Is the Panel Sequence Doing the Work?

If comics were merely containers for additional pixels, the number of panels, their allocation, and their order would be irrelevant. Figure 3 shows that all three factors contribute.

5.2.1 Four to six panels provide the best trade-off.

We sweep the number of generated panels ($N \in \{1, 2, 4, 6, 8\}$) in TwC and measure accuracy and token cost on MATH-500. Since $N = 1$ is equivalent to Thinking with Images, this experiment directly interpolates between static and sequential visual reasoning. As shown in Figure 3a, accuracy improves rapidly and then saturates: the gain decreases from $+9.0$ points (N=1→2) to $+2.0$ points (N=6→8), while token cost keeps increasing. Thus, $N = 4$–6 achieves the best accuracy-cost trade-off by providing sufficient intermediate states without excessive redundancy.

5.2.2 Panel budgets emerge adaptively.

We analyze the panel counts generated by TwC across GSM8K, MathVista, DocVQA, and CulturalBench-hard (Figure 3b). The empirical distributions align with the optimal range above. For example, GSM8K uses one panel for 33.28% of samples and four panels for 62.82%, while MathVista peaks at four panels and uses six panels for 30.41% of cases. This suggests that panel allocation behaves similarly to chain-of-thought length: simple queries require fewer states, whereas harder queries receive longer reasoning trajectories.

5.2.3 Panel order provides additional reasoning signals.

To test whether ordering matters, we perturb panel sequences in TwC on MATH-500. *Complete Shuffle* randomly permutes panel positions while keeping contents unchanged, and *Random Intermediate Deletion* removes intermediate panels while preserving the remaining order. Since all remaining panels are pixel-identical to the original sequence, performance changes reflect temporal structure alone. Accuracy drops from 90.0% to 85.8% (Figure 3c), with deletion causing larger degradation than shuffling. This indicates that models can partially recover disrupted order but cannot infer missing intermediate states. Overall, most gains come from decomposing reasoning into discrete states, while temporal order provides an additional measurable benefit.

5.3 Open-Source Pipeline Efficiency

To provide reproducible efficiency measurements, we build a fully open-source TwC pipeline using Stable Diffusion 3.5 Large (SD3.5) [4] and Qwen3.5-9B on a single A100 80G GPU, as TwC is constrained by the limited text generation capacity of open-source models. Baselines include:

[4]The model is from https://modelscope.cn/models/AI-ModelScope/stable-diffusion-3.5-large/

(i)*MLLM* is Qwen3.5-9B with direct prompting, (ii) *Text-COT LLM* is Qwen3.5-9B, (iii)*TwI* is DREAMLLM (Dong et al., 2023) (7B), a training-based Thinking-with-Image method, and (iv) *TwV* is Wan2.1 [5] + Qwen3.5-9B (v) our TwC pipeline (SD3.5 + Qwen3.5-9B).

Table 4: Open-source comparison at the 7B scale. TwC consistently outperforms direct, CoT, and training-based TwI baselines at the same parameter scale.

Category	Method	Reasoning (Acc %)			Context (Acc %)	
		MATH	GSM	M-V	Doc	CB-H
MLLM	Qwen3.5-9B	51.2	82.5	58.4	72.8	54.2
Text-CoT	Qwen3.5-9B (CoT)	54.8	85.1	60.2	71.3	55.6
TwI	DREAMLLM (7B)	12.6	18.4	35.9	65.5	42.8
TwV	Wan2.1 + Qwen3.5-9B	55.3	85.9	63.7	74.4	58.3
TwC	**SD3.5 + Qwen3.5-9B**	**58.1**	**88.6**	**68.5**	**78.9**	**62.4**

Table 5: Comparison of TwC with CoT-prompted frontier MLLMs.

Method	MATH	M-Vista	DocVQA	CB-H
GPT-5.2 (direct)	99.0	67.5	72.8	84.4
GPT-5.2 (CoT)	99.4	70.1	74.2	85.8
Gemini-3 Pro (direct)	100.0	71.5	94.5	90.0
Gemini-3 Pro (CoT)	100.0	73.8	95.1	90.2
TwC	92.3	**85.8**	**99.4**	**90.2**

As shown in Table 4, at the same 7B scale, TwC achieves substantial improvements over DREAMLLM across all benchmarks. Notably, TwC reaches 68.5% on MathVista compared to DREAMLLM's 35.9%, a +32.6% absolute gain. This confirms that the comic-based reasoning paradigm is effective even with open-source models, and that the performance advantage of TwC does not stem from larger model scales but from the structured multi-panel representation itself.

5.4 Comparison with CoT Baselines

To compare TwC against textual Chain-of-Thought reasoning under the same frontier models, we evaluate GPT-5.2 and Gemini-3 Pro with explicit CoT prompting on four benchmarks.

As shown in Table 5, even with CoT prompting, frontier MLLMs cannot match TwC on MathVista (85.8% vs. 73.8%) and DocVQA (99.4% vs. 95.1%). This confirms that the comic representation provides unique structural benefits beyond textual Chain-of-Thought. On text-centric tasks (MATH-500, CulturalBench-Hard) where frontier models already achieve near-perfect performance, TwC shows lower accuracy, reflecting the inherent trade-off when converting well-structured textual problems into a visual medium.

5.5 Cross-Model Generalization

This experiment evaluates the cross-model generalization in the TwC paradigm across diverse MLLMs architectures. We use the same TwC generated comic as a unified input and conduct large-scale evaluations on Claude 3.7 Sonnet, Qwen-VL-72B, GPT-5.2, Gemini 3 Pro, and GPT-4o[6].

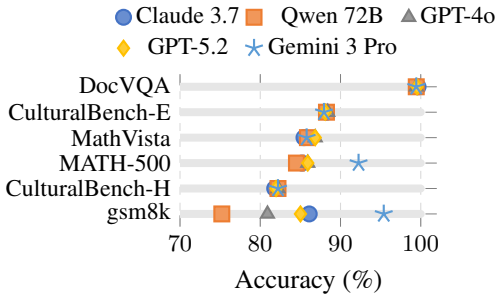


Figure 4: Architectural Robustness Analysis. The tight clustering of colored markers along the horizontal tracks demonstrates the high stability of the TwC paradigm across diverse MLLMs architectures.

Results are summarized in Figure 4. Results across different tasks show that TwC leads to largely consistent performance trends across models. On the DocVQA benchmark, all models maintain accuracy above 99.4%, indicating that emphasizing key visual regions in comics, together with accompanying textual prompts, provides reliable auxiliary information. Notably, Gemini 3 Pro achieves relatively stronger performance on several tasks, reaching 95.3% accuracy on GSM8K. Overall, comic panels function as a reusable intermediate representation that delivers stable performance gains across tasks and model configurations, demonstrating a certain degree of cross-model generalization.

[5]The version is wan2.1-t2v-1.3b-diffusers

[6]The versions of these models are respectively: claude-3-7-sonnet-20250219, qwen2.5-vl-72b-instruct, gpt-5.2-2025-12-11, gemini-3-pro-preview, and gpt-4o-2024-05-13.

5.6 Isolating the Comic Medium

To isolate the effect of the comic medium, we conduct controlled experiments on MathVista with the same Gemini-3 Pro Image model. We compare text-only, single-image TwI, unstructured multi-image sequences, and structured comics. We further compare TwC with a text-only 4-panel narrative to separate the effects of visual and textual structures.

Table 6: Accuracy across visual media on MathVista, all generated by Gemini-3 Pro Image.

Medium	MathVista (Acc %)
Direct (text-only)	71.5
Single Image (TwI)	63.6
Multi-Image Sequence	70.8
text-only 4-panel story	68.3
Comic	**85.8**

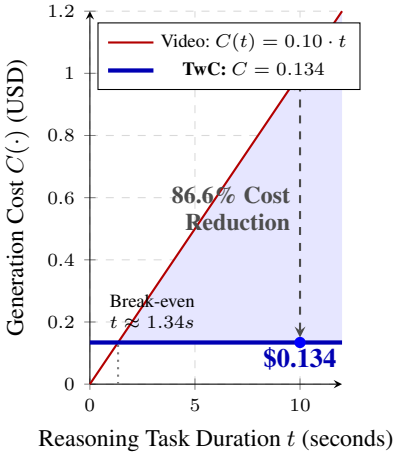
As shown in Table 6, the structured comic format significantly outperforms both single images, unstructured multi-image sequences, and text-only 4-panel stories under the same model, confirming that the comic medium itself—not merely additional visual content—drives the improvement.

5.7 Efficiency Analysis

We analyze TwC's efficiency from two complementary perspectives: closed-source API pricing and open-source computational cost. The key architectural insight is that TwC encodes all reasoning steps into a *single composite image*, making its generation cost constant and comparable to single-image methods (TwI), while video-based methods (TwV) incur cost proportional to temporal duration.

Closed-Source API Cost. We model cost as $C_{\text{video}}(t) = \alpha \cdot t$ (video, time-dependent) and $C_{\text{comic}} = \beta$ (TwC, fixed per image). With standard API pricing (α=\$0.10/s[7], β=\$0.134/img[8]), a 10-second task under Thinking with Video (Tong et al., 2025) costs \$1.00, versus only \$0.134 for TwC—an 86.6% reduction (Figure 5a). The break-even point is $t \approx 1.34\,\text{s}$.

Open-Source Computational Cost. On a single A100 80 GB GPU, we profile the generation stage of each open-source method (Figure 5b). Since the downstream reasoning model (Qwen3.5-9B) is shared across all pipelines, visual signal generation is the sole efficiency variable. TwC and TwI both produce a single image, yielding comparable generation TFLOPs (∼2.4k and ∼1.2k, respectively) and latency (14.5 s and 8.4 s, respectively). In contrast, TwV must synthesize a multi-frame video, incurring ∼4.8× more TFLOPs and ∼7.5× longer wall-clock time than TwC. These results confirm that TwC's efficiency advantage holds consistently across both commercial API and open-source deployment settings.



(a) Closed-source API cost comparison.

Method	Efficiency		
	TFLOPs↓	Lat. (s)↓	Mem.↓
TwI: DREAMLLM (7B)	∼1.2k	8.4	42.7
TwC: SD3.5 (8B) + Qwen3.5-9B	∼2.4k	14.5	54.2
TwV: Wan2.1 (14B) + Qwen-9B	∼11.6k	108.6	70.8

(b) Open-source generation cost, for one image (TwI, TwC) or one 5 s clip (TwV).

Figure 5: Efficiency analysis from both API pricing and computational cost perspectives. TwC maintains constant, low generation cost regardless of reasoning complexity—comparable to single-image methods and ∼20× cheaper than video generation.

[7] https://openai.com/api/pricing/

[8] https://ai.google.dev/gemini-api/docs/pricing

Table 7: Human evaluation comparing Global and Incremental generation. ACC denotes accuracy; Logic, State, and Quality are on a 1–5 scale.

Benchmark	Method	ACC ↑	Logic ↑	State ↑	Quality ↑
MATH-500	Incremental	80.0	4.17	3.72	3.58
	Global (Ours)	**95.0**	**4.86**	**4.67**	**4.61**
MathVista	Incremental	50.0	3.50	3.50	3.40
	Global (Ours)	**85.0**	**4.47**	**4.45**	**4.58**
Average	Incremental	65.0	3.83	3.61	3.49
	Global (Ours)	**90.0**	**4.67**	**4.56**	**4.59**

Table 8: TwC accuracy on MathVista with different comic generators. All generators outperform the single-image TwI baseline (63.6%).

Comic Generator	MathVista (Acc %)
G-IMG	63.6
Stable Diffusion 3.5 Large	65.5
DALL-E 3	78.2
Gemini-3 Pro Image	**85.8**

5.8 Comic Generation versus Incremental Image Chaining

This experiment compares *Global Comic* generation and *Incremental* image chaining under the same panel count. We evaluate 20 samples from MATH-500 and MathVista using human judgments of cross-panel coherence, state transitions, and text quality.

As shown in Table 7 and Figure 6, global generation improves cross-panel coherence, entity stability, and reasoning continuity, while incremental generation suffers from accumulated errors. These results demonstrate the importance of treating comics as holistic structured representations for multi-step reasoning.

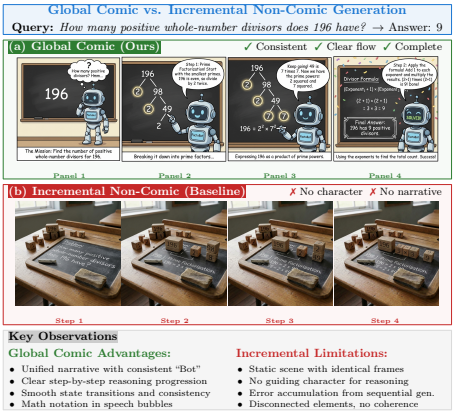


Figure 6: Qualitative comparison of Global Comic (Ours) and Incremental Non-Comic generation. Global generation maintains coherent characters and narratives, while incremental generation produces static, disconnected scenes.

5.9 Effect of Comic Generator

To verify that the benefits of TwC are not bound to a specific image generation model, we replace Gemini-3 Pro Image with DALL-E 3 and Stable Diffusion 3.5 Large while keeping the downstream VLM (Gemini-3 Pro) fixed. All methods are evaluated on MathVista using TwC.

As shown in Table 8, Gemini-3 Pro Image achieves the best results with higher-quality comics, while all generators outperform the single-image TwI baseline (63.6%), validating the effectiveness of the comic paradigm. The remaining gaps indicate the impact of visual generation quality, although even Stable Diffusion 3.5 Large improves over single images.

6 Conclusion

We introduce *Thinking with Comics* (TwC), a multimodal reasoning paradigm that externalizes intermediate reasoning into a sequence of comic panels with explicit temporal and causal structure, together with GutterBench, a 300-item benchmark that scores not only final answers but also whether intermediate states are correctly represented, updated, and used across panels. Across text-centric and vision-centric benchmarks, TwC outperforms Thinking with Images and Thinking with Video on most tasks and achieves the best accuracy on both GutterBench tracks, while encoding all reasoning steps in a single composite image keeps its generation cost constant and 86.6% lower than video generation. Our analyses show that narrative style acts as a visual system prompt, that four to six panels give the best accuracy-cost trade-off, that panel order and intermediate panels carry information the model cannot recover once removed, and that the gains stem from the structured comic medium itself and transfer across reasoning models and comic generators. TwC remains limited on tasks where frontier models are already near saturation and on abstract pattern induction such as ARC-AGI-2, and its quality is bounded by the comic generator. Future work includes finer control over panel content, verifying the faithfulness of generated panels to the underlying reasoning, and richer process-level evaluation of visual reasoning traces.

AI USE STATEMENT

Generative AI tools were used to assist with initial drafting, translation, language editing, and consistency checks. The authors take full responsibility for the final text, claims, and related artifacts.

REFERENCES

Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025.

Olivier Augereau, Motoi Iwata, and Koichi Kise. An overview of comics research in computer science. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 3, pp. 54–59. IEEE, 2017.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. OpenAI Technical Report, https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

Andong Chen, Yuchen Song, Kehai Chen, Xuefeng Bai, Muyun Yang, Liqiang Nie, Jie Liu, Tiejun Zhao, and Min Zhang. Make imagination clearer! stable diffusion-based visual imagination for multimodal machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26567–26583, 2025.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*, 2024.

François Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A new challenge for frontier AI reasoning systems. *CoRR*, abs/2505.11831, 2025. doi: 10.48550/ARXIV.2505.11831. URL https://doi.org/10.48550/arXiv.2505.11831.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9096–9105, 2024.

Google. Gemini ai video generator powered by veo 3.1. https://gemini.google/overview/video-generation/, 2025.

Google DeepMind. Gemini 3 pro. https://deepmind.google/models/gemini/pro/, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, pp. 1049–1065, 2023.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

OpenAI. Openai o3 and o4-mini system card, 2025a. URL https://api.semanticscholar.org/CorpusID:278283461.

OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, 2025b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *arXiv preprint arXiv:2511.04570*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

A Case Study

To provide an intuitive understanding of how TwC operates across different task modalities, we present qualitative examples from both text-centric and vision-centric tasks in Figure 7.



Figure 7: Qualitative case study of Thinking with Comics on text-centric (left) and vision-centric (right) tasks.

For the text-centric example, the model receives a mathematical reasoning problem and generates a comic that progressively works through the solution. Each panel corresponds to a distinct reasoning step—setting up equations, applying transformations, and arriving at the final answer—mirroring the structure of textual chain-of-thought but grounded in a visual narrative. For the vision-centric example, TwC transforms a visual question into a comic sequence that decomposes the required spatial or contextual reasoning into explicit visual steps. Rather than attempting to answer from a single glance, the model externalizes its reasoning by depicting relevant scene elements, relationships, and inferences across panels.

B Theoretical Justification of Comics as a High-Quality and Efficient Visual Reasoning Medium

B.1 Representation and Utility

Let q denote a question, a the ground-truth answer, and z an intermediate representation generated by a visual generator G_θ. Under Path II, the final prediction is $\hat{a} = F_\phi(q, z)$, consistent with Eq. (4–6) in the main paper.

We characterize an intermediate representation z by two orthogonal criteria: **(i) generation fidelity** (how well G_θ can produce z), and **(ii) task sufficiency** (how informative z is for predicting a).

Information-efficiency. We define the *information-efficiency* of z for task solving as

$$\eta(z) \triangleq \frac{I(a; z \mid q)}{C(z)}, \tag{4}$$

where $I(\cdot;\cdot \mid \cdot)$ is conditional mutual information and $C(z)$ is the media generation cost. Our main paper already instantiates $C(\cdot)$ for video and comics (constant per image vs. linear per second), providing an empirical cost rationale.

B.2 Comics Outperform Single Images Due to Temporal Structure and Textual Anchoring

A single image x is typically a snapshot of an underlying latent trajectory $s_{1:T}$ (temporal/causal process). If the answer a depends on multi-step temporal or causal relations in $s_{1:T}$, then any snapshot $x = h(s_t)$ may discard relevant states. Formally, whenever a is not conditionally independent of the latent trajectory given a snapshot, i.e.,

$$I(a; s_{1:T} \mid q, x) > 0, \tag{5}$$

we have a strict information gap:

$$\begin{aligned} I(a; s_{1:T} \mid q) &= I(a; s_{1:T}, q) - I(a \mid q) \\ &> I(a; x, q) - I(a \mid q) = I(a; x \mid q). \end{aligned} \tag{6}$$

Comics represent a structured summary $z_{\text{comic}} = (c_{1:K}, \tau)$ consisting of K panels $c_{1:K}$ (selected intermediate states) and embedded text τ (bubbles/narration). By the chain rule,

$$I(a; z_{\text{comic}} \mid q) = I(a; c_{1:K} \mid q) + I(a; \tau \mid q, c_{1:K}), \tag{7}$$

where the second term is *non-negative* and captures the additional semantic anchoring channel. Therefore, comics can strictly dominate pure-visual sequences whenever textual anchoring carries answer-relevant cues:

$$I(a; z_{\text{comic}} \mid q) \;\geq\; I(a; c_{1:K} \mid q), \tag{8}$$

and if $I(a; \tau \mid q, c_{1:K}) > 0$ then the inequality is strict. This matches our ablation that adding bubbles/narration improves robustness and accuracy.

B.3 Comics Are More Efficient Than Videos Under a Budget

Let a video be $v = (x_1, \ldots, x_T)$ with T frames. By the chain rule,

$$I(a; v \mid q) = \sum_{t=1}^{T} I(a; x_t \mid q, x_{<t}). \tag{9}$$

In realistic videos, consecutive frames are highly correlated, hence $I(a; x_t \mid q, x_{<t})$ quickly diminishes as t grows (temporal redundancy). Thus, $I(a; v \mid q)$ grows *sublinearly* with T while video cost grows *linearly* with T (or duration). Consequently, the efficiency $\eta(v) = I(a; v \mid q)/C(v)$ decreases with longer videos once redundancy dominates.

Comics can be seen as selecting $K \ll T$ *key states* (panels) from the latent trajectory to maximize task-relevant information:

$$c_{1:K} \approx \arg \max_{S \subseteq \{1,\ldots,T\},\, |S|=K} I(a; x_S \mid q)\,. \tag{10}$$

When the set function $f(S) = I(a; x_S \mid q)$ is approximately submodular (a standard diminishing-returns property for information measures), greedy selection achieves a $(1 - 1/e)$ approximation to the optimal subset. Hence, with far fewer visual tokens, comics retain most of the answer-relevant information while avoiding redundant frames, leading to higher $\eta(z_{\text{comic}})$ than $\eta(v)$ at the same budget. This aligns with our observed panel-scaling curve where accuracy saturates around $K \in [4, 6]$.

B.4 Why Comics Generate Better than Synthetic Sequential Images: A Domain-Shift Bound

We now justify the claim that comics (a real, widely observed visual genre) are generated with higher fidelity than ad-hoc "synthetic sequential images with logical relations" that do not correspond to a well-established visual manifold.

Let P_{train} be the (unknown) effective training distribution of the image generator. Let P_{comic} be the target distribution of real comics, and P_{syn} the distribution of synthetic sequential images. Consider a perceptual fidelity loss $\mathcal{L}(x)$ (e.g., measuring artifacts, inconsistency, or poor alignment with prompts). A standard domain adaptation bound (Ben-David type) implies that for any hypothesis class induced by the generator,

$$\mathbb{E}_{x\sim P_{\text{target}}}[\mathcal{L}(x)] \;\le\; \mathbb{E}_{x\sim P_{\text{train}}}[\mathcal{L}(x)] + \mathrm{Div}(P_{\text{train}}, P_{\text{target}}) + \lambda, \tag{11}$$

where $\mathrm{Div}(\cdot,\cdot)$ is a distribution divergence (e.g., $\mathcal{H}\Delta\mathcal{H}$-divergence or an IPM), and λ is the irreducible joint error term. If comics are a *real, frequent* genre, then P_{comic} is closer to P_{train} than an ad-hoc synthetic style:

$$\mathrm{Div}(P_{\text{train}}, P_{\text{comic}}) \;<\; \mathrm{Div}(P_{\text{train}}, P_{\text{syn}}). \tag{12}$$

Therefore, the expected fidelity loss is lower for comics:

$$\mathbb{E}_{x\sim P_{\text{comic}}}[\mathcal{L}(x)] \;<\; \mathbb{E}_{x\sim P_{\text{syn}}}[\mathcal{L}(x)], \tag{13}$$

i.e., the generator produces higher-quality outputs in the comic domain than in a less natural, distribution-shifted synthetic domain.

Comics simultaneously (i) reduce task uncertainty via structured temporal panels and textual anchoring, (ii) avoid the redundancy and high cost of video, and (iii) achieve higher generation fidelity due to smaller domain shift. Together, these provide a principled justification for Thinking with Comics as a high-density intermediate reasoning representation.

C Answer Extraction Protocol for Thinking with Comics

C.1 Path I: End-to-End Comic Reasoning

In Path I, the model generates a multi-panel comic as the complete reasoning result, where the final answer is visually embedded in the comic, typically appearing in the last panel as explicit text or a highlighted result. We perform answer extraction using **GPT-5.2** as an external answer reader (the detail of prompt is in Appendix D.2). The extractor is provided with the generated comic panels together with the original question, and is instructed to identify the final answer depicted in the comic. The extracted answer is matched against the ground-truth label to compute ACC.

Human Verification for Path I. To validate the reliability of model-based answer extraction, we randomly sample **20%** of the evaluation instances for manual inspection. For each sampled instance, a human annotator independently reads the answer from the comic and compares it with the answer extracted by GPT-5.2. We observe 100% agreement between automated extraction and human judgment, indicating that GPT-5.2 serves as a stable proxy for answer reading in comic-based reasoning. The detail of human verification is in Appendix E.2.

C.2 Path II: Comics-as-Context Reasoning

In Path II, comics are used solely as intermediate contextual representations, while the final answer is explicitly generated by a MLLM in textual form. Answer extraction in this setting is performed by directly parsing the final model output. The predicted answer is matched against the ground-truth label using standard normalization and exact-match rules to compute ACC.

D PROMPT EXAMPLES

D.1 PROMPT IN THE MAIN EXPERIMENT

Prompt for Gemini-3 Pro Image in MATH500 & GSM8K

Please help me draw a comic to solve this math problem: **{question}**

Prompt for Gemini-3 Pro Image in MathVista

Please draw a suitable comic strip to help solve this math/visual reasoning problem based on the provided image.
Context: This is a visual question answering problem at not applicable level, involving natural image.
Required skills: numeric commonsense, arithmetic reasoning.
Question: **{question}**
The answer should be a integer value.
Goal: Create a comic that illustrates the problem-solving process step-by-step, showing the complete reasoning from understanding the question to finding the answer.

Prompt for Gemini-3 Pro Image in CulturalBench-Easy

Please draw a comic strip to help solve this multiple-choice cultural knowledge question about **{country}**.
Question: **{question}**
Options: **{options}**
Your comic should:
1. Visually depict the cultural scenario described in the question.
2. Show the correct cultural practice/behavior/knowledge of **{country}**.
3. Through the comic story, clearly demonstrate which option (A, B, C, or D) is the correct answer.
4. The comic should lead the viewer to understand and identify the correct choice.
Goal: Help the viewer select the correct answer from the four options by illustrating the authentic cultural context of **{country}**.

Prompt for Gemini-3 Pro Image in CulturalBench-Hard

Please draw a comic strip to help determine if the following cultural statement about **{country}** is TRUE or FALSE.
Question: **{question}**
Statement to evaluate: **{statement_to_judge}**
Your comic should:
1. Visually depict the authentic cultural practice/behavior in **{country}**.
2. Show whether this statement accurately represents the real cultural norm.
3. Through the comic story, clearly demonstrate if this statement is TRUE or FALSE.
4. The comic should help the viewer judge the truthfulness of this cultural claim.
Goal: Help the viewer determine TRUE or FALSE by illustrating the actual cultural reality of **{country}**.

Prompt for Gemini-3 Pro Image in DocVQA

Please draw a suitable comic strip to help answer this document question based on the provided document image.
Task Type: Document Visual Question Answering **{question_types}**
Question: **{question}**
Goal: Create a comic that:
1. Shows the key information extraction process from the document.
2. Highlights the relevant parts of the document that contain the answer.
3. Illustrates the reasoning steps to find the correct answer.
4. Makes the final answer clear through visual storytelling.
The comic should help explain how to locate and extract the answer from the document.

Prompt for TWI method: G-Image

Instruction: You are an expert in writing prompts for text-to-image generation. Based on the following image and textual query, write a precise and detailed prompt to generate a image highly relevant to the query. This image will serve as an auxiliary tool to help resolve the task accurately. Consider composition, style, and detail to ensure practicality.

Reasoning Protocol: Based on the question and the additional synthesized image, let's think step by step, but avoid adding visual descriptions during the reasoning process!

Output Format: End your thinking process with the most appropriate answer in the format "ANSWER: (x)" followed by the choice.

Question: Q
Choices: C
Prompt Generated:
<Extra Image Input>
Your Response:

D.2 Prompt in the Analysis Experiment

Prompt for Role-playing Narrative Alignment Experiment

Please help me draw a **Slice-of-Life style** comic to solve this math problem: **{question}**
Please help me draw a Documentary **realistic style** picture to solve this math problem: **{question}**

Prompt for Textual Anchoring Ablation: CulturalBench-Easy

Please draw a comic strip to help solve this multiple-choice cultural knowledge question about **{country}**.
Question: **{question}**
Options: **{options}**
Your comic should:
1. Visually depict the cultural scenario described in the question.
2. Show the correct cultural practice/behavior/knowledge of **{country}**.
3. Through the comic story, clearly demonstrate which option (A, B, C, or D) is the correct answer.
4. The comic should lead the viewer to understand and identify the correct choice.
5. Relies solely on visual storytelling; DO NOT contain any text, speech bubbles, narration boxes, or onomatopoeia.
Goal: Help the viewer select the correct answer from the four options by illustrating the authentic cultural context of **{country}**.

Prompt for Textual Anchoring Ablation: CulturalBench-Hard

Please draw a comic strip to help determine if the following cultural statement about **{country}** is TRUE or FALSE.
Question: **{question}**
Statement to evaluate: **{statement_to_judge}**
Your comic should:
1. Visually depict the authentic cultural practice/behavior in **{country}**.
2. Show whether this statement accurately represents the real cultural norm.
3. Through the comic story, clearly demonstrate if this statement is TRUE or FALSE.
4. The comic should help the viewer judge the truthfulness of this cultural claim.
5. Relies solely on visual storytelling; DO NOT contain any text, speech bubbles, narration boxes, or onomatopoeia.
Goal: Help the viewer determine TRUE or FALSE by illustrating the actual cultural reality of **{country}**.

Prompt for Structural Coherence: Global Visual Reasoning

Please create a complete {num_panels}-panel comic strip that illustrates the step-by-step solution process for this math problem.
Problem: **{question}**
Requirements:
1. Create exactly {num_panels} panels arranged in a coherent sequence.

2. Panel 1: Introduce the problem and key information.
3. Panel 2–{num_panels−1}: Show the logical reasoning steps progressively.
4. Panel {num_panels}: Present the final solution and answer.
5. Maintain consistent characters/elements across all panels.
6. Include clear mathematical notation and explanations in each panel.
7. Ensure smooth visual transitions between panels.
8. Each panel should build logically on the previous one.

Important: Generate ALL {num_panels} panels as a single cohesive comic image with clear panel divisions.

Prompt for Structural Coherence: Incremental Visual Reasoning

Case 1: First Panel ($panel_num = 1$)

Create a realistic photo-style image (Step 1 of {total_panels}) for solving this math problem.
Problem: **{Question}**
Style: Realistic photo, NOT cartoon or comic style.
This is the FIRST step image. It should: 1) Introduce the problem scenario with realistic visual elements; 2) Set up real-world objects or scenes that represent the mathematical concepts; 3) Clearly present the mathematical question in a realistic context; 4) Use photorealistic rendering, natural lighting, and realistic textures.
Generate ONLY Step 1 as a single realistic photo-style image.

Case 2: Intermediate Panels ($1 < panel_num < total_panels$)

Create a realistic photo-style image (Step {panel_num} of {total_panels}) for solving this math problem.
Problem: **{Question}**
Style: Realistic photo, NOT cartoon or comic style.
This is Step {panel_num} of {total_panels}. It should: 1) Continue logically from the previous step; 2) Show the next step in the reasoning or calculation process with realistic visuals; 3) Maintain consistent visual elements from previous image; 4) Prepare for the next step in the solution; 5) Use photorealistic rendering, natural lighting, and realistic textures.
Based on the previous image shown, continue the problem-solving process. Generate ONLY Step {panel_num} as a single realistic photo-style image.

Case 3: Final Panel ($panel_num = total_panels$)

Create a realistic photo-style image (Step {panel_num} of {total_panels}, the FINAL step) for solving this math problem.
Problem: **{Question}**
Style: Realistic photo, NOT cartoon or comic style.
This is the LAST step. It should: 1) Continue from the previous step's reasoning; 2) Show the final calculation or conclusion with realistic visual elements; 3) Present the final answer clearly in a realistic context; 4) Provide a satisfying conclusion to the problem-solving journey.
Based on the previous image shown, maintain visual consistency and complete the solution. Generate ONLY Step {panel_num} as a single realistic photo-style image.

Prompt for GPT-5.2 Answer Extraction in Path I

You are an answer extraction model.

Context: You are given a multi-panel comic that visually depicts a complete reasoning process, together with the original question.

Task: Read the comic and identify the final answer shown in the last panel.

Rules:
- Only output the final answer.
- Do not explain the reasoning.
- If the answer is numeric, output the normalized numeric form.
- If the answer is a short phrase or option, output it verbatim.

Question: **{q}**

E HUMAN EVALUATION PROTOCOL

E.1 EVALUATION FOR GLOBAL AND INCREMENTAL VISUAL

We employed three expert annotators to conduct human evaluations for all experiments described in this paper. All annotators hold a master's degree or higher and have prior experience with vision–language evaluation tasks.

Before annotation, we provided a detailed training session covering task definitions, scoring rubrics, and representative examples. The annotators then completed a pre-annotation phase, during which we aligned interpretations of the evaluation criteria (Accuracy, Logic, State, and Quality) and resolved ambiguities in the scoring guidelines.

Each sample was independently rated by all three annotators. We used the averaged score across annotators as the final reported value. Inter-annotator agreement was monitored throughout the process, and inconsistencies were discussed and resolved according to the established rubric.

E.2 EVALUATION FOR EXTERNAL ANSWER READER

To verify the reliability of model-based answer extraction in Path I, we conduct a manual cross-validation study involving three independent human annotators. A shared subset comprising 20% of the evaluation instances is randomly sampled across benchmarks. For each sampled instance, all three annotators are provided with the original question and the generated multi-panel comic, and independently identify the final answer depicted in a panel. The annotations are then compared across annotators to ensure consistency, and any discrepancies are resolved through discussion to reach a consensus. The consensus human answer is finally compared against the answer extracted by GPT-5.2 under identical normalization rules. We observe complete agreement between the consensus human judgments and the automated extraction, supporting the reliability of GPT-5.2 as an answer reader in comic-based reasoning.

E.3 HUMAN VERIFICATION AND CONTAMINATION ANALYSIS

Verification protocol and inter-annotator agreement We drew a stratified subset of 100 items (20 per family, covering all four depth strata and three query types; the sampling script and seed are released with the benchmark) and had it independently verified by three annotators who were blind to one another's judgements. Each item was rated on three binary dimensions: (i) *answerability*—whether the stem alone determines a unique answer; (ii) *option validity*—whether all four options are physically possible and exactly one is correct; and (iii) *storyboard faithfulness*—whether every state of the gold storyboard follows from the stated rules step by step. Annotators completed a one-hour training session and a 15-item pilot, during which two guideline ambiguities were identified and resolved. Items receiving unanimous passes were retained; 2:1 splits were adjudicated by the first author, and any item still judged invalid was discarded and replaced by a freshly generated item from the same (family, depth, query) cell, which then went through the same verification, leaving the stratification intact. The remaining 200 items were checked in full by one annotator under the same checklist, with a 50-item second pass by a different annotator. Fleiss' κ across the three annotators on the 100-item subset was 0.87 for answerability, 0.91 for option validity, and 0.82 for storyboard faithfulness (macro-average $\kappa = 0.87$; all pairwise Cohen's $\kappa \geq 0.81$); the full-set retention decision was unanimous for 94% of items, and 3 items were replaced after adjudication.

Human ceiling The human baseline was measured with five participants who had *not* taken part in verification, under exactly the instructions and presentation given to the models (same stem, same four options, untimed but with completion time recorded). Accuracy by depth was 99.3% at D=2, 97.1% at D=4, 93.8% at D=6, and 89.5% at D=8 (majority vote over the five participants; mean $\pm$ s.d. across participants $94.9 \pm 2.1\%$ pooled over all depths), confirming that every stratum is comfortably solvable by humans while remaining far above the best system at D=8. In addition, the MLLM judge used for Trace-F1, Order-τ, and Anchor was calibrated against human state-matching labels on a 100-trace subset, reaching $\kappa = 0.85$ (judge precision 0.89, recall 0.86).

Template diversity The 300 items instantiate 23 surface templates across the five families (paper folding 4, die tipping 4, egocentric navigation 5, tile transformation 4, gossip 6); no single template accounts for more than 9.7% of the benchmark, and templates are balanced within each (family, depth) cell. Surface form is deliberately decoupled from the state machine: names, colours, coordinates, and fold sequences are sampled from a seeded RNG per item; the answer position is shuffled per item, yielding a near-uniform distribution over the four option slots (25.3 / 24.7 / 25.0 / 25.0%); and the correct answer's surface length is statistically indistinguishable from that of the distractors (two-sided t-test, $p = 0.61$), so no length or position shortcut identifies the correct option.

Leakage control All items were programmatically generated from fixed seeds and had not been released, posted, or used in any public repository beforehand; because the generator and seeds ship with the benchmark, the full set is reproducible by construction, and no evaluation-time channel to the items exists other than the API calls we report. All models were evaluated through their public APIs with recorded sampling parameters (temperature 0.0 where supported); no benchmark item was used for prompt tuning, which was done exclusively on a disjoint development split generated from a different seed.

F EXAMPLES OF TWC

This section provides qualitative examples of Thinking with Comics (TwC). It begins with a comparison of different comic styles, followed by illustrative examples from Reasoning Tasks and (Long) Context Understanding Tasks. In total, five benchmarks are included to demonstrate the use of TwC under different task formulations and contextual requirements.

F.1 COMPARISON OF DIFFERENT COMIC STYLES

We provide qualitative examples of different comic-style visualizations for problem solving. The Documentary style mainly relies on realistic images to directly present the problem context and relevant information. The Role-playing style introduces explicit characters or professional roles, through which the reasoning process is narrated and unfolded in a role-driven manner. In contrast, the Slice-of-life style embeds the reasoning process within everyday scenarios, illustrating problem solving through familiar daily-life activities.



Figure 8: Examples of different comic-style visualizations for problem solving.

F.2 Reasoning Tasks

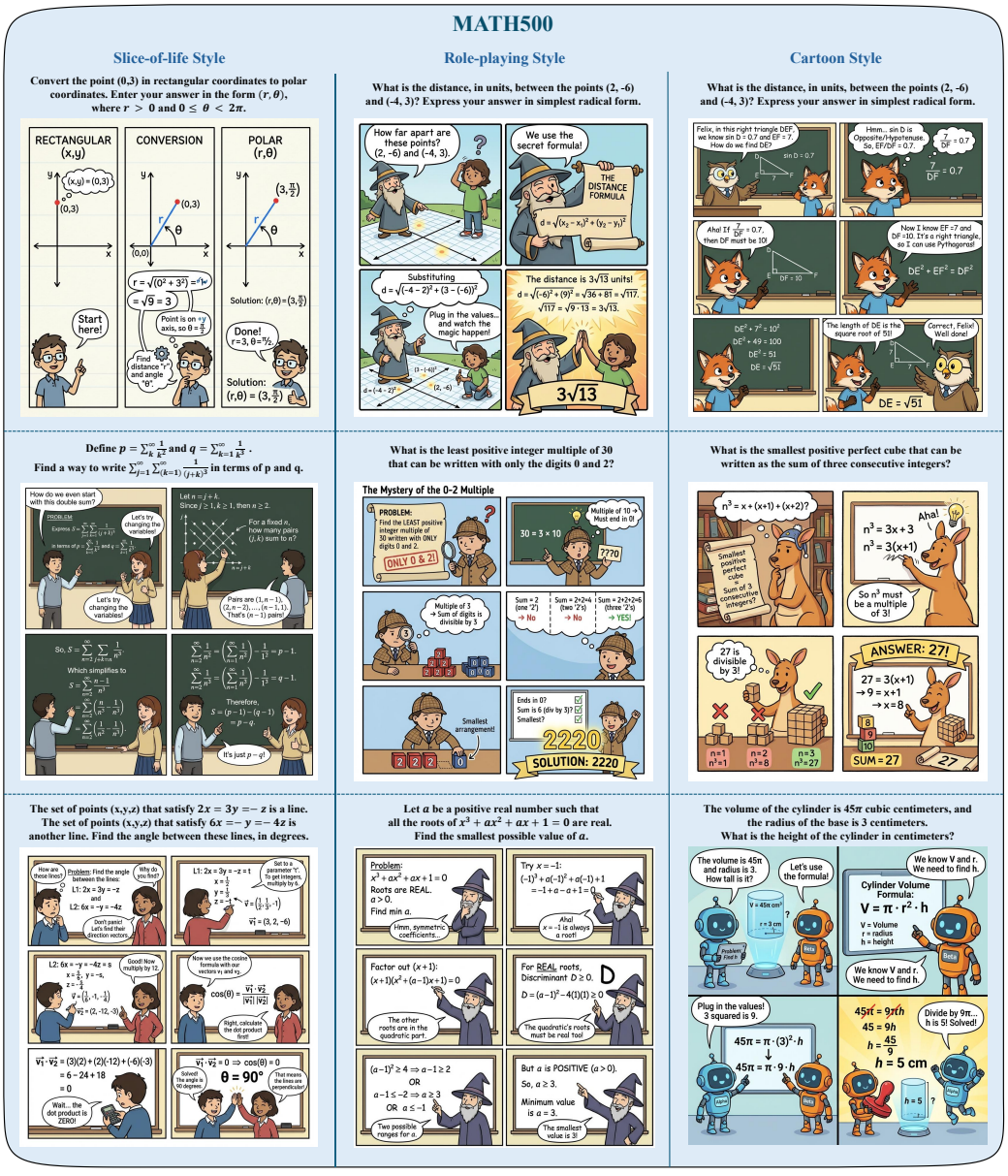


Figure 9: Examples of TwC on MATH-500.



Figure 10: Examples of TwC on MathVista.



Figure 11: Examples of TwC on GSM8K.

F.3 (LONG) CONTEXT UNDERSTANDING TASKS



Figure 12: Examples of TwC on CulturalBench.



Figure 13: Examples of TwC on DocVQA.

G Failure Case Analysis

While TwC demonstrates strong performance on text-centric and several vision-centric benchmarks, it exhibits systematic failures on tasks that demand pixel-level spatial precision. We identify three representative failure modes through qualitative analysis of Vision-Centric benchmarks where TwC underperforms, and discuss the underlying causes and implications for future work.

G.1 Type I: Pixel-Level Precision Failure (ARC-AGI-2)

ARC-AGI-2 requires models to discover abstract transformation rules from input–output grid pairs and apply them to unseen test grids. As shown in Figure 14, TwC generates a well-structured comic narrative: the character correctly identifies high-level rules (e.g., "colored pixels move to the nearest 2×2 gray block") and articulates a plausible hypothesis. However, the final output grid diverges from the ground truth because the image generation model cannot perform exact pixel-level grid operations. The comic medium excels at conveying *qualitative* reasoning chains but lacks the *quantitative* precision needed to place individual colored cells at exact coordinates. This explains why all methods—including frontier MLLMs—achieve near-zero accuracy ($\leq$5.3%) on ARC-AGI-2: the bottleneck lies not in understanding the rule but in executing it with pixel-exact fidelity.

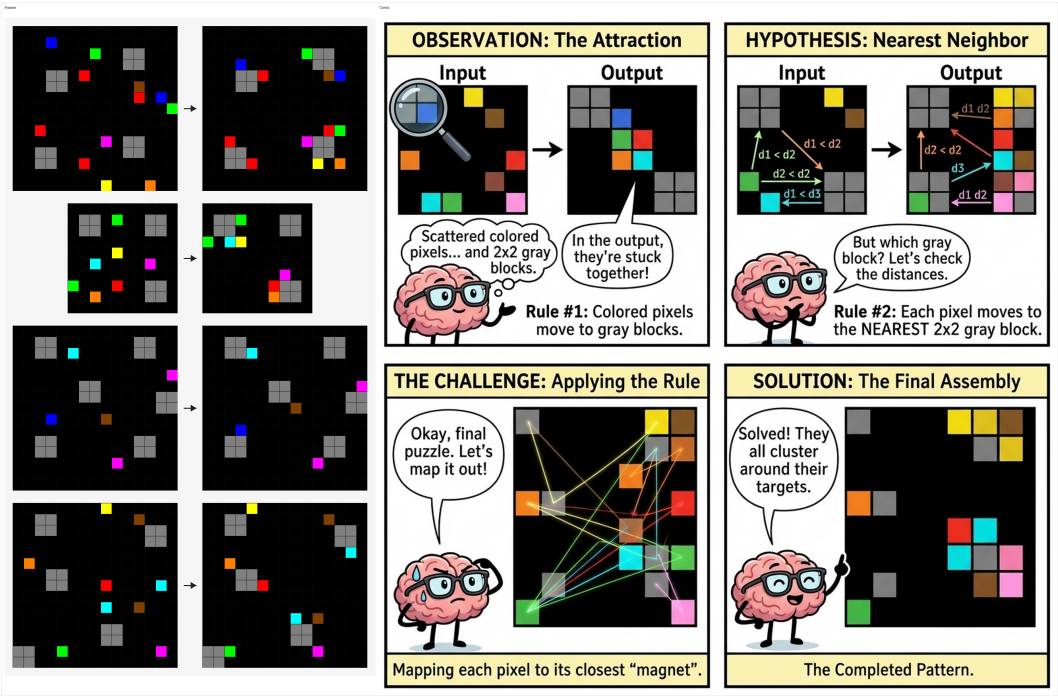


Figure 14: Failure case on ARC-AGI-2. **Left:** ground-truth input–output pairs. **Right:** TwC-generated comic. The model correctly identifies the "nearest neighbor" transformation rule and constructs a coherent reasoning narrative, but the final output grid contains incorrect pixel placements due to the image generator's inability to perform exact spatial computations.

G.2 Type II: Spatial Fidelity Loss in Non-Standard Topologies (Mazes)

Maze-solving requires tracing a valid path that respects wall constraints. Figure 15 illustrates a hexagonal maze where TwC generates a four-panel comic depicting progressive path exploration. While the narrative structure is reasonable—panels show incremental path extensions—the generated paths *pass through maze walls* and fail to satisfy topological constraints. Two factors contribute to this failure: (1) the image generation backbone treats maze walls as approximate visual textures rather than hard constraints, and (2) hexagonal grids introduce irregular adjacency relations that are difficult to encode in the comic's discrete panel structure. On standard square mazes, TwC achieves 41.0% by leveraging the simpler four-directional topology, but accuracy drops to 0% on hexagonal and circular variants where spatial fidelity demands are higher. This suggests that TwC's panel-based decomposition is effective when spatial reasoning can be approximated by coarse directional moves, but breaks down when the task requires geometrically faithful path tracing.

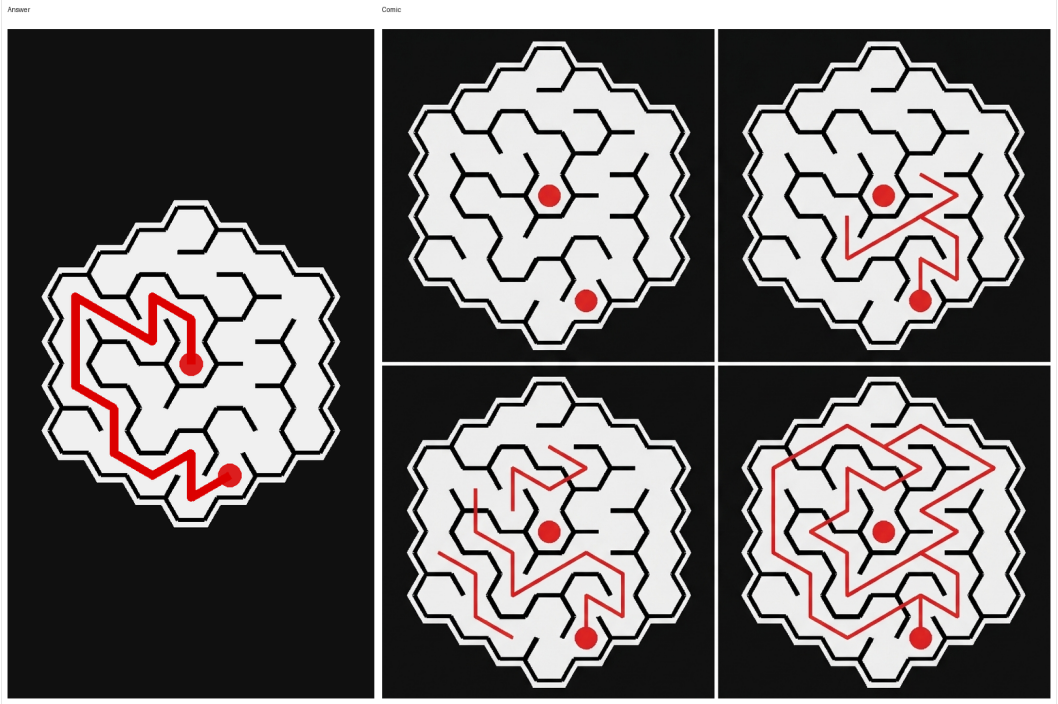


Figure 15: Failure case on hexagonal Mazes. **Left:** ground-truth solution path. **Right:** TwC-generated comic panels. The comic depicts a progressive path-finding narrative, but the generated paths violate wall constraints—lines cross through hexagonal cell boundaries—because the image generator treats maze walls as soft visual features rather than hard topological constraints.

G.3 Type III: Geometric Estimation Error (Eyeballing Puzzles)

Eyeballing Puzzles require fine-grained geometric judgments such as estimating which arc on the right continues a curve on the left. Figure 16 shows TwC generating a comic where a robot character "extends the curve" using a dashed red line and concludes the answer is C. However, the correct answer is B—the comic's visual approximation of the curve extension lacks the sub-pixel accuracy needed to distinguish between closely spaced arc candidates. Unlike the previous two failure modes, the reasoning process here is *almost* correct: the model understands *what* to do (extend the curve and match) but the generated visual lacks sufficient geometric precision to do it accurately. TwC still achieves competitive results on coarser Eyeballing subtasks (Point: 47.0%, Line: 35.5%), indicating that the comic format provides useful spatial scaffolding when the discrimination threshold is not too fine.

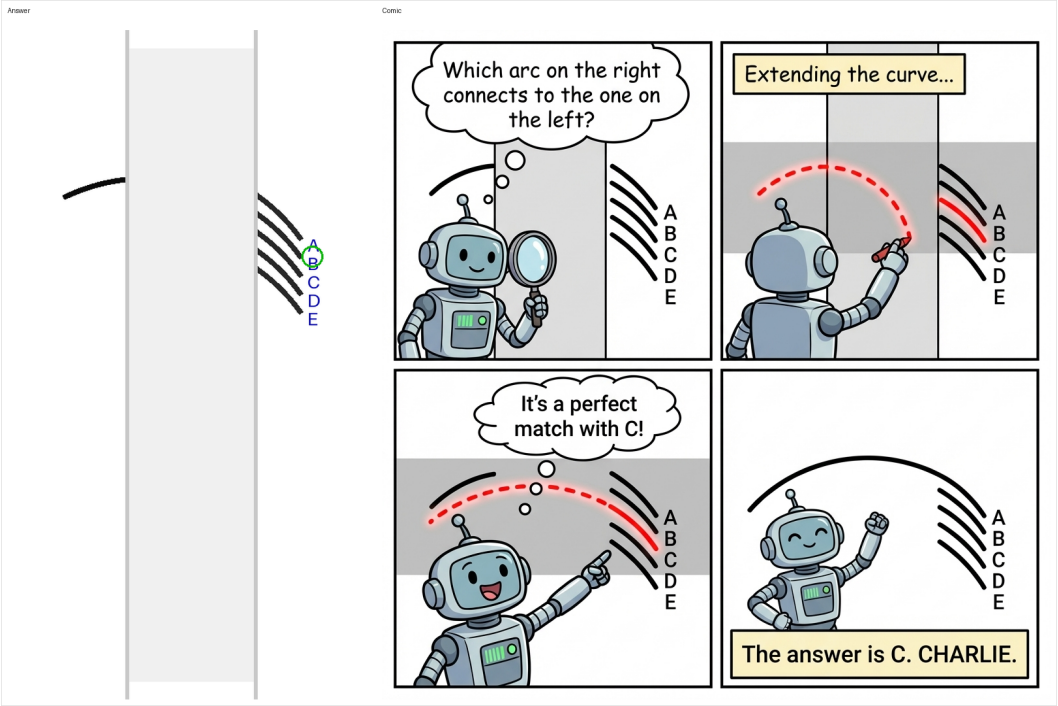


Figure 16: Failure case on Eyeballing Puzzles (arc connection). **Left:** original puzzle with correct answer B. **Right:** TwC-generated comic. The model correctly identifies the task (extend the left curve to find the matching right arc) and visualizes the extension, but the geometric precision of the generated curve is insufficient to distinguish between adjacent options, leading to the incorrect answer C.

G.4 Summary and Implications

The three failure modes share a common root cause: **the generation fidelity ceiling of current image generation models**. TwC's comic-based reasoning pipeline consistently produces correct high-level reasoning narratives—identifying transformation rules, decomposing spatial navigation, and articulating geometric strategies—but the image generation backbone cannot translate these narratives into spatially precise visual outputs. This decoupling between "understanding" and "rendering" represents the primary bottleneck for vision-centric tasks.

These findings suggest two promising directions for future work: (1) integrating structured output modules (e.g., grid renderers or vector graphics engines) as post-processing steps for tasks requiring pixel-exact outputs, and (2) developing hybrid architectures that combine TwC's narrative reasoning with programmatic execution for spatial constraints. The consistent success of TwC on tasks with coarser spatial granularity (Visual Puzzles: 86.0%, DocVQA: 99.4%) confirms that the comic paradigm itself is sound—the limitation is in the rendering precision, not the reasoning framework.