EXTENDED ABSTRACT: ThinkMorph: Emergent Properties in Multimodal Interleaved Chain-of-Thought Reasoning

Anonymous CVPR submission

Paper ID 16

Abstract

Multimodal reasoning requires iterative coordination between language and vision, yet it remains unclear what constitutes a meaningful interleaved chain of thought. We posit that text and image thoughts should function as complementary*, rather than isomorphic, modalities that mutually advance reasoning. Guided by this principle, we build* ***ThinkMorph****, a unified model fine-tuned on ∼24K high-quality interleaved reasoning traces spanning tasks with varying visual engagement. ThinkMorph delivers large gains on vision-centric benchmarks (averaging* ***34.7%*** *over the base model) and generalizes to out-of-domain tasks, matching or surpassing larger and proprietary VLMs. Beyond performance, ThinkMorph exhibits three emergent properties indicative of higher-level multimodal intelligence:* ***(1) unseen visual manipulation skills****,* ***(2) adaptive reasoning mode switching****, and* ***(3) better test-time scaling*** *through diversified multimodal thoughts. These findings suggest promising directions for characterizing emergent capabilities in unified models for multimodal reasoning.*

1. Introduction

Multimodal reasoning is not a single-pass perception task but an iterative process that interweaves language and vision. While textual Chain-of-Thought (CoT) [7] has advanced verbal reasoning, it contributes little when problems demand visual manipulation beyond description [3, 5]. These limitations motivate a shift toward *genuinely cross-modal reasoning*—mirroring the human "think-and-sketch" strategy.

Existing interleaved CoT approaches remain limited: tool-augmented designs rely on external visual modules [4, 8], while unified models [6] show little generalization beyond training domains, with their textual and visual components often *isomorphic* [2] rather than complementary.

We introduce **ThinkMorph**, built on the principle that text and image thoughts must be *complementary*—each advancing reasoning in ways the other cannot. Fine-tuned on ∼24K interleaved traces across four tasks with varying visual engagement, ThinkMorph achieves an average improvement of **34.74%** over the base model on vision-centric benchmarks. Despite its modest data scale, it generalizes robustly to out-of-domain settings, surpassing InternVL3.5-38B on SAT and matching Gemini 2.5 Flash on MMVP. Our contributions are as follows:

- **Systematic study of interleaved multimodal reasoning.** We present ThinkMorph as a unified framework for investigating when and how multimodal interleaving surpasses text-only and image-only reasoning modes, achieving substantial improvements across diverse benchmarks.
- **Emergent properties in interleaved reasoning.** We identify three distinctive emergent behaviors—unseen visual manipulations, autonomous mode switching, and diversified multimodal exploration—that arise naturally from interleaved training.
- **New avenues for test-time scaling.** We show that interleaved reasoning enables superior test-time scaling by exploring broader multimodal solution spaces, with gains amplifying under distribution shifts.

2. ThinkMorph

ThinkMorph models reasoning as an interleaved sequence $\mathcal{T} = (\hat{m}_1, \ldots, \hat{m}_n)$, where each token $\hat{m}_i \in \{\hat{t}_i, \hat{v}_i\}$ is either a text token or an image token generated by a unified VLM. Modality transitions are controlled via delimiter tokens (<image_start>/<image_end>), enabling seamless switching between verbal and visual reasoning.

Interleaved Thought Collection. We construct ∼24K high-quality interleaved traces across four tasks (Figure 1): **Jigsaw Assembly**—text describes patches, image visualizes re-arrangements; **Spatial Navigation**—text abstracts the map, image renders the trajectory; **Visual Search**—text hypothesizes regions, image draws bounding boxes; **Chart Refocus**—text identifies data elements, image highlights relevant regions. Each trace is carefully designed so text and visual steps are *complementary*: neither modality alone could replace the other's contribution. Quality is ensured

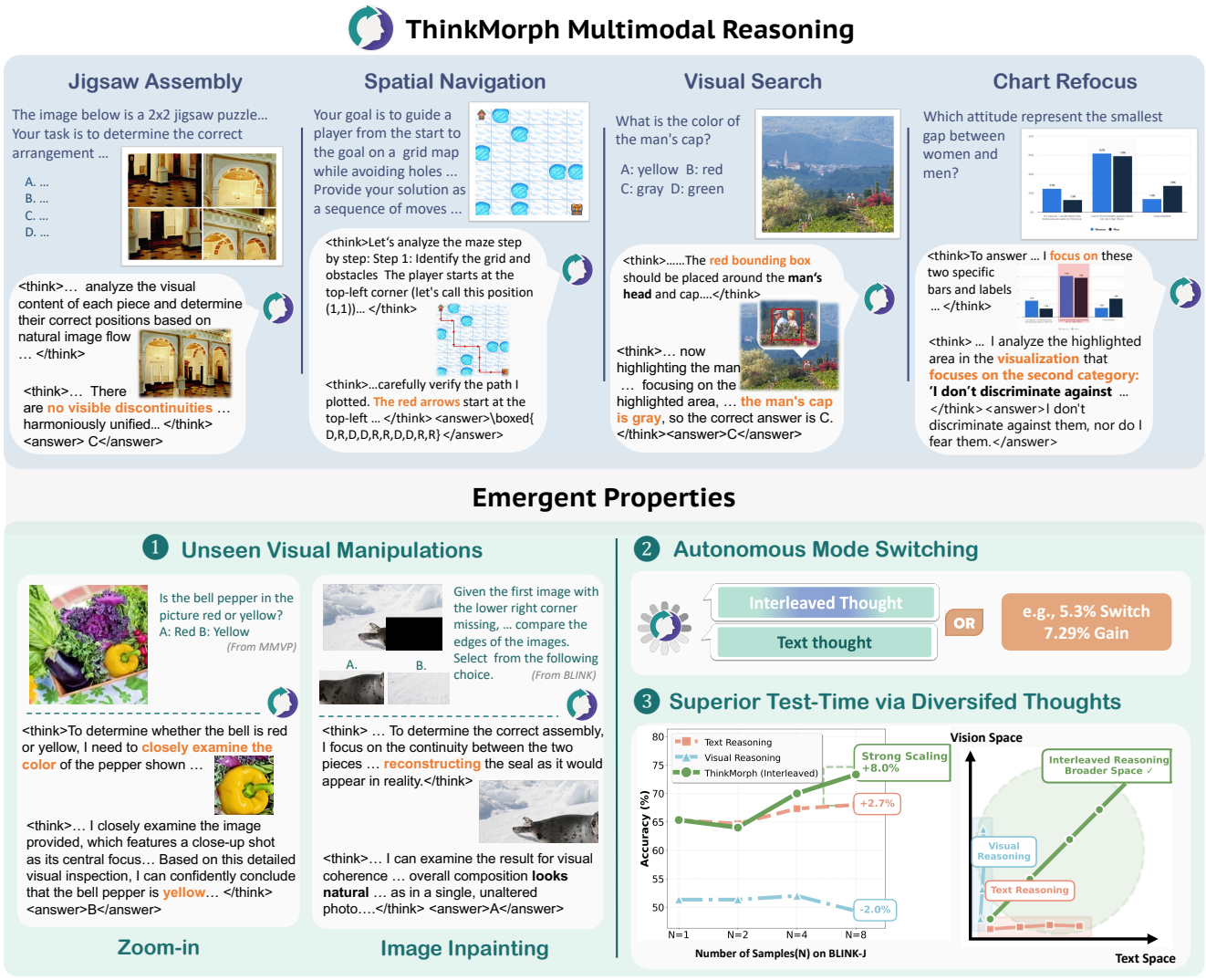

Figure 1. **ThinkMorph overview.** Top: ThinkMorph generates interleaved text–image reasoning steps across four tasks with varying visual engagement. Bottom: Emergent properties observed at inference time.

through human-in-the-loop MLLM filtering; for instance, in Visual Search, we filter from 144K to 6,990 questions by constraining target bounding boxes to 1–30% of image area.

Training. We adopt Bagel-7B [1] as our base model and optimize dual objectives: MSE loss $\mathcal{L}_{\text{img}}$ for image tokens and negative log-likelihood loss $\mathcal{L}_{\text{text}}$ for text tokens.

Results. As shown in Table 1, ThinkMorph delivers an average improvement of **20.74%** over Bagel-7B across nine benchmarks. Despite being fine-tuned on only 24K samples, it rivals models an order of magnitude larger: outperforming Qwen2.5-VL-72B by 34% on VSP, surpassing InternVL3.5-38B on SAT (52.67% vs. 49.33%), and matching Gemini 2.5 Flash on MMVP (80.33%).

3. Emergent Properties in Interleaved Reasoning

Beyond performance improvements, ThinkMorph exhibits three emergent properties that arise naturally from interleaved training, reflecting higher-level multimodal intelligence.

> **PROPERTY ①: Unseen Visual Manipulations**
> The model develops accurate visual manipulations *unseen* in training when generalizing to out-of-domain tasks.

We identify eight distinct types of unseen manipulations, including *zoom-in*, *inpainting*, *multi-box generation*, motion *forecasting*, and region *cropping*. On some benchmarks,

	Size	VSP	VisPuzzle	ChartQA	VStar★	BLINK-J★	MMVP★	SAT★	BLINK★	CV-Bench★
Visual Understanding-only VLM										
GPT-4o	-	33.50	43.75	76.34	61.78	72.67	84.67	28.00	60.28	75.61
GPT-5	-	57.33	78.00	80.85	71.73	77.33	86.33	73.30	69.86	85.46
Gemini 2.5 Flash	-	59.33	47.00	83.79	70.68	66.00	80.33	56.00	67.49	85.07
InternVL3.5	*8B*	8.17	34.75	76.26	68.59	71.33	76.33	45.33	59.60	81.99
	38B	20.16	36.50	80.44	76.96	80.67	80.33	49.33	62.65	85.96
Qwen2.5-VL	*7B*	2.16	34.75	78.12	76.44	59.33	77.33	51.33	55.92	75.20
	72B	41.83	40.00	82.03	85.86	61.33	82.00	64.67	61.91	82.54
Unified Models										
Janus-pro	*7B*	00.00	33.50	43.08	38.22	50.67	63.33	22.00	38.51	67.83
Chameleon	*7B*	00.83	30.50	5.74	28.27	00.67	47.67	10.67	16.52	36.52
Bagel	*7B*	00.83*	35.00*	61.82	55.49	67.33	70.33	44.67	47.66	76.03*
ThinkMorph	*7B*	**75.83**	**79.00**	**78.10**	**67.02**	**72.00**	**80.33**	**52.67**	**60.07**	**80.82**
Δ (vs Bagel)		**+75.00**	**+44.00**	**+16.28**	**+11.53**	**+4.67**	**+10.00**	**+8.00**	**+12.41**	**+4.79**

Table 1. **Comparison of ThinkMorph with Other Models.** Bagel-7B is tested under think mode (*: no-think mode for tasks where thinking prevents Bagel from generating answers). ★: out-of-domain benchmarks.

these account for up to 10% of all visual operations at inference. Crucially, they are not artifacts—they are **precise** and **task-effective**: e.g., when asked "Is the bell pepper red or yellow?", the model autonomously generates a zoomed-in view, mirroring human visual inspection.

Statistical analysis reveals systematic patterns: phrases like *"examine closely"* consistently elicit zoom-in operations; *"restore"* triggers inpainting. This capability emerges from the interplay between Bagel's large-scale multimodal pretraining (raw manipulation ability) and ThinkMorph's interleaved fine-tuning (reasoning-oriented activation of those skills).

PROPERTY ②: Autonomous Mode Switching
Despite training exclusively on interleaved data, the model adaptively switches to text-only reasoning based on task complexity.

In 5.3% of inference cases, the model spontaneously switches to pure text reasoning (Figure 2). These switched instances achieve **81.25% accuracy**—a **+7.29%** improvement over the same samples solved with interleaved reasoning (73.96%).

As shown in Figure 2, mode switching is task-adaptive: the model maintains interleaved reasoning when fine-grained visual details are critical, but switches to text-only when verbal inference suffices. This reflects *front-loaded visual engagement*—after processing the image, the model implicitly decides whether text alone can complete the reasoning, dynamically allocating effort based on task demands.

PROPERTY ③: Better Test-Time Scaling via Diversified Thoughts
Interleaved reasoning enables superior test-time scaling by exploring broader multimodal solution spaces.

We compare interleaved and unimodal reasoning under Best-of-N sampling across four benchmarks representing a *continuum of distribution shifts* (Figure 3), from in-domain VSP to the most challenging out-of-domain BLINK-J.

Interleaved reasoning scales more effectively, with gains amplifying under distribution shifts. Across all benchmarks, interleaved reasoning maintains consistent improvements: +5.2% on VSP, +1.0% on VStar, +0.7% on MMVP, and a substantial **+8.0%** on BLINK-J. On BLINK-J, ThinkMorph improves from 65.33% to 73.33%, while visual reasoning drops by 2.0% and text reasoning rises only 2.67%. The 10-point gap between interleaved and visual modes highlights that multimodal exploration becomes most critical when single modalities cannot generalize effectively.

Scaling patterns vary across task types. Two distinct patterns emerge. For *reasoning-intensive* tasks, performance improves **monotonically** with larger N: VStar shows +5.89%

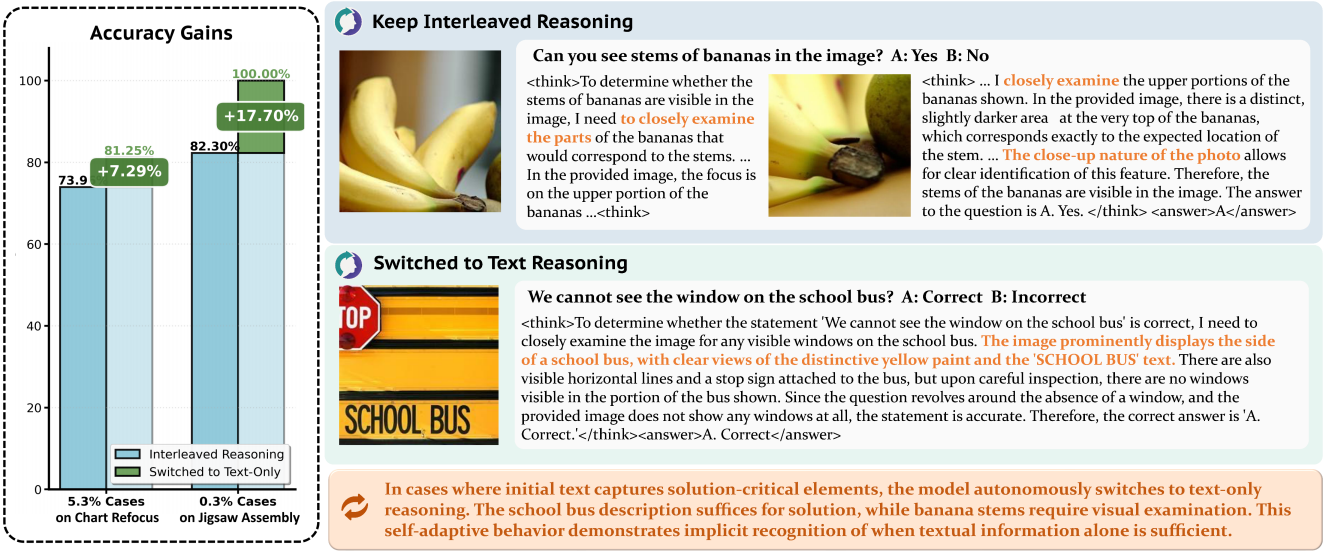

Figure 2. **Autonomous Mode Switching.** The model maintains interleaved reasoning when fine-grained visual details are critical (left), and switches to text-only reasoning when verbal inference suffices (right).

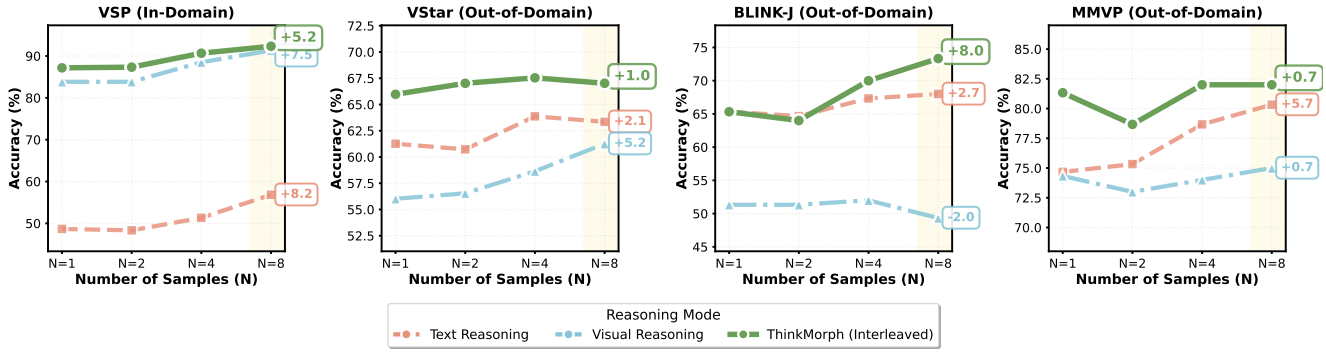

Figure 3. **Test-Time Scaling (Best-of-N) Across Reasoning Modes.** Interleaved reasoning demonstrates robust scaling advantages, especially under distribution shift.

at N=8, and CV-Bench follows a similar trend. In contrast, *perception-focused* benchmarks exhibit **U-shaped scaling**: BLINK-J initially drops 2.91% from N=2 to N=4 before recovering at N=8. These patterns indicate that reasoning-oriented tasks gain steadily from expanded multimodal exploration, whereas perception-heavy tasks require larger sample sizes to escape local optima.

The scaling advantage arises from richer trajectory diversity: unimodal chains are confined to single representational spaces, whereas interleaved reasoning spans both modalities simultaneously, producing diverse trajectories that cover complementary subsets of the solution space. As N increases, this diversity becomes crucial, greatly improving the likelihood that at least one trajectory reaches the correct answer.

4. Conclusion

We present ThinkMorph, a unified model for interleaved multimodal CoT reasoning built on the principle that text and image thoughts must be *complementary*, not isomorphic. With ~24K carefully curated interleaved traces, ThinkMorph achieves substantial gains on vision-centric benchmarks (+34.7% average) and generalizes robustly out-of-domain, rivaling models an order of magnitude larger. More importantly, it reveals three emergent properties—unseen visual manipulations, autonomous mode switching, and superior test-time scaling—that together characterize a new frontier of adaptive multimodal intelligence. We hope these findings inspire future research on emergent capabilities and test-time scaling in unified multimodal reasoning models.

References

[1] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[2] Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*, 2024.

[3] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[4] 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.

[5] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[6] 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.

[7] 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.

[8] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024.