

Let's Think in Multimodality

Jiawei Gu

Outline

1. WHAT does it mean to Think in multimodality?

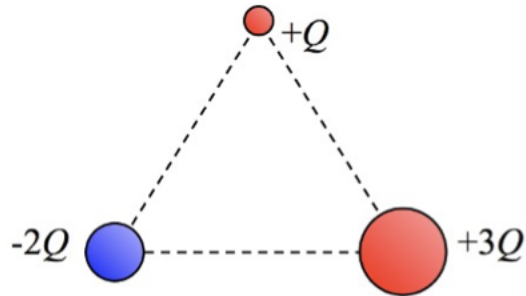
( : multimodal reasoning)

2. HOW to truly Think in multimodality?

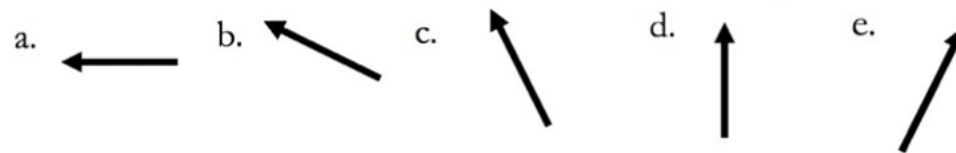
( ThinkMorph)

What is MM Reasoning?

Multimodal reasoning question in EMMA



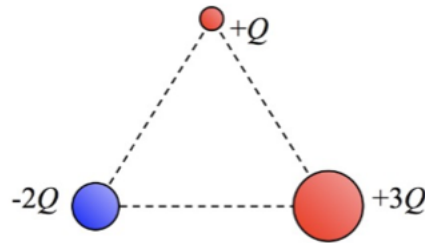
Question: Three point charges, of charge $+Q$, $-2Q$, and $+3Q$, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the $+Q$ charge?



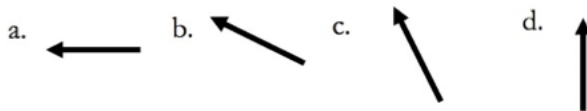
A physics problem that requires force analysis.

What is MM Reasoning?

Multimodal reasoning question



Question: Three point charges, of charge $+Q$, are placed equidistant as shown. Which vector the net direction of the electric force acting on



Human reasoning

The direction of the electric force due to $-2Q$ and $+3Q$ can be visualized with a quick free-body sketch of the electric forces:

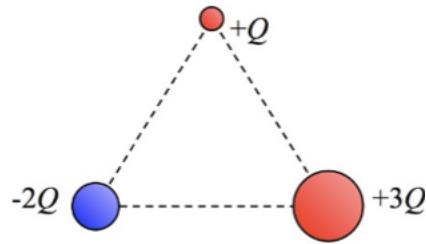


Try a quick sketch

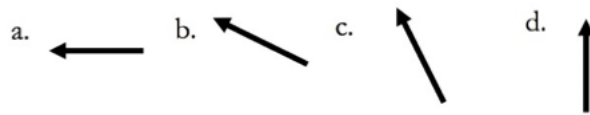
Force analysis between two positive and one negative

What is MM Reasoning?

Multimodal reasoning question



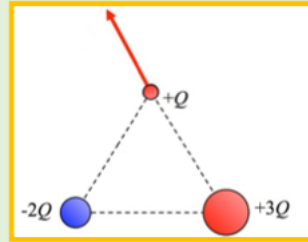
Question: Three point charges, of charge $+Q$, are placed equidistant as shown. Which vector the net direction of the electric force acting on



Human reasoning

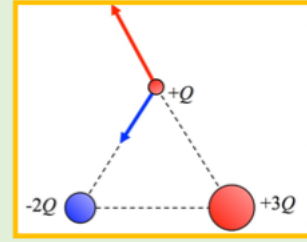
The direction of the electric force due to $-2Q$ and $+3Q$ can be visualized with a quick free-body sketch of the electric forces:

1

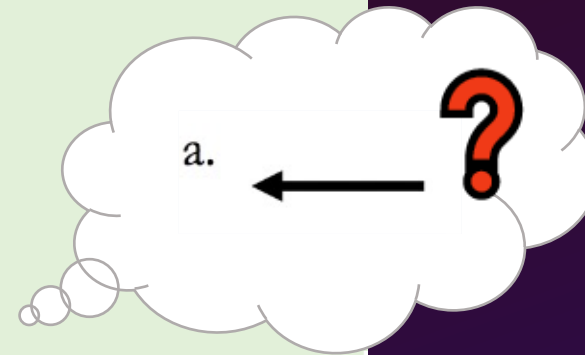
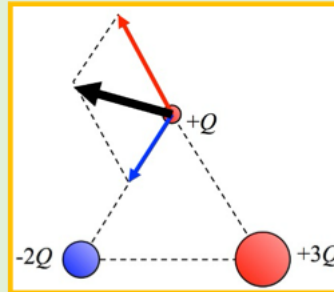


repel

2



3



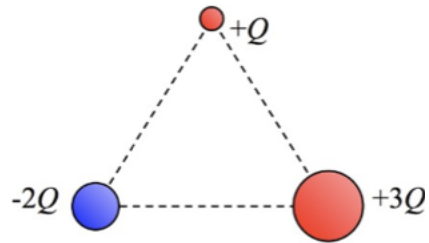
“Alike repel”

“Opposites attract”

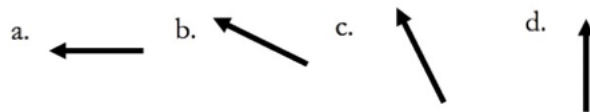
The direction of joint force is clear.

What is MM Reasoning?

Multimodal reasoning question



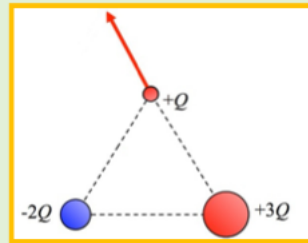
Question: Three point charges, of charge $+Q$, are placed equidistant as shown. Which vector the net direction of the electric force acting on



Human reasoning

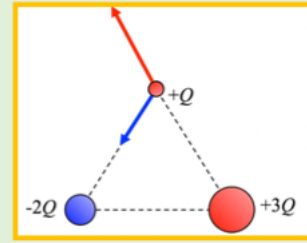
The direction of the electric force due to $-2Q$ and $+3Q$ can be visualized with a quick free-body sketch of the electric forces:

1



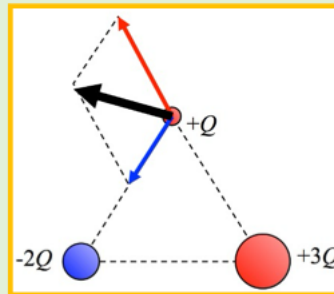
repel

2



attract

3



Combining the forces:

$$\because |+3Q| > |-2Q|$$

$$\therefore \text{repel} > \text{attract}$$

b.



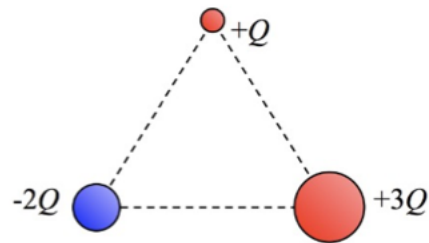
A back-and-forth thinking process between Multimodality



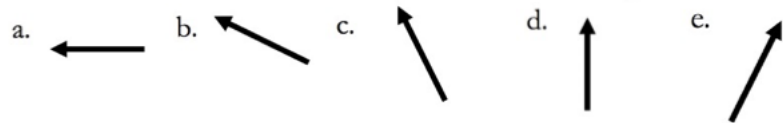
Think in Multimodality

Challenge

Multimodal reasoning question in EMMA

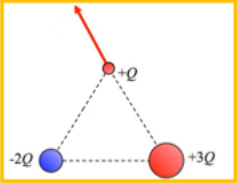


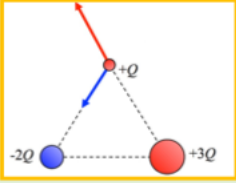
Question: Three point charges, of charge $+Q$, $-2Q$, and $+3Q$, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the $+Q$ charge?

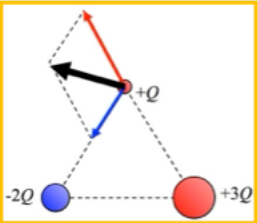


Human reasoning


The direction of the electric force due to $-2Q$ and $+3Q$ can be visualized with a quick free-body sketch of the electric forces:

1 
repel

2 
attract

3 

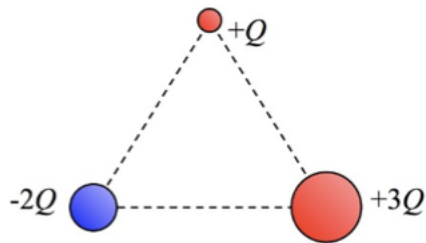
Combining the forces:
 $\because | +3Q | > | -2Q |$
 \therefore **repel** > **attract**

b. 

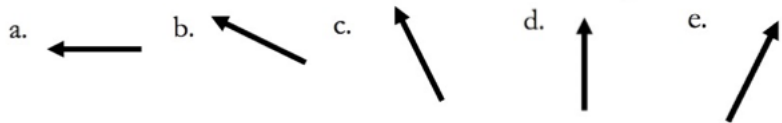
Can MLLMs perform this type of "cross-modal" reasoning?

Challenge

Multimodal reasoning question in EMMA

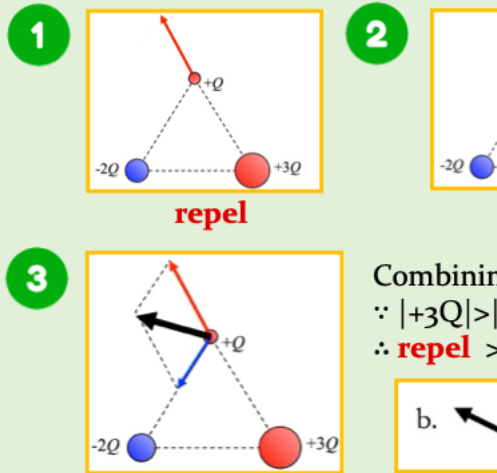


Question: Three point charges, of charge $+Q$, $-2Q$, and $+3Q$, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the $+Q$ charge?



Human reasoning

The direction of the electric force due to $+3Q$ can be visualized with a quick free sketch of the electric forces:



Model reasoning (GPT-4o)

1. Forces Due to the Charges:

- The $-2Q$ charge attracts the $+Q$ charge. (Force points downward-left).
- The $+3Q$ charge repels the $+Q$ charge. (Force points downward-right). ❌

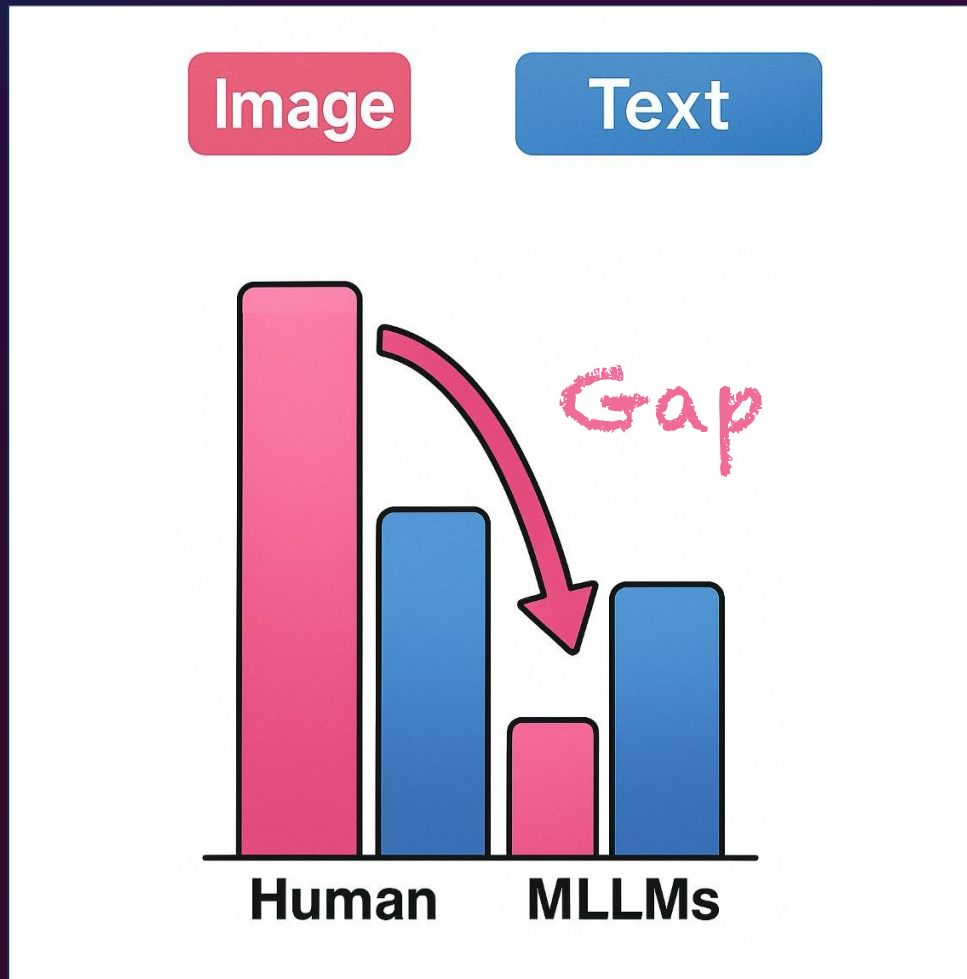
2. Force Magnitudes (Coulomb's Law):

- Force due to $-2Q$: $F_{-2Q} = \frac{2kQ^2}{r^2}$.
- Force due to $+3Q$: $F_{+3Q} = \frac{3kQ^2}{r^2}$.
-

Error: The issue arose from not correctly identifying the force direction from $+3Q$ to $+Q$, which causes the net force to point ↙, not downward-right ↘.

The answer seems to be negative.

Cross-modal Reasoning is Challenging



MLLMs

- Language-dominant reasoning
- Vision bottleneck exists
- 80% vs 20%

What can we do NEXT?

**New paradigms
needed!**





THINKMORPH: EMERGENT PROPERTIES IN MULTIMODAL INTER- LEAVED CHAIN-OF-THOUGHT REASONING

**Jiawei Gu^{*,1}, Yunzhuo Hao^{*,2}, Huichen Will Wang^{*,3}, Linjie Li^{*,3}, Michael Qizhe Shieh¹,
Yejin Choi⁴, Ranjay Krishna³, Yu Cheng⁵**

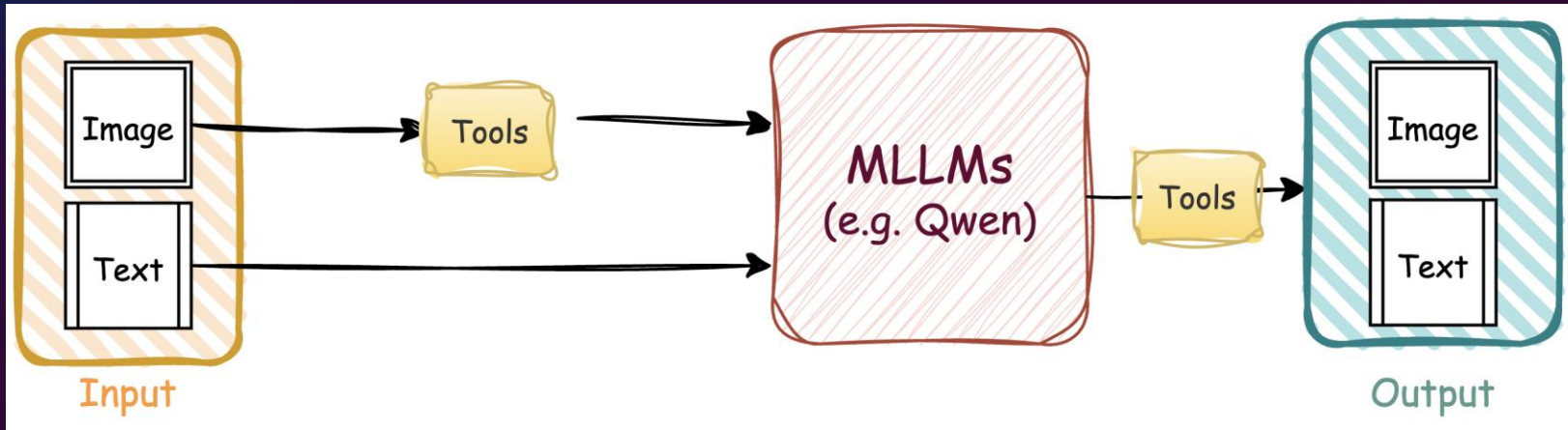
¹National University of Singapore, ²Zhejiang University, ³University of Washington,
⁴Stanford University, ⁵The Chinese University of Hong Kong

🏠 Homepage: <https://thinkmorph.github.io>

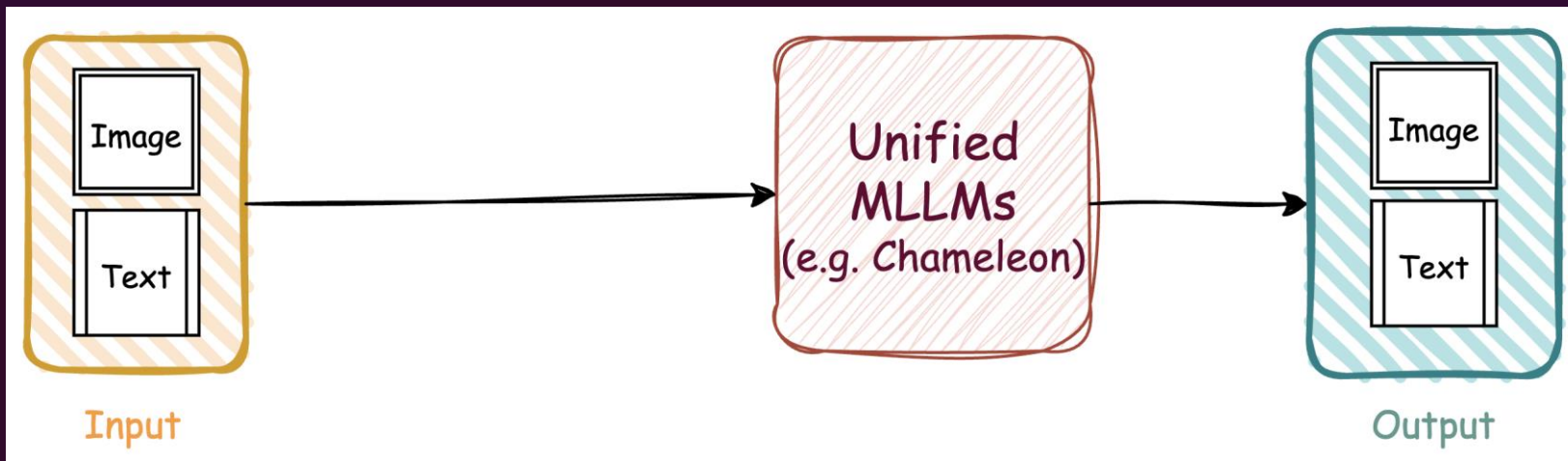
🔗 Code: <https://github.com/ThinkMorph/ThinkMorph>

😊 Models and Datasets: <https://huggingface.co/ThinkMorph>

Related Work



indirect

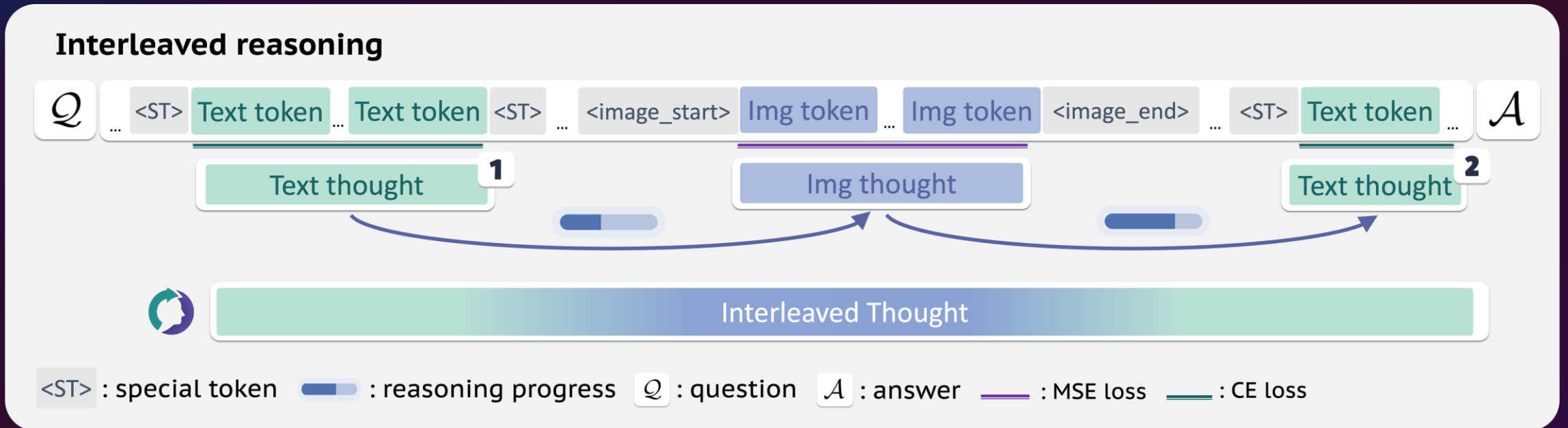


Natural
Truly



(Native Any-to-Any)

ThinkMorph (base unified model)



Text and vision contribute complementary information that progressively drives the reasoning process toward the solution.

Four Vision-centric Tasks

Jigsaw Assembly

The image below is a 2x2 jigsaw puzzle...
Your task is to determine the correct arrangement ...

- A. ...
- B. ...
- C. ...
- D. ...



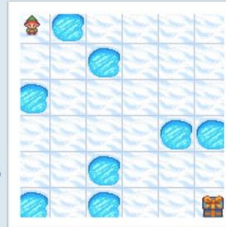
<think>... analyze the visual content of each piece and determine their correct positions based on natural image flow ... </think>



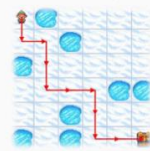
<think>... There are **no visible discontinuities** ... harmoniously unified... </think>
<answer> C</answer>

Spatial Navigation

Your goal is to guide a player from the start to the goal on a grid map while avoiding holes ...
Provide your solution as a sequence of moves ...



<think>Let's analyze the maze step by step: Step 1: Identify the grid and obstacles. The player starts at the top-left corner (let's call this position (1,1))... </think>



<think>...carefully verify the path I plotted. **The red arrows** start at the top-left ... </think> <answer>\boxed{D,R,D,D,R,R,D,D,R,R}</answer>

Visual Search

What is the color of the man's cap?

- A: yellow B: red
- C: gray D: green



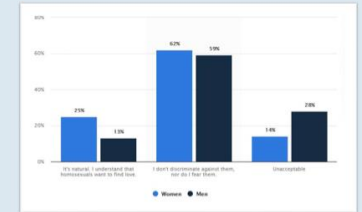
<think>.....The **red bounding box** should be placed around the **man's head and cap**...</think>



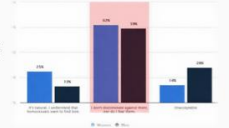
<think>... now highlighting the man ... focusing on the highlighted area, ... **the man's cap is gray**, so the correct answer is C. </think> <answer>C</answer>

Chart Refocus

Which attitude represent the smallest gap between women and men?



<think>To answer ... I **focus on** these two specific bars and labels ... </think>



<think> ... I analyze the highlighted area in the **visualization** that **focuses on the second category: 'I don't discriminate against ...** </think> <answer>I don't discriminate against them, nor do I fear them.</answer>

Interleaved thought format: text-image-text

Jigsaw Assembly

The image below is a 2x2 jigsaw puzzle... Your task is to determine the correct arrangement ...



A. The top-left part should be Part 1; the top-right part should be Part 4; the bottom-left part should be Part 3; and the bottom-right part should be Part 2.

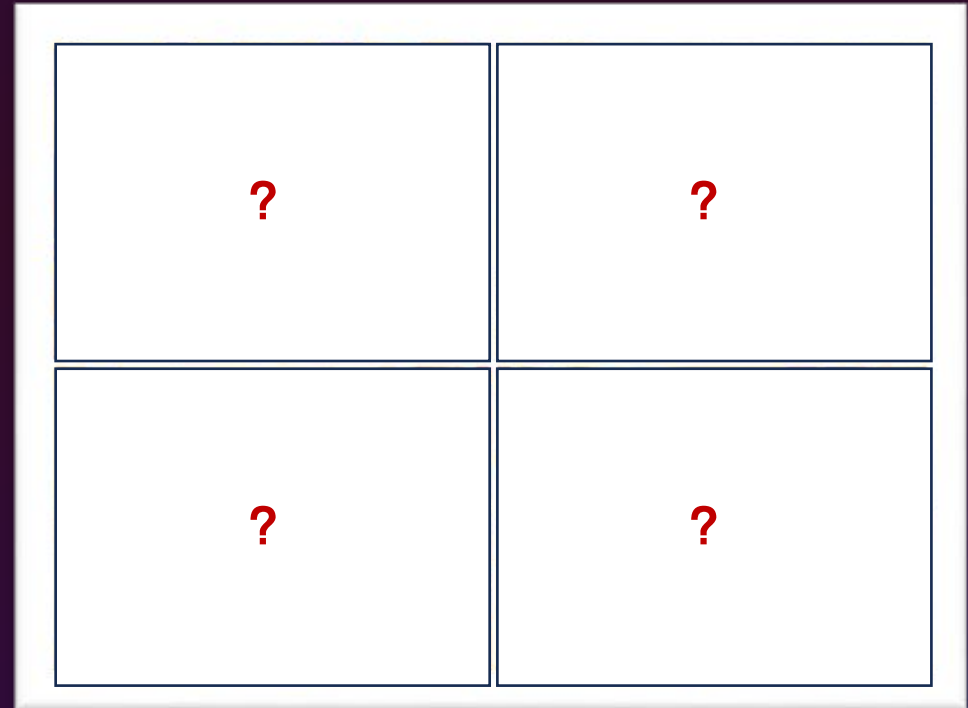
B. The top-left part should be Part 2; the top-right part should be Part 1; the bottom-left part should be Part 4; and the bottom-right part should be Part 3.

C. The top-left part should be Part 2; the top-right part should be Part 4; the bottom-left part should be Part 3; and the bottom-right part should be Part 1.

D. The top-left part should be Part 4; the top-right part should be Part 2; the bottom-left part should be Part 3; and the bottom-right part should be Part 1.

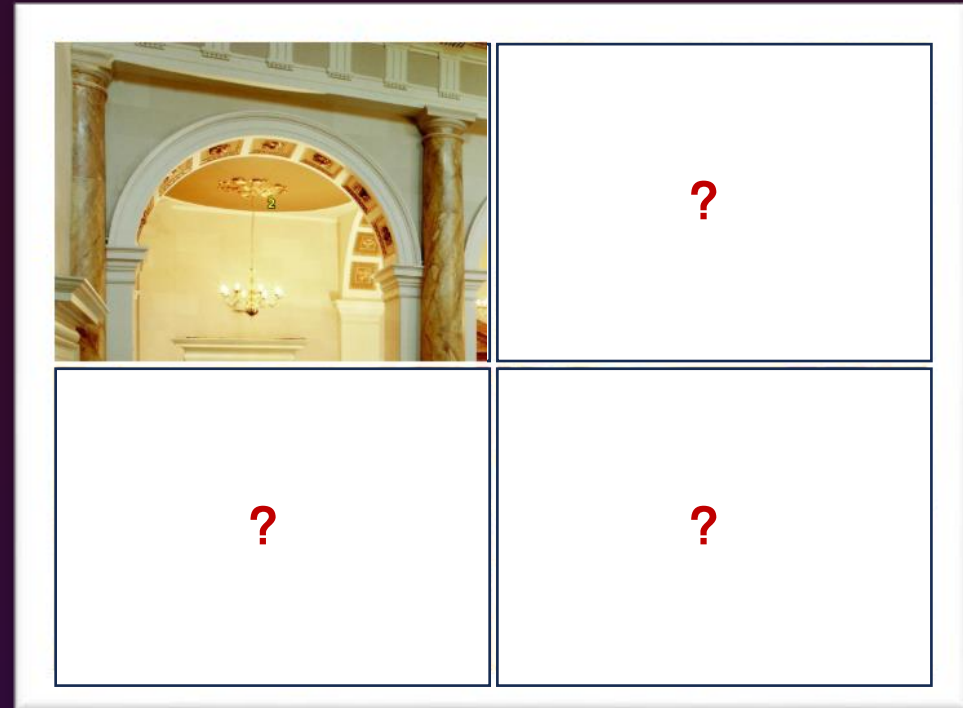
Jigsaw Assembly

The image below is a 2x2 jigsaw puzzle... Your task is to determine the correct arrangement ...



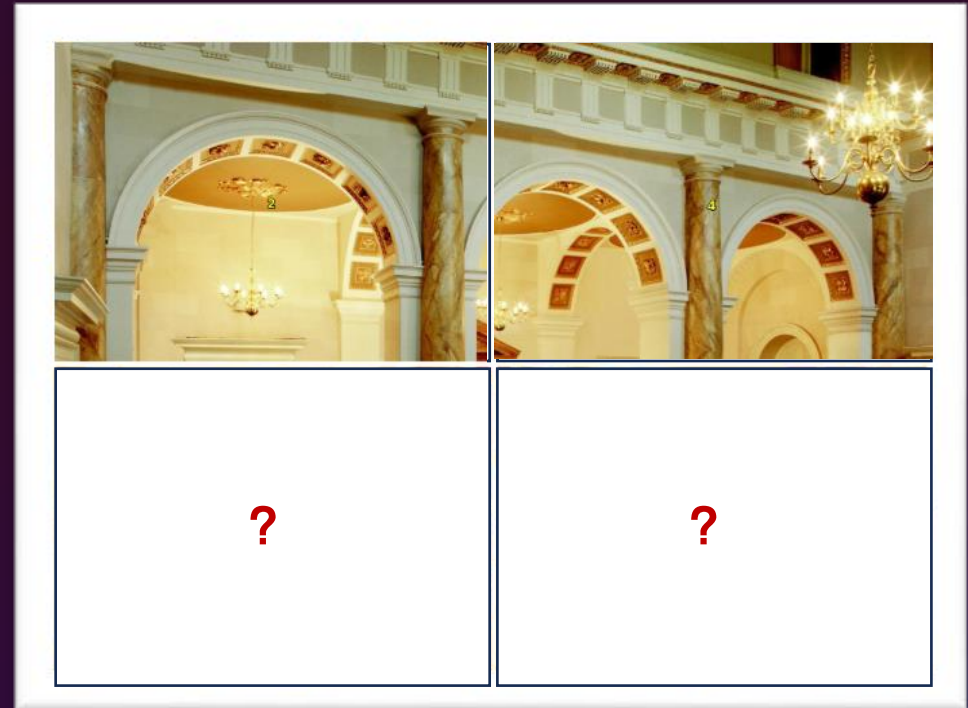
Jigsaw Assembly

The image below is a 2x2 jigsaw puzzle... Your task is to determine the correct arrangement ...



Jigsaw Assembly

The image below is a 2x2 jigsaw puzzle... Your task is to determine the correct arrangement ...



Jigsaw Assembly

The image below is a 2x2 jigsaw puzzle... Your task is to determine the correct arrangement ...



<think>... analyze the visual content of each piece and determine their correct positions based on natural image flow... </think>



<think>... There are **no visible discontinuities** ... harmoniously unified... </think> <answer> C </answer>

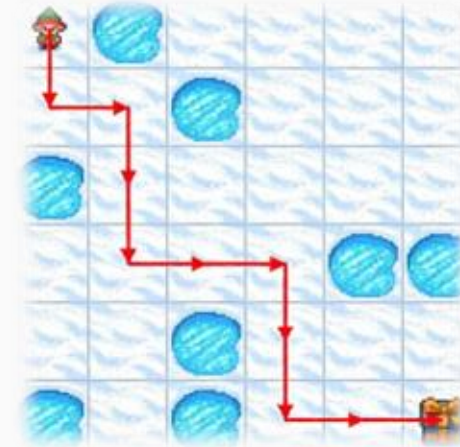


Spatial Navigation

Your goal is to guide a player from the start to the goal on a grid map while avoiding holes ...
Provide your solution as a sequence of moves ...



<think>Let's analyze the maze step by step:
Step 1: Identify the grid and obstacles The player starts at the top-left corner (let's call this position (1,1))
... </think>



<think>...carefully
verify the path I plotted. **The red arrows** start at the top-left ... </think> <answer>\boxed{D,R,D,D,R,R,D,D,R,R}</answer>



Data for SFT

Task	Data Source	Count	Visual Manipulation	Curation Steps
Jigsaw Assembly	SAT (Ray et al., 2024), ADE20K (Zhou et al., 2017), Omni3D (Brazil et al., 2023)	6,000	Visualizing re-arranged pieces	Newly generate questions from a customized pipeline
Spatial Navigation	N/A	6,000	Overlaying mazes with paths highlighted with red lines and arrows	Newly generated questions from a customized pipeline
Visual Search	Visual CoT (Shao et al., 2024), GQA (Hudson & Manning, 2019), VSR (Liu et al., 2023)	6,990	Highlighting Regions with Red Bounding Boxes	Filtering for valid (question, answer) with MLLMs + other criteria
Chart Refocus	ChartQA (Masry et al., 2022), Refocus (Fu et al., 2025)	6,000	Highlighting Regions with Red Bounding Boxes or Overlays	Filtering for valid (question, answer) with MLLMs + other criteria

Table 1: Summary of Questions Used for Training ThinkMorph.

Bagel

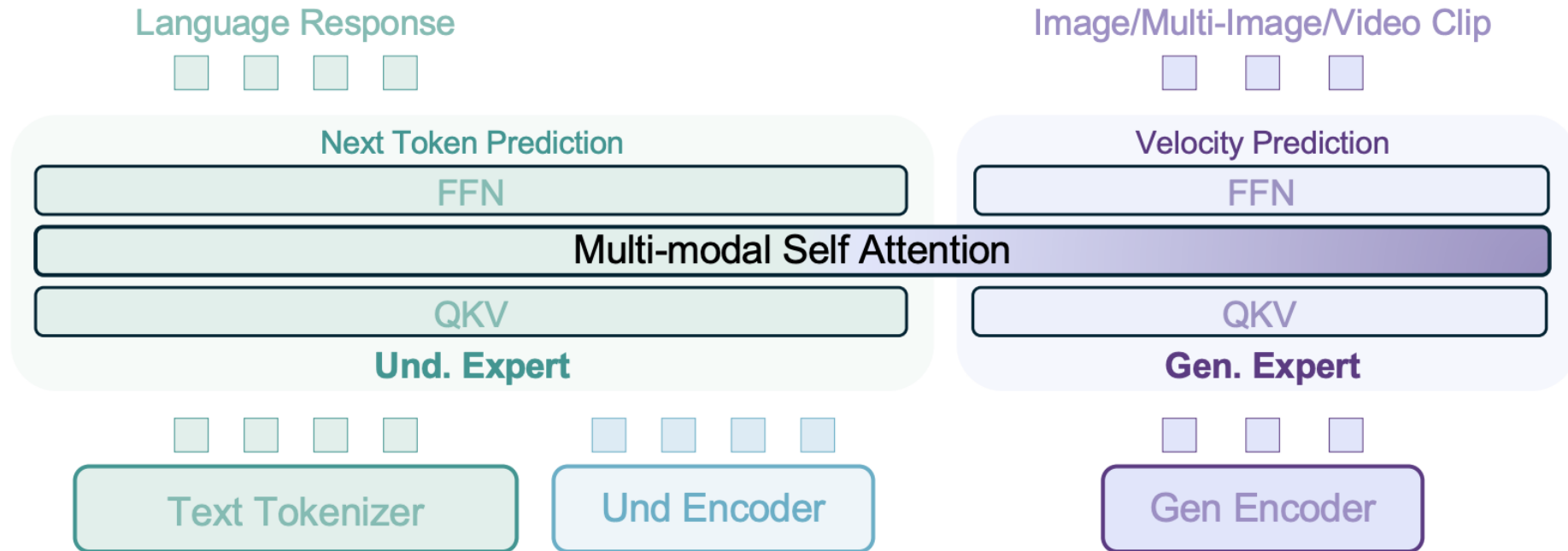
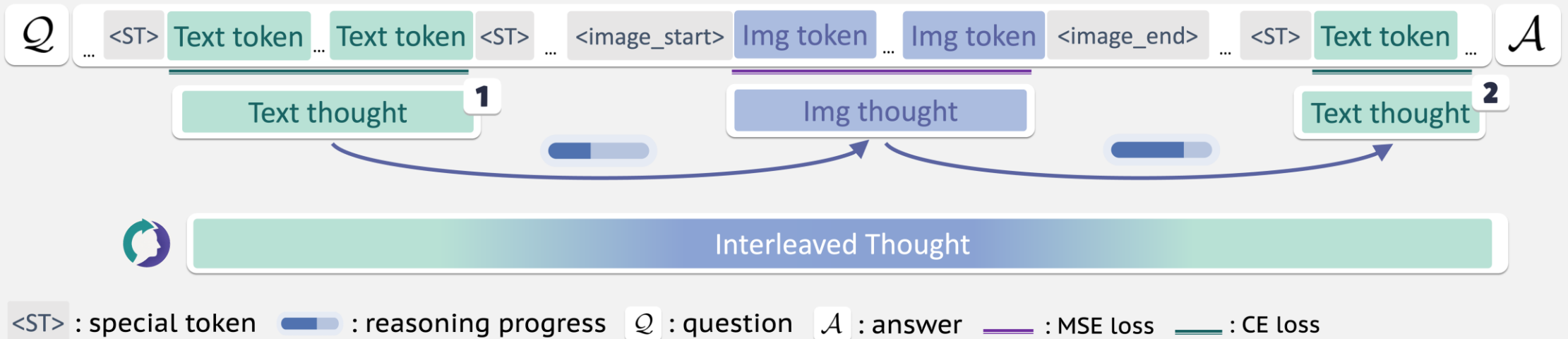


Figure 2 We use **two Transformer experts** to process understanding and generation information, and all tokens do shared multi-modal self attention in each Transformer block. We adopt two distinct encoders to separately capture semantic content and low-level pixel information for image understanding and generation tasks.

ThinkMorph (base Bagel)

Interleaved reasoning



CE loss for text and MSE loss for vision

1 (text) + 1 (image) > 1 (text/image-only) ?

ThinkMorph


	Spatial Navigation	Visual Search	Jigsaw Assembly		Chart Refocus	
	VSP	VStar [★]	VisPuzzle	BLINK-J [★]	ChartQA	MMVP [★]
Bagel-7B	0.83*	55.49	35.00*	67.33	62.05	70.33
Text Reasoning	49.17	56.02	<u>63.50</u>	<u>68.67</u>	81.66	<u>76.33</u>
Visual Reasoning	<u>85.50</u>	<u>58.63</u>	61.25	47.33	73.08	73.00
 Interleaved Reasoning	86.67	63.87	73.75	73.33	<u>79.78</u>	82.66

Table 2: Reasoning Mode Comparison. Bagel-7B is tested under think mode (*: no-think mode for tasks where thinking prevents Bagel from generating answers). ChartQA results are the average performance on horizontal and vertical bar chart questions. [★]: out-of-domain benchmarks. **Best**, second-best.

- Interleaved reasoning excels on vision-centric tasks.
-  Interleaved vs. Others: +5.33% (vs **Text**-only & **vision**-only)

Generalization


	Size	VSP	VisPuzzle	ChartQA	VStar [★]	BLINK-J [★]	MMVP [★]	SAT [★]	BLINK [★]	CV-Bench [★]
<i>Visual Understanding-only VLMs</i>										
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
	33B	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
<i>Unified Models</i>										
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 3: 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.

ThinkMorph with 24K samples vs. Large VLMs: **MMVP, SAT**

Beyond Performance ...

We started noticing something unexpected



Emergent Properties

1 Unseen Visual Manipulations



Is the bell pepper in the picture red or yellow?
A: Red B: Yellow

(From MMVP)

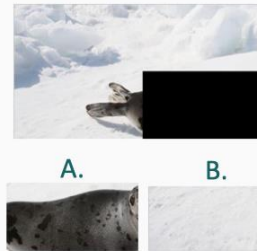


<think>To determine whether the bell is red or yellow, I need to **closely examine the color** of the pepper shown ...



<think>... I closely examine the image provided, which features a close-up shot as its central focus... Based on this detailed visual inspection, I can confidently conclude that the bell pepper is **yellow**... </think>
<answer>B</answer>

Zoom-in



Given the first image with the lower right corner missing, ... compare the edges of the images. Select from the following choice. (From BLINK)



<think> ... To determine the correct assembly, I focus on the continuity between the two pieces ... **reconstructing** the seal as it would appear in reality.</think>



<think>... I can examine the result for visual coherence ... overall composition **looks natural** ... as in a single, unaltered photo....</think> <answer>A</answer>

Image Inpainting

2 Autonomous Mode Switching



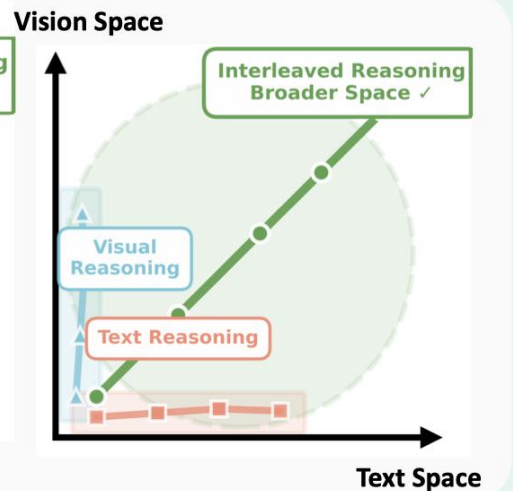
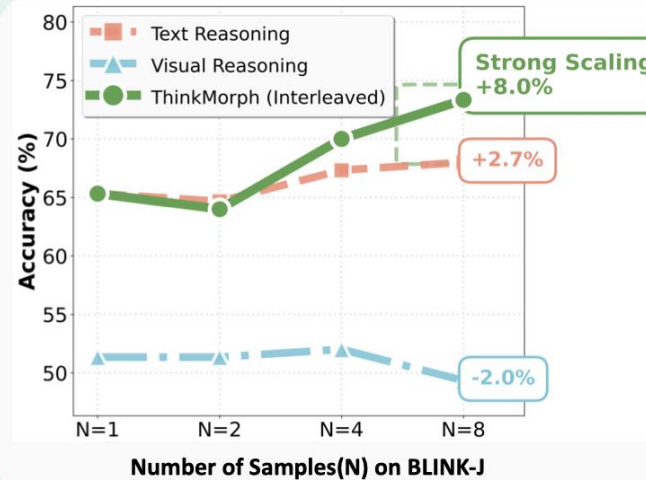
Interleaved Thought

Text thought

OR









e.g., 5.3% Switch
7.29% Gain

3 Superior Test-Time via Diversified Thoughts



EMERGENT PROPERTY ① : Unseen Visual Manipulations The model develops accurate and meaningful visual manipulations unseen in training data when generalizing to out-of-domain multimodal tasks, actively advancing the reasoning process.

8 Types :

Multiple bboxes	Motion Forecasting	Perspective Shift	Crop
 <p>Is the blue mask on the left or right side of the black mask? A: right B: left</p>	 <p>The image captures a woman running from: A: Back view B: Side view</p>	 <p>Is there a knee visible in the image? A: Yes B: No</p>	 <p>What is the color of the telephone? A. green B. white C. orange D. blue</p>
<p><think> To determine where the red bounding box should go, first look for the faces of the people in the image...</think></p>  <p><think>...it's clear he is wearing a blue mask. The other visible worker, positioned to his right...</think> answer>left</answer></p>	<p><think> ...I need to focus on the perspective ... it makes sense to highlight the area that clearly shows the side profile, such as the alignment of her body, and the profile of her running form ...</think></p>  <p><think>... In the edited visualization... showing her side silhouette, the alignment of legs, and the profile of her face in motion...</think><answer>B</answer></p>	<p><think> ... to highlight the visible knee(s) in the image ... the manipulation involves drawing attention to the area where the knee is visible ...</think></p>  <p><think>... In the edited visualization, attention is drawn to the area where the thigh meets the shin, which is the location of the knee. Both individuals have their knees slightly bent.</think> <answer>A</answer></p>	<p><think> ... the red bounding box should go...The bounding box should be drawn around this wall-mounted device to highlight its color..</think></p>  <p><think>... The object's color is a distinct orange...as shown in the highlighted region, is orange.</think><answer>C</answer></p>

- Zoom-in (most)
- Inpainting
- Multi-box generation
- Motion forecasting
- Perspective Shifting
- Region cropping
- ... and more

Emergent Property 1

- What Scale? Up to 10%

1 Unseen Visual Manipulations



Is the bell pepper in the picture red or yellow?

A: Red B: Yellow

(From MMVP)

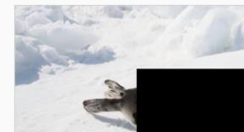


<think>To determine whether the bell is red or yellow, I need to **closely examine the color** of the pepper shown ...



<think>... I closely examine the image provided, which features a close-up shot as its central focus... Based on this detailed visual inspection, I can confidently conclude that the bell pepper is **yellow**... </think>
<answer>B</answer>

Zoom-in



Given the first image with the lower right corner missing, ... compare the edges of the images.

Select from the following choice. (From BLINK)

A.

B.



<think> ... To determine the correct assembly, I focus on the continuity between the two pieces ... **reconstructing** the seal as it would appear in reality.</think>



<think>... I can examine the result for visual coherence ... overall composition **looks natural** ... as in a single, unaltered photo....</think> <answer>A</answer>

Image Inpainting

Example

➤ "Is the bell pepper red or yellow?"

→ Model automatically zooms in to distinguish colors

✓ Mirrors human visual inspection

Emergent Property 1

- Why This Happens?

1 Unseen Visual Manipulations



Is the bell pepper in the picture red or yellow?

A: Red B: Yellow

(From MMVP)

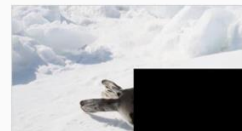


<think>To determine whether the bell is red or yellow, I need to **closely examine the color** of the pepper shown ...



<think>... I closely examine the image provided, which features a close-up shot as its central focus... Based on this detailed visual inspection, I can confidently conclude that the bell pepper is **yellow**... </think>
<answer>B</answer>

Zoom-in



Given the first image with the lower right corner missing, ... compare the edges of the images.

Select from the following choice. (From BLINK)



<think> ... To determine the correct assembly, I focus on the continuity between the two pieces ... **reconstructing** the seal as it would appear in reality.</think>



<think>... I can examine the result for visual coherence ... overall composition **looks natural** ... as in a single, unaltered photo....</think> <answer>A</answer>

Image Inpainting

➤ Text Cues Trigger Visual Actions:

Textual Cue

Visual Actions

"examine closely" / "focus on"

Zoom-in

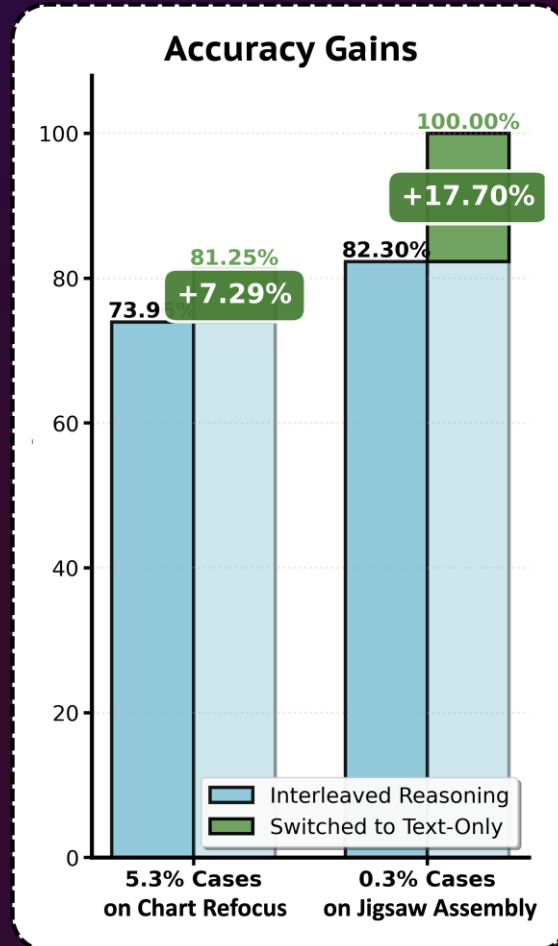
"restore" / "reconstruct"

Inpainting

➤ Pretraining: supplies ability

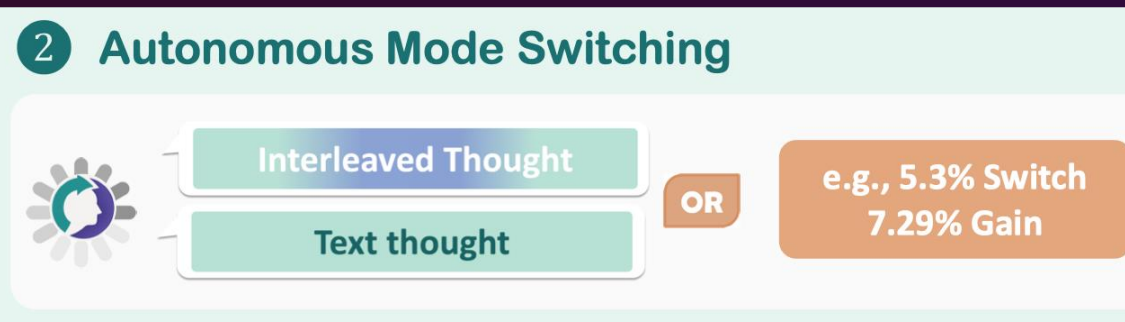
➤ Fine-tuning: directs it for reasoning

EMERGENT PROPERTY ②: Autonomous Mode Switching The model adaptively switches from interleaved to text-only reasoning based on task complexity, despite being trained exclusively on interleaved data.



The Surprise:

- Trained **exclusively** on interleaved data
- Model switches to text-only reasoning
- **5.3%** of inference cases in MMVP



Emergent Property 2

- Why This Happens?
 - ✓ Visual redundant → Text suffices
 - ✓ Visual essential → Stay Interleaved

Keep Interleaved Reasoning



Can you see stems of bananas in the image? A: Yes B: No

<think>To determine whether the stems of bananas are visible in the image, I need **to closely examine the parts** of the bananas that would correspond to the stems. ... In the provided image, the focus is on the upper portion of the bananas ...</think>



<think> ... I **closely examine** the upper portions of the bananas shown. In the provided image, there is a distinct, slightly darker area at the very top of the bananas, which corresponds exactly to the expected location of the stem. ... **The close-up nature of the photo** allows for clear identification of this feature. Therefore, the stems of the bananas are visible in the image. The answer to the question is A. Yes. </think> <answer>A</answer>

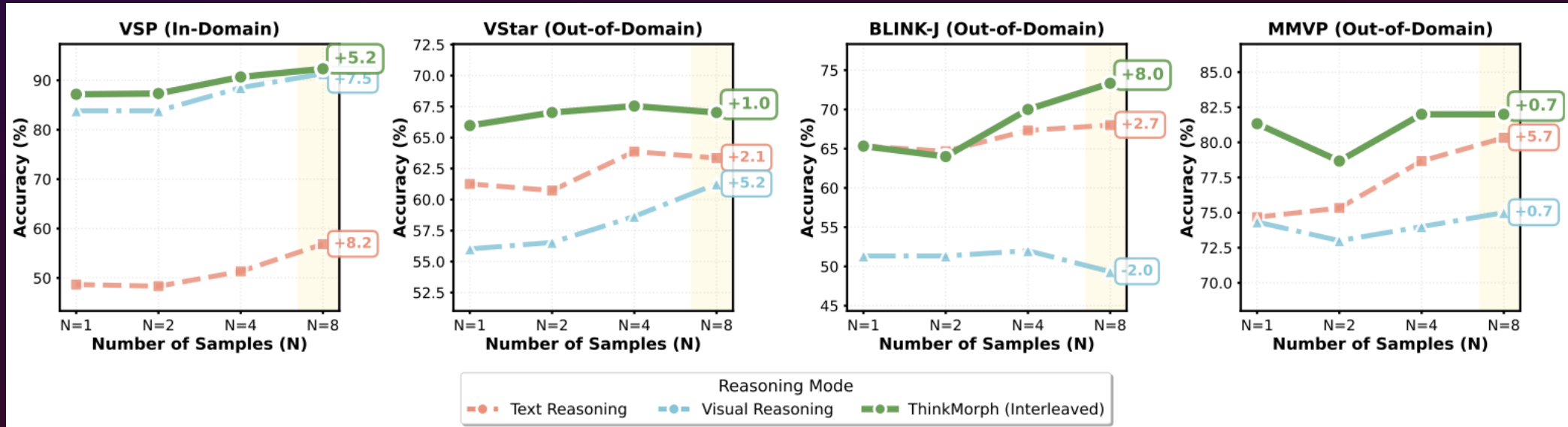
Switched to Text Reasoning




We cannot see the window on the school bus? A: Correct B: Incorrect

<think>To determine whether the statement 'We cannot see the window on the school bus' is correct, I need to closely examine the image for any visible windows on the school bus. **The image prominently displays the side of a school bus, with clear views of the distinctive yellow paint and the 'SCHOOL BUS' text.** There are also visible horizontal lines and a stop sign attached to the bus, but upon careful inspection, there are no windows visible in the portion of the bus shown. Since the question revolves around the absence of a window, and the provided image does not show any windows at all, the statement is accurate. Therefore, the correct answer is 'A. Correct.'</think><answer>A. Correct</answer>

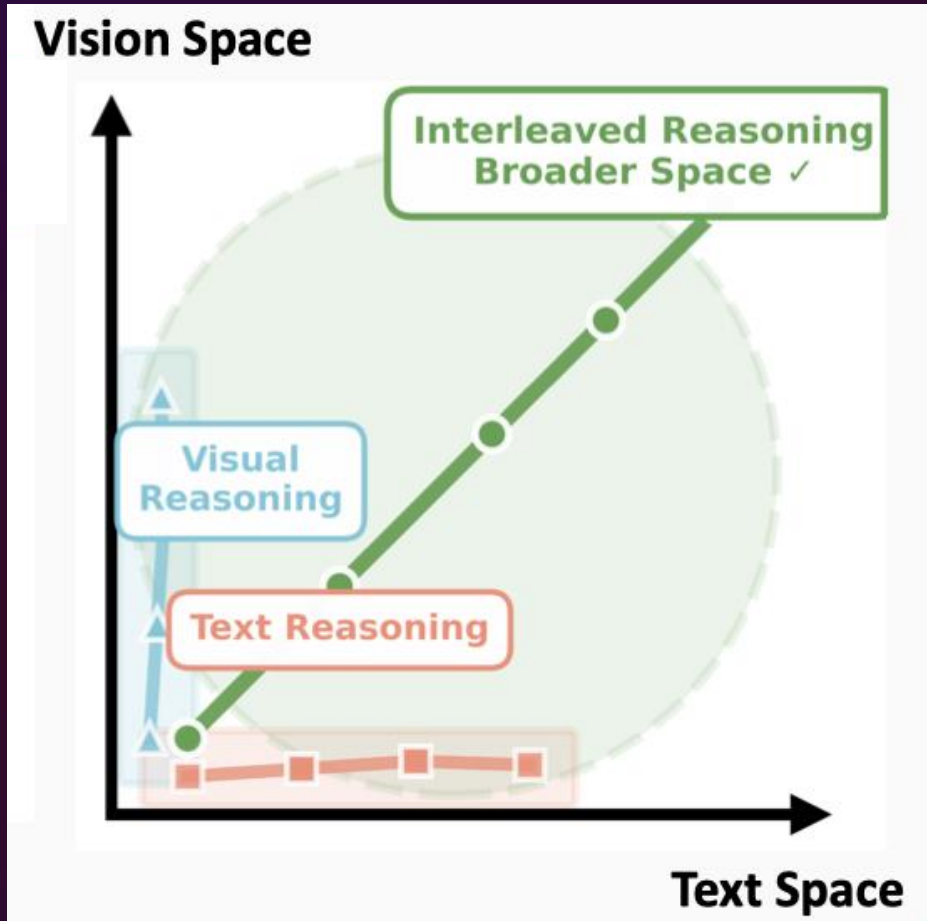
EMERGENT PROPERTY ③: Better Test-Time Scaling via Diversified Thoughts Interleaved reasoning enables superior test-time scaling by generating diversified thoughts that explore broader multimodal solution spaces, delivering stable accuracy gains that consistently outperform unimodal approaches.



 Interleaved reasoning demonstrates robust scaling advantages, particularly on challenging benchmarks where unimodal approaches plateau or decline.

Emergent Property 3

- Why This Happens?



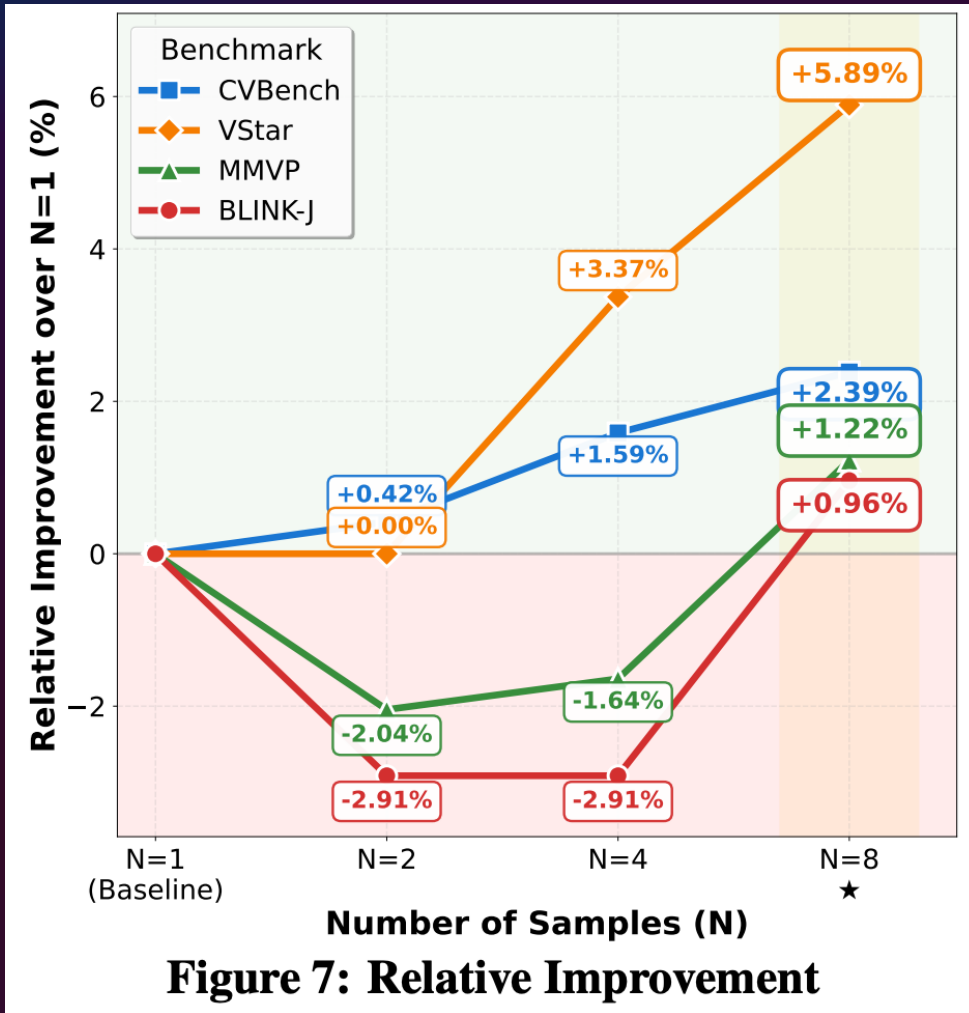
➤ Interleaved explores broader space

- Unimodal: Text OR Vision (confined)
- Interleaved: Text AND Vision (diverse)

➤ Robust scaling under distribution shifts

- OOD results

Generalization with Emergent Properties



- ✓ Unseen manipulations continue on OOD
- ✓ Autonomous mode switching stays adaptive

Property ③: Task-Dependent Scaling Patterns

- Reasoning-Intensive
 - VStar: +5.89%
 - CV-Bench: +2.39%
 - Monotonic growth
- Perception-Focused
 - MMVP: U-shaped
 - BLINK-J: drops then up
 - Needs larger N

Thanks! Happy to take any questions you have!

Jiawei Gu