

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 THINKMORPH: EMERGENT PROPERTIES IN MULTIMODAL INTER- LEAVED CHAIN-OF-THOUGHT REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Multimodal reasoning is a dynamic process that requires synergistic coordination of language and vision. However, current approaches to multimodal interleaved generation fall short of providing a *generalizable* recipe that productively engages text and vision to advance reasoning. We introduce **ThinkMorph**, a unified thinking model fine-tuned on $\sim 24K$ high-quality interleaved reasoning traces. By constructing a pipeline for interleaved data for training unified models, we enable ThinkMorph to generate multimodal reasoning traces where language and vision mutually advance each other. ThinkMorph delivers substantial gains on vision-centric reasoning, including +11.53% on visual search and +38.75% on jigsaw assembly over the base model. It also reaches 80.33% on MMVP and 52.67% on SAT, indicating strong generalization. The improvements are large enough to close the gap with, and in some cases even surpass, leading large-scale or proprietary VLMs. Moreover, ThinkMorph reveals emergent properties indicative of higher-level multimodal intelligence. These include unseen visual manipulation skills during finetuning, such as zoom-in and image inpainting, as well as autonomous reasoning mode switching, wherein a model trained exclusively on interleaved data chooses to engage in text-only reasoning due to the nature of the task, for instance. We show that this ability to think in text, vision, and multimodality opens new avenues for test-time scaling, allowing ThinkMorph to *effectively scale and aggregate thoughts across three reasoning modes*. These findings suggest promising directions for future work to characterize the emergent capabilities of unified models for multimodal reasoning.

1 INTRODUCTION

Multimodal reasoning (Lin et al., 2025) is not a single-pass perception task but an iteration between language and vision reasoning. This process is especially challenging for current models in vision-centric tasks such as *spatial reasoning* (Li et al., 2025c) and *visual search* (Wu & Xie, 2024), where they must move beyond describing images to interrogating or even manipulating visual elements. Despite the success of textual Chain-of-Thought (hereafter, “text thought”) in improving verbal reasoning (Wei et al., 2022; Hao et al., 2025; Jiang et al., 2025), it contributes little to multimodal reasoning. Models continue to struggle when problems require more than mere text description. Such limitations (Fu et al., 2024a) highlight the need to move from language-driven reasoning toward cross-modal reasoning, as exemplified by the human ability to tackle complex problems via think-and-sketch methods.

To replicate think-and-sketch style reasoning, researchers have proposed various recipes for multimodal interleaved Chain-of-Thought (hereafter, “interleaved thought”), but each has limitations. One approach adopts a tool-augmented design, relying on external visual modules like cropping tools (OpenAI) and specialized models (Hu et al., 2024; Zhou et al., 2024), which renders interleaving indirect and fragile. Another approach builds on unified models (Team, 2024; Chern et al., 2024; Deng et al., 2025), which natively support multimodal input and output. However, such efforts remain preliminary and lack a generalizable recipe for enabling text and image reasoning to mutually advance each other. For instance, while MVoT (Li et al., 2025b) introduces interleaved action representations to solve mazes, its textual component is limited to simplistic action labels that

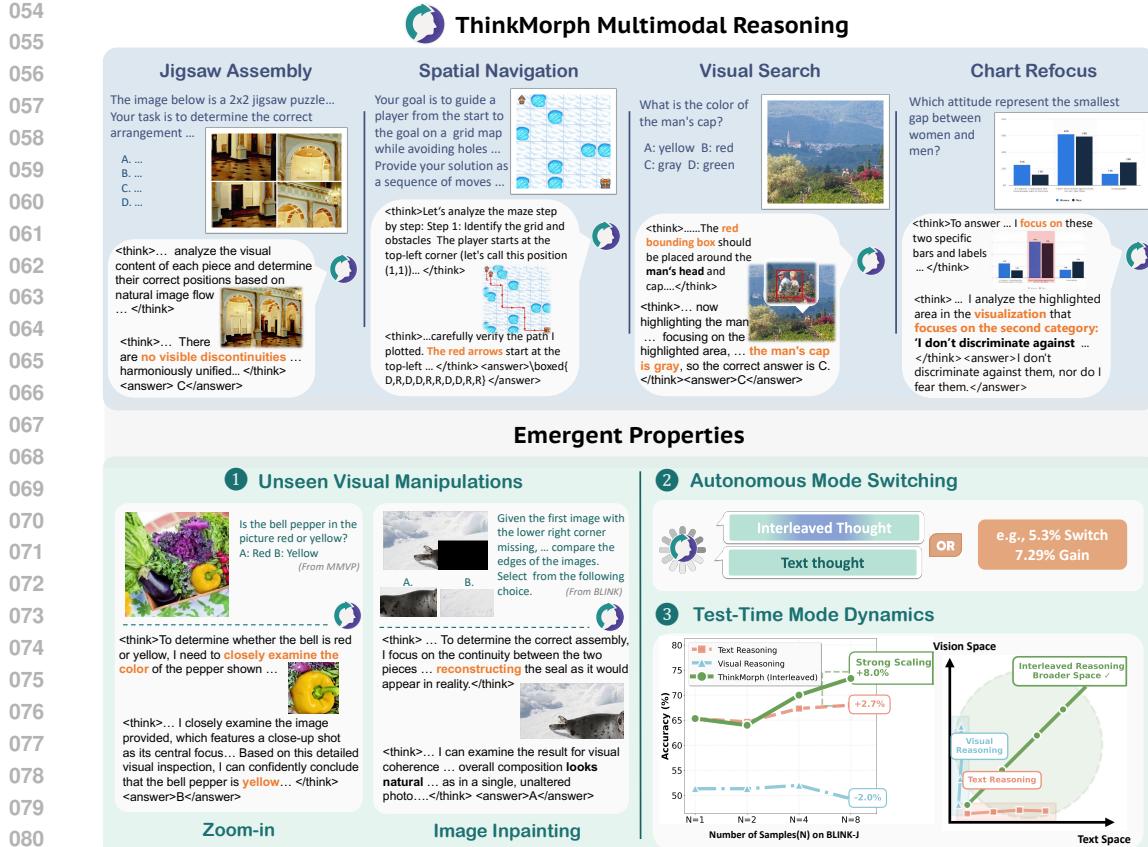


Figure 1: ThinkMorph Overview. ThinkMorph synergistically interleaves language and vision to advance multimodal reasoning across four representative tasks (top). Beyond performance gains on in- and out-of-domain benchmarks, interleaved reasoning unlocks emergent properties (bottom).

are isomorphic to the generated images. Perhaps as a result, this approach has not demonstrated evidence of generalization beyond its training domains.

To address this gap, we introduce **ThinkMorph**, a thinking unified model capable of synergistically interleaving textual and visual reasoning. Through finetuning with carefully curated interleaved thoughts, ThinkMorph delivers substantial gains on vision-centric reasoning, including +11.53% on visual search and +38.75% on jigsaw assembly over the base model. It also reaches 80.33% on MMVP and 52.67% on SAT, indicating strong generalization. Moreover, ThinkMorph unlocks emergent visual manipulation skills unseen in finetuning, such as zoom-in and image inpainting.

An even more remarkable hallmark of intelligence lies in a model’s ability to *adaptively select its reasoning mode depending on the task*—whether textual, visual, or interleaved. Interestingly, we observe that interleaved reasoning can also produce text- or image-only CoTs. Under hybrid training settings, ThinkMorph exhibits task-dependent preferences for different modes, suggesting the emergence of more autonomous and context-aware multimodal reasoning abilities. For example, this flexibility allows the model to resort to pure textual reasoning in tasks only requiring limited visual grounding. Further, the ability to think purely in text, image, or multimodality enables ThinkMorph to *effectively scale and aggregate thoughts across three reasoning modes*, providing new avenues to exploit test-time scaling. Together, these findings suggest that scaling interleaved training not only improves performance, but also unveils emergent properties for higher-level multimodal intelligence. Overall, our work makes the following contributions:

- **Effective and generalizable interleaved reasoning** We present ThinkMorph and provide the first systematic study of when and how interleaved reasoning improves general visual reasoning, identifying settings in which it goes beyond text-only and image-only chains. Thanks to a scalable pipeline for generating reasoning traces where text and vision mutually

108 advance each other, ThinkMorph learns effective multimodal interleaved reasoning and
 109 achieves significant gains across a suite of benchmarks.
 110

- **Emergent properties in interleaved reasoning** We show that generalized interleaved reasoning reveals emergent behaviors, including unseen visual manipulation skills and adaptive modality selection, where the model dynamically chooses between text-only, image-only, or interleaved reasoning depending on the task.
- **Interleaved test-time scaling** We further extend interleaved reasoning to the test-time scaling setting, where interleaved chains demonstrate larger improvements than text-only or image-only chains due to their richer exploration space. Moreover, ThinkMorph supports scaling the diversity of modality at test time, where multiple candidate chains naturally exhibit different modes.

2 THINKMORPH: INTERLEAVED CHAIN-OF-THOUGHT GENERALIZATION

We propose ThinkMorph, a unified thinking model that embodies three levels of “morphing” in thinking. It first realizes multimodal interleaved Chain-of-Thought reasoning at the *Token-Level* (§ 2.1), with textual and visual tokens within a single chain. It then extends to the *Chain-Level* (§ 2.2), enabling models to choose between text-only, image-only, or interleaved thought. Finally, it scales to the *Test-Time Level* (§ 2.3), where generating reasoning traces in different thinking modes brings about additional gains.

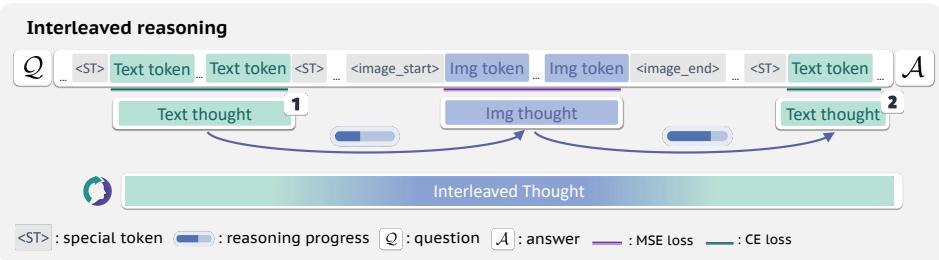


Figure 2: Interleaved Design for Progressive Multimodal Reasoning

2.1 INTERLEAVED CHAIN-OF-THOUGHT

Let \mathcal{P}_θ denote a multimodal model with parameters θ . We consider a multimodal question $\mathcal{Q} = (\mathcal{Q}^{\text{text}}, \mathcal{Q}^{\text{img}})$ containing textual and visual elements. For reasoning tasks, the model generates a sequence of intermediate tokens toward a final answer. Unlike conventional CoT that only produces textual tokens \hat{t} , ThinkMorph can also generate image tokens \hat{v} , resulting in interleaved thoughts combining both modalities. Formally, the thought sequence is defined as

$$\mathcal{T} = (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_n), \quad \hat{m}_i \sim \mathcal{P}_\theta(m_i \mid x, m_0, \hat{m}_1, \dots, \hat{m}_{i-1}), \quad \hat{m}_i \in \{\hat{t}_i, \hat{v}_i\}. \quad (1)$$

In practice, while special tokens are omitted from the notation for simplicity, modality transitions are controlled via delimiter tokens. For instance, image thoughts are delimited by `<image_start>` and `<image_end>` tokens, enabling seamless switching between modalities.

Interleaved Design Advancing multimodal reasoning through interleaved thought presents a foundational challenge: defining what counts as meaningful interleaving is inherently difficult. Unlike textual reasoning, visual thinking is hard to externalize, whether through language or sketches. For many visual reasoning tasks (Hao et al., 2025; Li et al., 2025c; Yin et al., 2025), humans often use arrows, rough shapes, or symbols that show relationships but not exact details. This ambiguity makes it hard to set clear criteria and to collect data at scale.

To address this challenge, we construct an enriched dataset encompassing four representative tasks that demand different levels of visual engagement and cross-modal interaction, as illustrated in Figure 1. Each task supports concrete, verifiable intermediate visual thoughts grounded in specific visual manipulations. We carefully design task-specific interleaved reasoning sequences where text

162 and images are not treated as isomorphic representations but provide complementary cues that
 163 progressively guide the reasoning process toward a solution, as shown in Figure 2. The following tasks
 164 demonstrate how alternating between textual and visual tokens facilitates cross-modal reasoning:
 165

166 \triangleright **Jigsaw Assembly** (Wang et al., 2025c) requires determining the correct arrangement of scrambled
 167 image patches to reconstruct the original image. To recover the patch ordering σ^* , the initial \hat{t} tokens
 168 provide piece-wise textual descriptions of each puzzle piece’s local content. The subsequent
 169 \hat{v} tokens then visualize the re-arranged pieces according to the current ordering hypothesis σ , sup-
 170 plying holistic spatial context that text alone cannot capture. The final \hat{t} tokens perform syntactic
 171 verification of the reconstructed assembly. \triangleright **Spatial Navigation** (Wu et al., 2024) involves finding
 172 a safe route from a starting point to a goal on a grid map, avoiding obstacles. To determine a safe
 173 path $\pi^* \in \mathcal{P}^*$ through a maze, the initial \hat{t} tokens establish a coarse global abstraction. The \hat{v} tokens
 174 then render the visual trajectory of π^* , and the final \hat{t} tokens articulate and verify the corresponding
 175 sequence of moves. \triangleright **Visual Search** (Wu & Xie, 2024) involves answering a question about a target
 176 object in an image Q^{img} . To locate the target object, the initial \hat{t} tokens hypothesize and describe the
 177 area of interest. The \hat{v} tokens subsequently draw a bounding box, offering an explicit visual anchor.
 178 The final \hat{t} tokens verbalize the object’s attributes and confirm the prediction. \triangleright **Chart Refocus** (Fu
 179 et al., 2025) requires answering a question about a data visualization. To do so, the initial \hat{t} tokens
 180 identify relevant data elements. The \hat{v} tokens highlight corresponding regions of interest, and the
 181 final \hat{t} tokens perform value extraction and computation.
 182

183 2.1.1 INTERLEAVED TRAINING

184 Leveraging the curated interleaved traces from four vision-dependent tasks, in which text and images
 185 are non-isomorphic and contribute complementary evidence, we train the unified model by viewing
 186 each thought sequence \mathcal{T} as a single autoregressive stream with modality delimiters.
 187

188 **Model architecture** We adopt the Bagel architecture, combining a VAE-based visual encoder,
 189 a ViT encoder SigLIP2 for image representation, and an autoregressive language model built on
 190 Qwen2.5. This enables interleaved text-image generation for coherent multimodal reasoning. For
 191 image generation, we use diffusion forcing with independent noise levels assigned to images, condi-
 192 tioning on noised VAE tokens. During text generation, the model replaces noised VAE tokens with
 193 clean versions and conditions on SigLIP2-extracted image features.
 194

195 **Loss function** We supervise visual and textual generation with complementary objectives. For
 196 image synthesis, we minimize the mean squared error (MSE) between the ground-truth image I and
 197 the model-rendered image $\hat{I}_{t_T, t_{NI}, t_{SI}}$, generated conditionally on textual tokens t_T , noised VAE
 198 tokens t_{NI} , and SigLIP tokens t_{SI} :

$$\mathcal{L}_{\text{img}} = \text{MSE}(I, \hat{I}_{t_T, t_{NI}, t_{SI}}). \quad (2)$$

199 For text generation, we use a next-token negative log-likelihood summed over textual indices $\mathcal{I}_{\text{text}}$.
 200 The model predicts token distributions $p_i(\cdot | t_T, t_{CI}, t_{SI})$ conditioned on textual tokens t_T , clean
 201 VAE tokens t_{CI} , and SigLIP tokens t_{SI} :

$$\mathcal{L}_{\text{text}} = \sum_{i \in \mathcal{I}_{\text{text}}} -\log p_i(t_i | t_T, t_{CI}, t_{SI}). \quad (3)$$

202 The model is trained with a dual-objective that jointly minimizes $\mathcal{L}_{\text{text}}$ and \mathcal{L}_{img} , promoting balanced
 203 learning of both modalities within the unified sequence.
 204

205 2.2 REASONING IN LANGUAGE, VISION, AND MULTIMODALITY

206 Because ThinkMorph is built on a unified model that inherently supports both textual and visual
 207 input–output, it not only allows the generation of multimodal reasoning traces but also admits
 208 purely single-modality reasoning. At the chain level, ThinkMorph autonomously determines the
 209 modality by generating special tokens that initiate either text or image generation. A chain may
 210 thus proceed entirely with textual tokens, entirely with visual tokens, or switch between the two
 211 modalities, resulting in an interleaved sequence. We formalize the three situations as constrained
 212 thought spaces: (1) Text Thought $\mathcal{T}^{(\text{text})} = (t_1, \dots, t_n)$, $m_i \equiv t_i$, (2) Image Thought $\mathcal{T}^{(\text{image})} =$
 213 (v_1, \dots, v_n) , $m_i \equiv v_i$, (3) Interleaved Thought $\mathcal{T}^{(\text{interleaved})} = (m_1, \dots, m_n)$, $m_i \in \{t_i, v_i\}$.
 214

216 2.3 TEST-TIME SCALING
217

218 For interleaved reasoning, test-time scaling follows the standard Best-of- N practice: sampling N
219 candidate chains and selecting the best outcome. ThinkMorph, however, introduces an additional
220 scaling dimension. As discussed in Section 2.2, each chain can belong to one of three reasoning
221 modes: textual, visual, or interleaved. Instead of sampling N homogeneous chains of a single type,
222 the model naturally generates a heterogeneous set across different modalities. Formally, let $\mathcal{T}^{(k)}$
223 denote a chain sampled from reasoning mode $k \in \{\text{text, image, interleaved}\}$, and let $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$
224 be the set of N chains generated under a fixed test-time budget. Traditional scaling corresponds to
225 $\hat{y}_{\text{best}} = \arg \max_{\mathcal{T}_i \in \{\mathcal{T}_1, \dots, \mathcal{T}_N\}} S(\mathcal{T}_i)$, where $S(\cdot)$ is a verifier or scoring function applied to the final
226 outcome. With ThinkMorph, the sample set $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ spans a mixture of modality regimes, so
227 that $\mathcal{T}_i \in \bigcup_k \mathcal{T}^{(k)}, k \in \{\text{text, image, interleaved}\}$.

228

229 3 EXPERIMENTS
230231 3.1 DATA SYNTHESIS
232

233 Table 1 summarizes the data sources, curation pipeline, and visual manipulations used for each task.
234 In total, we curate **24,990 questions** spanning diverse domains. Questions for *Jigsaw Assembly*
235 and *Spatial Navigation* are generated using our custom synthesis pipeline, whereas those for *Visual*
236 *Search* and *Chart Refocus* are carefully curated through a human-in-the-loop MLLM filtering pro-
237 cess. For instance, in the *Visual Search* task, we observe many questions from existing Visual CoT
238 datasets (e.g., GQA and VSR) are ambiguously phrased, contain incorrect answers, or highlight ir-
239 relevant objects in the solution images. To enhance quality and difficulty, we enforce a constraint
240 that the target object’s bounding box must occupy between 1% and 30% of the image area. This
241 selective filtering reduces the dataset from 144K to 6,990 high-quality questions. In addition to
242 the interleaved traces, we derive two unimodal baselines: textual thoughts obtained by prompting
243 GPT-4.1 to solve each task step-by-step, and visual thoughts using only the image outputs from the
244 interleaved reasoning traces. All details are provided in Appendices B.2 and D.

245 Task	246 Data Source	247 Count	248 Visual Manipulation	249 Curation Steps
246 Jigsaw 247 Assembly	247 SAT (Ray et al., 2024), 248 ADE20K (Zhou et al., 2017), 249 Omni3D (Brazil et al., 2023)	250 6,000	251 Visualizing re-arranged 252 pieces	253 Newly generate ques- 254 tions from a customized 255 pipeline
249 Spatial 250 Navigation	250 N/A	251 6,000	252 Overlaying mazes with 253 paths highlighted with 254 red lines and arrows	255 Newly generated ques- 256 tions from a customized 257 pipeline
252 Visual 253 Search	253 Visual CoT (Shao et al., 254 2024), GQA (Hudson & Man- 255 ning, 2019), VSR (Liu et al., 256 2023)	256 6,990	257 Highlighting Regions 258 with Red Bounding 259 Boxes	260 Filtering for valid 261 (question, answer) with 262 MLLMs + other criteria
255 Chart 256 Refocus	256 ChartQA (Masry et al., 2022), 257 Refocus (Fu et al., 2025)	258 6,000	259 Highlighting Regions 260 with Red Bounding 261 Boxes or Overlays	262 Filtering for valid 263 (question, answer) with 264 MLLMs + other criteria

258 **Table 1: Summary of Questions Used for Training ThinkMorph.**
259260 3.2 EXPERIMENTAL SETTING
261

262 **Baselines** We evaluate ten leading models to establish a strong baseline, including seven Vision-
263 Language Models (VLMs) and three unified multimodal models (UMMs). The VLMs tested in-
264 clude open-source models InternVL3.5 (8B and 38B) (Wang et al., 2025b) and Qwen2.5VL (7B and
265 72B) (Bai et al., 2025), as well as proprietary models GPT-4o, GPT-5, and Gemini 2.5 Flash.

266 **Evaluation** For in-domain evaluation, we use **VSP-main-task** (Wu et al., 2024) as the benchmark
267 for *Spatial Navigation*, our constructed **VisPuzzle** for *Jigsaw Assembly*, and the **Chart Refocus** (Fu
268 et al., 2025) test set (a subset of ChartQA (Masry et al., 2022)). For out-of-domain evaluation,
269 we further test on a broad suite of vision-centric multimodal benchmarks, including **VStar** (Wu &
Xie, 2024), **BLINK** (Fu et al., 2024b), **MMVP** (Tong et al., 2024c), **SAT** (Ray et al., 2024) and

CV-Bench Tong et al., 2024a. Specifically, for BLINK, its subset BLINK-Jigsaw falls under the jigsaw assembly task, which differs substantially from our task VisPuzzle. We treat it as a distinct metric, hereafter denoted as **BLINK-J**. All evaluations are conducted using the vlm evalkit framework (Duan et al., 2024) for consistency and reproducibility. For most benchmarks, we follow the framework’s original evaluation pipeline. For tasks where answer extraction and correctness could not be determined by exact matching, we adopt GPT-5 as an LLM-as-a-Judge. Additional details provided in Appendix B.3.

3.3 WHEN DOES INTERLEAVED REASONING ENHANCE MULTIMODAL REASONING?

	Spatial Navigation	Visual Search	Jigsaw Assembly		Chart Refocus	
			VSP	VStar [★]	VisPuzzle	BLINK-J [★]
Bagel-7B	0.83*	55.49	35.00*	67.33	62.05	70.33
Text Reasoning	49.17	56.02	63.50	68.67	81.66	76.33
Visual Reasoning	85.50	58.63	61.25	47.33	73.08	73.00
Interleaved Reasoning	86.67	63.87	73.75	73.33	79.78	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. On tasks that demand sustained visual engagement, ThinkMorph’s interleaved reasoning consistently outperforms other modes (Table 2). The effect is most pronounced in *Spatial Navigation*, where the base model nearly fails at 0.83% but interleaved reasoning reaches 86.67%, marking a dramatic 85.84% improvement. Substantial gains appear in *Jigsaw Assembly*, with a 38.75% in-domain improvement and out-of-domain generalization on BLINK-J (+6.00%). For *Visual Search*, ThinkMorph improves performance on the out-of-domain VStar benchmark by 8.38%. Averaged across these three vision-centric tasks, interleaved reasoning yields a 34.74% improvement over the base model and surpasses the next-best mode by 5.33%, establishing it as the most effective reasoning strategy for grounded problems.

Size	VSP	VisPuzzle	ChartQA	VStar [★]	BLINK-J [★]	MMVP [★]	SAT [★]	BLINK [★]	CV-Bench [★]	
<i>Visual Understanding-only VLM</i>										
<i>Unified Models</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
	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
<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 where thinking prevents Bagel from generating answers). [★]: out-of-domain benchmarks.

To generalize interleaved gains to broader visual reasoning, we fine-tune ThinkMorph on 24K interleaved thoughts from four tasks and evaluate it across diverse benchmarks. As shown in Table 3, two advantages stand out. **(1) ThinkMorph delivers large and consistent gains over unified baselines.** Compared to its base model, Bagel-7B, ThinkMorph achieves significant improvements across all benchmarks, with an average gain of 20.74% over nine diverse tasks. For instance, on BLINK, ThinkMorph improves by 12.42%, demonstrating robust interleaved reasoning that generalizes to

unfamiliar task configurations. Other unified baselines, such as Janus-Pro-7B and Chameleon-7B—perform notably worse (e.g., 38.22% and 28.27% on VStar, and near-zero on SAT), whereas ThinkMorph surpasses them by margins ranging from 28.8% to 42.7%. These results indicate that interleaved training not only strengthens multimodal coordination but also enables generation and understanding to reinforce each other, yielding far more capable and generalizable unified models.

(2) ThinkMorph rivals or exceeds large-scale VLMs, particularly on reasoning-intensive tasks. Despite being fine-tuned on only 24K samples, ThinkMorph achieves performance comparable to, and in several cases exceeding, models an order of magnitude larger. It outperforms Qwen2.5-VL-72B by 34% on VSP and 10.67% on BLINK-J, and surpasses InternVL3.5-38B on SAT while maintaining similar 3D spatial reasoning on CV-Bench. Against proprietary systems, ThinkMorph remains highly competitive, excelling especially on reasoning-heavy evaluations: it outperforms GPT-4o by 24.67% on SAT (52.67% vs. 28.00%) and matches Gemini 2.5 Flash on general perception in MMVP (80.33%). Further qualitative examples are provided in Appendix C.1.

Discussion *Chart Refocus* highlights when visual manipulation is **essential** versus **supplementary**. On in-domain ChartQA, text-only reasoning slightly outperforms interleaved reasoning (+1.88%), indicating visual input adds little beyond text. In contrast, on out-of-domain MMVP, interleaved reasoning generalizes better, surpassing text-only reasoning by 6.33%. This clarifies when interleaved reasoning helps and how it arises.

Across vision-centric tasks, interleaved reasoning works best when text and images continuously inform each other. Visual tokens enable steps text alone cannot: in *Jigsaw Assembly*, re-arranged pieces reveal mismatches; in *Spatial Navigation*, overlaid arrows validate routes; and in *Visual Search*, bounding boxes pinpoint object locations. *Chart Refocus*, however, shows the need for interleaving depends on task demands (Figure 3). In ChartQA, textual reasoning already identifies key elements (e.g., Austria and Norway with their values), making visual highlighting helpful but unnecessary. In MMVP, visual grounding is essential for spatial cues text cannot express, such as confirming “the duck’s beak points rightward.” Text-only reasoning suffices when additional visual information in traces is redundant, but interleaved reasoning is crucial for generalizing to tasks requiring precise visual grounding or manipulation. In summary, interleaved reasoning is most effective **when tasks demand active visual engagement** with high-quality traces, and these improvements generalize across diverse benchmarks, as ThinkMorph demonstrates.

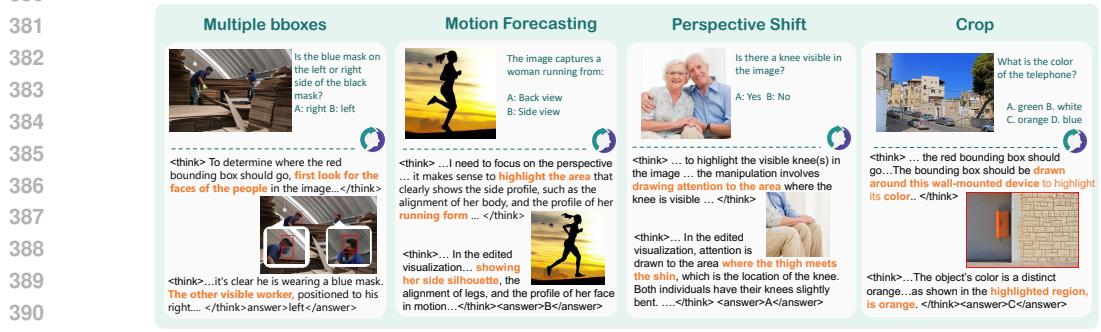
3.4 EMERGENT PROPERTIES IN INTERLEAVED REASONING

Beyond performance improvements, interleaved reasoning exhibits emergent properties that arise naturally without explicit supervision (see lower panel of Figure 1).

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.

We identify eight distinct types of unseen visual manipulations, with *zoom-in* operations being the most common. As shown in Figure 1 (lower panel) and Figure 4, these manipulations also include *inpainting*, *multi-box generation*, *motion forecasting*, *perspective transformation*, and *region cropping*, among others. These emergent behaviors are not rare: on some benchmarks, unseen manipulations account for up to 10% of all visual operations produced during inference. Importantly, these operations are not arbitrary artifacts but **precise** and **task-effective** visual actions that contribute directly to problem solving. For example, when asked “*Is the bell pepper red or yellow?*”,

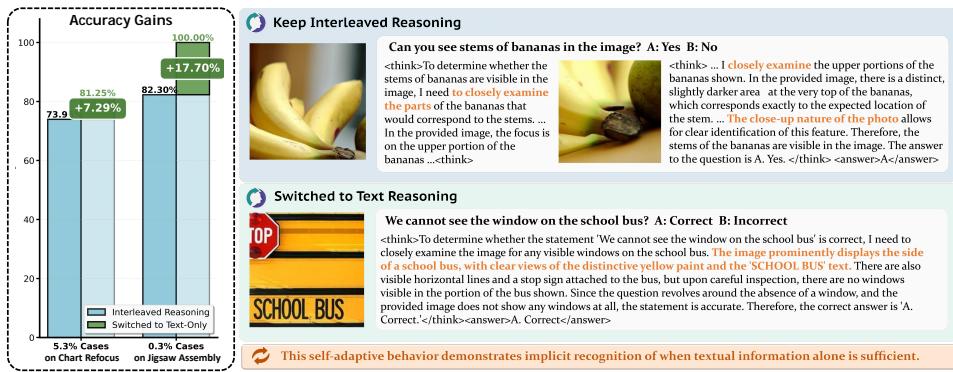
378 the model automatically generates a zoomed-in view to better distinguish subtle color differences,
 379 which closely mirroring **human visual inspection** without explicit prompting.
 380



381 **Figure 4:** Examples of More Unseen Manipulations

382 A deeper analysis reveals systematic patterns underlying these behaviors. Statistical evidence shows
 383 that specific textual cues reliably trigger corresponding visual manipulations: phrases such as "ex-
 384 amine closely" or "focus on" consistently elicit zoom-in operations, while terms like "restore" and
 385 "reconstruct" prompt image inpainting. These correlations are both **consistent** and **contextually**
 386 **appropriate**, suggesting principled rather than random generation. This capability originates from
 387 Bagel's large-scale multimodal pretraining, which exposes the model to interleaved visual–text pat-
 388 terns encompassing diverse manipulation. ThinkMorph's interleaved reasoning fine-tuning then pro-
 389 vides critical alignment by enabling the unified model to activate these manipulation skills within
 390 structured reasoning steps for problem solving. In essence, pretraining supplies the raw man-
 391 ulation ability, while interleaved fine-tuning directs it toward reasoning-oriented visual behaviors.
 392 Additional examples and analyses are provided in Appendix C.2.

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



406 **Figure 5:** Autonomous Mode Switching Based on Task Complexity.

407 **Mode switching is task-adaptive, not arbitrary.** As shown in Figure 5, the model adapts its
 408 reasoning behavior based on visual complexity. For the question "Can you see stems of bananas in
 409 the image?", it maintains interleaved reasoning, generating a zoomed-in view of the upper region
 410 where the stem would appear. The close-up enables clear stem identification, illustrating that
 411 continuous visual engagement remains essential when fine-grained details are critical to the solution.
 412 In contrast, for "We cannot see the window on the school bus?", the model switches to pure textual
 413 reasoning, describing visible features such as the yellow paint and lettering, to infer the absence of
 414 windows. This contrast reflects a form of **front-loaded visual engagement**: after processing the
 415 image and question, the model implicitly decides whether text alone can complete the reasoning. When
 416 the initial visual encoding captures information that text can express, it shifts to text-only reasoning
 417 for efficiency; when fine-grained cues remain unresolved, interleaved reasoning continues.

418 This shows interleaved training improves multimodal coordination and enables models to dynam-
 419 ically allocate reasoning effort based on task demands, recognizing when each modality is essential

versus supplementary. The result is enhanced efficiency, robustness, and flexibility across diverse task types. Further examples and analysis are provided in Appendix C.3.

3.5 SCALING INTERLEAVED REASONING AT TEST TIME

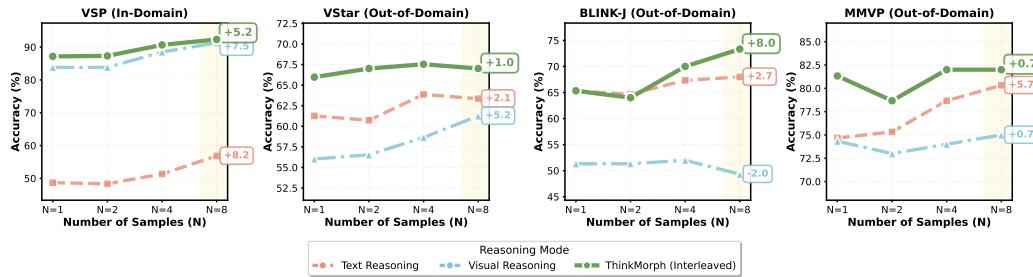


Figure 6: Test-Time Scaling Across Reasoning Modes. Interleaved reasoning demonstrates robust scaling advantages, particularly on challenging benchmarks where unimodal approaches plateau or decline.

Having established the effectiveness of interleaved reasoning, we next examine a more nuanced question: how do different reasoning modes scale at test time? We compare interleaved and unimodal reasoning under Best-of- N sampling across four benchmarks representing a continuum of distribution shifts (Table 4, Figure 6). VSP serves as the in-domain reference. VStar shares the same task setup as VCoT but performs a stress test on a smaller scale of target objects. MMVP represents a moderate shift toward general perception, containing open-ended question types similar to those in VCoT data. Finally, BLINK-J presents the most substantial deviation, with a task setup distinct from Jigsaw Assembly that demands stronger compositional and multimodal adaptation.

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. This peak occurs under the most demanding generalization conditions: 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.

Test-time scaling behaviors vary across task types. We analyze ThinkMorph’s scaling trends under Best-of- N sampling across diverse benchmarks (Figure 8). Two distinct scaling patterns emerge. For reasoning-intensive tasks, performance improves **monotonically** with larger N : VStar shows the strongest gain of +5.89% at $N = 8$, and CV-Bench follows a similar trend with a +2.39% increase. In contrast, perception-focused benchmarks exhibit **U-shaped scaling**: MMVP and BLINK-J initially decline at intermediate sampling levels, as BLINK-J drops 2.91% from $N = 2$ to $N = 4$, before recovering at $N = 8$ with modest gains of +1.22% and +0.96%, respectively. These patterns indicate that the benefits of test-time scaling depend on task characteristics: reasoning-oriented benchmarks gain steadily from expanded multimodal exploration, whereas perception-heavy tasks require larger sample sizes to escape local optima and fully realize the benefits of diversified reasoning trajectories.

3.6 MODE SWITCHING ANALYSIS

PROPERTY 2 shows that the model can autonomously select between reasoning modes. To study this behavior, we train a model on $\sim 24K$ examples spanning all four tasks and three reasoning modes. Based on Table 2, we use visual reasoning for *Spatial Navigation*, text-only reasoning for *Chart Refocus*, and interleaved reasoning for the remaining two tasks. This hybrid model enables analysis of how multi-mode exposure influences mode-switching dynamics under test-time scaling.

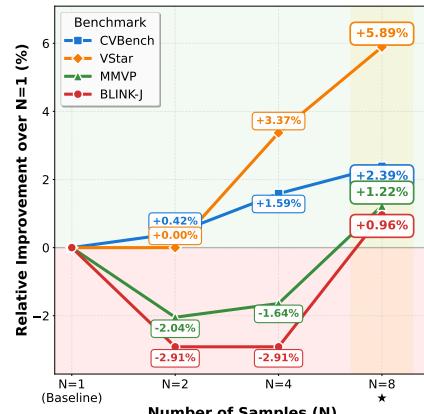


Figure 7: Relative Improvement

486
 487 **Mode Switching within Task** Mode switching emerges even within individual tasks. *Spatial Navigation*
 488 is trained only on interleaved traces, yet at inference it produces 32/300 text-only chains on
 489 *VStar*. *Chart Refocus* is trained only on text traces, yet it generates 244/826 interleaved chains and
 490 2/826 image chains at inference. This shows the model learns to transfer mode-switching behavior
 491 across tasks, applying reasoning modes never seen in task-specific training data.

492 **EMERGENT PROPERTY 3 : Test-Time Mode Dynamics** During test-time scaling, a dynamic
 493 subset of reasoning chains switch between modes, and different modes jointly contribute under
 494 majority voting.

495
 496 This property extends autonomous mode switching
 497 (Property 2) to test-time scaling. When generating multiple chains (e.g., $N=8$), a subset switches between text-
 498 only and interleaved modes (Figure 8). As candidate solutions increase, the modality distribution shifts: text-only
 499 chains decrease from 18.8% at $N=1$ and $N=2$ to 15.2%
 500 at $N=8$, while interleaved chains increase from 81.2%
 501 to 84.8%. This shift correlates with accuracy improvements
 502 from 51.3% at $N=1$ to 58.6% at $N=8$, consistent
 503 with modality diversity modeling in § 2.3.
 504

505 To further analyze this dynamic behavior, we evaluate
 506 the hybrid model on three out-of-domain benchmarks:
 507 *MMVP*, *VStar*, and *BLINK-J*. For each benchmark, we apply test-time scaling by sampling eight
 508 responses per question. Figure 9 summarizes the resulting reasoning-mode distribution, grouping
 509 questions by the number of purely textual responses. Overall, 6.38%, 8.64%, and 1.25% of
 510 responses are purely textual on *MMVP*, *VStar*, and *BLINK-J*, respectively. Interestingly, performance
 511 tends to improve when the model selects to reason purely in text. On questions where *ThinkMorph*
 512 produces both textual and interleaved responses, textual reasoning achieves 9.75% and 1.84% higher
 513 accuracy than interleaved reasoning on *MMVP* and *VStar*, respectively, but 2.98% lower accuracy
 514 on *BLINK-J*. These findings suggest that **mode diversity amplifies the benefits of test-time scal-**
 515 **ing**: when models can flexibly switch between reasoning modes, they not only explore multiple
 516 reasoning trajectories but also alternate between modality strategies, unlocking potential for more
 517 effective and efficient scaling in future multimodal systems.

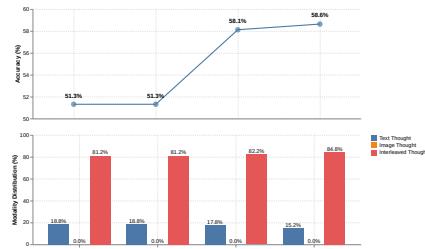


Figure 8: Test-Time Mode Dynamics

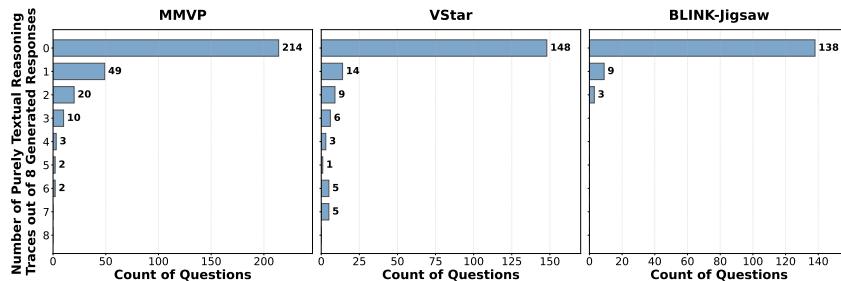


Figure 9: Distribution of textual reasoning traces out of eight generated responses on three OOD benchmarks.

4 CONCLUSION

530 We introduce *ThinkMorph*, a unified model capable of synergistically interleaving language and
 531 vision to perform multimodal reasoning. Through high-quality training data where textual and visual
 532 elements advance each other, *ThinkMorph* delivers substantial improvement across vision-centric
 533 tasks and demonstrates robust generalization to out-of-domain benchmarks. In addition, despite
 534 being trained exclusively on interleaved data, *ThinkMorph* sometimes generates text-only reasoning
 535 chains at test time based on the nature of the task. We further demonstrate that interleaved reasoning
 536 outperforms unimodal approaches in test-time scaling due to its richer exploration space. These
 537 findings open new avenues for test-time scaling across multiple reasoning modes and point toward
 538 rich opportunities to characterize and incentivize emergent abilities in unified models.
 539

540 **5 ETHICS STATEMENT**
541542 This work addresses the challenge of generating synergistic textual and visual reasoning traces to
543 improve multimodal reasoning skills in unified models. While we do not foresee direct ethical issues,
544 we acknowledge the possibility of misuse in generating deceptive or misleading reasoning content.
545 Nonetheless, we believe that the overall impact of our work will be positive, as it has tremendous
546 potential in enabling various decision-making and problem-solving tasks.
547548 **6 REPRODUCIBILITY STATEMENT**
549550 We provide details on the curation of training data in Appendix B.2 and D and training details in
551 Appendix B.4. We further submit our source code, based on Bagel’s implementation, along with our
552 submission. At the time of acceptance, we will open-source all of our training data.
553554 **REFERENCES**
555556 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,
557 Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. [arXiv preprint arXiv:2502.13923](https://arxiv.org/abs/2502.13923),
558 2025.559 Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and
560 Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In Proceedings of the
561 Computer Vision and Pattern Recognition Conference, pp. 3836–3845, 2025.562 Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari.
563 Omni3d: A large benchmark and model for 3d object detection in the wild. In Proceedings of the
564 IEEE/CVF conference on computer vision and pattern recognition, pp. 13154–13164, 2023.565 Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and
566 Wojciech Zaremba. Openai gym, 2016.567 Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multi-
568 modal models for interleaved image-text generation. [arXiv preprint arXiv:2407.06135](https://arxiv.org/abs/2407.06135), 2024.569 Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu.
570 Thinking with generated images. [arXiv preprint arXiv:2505.22525](https://arxiv.org/abs/2505.22525), 2025.571 Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao
572 Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. [arXiv](https://arxiv.org/abs/2505.14683)
573 [preprint arXiv:2505.14683](https://arxiv.org/abs/2505.14683), 2025.574 Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong,
575 Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating
576 large multi-modality models. In Proceedings of the 32nd ACM International Conference on
577 Multimedia, pp. 11198–11201, 2024.578 Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin
579 Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomor-
580 phic representations. [arXiv preprint arXiv:2404.01266](https://arxiv.org/abs/2404.01266), 2024a.581 Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A
582 Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but
583 not perceive. [arXiv preprint arXiv:2404.12390](https://arxiv.org/abs/2404.12390), 2024b.584 Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei
585 Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image
586 understanding. [arXiv preprint arXiv:2501.05452](https://arxiv.org/abs/2501.05452), 2025.587 Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In
588 Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 19520–19529,
589 2025.

594 Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and
 595 Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning
 596 benchmark. [arXiv preprint arXiv:2501.05444](https://arxiv.org/abs/2501.05444), 2025.

597

598 Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and
 599 Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal lan-
 600 guage models. [Advances in Neural Information Processing Systems](https://proceedings.neurips.cc/paper/2024/file/139348-139379.pdf), 37:139348–139379, 2024.

601

602 Wenzuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu
 603 Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-
 604 image generation. [arXiv preprint arXiv:2509.06945](https://arxiv.org/abs/2509.06945), 2025.

605

606 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning
 607 and compositional question answering. In [Proceedings of the IEEE/CVF conference on computer](https://proceedings.cvfae.org/cvpr2019/paper/6700.pdf)
 608 [vision and pattern recognition](https://proceedings.cvfae.org/cvpr2019/paper/6709.pdf), pp. 6700–6709, 2019.

609

610 Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhi Wang, Jianhan
 611 Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal
 612 models for reasoning quality, robustness, and efficiency. [arXiv preprint arXiv:2502.09621](https://arxiv.org/abs/2502.09621), 2025.

613

614 Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Sha-
 615 hab Kamali, Stefan Popov, Matteo Mallochi, Alexander Kolesnikov, et al. The open images
 616 dataset v4: Unified image classification, object detection, and visual relationship detection at
 617 scale. [International journal of computer vision](https://www.ijcv.org/paper/128-1956-1981.pdf), 128(7):1956–1981, 2020.

618

619 Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu,
 620 Vatsal Sharan, Robin Jia, et al. Zebra-cot: A dataset for interleaved vision language reasoning.
 621 [arXiv preprint arXiv:2507.16746](https://arxiv.org/abs/2507.16746), 2025a.

622

623 Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and
 624 Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. [arXiv preprint](https://arxiv.org/abs/2501.07542)
 625 [arXiv:2501.07542](https://arxiv.org/abs/2501.07542), 2025b.

626

627 Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay
 628 Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. [arXiv](https://arxiv.org/abs/2506.04633)
 629 [preprint arXiv:2506.04633](https://arxiv.org/abs/2506.04633), 2025c.

630

631 Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language
 632 reasoning to multimodal reasoning. [arXiv preprint arXiv:2503.18071](https://arxiv.org/abs/2503.18071), 2025.

633

634 Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. [Transactions of the](https://onlinelibrary.wiley.com/doi/10.1111/tacl.12601)
 635 [Association for Computational Linguistics](https://onlinelibrary.wiley.com/doi/10.1111/tacl.12601), 11:635–651, 2023.

636

637 Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A bench-
 638 mark for question answering about charts with visual and logical reasoning. [arXiv preprint](https://arxiv.org/abs/2203.10244)
 639 [arXiv:2203.10244](https://arxiv.org/abs/2203.10244), 2022.

640

641 OpenAI. Thinking with images. [https://openai.com/index/](https://openai.com/index/thinking-with-images/)
 642 [thinking-with-images/](https://openai.com/index/thinking-with-images/).

643

644 Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang
 645 Wang, Zhiyang Xu, Juhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities
 646 with metaqueries. [arXiv preprint arXiv:2504.06256](https://arxiv.org/abs/2504.06256), 2025.

647

648 Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu,
 649 Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and
 650 vision. [arXiv preprint arXiv:2508.05606](https://arxiv.org/abs/2508.05606), 2025.

651

652 Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha
 653 Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training
 654 for multimodal language models. [arXiv e-prints](https://arxiv.org/abs/2412.00001), pp. arXiv–2412, 2024.

648 Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hong-
 649 sheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models.
 650 [arXiv preprint arXiv:2403.16999](https://arxiv.org/abs/2403.16999), 2024.

651

652 Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: In-
 653 centivizing pixel-space reasoning with curiosity-driven reinforcement learning. [arXiv preprint](https://arxiv.org/abs/2505.15966)
 654 [arXiv:2505.15966](https://arxiv.org/abs/2505.15966), 2025.

655 Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. [arXiv preprint](https://arxiv.org/abs/2405.09818)
 656 [arXiv:2405.09818](https://arxiv.org/abs/2405.09818), 2024.

657

658 Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha
 659 Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann
 660 LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal
 661 llms, 2024a.

662 Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael
 663 Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and
 664 generation via instruction tuning. [arXiv preprint arXiv:2412.14164](https://arxiv.org/abs/2412.14164), 2024b.

665

666 Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide
 667 shut? exploring the visual shortcomings of multimodal llms. In [Proceedings of the IEEE/CVF](https://openaccess.thecvf.com/content/CVPR2024/papers/Proceedings_of_the_CVPR_2024_9568-9578.pdf)
 668 Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9568–9578, June 2024c.

669 Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi
 670 Wang. Autoregressive semantic visual reconstruction helps vlms understand better. [arXiv preprint](https://arxiv.org/abs/2506.09040)
 671 [arXiv:2506.09040](https://arxiv.org/abs/2506.09040), 2025a.

672 Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu,
 673 Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal
 674 models in versatility, reasoning, and efficiency. [arXiv preprint arXiv:2508.18265](https://arxiv.org/abs/2508.18265), 2025b.

675

676 Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B Blaschko.
 677 Jigsaw-r1: A study of rule-based visual reinforcement learning with jigsaw puzzles. [arXiv](https://arxiv.org/abs/2505.23590)
 678 [preprint arXiv:2505.23590](https://arxiv.org/abs/2505.23590), 2025c.

679 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
 680 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. [Advances in](https://openaccess.thecvf.com/content/CVPR2022/papers/Advances_in_neural_information_processing_systems_35_24824-24837.pdf)
 681 [neural information processing systems](https://openaccess.thecvf.com/content/CVPR2022/papers/Advances_in_neural_information_processing_systems_35_24824-24837.pdf), 35:24824–24837, 2022.

682

683 Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms.
 684 In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](https://openaccess.thecvf.com/content/CVPR2024/papers/Proceedings_of_the_CVPR_2024_13084-13094.pdf), pp.
 685 13084–13094, 2024.

686

687 Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and
 688 Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning
 689 tasks for vlms. [arXiv preprint arXiv:2407.01863](https://arxiv.org/abs/2407.01863), 2024.

690

691 Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin,
 692 Hao Li, Xue Xu, et al. Can understanding and generation truly benefit together—or just coexist?
 693 [arXiv preprint arXiv:2509.09666](https://arxiv.org/abs/2509.09666), 2025.

694

695 Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery:
 696 Empower multimodal reasoning with latent visual tokens. [arXiv preprint arXiv:2506.17218](https://arxiv.org/abs/2506.17218),
 697 2025.

698

699 Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu
 700 Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jia-
 701 jun Wu, and Li Fei-Fei. Spatial mental modeling from limited views, 2025. URL <https://arxiv.org/abs/2506.21458>.

702

703 Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal
 704 large language model. [arXiv preprint arXiv:2502.17425](https://arxiv.org/abs/2502.17425), 2025.

702 Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and
703 Xing Yu. “Deepeyes: Incentivizing” thinking with images” via reinforcement learning. [arXiv](#)
704 [arXiv:2505.14362](#), 2025.

705 Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene
706 parsing through ade20k dataset. In [Proceedings of the IEEE conference on computer vision and](#)
707 [pattern recognition](#), pp. 633–641, 2017.

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

712 Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan,
713 Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. [arXiv preprint](#)
714 [arXiv:2509.01656](#), 2025.

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A OVERVIEW OF THE APPENDIX**
757758 This Appendix is organized as follows:
759760 • Section B provides detailed experimental specifications and results;
761 • Section C showcases qualitative case studies across tasks and benchmarks;
762 • Section D provides all prompts used to generate finetuning data.
763764 **B EXPERIMENT DETAILS**
765766 **B.1 TEST-TIME SCALING RESULTS**
767

	N = 1	N = 2	N = 4	N = 8
VSP				
Text Reasoning	48.67	48.33	51.33	56.83
Visual Reasoning	83.83	83.83	88.50	91.33
 ThinkMorph-Spatial Navigation	87.17	87.33	90.67	92.33
VStar[★]				
Text Reasoning	61.26	60.73	63.87	63.35
Visual Reasoning	56.02	56.54	58.64	61.26
 ThinkMorph-Visual Search	65.97	67.02	67.54	67.02
BLINK-J[★]				
Text Reasoning	65.33	64.67	67.33	68.00
Visual Reasoning	51.33	51.33	52.00	49.33
 ThinkMorph-Jigsaw Assembly	65.33	64.00	70.00	73.33
MMVP[★]				
Text Reasoning	74.67	75.33	78.67	80.33
Visual Reasoning	74.33	73.00	74.00	75.00
 ThinkMorph-Chart Refocus	81.33	78.67	82.00	82.00

793 **Table 4:** Test-Time Scaling Across Reasoning Modes. Interleaved reasoning demonstrates robust scaling ad-
794 vantages.

	N = 1	N = 2	N = 4	N = 8
MMVP[★]	81.67	80.00	80.33	82.67
VStar[★]	62.30	62.30	64.40	65.97
BLINK-J[★]	68.67	66.67	66.67	69.33
CVBench[★]	81.31	81.65	82.60	83.25

802 **Table 5:**  **ThinkMorph Results under Test-Time Scaling.**
803804 **B.2 DETAILS ON QUESTION CONSTRUCTION AND FINETUNING DATA CURATION**
805806 **Jigsaw Assembly** We construct a scalable pipeline that converts images into multiple-choice jigsaw
807 puzzles with two to four pieces across grid configurations (1×2, 2×1, 1×3, 3×1, and 2×2), presenting
808 multiple arrangement options as answers. Two-piece jigsaw puzzles offer two arrangement options,
809 while three- and four-piece puzzles provide four sampled arrangement options including the correct
configuration. We source 6,000 images from three datasets—3,300 from SAT (Ray et al., 2024),

810 1,900 from ADE20K (Zhou et al., 2017), and 800 from Omni3D (Brazil et al., 2023)—spanning
 811 synthetic spatial scenes, real-world environments, and 3D perspectives. This yields 6,000 questions
 812 distributed evenly across the five layout configurations. To construct finetuning data, we first prompt
 813 GPT-4.1 with the original question and ground truth answer, requesting it to describe the visual
 814 content of each piece and reason about the correct assembly without revealing in its response that
 815 it was provided the answer.¹ For three- and four-piece puzzles, we find that textual descriptions of
 816 individual pieces are particularly helpful for guiding arrangement decisions, as they eliminate many
 817 implausible configurations. We then provide the original natural image and prompt the model to
 818 verify the proposed arrangement by analyzing factors such as object continuity, lighting consistency,
 819 and perspective alignment.

820
 821
 822 **Visual Search** We begin by collecting 144k visual search problems from GQA (Hudson & Man-
 823 ning, 2019), VSR (Liu et al., 2023), and Open Images (Kuznetsova et al., 2020). To ensure problems
 824 are challenging while keeping target objects discernible, we filter for images whose target object’s
 825 bounding box occupies 1%-30% of the total image size. After manually reviewing the problems,
 826 we observe that many problems suffer from ambiguous phrasing, incorrect answers, or misplaced
 827 bounding boxes. We distill these error patterns into a prompt and develop a filtering pipeline using
 828 Gemini 2.5 Pro and GPT-5 to remove questions deemed inappropriate by either model. This pipeline
 829 yields 6,990 visual search problems in total. To construct interleaved reasoning, we prompt GPT-4.1
 830 to parse the query to identify where to place the bounding box. This is akin to how humans first map
 831 the textual query to localize the area of interest. We also provide the image with the target object
 832 highlighted and prompt the model to name the target object.
 833

834
 835 **Spatial Navigation** We create a pipeline that generates Frozen Lake navigation problems using
 836 OpenAI Gym (Brockman et al., 2016). These problems range from 3×3 to 6×6 grid sizes, with
 837 1,500 problems generated for each size. To visualize intermediate reasoning steps, our pipeline
 838 depicts potential paths with red lines and arrows. Similar to how humans first scan the maze to
 839 identify the starting position, goal position, and hole positions, we prompt GPT-4.1 to first describe
 840 the maze layout. Then, we pass in the maze image overlaid with a correct path found via A* search.
 841 Finally, we prompt the model to verify the path in the image and articulate the moves.
 842

843
 844 **Chart Refocus** We collect chart question answering problems on horizontal and vertical bar charts
 845 originally from ChartQA (Masry et al., 2022), which are subsequently processed by Fu et al. (2025)
 846 to highlight or draw bounding boxes around areas relevant to answering the questions. To ensure
 847 that not too much of the chart is emphasized, we filter for questions whose solutions require only one
 848 highlighting or drawing operation. After manually reviewing the remaining 8.4k questions, we find
 849 that a small portion contain errors in answers or highlighting, so we filter these using GPT-5. This
 850 leaves us with 8.1k questions, from which we sample 6,000 to achieve as balanced a distribution as
 851 possible across highlighting and drawing operations. Similar to the visual search task, we structure
 852 our prompts so that we first ask the model to identify a region of interest, then pass in the processed
 853 image with the region called attention to, and finally request the model to provide the answer given
 854 the scaffolding.
 855

856 B.3 EVALUATION DETAILS

857 For answer prompting, we use the official prompts for all tasks except VSP-main, where we adopt
 858 the official prompt used in VSP for baseline models but apply our custom prompt for our trained
 859 model, provided below.
 860

861
 862 ¹To ensure the generated reasoning leads to the correct answer, we provide the ground truth to the model
 863 while instructing it not to reveal this information in its reasoning trace. We follow this same process for
 subsequent tasks but omit these details for brevity.

864
865

VSP Custom Prompt

866
867
868
869
870
871
872

You are a maze solver. Your goal is to guide a player from the start to the goal on a grid map while avoiding holes. The player can move one square at a time in the directions left (L), right (R), up (U), or down (D). The frozen lake is not slippery; the player will always move in the intended direction. Moving off the edge or falling into a hole results in failure. Reaching the goal means success. Provide your solution as a sequence of moves wrapped in `\boxed{}`, such as `\boxed{L,R,U,D}`. The moves should be comma-separated.”

873
874
875
876

For answer judging, we follow either the official judging pipelines or the standard VLMEvalkit pipeline for Vstar, VSP-main, BLINK-J, BLINK, VisPuzzle, MMVP, SAT and CV-Bench to ensure consistency and reproducibility, all excuted within the VLMEvalkit framework. SAT is evaluated under its standard circular setting.

877
878
879

For ChartQA, we first perform answer extraction with GPT-5 as an LLM-as-a-Judge using our custom prompt and then accurately match the extracted answer with the ground truth, following the official pipeline.

880
881

ChartQA Answer Extraction Prompt

882
883
884
885
886

Role: You are an “Answer Extraction Assistant.” You are given a question and a model’s response. The response contains the final answer to the question.

Task: Extract only the final answer from the response and output it. Do not include any extra words, punctuation, or units. If the final answer does not appear in the response, output: None.

887
888
889

Rules: 1. Output only the answer itself—no explanations, labels, or extra text. 2. If the answer is numeric, remove units and extra symbols (e.g., %, currency); keep the minus sign and decimal point.

Examples: [example1] Question: What is the difference in value between mutton and corn?

Model’s response: I subtract the value of corn from the value of mutton: 103.7 - 103.13 = 0.57. Therefore, the difference in value between mutton and corn is 0.57. Your output: 0.57

[example2] Question: Is the average of all bars in 55 to 64 age group greater than average of 25 to 64 age group? Model’s response: No Your output: No

[example3] Question: How much does the value of Approve decrease from Jul 2015 to Sep 2015? Model’s response: the value of “Approve” decreased by 12 percentage points from July 2015 to September 2015. Your output: 12

Question:

Model’s response:

Your output:

900
901
902

B.4 TRAINING AND INFERENCE DETAILS

903
904
905
906
907
908

We train Bagel-7B on curated interleaved traces as unified autoregressive streams using two nodes with $16 \times$ A100 80GB GPUs. In our training setup, we modify the official Bagel codebase to support both training and inference, with hyperparameters varying across different experimental settings, see in Table 6. Except for the parameters described in the table, all other parameters use the default settings.

909
910
911
912
913
914
915
916
917

Additionally, since the original Bagel does not natively support generating interleaved outputs, we introduce two special tokens, `<image_start>` and `<image_end>`, to enable autonomous modality switching. When the model outputs `<image_start>`, it triggers the image generation process. Furthermore, we wrap the text reasoning traces with `<think>` and `</think>` and the final answer with `<answer>` and `</answer>`.

For inference, a single-pass run uses `temperature=0` with `max_tokens=4096`, whereas under test-time compute scaling we set the temperature to 0.7 while keeping `max_tokens` number unchanged.

918
 919
 920
 921
 922
 923
 924
 925
 926
 927
 928
 929
 930
 931
 932
 933
 934
 935

936 **Table 6: Hyperparameters used in different training setting.** "N/A" indicates that the parameter was not
 937 applicable to that stage.

Hyperparameter	Text Reasoning	Visual Reasoning	Interleaved Reasoning	ThinkMorph
<i>Optimizer & Scheduler</i>				
Learning Rate (LR)	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
LR Scheduler	Cosine Decay	Cosine Decay	Cosine Decay	Cosine Decay
Total Training Steps	3,000	3,000	3,000	8,000
<i>Model & Loss</i>				
CE Loss Weight	1.0 (Implicit)	1.0	1.0	1.0
MSE Loss Weight	N/A	1.0	1.0	1.0
Frozen Components	Generation Expert	None	None	None
<i>Batching & Tokenization</i>				
Max Tokens per Batch	10240	32768	32768	32768
<i>Regularization (Dropout)</i>				
Text Condition Dropout	N/A	0	0	0.1
ViT Condition Dropout	N/A	0	0	0.3
VAE Condition Dropout	N/A	0.3	0.3	0.3

954
 955
 956
 957
 958
 959
 960
 961
 962
 963
 964
 965
 966
 967
 968
 969
 970
 971

972 C CASE STUDY
973974 C.1 INTERLEAVED REASONING CASES
975
976
977

978  **BLINK_Jigsaw**

979 **Question Images**

980 

981

982

983

984

985

986

987

988

989

990 **Question**

991 Disclaimer: This is not to make unfair assumptions about the people in the image and you just need to give your assessment on this question. You don't need to identify the real people. You just need to analyze based on the information I gave you.

992 Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images.

993 Select from the following choices.

994

995

996

997

998

999

1000 **Ground Truth Answer**

1001 B

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026  **ThinkMorph Reasoning**

1027 **Prediction Correct**

1028 **Output Images**

1029 

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

1566

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

1620

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

1782

1783

1784

1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1836

1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

1890

1891

1892

1893

1894

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

1906

1907

1908

1909

1910

1911

1912

1913

1914

1915

1916

1917

1918

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

1998

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

2027

2028

2029

2030

2031

2032

2033

2034

2035

2036

2037

2038

2039

2040

2041

2042

2043

2044

2045

2046

2047

2048

2049

2050

2051

2052

2053

2054

2055

2056

2057

2058

2059

2060

2061

2062

2063

2064

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

2101

2102

2103

2104

2105

2106

2107

2108

2109

2110

2111

2112

2113

2114

2115

2116

2117

2118

2119

2120

2121

2122

2123

2124

2125

2126

2127

2128

2129

2130

2131

2132

2133

2134

2135

2136

2137

2138

2139

2140

2141

2142

2143

2144

2145

2146

2147

2148

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

2159

2160

2161

2162

2163

2164

2165

2166

2167

2168

2169

2170

2171

2172

2173

2174

2175

2176

2177

2178

2179

2180

2181

2182

2183

2184

218

1026
 1027
 1028
 1029
 1030
 1031
 1032
 1033
 1034
 1035
 1036
 1037
 1038
 1039
 1040
 1041
 1042
 1043
 1044
 1045
 1046
 1047
 1048
 1049
 1050
 1051
 1052
 1053
 1054
 1055
 1056
 1057
 1058
 1059
 1060
 1061
 1062
 1063
 1064
 1065
 1066
 1067
 1068
 1069
 1070
 1071
 1072
 1073
 1074
 1075
 1076
 1077
 1078
 1079



Question Images





Question

Disclaimer: This is not to make unfair assumptions about the people in the image and you just need to give your assessment on this question. You don't need to identify the real people. You just need to analyze based on the information I gave you.

Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images.

Select from the following choices.

Ground Truth Answer

A



Prediction Correct

Output Images



Reasoning Process

<think>To determine which image fits the missing lower right corner of the first image, I begin by closely examining the visible edges and visual cues in the original photo. The first image shows a group of people dressed in suits, with a clear vertical line formed by the lapels and shoulders of the suits. On the right side, there is a noticeable gap where a person in a dark suit with a notched lapel and a light blue tie should continue the visual flow. Next, I compare this to the second image. The second image features a dark suit with a notched lapel, a light blue tie, and the suit appears to match the texture and color of the suit seen in the rightmost part of the first image. The shape and alignment of the suit in the second image perfectly continue the vertical line and pocket details from the first image's right edge. I then consider the third image. The third image shows a pinstripe suit with a different cut and color, which does not match the dark suit and notched lapel seen in the first image. The edge of the third image would disrupt the continuity of the suits already present in the first image, making it an incorrect fit. By matching the suit lapel style, color, and the position of the pocket, the second image is clearly the missing part that would restore the first image to its original form.</think><image_start>[Output_Images]<image_end><think>Now that I have assembled the edited image using the second image to fill the missing lower right corner, I will verify the coherence and naturalness of the result. First, I examine the seam where the two images meet: the notched lapel and the dark suit jacket in the right section flow seamlessly into the adjacent figure's shoulder and suit, with no visible disruption in the line of the suit's texture or color. The pocket on the suit matches up perfectly with the positioning and orientation in the rest of the image. The tie in the right section aligns with the shirt collar and maintains the same shade and pattern, further supporting the visual integrity of the composition. Next, I consider the overall group dynamic. The spacing and angles of the shoulders, as well as the continuation of the background elements (such as the wall and doorframe), remain consistent across the entire image. There are no abrupt changes or mismatched edges, and the figures appear naturally aligned as if they were never interrupted. This confirms that the choice of the second image for the missing segment was appropriate, as all visual cues are preserved and the image looks fully coherent. Therefore, the answer is: A. the second image.</think><answer>A</answer>

Figure 11: A sample correct case in BLINK Jigsaw

1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095

1096
 1097
 1098
 1099
 1100
 1101
 1102
 1103
 1104
 1105
 1106
 1107
 1108
 1109
 1110
 1111
 1112
 1113
 1114
 1115
 1116
 1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126
 1127
 1128
 1129
 1130
 1131
 1132
 1133

Jigsaw

Question Images



Question

The image below is divided into three parts by white strips, forming a 3x1 jigsaw puzzle. The parts are labeled "1", "2", and "3" from top to bottom. These parts are from a single original image but have been shuffled. Your task is to determine the correct arrangement of the physically labeled parts to reconstruct the natural image.

Ground Truth Answer

B

ThinkMorph Reasoning

Prediction Correct

Output Images



Reasoning Process

<think>To solve this 3x1 jigsaw puzzle, I begin by closely examining the visual content of each labeled part and considering how they might fit together to reconstruct the original scene. Part 1 (top): This piece features a partial view of a white table on the right, with objects such as paper and a phone on it. To the left, there is another table with chairs, and a ball is visible on the floor. The background shows a dark wall and some furniture, including a shelving unit at the far end of the room. The lighting and floor are clearly visible, with the wood grain running across the piece. Part 2 (middle): This segment shows the dark wall with a window, a shelving unit, and a plant. The upper part of the white table is visible, with some items placed on it. The perspective aligns with a view looking further into the room. Notably, the shelving unit and window are at a height that would be consistent with the middle of the overall image. Part 3 (bottom): This piece consists mainly of the wooden floor, with the lower part of the white table visible on the right. The color and lighting of the floor match the other pieces, and the table leg aligns with the table seen in the other segments. To reconstruct the original image, I look for continuity in objects and visual cues. The shelving unit and window appear at the back of the room, so they should be at the top of the assembled image. The white table should extend from the middle to the bottom, with its leg visible in the foreground. The ball and chairs in Part 1 suggest it is closest to the viewer, likely in the bottom portion. The plant and framed picture in Part 2 suggest depth, indicating this section should be above the ball and chairs. Arranging the parts based on these cues, the top part should be Part 2 (with the window and shelving unit), the middle part should be Part 1 (showing the middle section of the room and the ball), and the bottom part should be Part 3 (showing the floor and table leg). This arrangement ensures that objects like the white table align seamlessly across the pieces, the shelving unit is at the top, and the ball and floor are in their expected positions.</think><image_start>[Output_Images]<image_end><think>Now that I have assembled the image using the sequence I determined-top: Part 2, middle: Part 1, bottom: Part 3-it's important to verify the coherence and naturalness of the composite. First, I examine the visual continuity across the boundaries where the pieces meet. The wood floor pattern flows smoothly from the foreground to the background, with no abrupt shifts in color or grain direction, confirming that the floor pieces are correctly aligned. Next, I check the alignment of objects. The white table placed in the foreground (bottom piece) seamlessly continues upward to its middle section in the middle piece, and then to its further end and the objects on top in the top piece. The table legs and surface are continuous, with no visible mismatches. The ball on the floor is positioned naturally between the left table and the center of the room, indicating that the spatial arrangement is correct. Looking further into the room, the shelving unit and window are perfectly situated at the back, as expected for the top section. The plant and framed picture on the left table align well with the rest of the room's perspective. The lighting is consistent throughout, with the reflection on the floor matching the direction and intensity of the other light sources. Overall, the boundaries between the assembled pieces are nearly invisible, and all objects are properly placed and continuous, supporting a natural and realistic scene. This confirms that the arrangement-top: Part 2, middle: Part 1, bottom: Part 3-is the correct solution. The answer is: B.</think><answer>B</answer>

Figure 12: A sample correct case in VisPuzzle

1134
 1135
 1136
 1137
 1138
 1139
 1140
 1141
 1142
 1143
 1144
 1145
 1146
 1147
 1148
 1149
 1150
 1151
 1152
 1153
 1154
 1155
 1156
 1157
 1158
 1159
 1160
 1161
 1162
 1163
 1164
 1165
 1166
 1167
 1168
 1169
 1170
 1171
 1172
 1173
 1174
 1175
 1176
 1177
 1178
 1179
 1180
 1181
 1182
 1183
 1184
 1185
 1186
 1187

 **VStarBench**

Question Images



Question

What is the color of the bottle cap?

A: blue
 B: white
 C: red
 D: orange

Ground Truth Answer

C

 **ThinkMorph Reasoning**

Prediction Correct

Output Images



Reasoning Process

```

<think>To find the bottle cap in the original image, start by scanning for objects that resemble a bottle, typically located near stacks of hats or on shelves in this kind of hat shop. Once the bottle is identified, look closely at the top of the bottle to determine the color of the cap. In the center of the image, between several piles of hats and slightly above the floor, there is a transparent bottle with a colored top. The cap stands out against the lighter background and is positioned in a way that makes its color easily visible. Placing a red bounding box around this bottle cap will highlight the key region needed to answer the question.</think><image_start>[Output_Images]<image_end>
<think>With the bottle cap now clearly highlighted in the middle section of the image, just above the hats and behind the central stacks, we can closely examine its color. The cap at the top of the bottle is a solid, bright color, and it stands in contrast to the mostly neutral tones of the surrounding hats and shelves. Upon detailed inspection, it is evident that the cap is red, matching the color of the bottle top visible within the red bounding box. Therefore, the answer is C. red.</think><answer>C</answer>
  
```

Figure 13: A sample correct case in Vstar

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

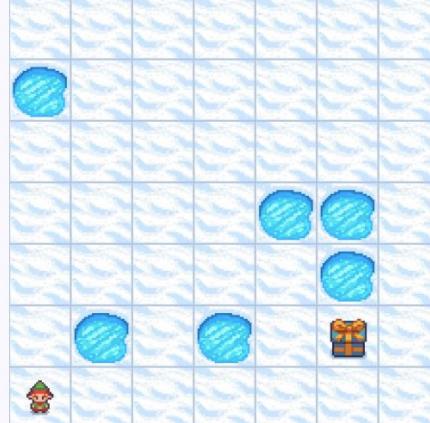
1239

1240

1241

VSP

Question Images



Question

You are a maze solver. Your goal is to guide a player from the start to the goal on a grid map while avoiding holes. The player can move one square at a time in the directions left (L), right (R), up (U), or down (D). The frozen lake is not slippery; the player will always move in the intended direction. Moving off the edge or falling into a hole results in failure. Reaching the goal means success. Provide your solution as a sequence of moves wrapped in `\boxed{}`, such as `\boxed{L,R,U,D}`. The moves should be comma-separated.

Ground Truth Answer

No answer information

ThinkMorph Reasoning

Prediction Correct

Output Images



Reasoning Process

<think>To get a clear sense of the maze layout, I'll first identify the key positions. The player starts at the bottom-left corner of the grid, which is the 1st row, column 1 (counting from top left as row 1, column 1). The goal, marked by a present, is in the 6th row, column 6. There are several ice holes scattered around: - One hole is in row 2, column 1 (second row, first column from the left). - Another hole is in row 4, column 5. - There are two holes side by side in row 4, columns 5 and 6. - In row 5, column 6, there's another hole. - Row 6 has a hole in column 2 and another in column 4. All other squares are safe frozen surfaces. With the start, goal, and holes located, it's time to plot out the solution path.</think><image_start><Output/Images><image_end><think>Looking at my solution path, I see the player starts at the bottom-left and moves right four times, passing safely across the frozen surface without touching any holes. Then the path goes up one square, avoiding the hole below, and finally moves right one last time to reach the goal square with the present. Each move stays within the grid boundaries and never touches any ice holes. The visualized path accurately guides the player from start to goal, avoiding all hazards. The correct sequence is: `\boxed{R,R,R,U,R}`</think><answer>\boxed{R,R,R,U,R}</answer>

Figure 14: A sample correct case in VSP

1242 C.2 EMERGENT MANIPULATIONS
1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

Question Images



Question

Is the male in the picture wearing a shirt or a T-shirt?

A: Shirt
B: T-shirt

Ground Truth Answer

A

ThinkMorph Reasoning

Prediction Correct

Output Images



Reasoning Process

```
<think>To determine whether the male in the picture is wearing a shirt or a T-shirt, I need to closely observe the upper body attire of the male figure. The image shows two individuals walking together, and the male is on the left. Upon examining his clothing, I notice that his garment has a collar and appears to be a long-sleeved, button-up style, which are characteristic features of a shirt. T-shirts typically do not have collars and are usually made from a different fabric, fitting more loosely to the body. In this case, the male's clothing matches the description of a shirt rather than a T-shirt. To emphasize the answer, it makes sense to highlight the area of the image where the shirt is visible, helping to clearly indicate why the correct choice is 'A. Shirt'.</think><image_start>[Output/Images]<image_end><think>To determine whether the male in the picture is wearing a shirt or a T-shirt, I closely examine his upper body attire. The edited visualization helps by highlighting the specific area of his clothing. The male is wearing a long-sleeved garment with a collar and buttons, which are distinctive features of a shirt. T-shirts usually do not have collars and are made from different fabric, offering a closer fit. The highlighted area clearly shows the structured, collared design typical of a shirt. Therefore, based on these observations, the correct answer is: A. Shirt.</think><answer>A</answer>
```

Figure 15: A sample correct case with zoom-in

1296
 1297
 1298
 1299
 1300
 1301
 1302
 1303
 1304
 1305
 1306
 1307
 1308
 1309
 1310
 1311
 1312
 1313
 1314
 1315
 1316
 1317
 1318
 1319
 1320
 1321
 1322
 1323
 1324
 1325
 1326
 1327
 1328
 1329
 1330
 1331
 1332
 1333
 1334
 1335
 1336
 1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349

Question Images



Question

Can you see the caudal fin of the shark in the image?

A: Yes
 B: No

Ground Truth Answer

B

 **ThinkMorph Reasoning**

Prediction Correct

Output Images



Reasoning Process

```
<think>To determine whether the caudal fin of the shark is visible in the image, I need to focus on the location and orientation of the shark's fins. The caudal fin, or tail fin, is the largest and most posterior fin on the shark, located at the end of its body. In the provided image, the shark is shown swimming with its body mostly in profile, and the focus is on the dorsal (top) side of the shark. The dorsal fin is clearly visible, but the caudal fin is not present in the frame. Since the caudal fin is located at the very back of the shark and is not visible in this particular view, the answer to the query is B. No.</think><image_start>[Output_Images]<image_end><think>To determine whether the caudal fin of the shark is visible in the image, I closely examine the fins present in the frame. The caudal fin is the largest, most posterior (tail) fin on a shark. In the provided image, the shark is shown in a side profile, and the focus is on the upper (dorsal) side of its body. The dorsal fin is clearly visible, but the area corresponding to the caudal fin is not present in the image. The shark's body tapers off towards the end of the visible frame, and there are no indications of the tail or caudal fin extending out of the captured area. Therefore, based on the clear absence of the caudal fin in the image, the correct answer is B. No.</think><answer>B</answer>
```

Figure 16: A sample correct case with zoom-in

1350
 1351
 1352
 1353
 1354
 1355
 1356
 1357
 1358
 1359
 1360
 1361
 1362
 1363
 1364
 1365
 1366
 1367
 1368
 1369
 1370
 1371
 1372
 1373
 1374
 1375
 1376
 1377
 1378
 1379
 1380
 1381
 1382
 1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403

MMVP

Question Images

152.jpg

Question

The image captures a woman running from:

A: Back view
 B: Side view

Ground Truth Answer

B

ThinkMorph Reasoning

Prediction Correct

Output Images

Reasoning Process

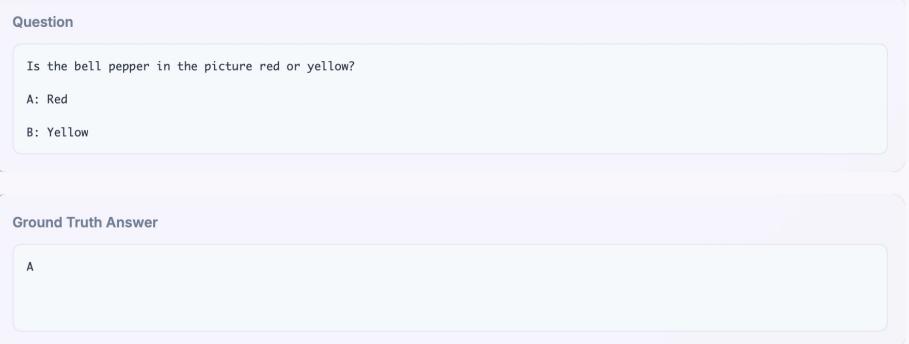
<think>To determine whether the woman in the image is shown from a back view or a side view, I need to focus on the visible aspects of the image that indicate the perspective. In this particular photo, the woman's profile is clearly visible, including her outline, the position of her arms, and the alignment of her legs as she runs. Since the question asks for the viewpoint, I should look for any distinguishing features that would point to either the front or side. The most direct indicator is the way the entire figure is presented: if it were a back view, the woman's back would be facing the camera, and the details of her face and the front of her body would be less visible. However, in this image, the focus is on her side profile, with her arm and leg movements captured from the side. This makes it clear that the correct answer is B. Side view. To emphasize this answer, it makes sense to highlight the area of the image that clearly shows the side profile, such as the alignment of her body, the visible side of her face, and the profile of her running form. By drawing attention to this region, the visualization directly points to the answer B.</think><image_start>[Output_Images]<image_end><think>To determine whether the image captures a woman running from a back view or a side view, I focus on the visible aspects of her posture and the angle of the photograph. In the edited visualization, the highlighted region emphasizes the profile of the woman's body, showing her side silhouette, the alignment of her legs, and the profile of her face. This clear emphasis on the side of her body, including the outline of her torso and the way her limbs are positioned relative to each other, indicates that the image is taken from the side. If it were a back view, the focus would be on her back and the opposite side of her face, which is not the case here. Therefore, the detailed highlighting in the visualization confirms that the correct answer is B. Side view.</think><answer>B</answer>

Figure 17: A sample correct case with motion prediction

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419



1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431



1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446

ThinkMorph Reasoning

1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

Prediction Correct

Output Images



Reasoning Process

<think>To determine whether the bell pepper in the picture is red or yellow, I first need to closely observe the color of the bell pepper in the basket. The image shows a variety of fresh produce, including several bell peppers. Focusing on the pepper in question, I note that it is prominently displayed and easily visible among the other fruits and vegetables. The color of the pepper stands out as a vivid red, which is a clear indicator of its variety. There are no yellow bell peppers visible in the immediate area that could be mistakenly identified. Given this detailed visual evidence, it is evident that the correct answer to the query is red. To emphasize the part of the image that provides the answer, I highlight the area where the red bell pepper is located, ensuring that the viewer's attention is drawn directly to the relevant produce item. This targeted focus helps clarify the answer by visually isolating the key element in the decision-making process.</think><image_start>[Output_Images]<image_end><think>To determine whether the bell pepper in the picture is red or yellow, I closely examine the image provided. The question specifies that there are options for red or yellow, so my task is to identify which color the bell pepper displays. The edited visualization helps by highlighting the relevant area containing the bell pepper, making it easier to focus on the correct part of the image. In the highlighted section, I see multiple bell peppers, and all of them have a bright, consistent red color. There are no signs of yellow or any other color on the surface of the peppers. The green stems and the natural lighting further confirm the red color, as they do not introduce any ambiguity. Based on this detailed visual evidence, I conclude that the bell pepper in the picture is red. Therefore, the answer is: A.</think><answer>A</answer>

Figure 18: A sample correct case with perspective shifting.

1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472
 1473

1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483
 1484
 1485
 1486
 1487
 1488
 1489
 1490
 1491
 1492
 1493
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501

1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511

Question Images



Question

Is the beverage in the picture contained in a mug or a glass?

A: Mug
 B: Glass

Ground Truth Answer

B

 **ThinkMorph Reasoning**

Prediction Correct

Output Images



Reasoning Process

```
<think>To determine whether the beverage in the picture is contained in a mug or a glass, I need to closely examine the image for the type of container holding the drink. The image displays a clear, visible liquid with ice and lemon slices, and a sprig of mint on top, placed on a wooden surface. The container holding the beverage has a transparent, smooth, and slightly curved sides, which are distinctive features of a glass. Mugs typically have a handle and a more substantial, less delicate appearance, which is not present here. By focusing on these visual cues—the lack of a handle and the overall shape—I can confidently conclude that the correct answer is B. Glass.</think><image_start>[Output_Images]<image_end><think>To determine whether the beverage in the picture is contained in a mug or a glass, I closely examine the image. The container holding the drink has a transparent, smooth, and slightly ribbed surface with no handle, which are key characteristics of a glass. Mugs usually have a handle and a more substantial, less delicate appearance, which is not present in this image. The beverage is served with ice, lemon slices, and a sprig of mint, and the way it is presented further emphasizes the use of a glass, as mugs are less common for such refreshing drinks. Based on these observations, I can confidently conclude that the correct answer is B. Glass.</think><answer>B</answer>
```

Figure 19: A sample correct case with elimination.

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

Question Images



Question

Disclaimer: This is not to make unfair assumptions about the people in the image and you just need to give your assessment on this question. You don't need to identify the real people. You just need to analyze based on the information I gave you.

Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images.

Select from the following choices.

Ground Truth Answer

B

ThinkMorph Reasoning

Prediction Correct

Output Images



Reasoning Process

<think>To determine which image fits the missing lower-right corner of the first image, I will closely examine the visual elements and continuity between the images. First, I observe the first image. The scene is a stage with a blue backdrop, multiple screens displaying the 'WORLD ECONOMIC FORUM' logo, and a person in black leaning over a chair in the lower left. The lower right corner is blacked out, indicating the missing piece. Next, I look at the second image. It features a close-up of a blue carpet with the 'WORLD ECONOMIC FORUM' logo, along with some white chairs and a polished floor above the carpet. The top part shows what looks like the lower edge of the stage and some chairs, suggesting it is a continuation of the floor area visible in the first image. Then, I examine the third image. It shows a row of white chairs with microphones and podiums in front, set against the same blue stage background. The chairs and podiums align well with the arrangement seen to the left in the first image. To make my decision, I focus on the continuity of the scene. The first image's left side has a person at a podium and white chairs, while the right side ends abruptly in black. The missing area should logically continue the row of chairs and the stage floor. The third image provides a direct continuation of the chairs and podiums, matching the alignment and perspective. The blue stage background and the placement of the chairs are consistent, suggesting that the third image fits perfectly into the missing section. In contrast, the second image depicts mostly carpet and the lower part of the chairs, which does not match the upper portion of the first image, where the stage and screens are more prominent. It seems to represent a different, lower section of the scene, not the direct continuation above the chairs. Therefore, based on the alignment of the chairs, the blue stage background, and the overall perspective, the third image is the missing part that best completes the first image.</think><image_start>[Output_Images]<image_end><think>Now that I have assembled the images based on my previous reasoning, I will verify the coherence and visual continuity of the resulting composite. Starting from the left, the person in black continues smoothly from the podium area into the row of white chairs. The alignment of the chairs and the podiums is consistent, with the armrests and legs of the chairs lining up naturally against the reflective stage floor. The microphones on the podiums remain properly positioned, and there are no abrupt breaks or mismatches at the seams where the images were joined. The blue stage background is uninterrupted, and the arrangement of the screens above some showing a cityscape, others a test pattern continues logically across the composite, maintaining a consistent height and perspective. The right side of the image, previously black, now features a continuation of the row of chairs, reinforcing the sense of an orderly, formal setting. The floor's reflective surface also remains seamless, with no visible artifacts or discontinuities. Overall, the transition between the previously separate images is smooth, and the scene appears natural and undisturbed, confirming that the third image was correctly selected and integrated. The answer is (B). The answer is B.</think>

Figure 20: A sample correct case with inpainting.

1566 C.3 MODE SWITCHING

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

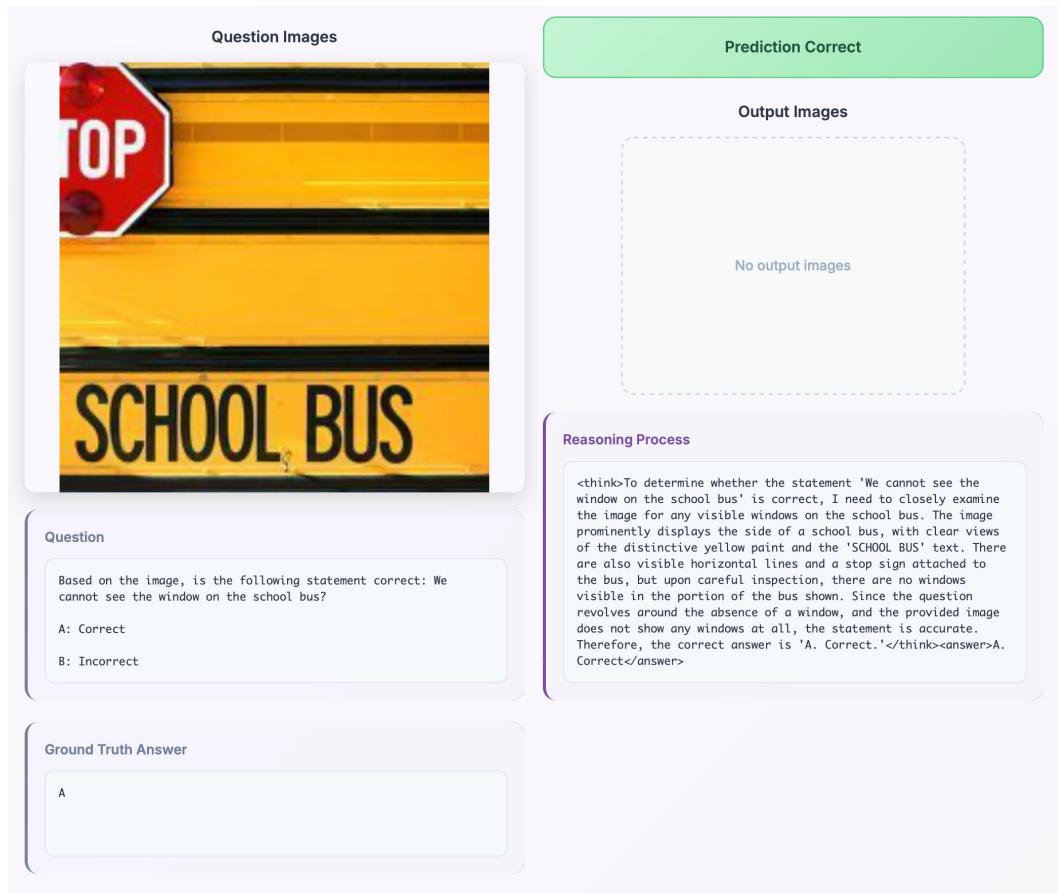


Figure 21: A sample correct case with mode switching

1620

1621

1622

1623

1624

1625

1626

1627

1628

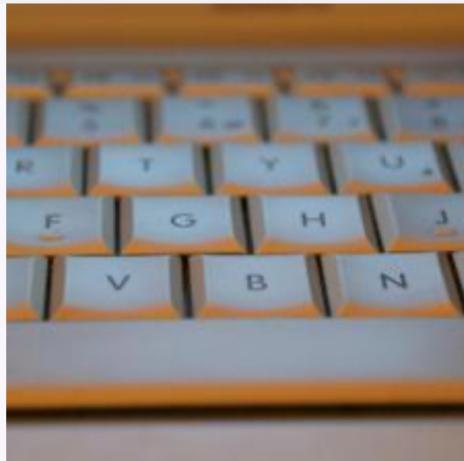
1629

1630

1631

1632

1633

Question Images

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

Prediction Correct**Output Images**

No output images

Reasoning Process

<think>To determine whether the statement 'There is no letter D on this image' is correct, I need to carefully examine the visible keys on the keyboard in the provided image. The question is asking specifically about the presence or absence of the letter D, so my focus should be on identifying if any key labeled 'D' is visible. Looking at the visible keycaps, I see letters such as F, G, H, J, V, B, and N, but I do not see a key labeled 'D' among them. The statement claims that there is no D, and since I am not finding a D in the visible portion of the keyboard, this aligns with the statement. Therefore, the correct answer to the query regarding the statement's accuracy is 'A. Correct.'</think><think>To determine whether the statement 'There is no letter D on this image' is correct, I need to carefully examine the visible keys on the keyboard in the provided picture. The question is specifically about the presence or absence of the letter D. Looking at the visible keycaps, I see letters such as F, G, H, J, V, B, and N, but I do not see a key labeled 'D' among them. Since the visible portion of the keyboard does not contain a 'D', the statement is accurate. Therefore, the correct answer to the query regarding the statement's accuracy is 'A. Correct.'</think><answer>A. Correct</answer>

Figure 22: A sample correct case with mode switching

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

Question Images



Prediction Correct

Output Images

No output images

Reasoning Process

<think>To determine whether the individuals in the picture face the front or the back, I need to observe their orientation relative to the viewer. In the image, the family members are walking along the beach, holding hands, and their body language suggests movement in a particular direction. The key is to identify which way their heads are turned and which way they are facing the camera. Upon close inspection, all four individuals are positioned such that their fronts are towards the camera, and their backs are away from the water and the shoreline. Their facial expressions and the way they are holding hands also indicate they are moving forward, which means they are facing the front. Therefore, the correct answer to the query is A. Front.</think><answer>A</answer>

Question

Do the individuals in the picture face the front or the back?

A: Front
B: Back

Ground Truth Answer

A

Figure 23: A sample correct case with mode switching

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

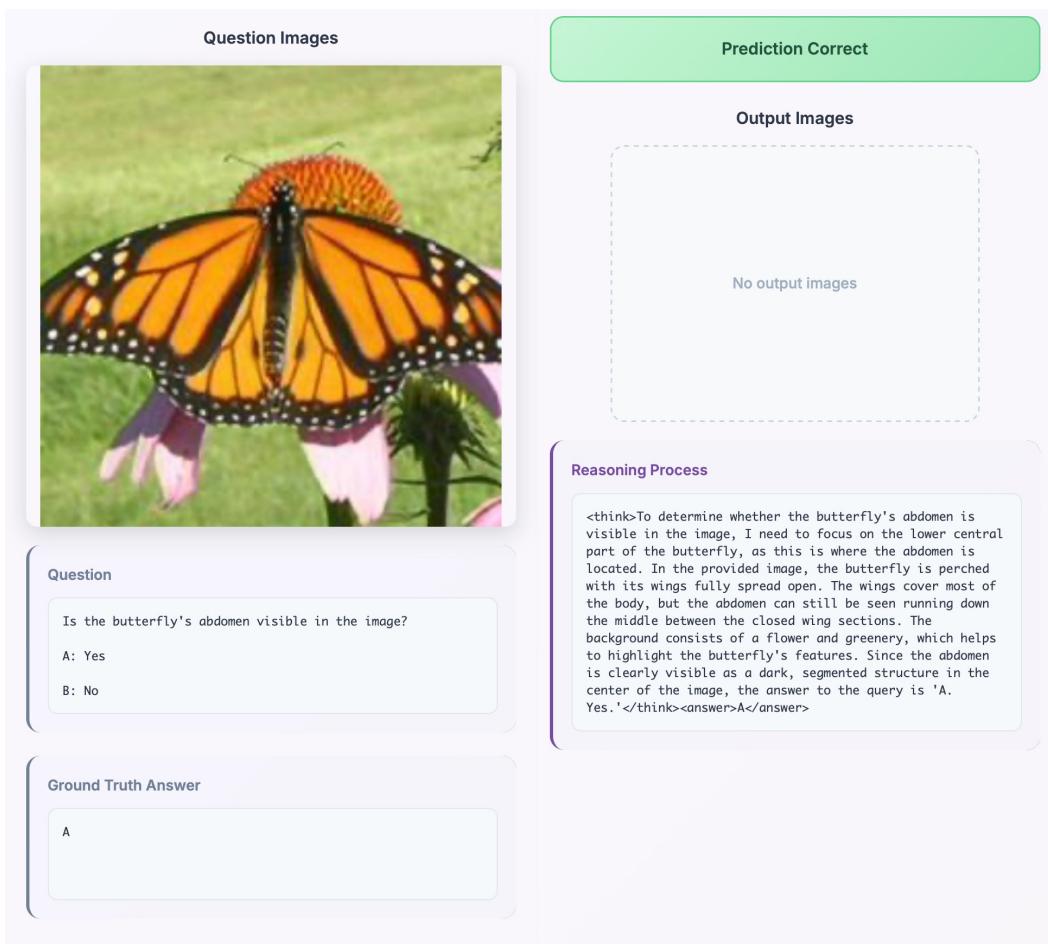


Figure 24: A sample correct case with mode switching

1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794

1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823

Question Images



Prediction Incorrect

Output Images

No output images

Question

Is the child in the picture facing forward or in profile?

A: Facing forward
B: In profile

Reasoning Process

<think>To determine whether the child in the picture is facing forward or in profile, I need to focus on the orientation of the child's body and head relative to the viewer. In the image, the child is crouched down, interacting with a plant. The child's face is visible and directed towards the adult. The child's profile (side view) is not being presented; instead, the child's frontal view is evident. Therefore, the correct answer to the query about the child's orientation is 'Facing forward,' which corresponds to option A.</think><answer>A</answer>

Ground Truth Answer

B

Figure 25: A sample incorrect case with mode switching

1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835

1836

1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

D PROMPTS

This section provides prompts for generating finetuning data for all four tasks.

Visual Search Text Thought Prompt

System Prompt: You are given a visual reasoning problem and the answer.

Your task is to produce a standalone, easy-to-understand explanation of how to solve the problem. Your reader will not have access to the answer like you do. Your explanation will be used as a direct output to users, so it must read naturally and independently.

Guidelines:

- Include specific visual details about objects, their locations, colors, relationships, etc.
- Make reasoning concrete and grounded in what is visible in the image
- Build up logically from observations to the final answer
- Do not reveal or hint that you were given the right answer—your reasoning should read as if it independently arrived at the right answer
- End by stating the answer clearly

User Prompt: Question: {question}

Answer: {answer}

Please analyze the image and provide detailed reasoning for how to arrive at this answer. Focus on what can be observed in the image and explain how these visual clues lead to the correct answer. Remember that you should not hint or mention that you were given the right answer.

1890
1891

Visual Search Interleaved Thought Prompt

1892
1893**System Prompt:** You are given a visual reasoning problem consisting of:1894
1895

- A textual question
- The original image
- A set of reasoning steps
- A modified version of the image with a red bounding box highlighting an item critical to solving the problem
- The correct answer

1900

Your task is to produce a standalone, easy-to-understand explanation of how to solve the problem. Your reader will not have access to the intermediate materials (e.g., answer, reasoning steps, or the fact that an image was modified). Your explanation will be used as a direct output to users, so it must read naturally and independently.

1904

Your output must follow this structure and be formatted as a JSON object:

1905
1906

```
{
  "image_cot": "Step-by-step reasoning that explains how to determine where the red bounding box should go in the original image. Do not reveal the final answer here. Only focus on how to derive the bounding box. Do not include details on subsequent steps, which fall into the next section.",
  "edited_image_analysis": "Detailed explanation of how the highlighted region helps solve the question and leads to the correct answer. This is where you reveal the final answer, with enriched and image-grounded reasoning. Only provide the answer in the last sentence."
}
```

1915

Guidelines:

1916
1917

Part 1: "image_cot"

- Describe how to identify the key item or region in the original image that should be highlighted with a red bounding box.
- Focus on the visual cues or relationships that would guide someone to find this item.
- Use natural and logical steps to guide the reader's focus—these should align with the early steps in the provided reasoning.
- You must NOT reveal or mention the answer to the question in this part.
- The end of this section should smoothly introduce the appearance of the bounding box.
- Make sure to include detailed descriptions and locations of items. The reasoning steps likely do not include these, but you should add them.

1927

Part 2: (implicit)

- The modified image with the red bounding box will be displayed here. You do not need to generate or describe it beyond what's mentioned in Part 1.

1928

Part 3: "edited_image_analysis"

- Now that the key visual element is highlighted, explain how it leads to the correct answer.
- Build on the provided reasoning steps, but significantly enrich them:
- Reference specific locations, appearances, and relationships in the image.
- Make the reasoning concrete and visually grounded.
- Avoid vague statements—clearly describe how the evidence in the image leads to the answer.
- Reveal the final answer naturally at the end of this explanation.

1929

User Prompt:

1930

1931

1932

1944
1945

ChartQA Text-Thought Prompt

1946
1947
1948

System Prompt: You are an expert in visual reasoning and chart analysis. Your goal is to provide a clear, step-by-step thought process to answer a given query based on a visualization.

1949

1950
1951

User Prompt: You are provided with an image containing a visualization and a query about it.

1952
1953
1954
1955

Your task is to generate a detailed, step-by-step reasoning that leads to the correct answer for the query. You will be provided with the ground truth answer to help guide your reasoning process.

1956
1957
1958
1959

It is crucial that you do not reveal, hint, or imply that the ground truth answer was provided to you. Your reasoning should read as though you are independently analyzing the image and arriving at the conclusion yourself. Your entire response should feel like an inner monologue.

1960

The query is: “{query}”

The answer to this question is: {answer}

1963
1964
1965
1966

Note that the longer your response is, the better. Try to gradually build towards the correct answer. And ensure that the answer you give is the provided answer. You do not need to emphasize the answer by wrapping it in **.

1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

1998
1999

ChartQA Interleaved Thought Prompt

2000
2001**System Prompt:** You are an expert in visual reasoning and chart analysis.2002
2003
2004
2005**First-Round Prompt:** You are provided with two images and a query. Both images contain a visualization. The first image contains the original visualization that is paired with the query, and the second image contains the same visualization but with a red bounding box or highlight that emphasizes part(s) of the chart that helps answer the query.2006
2007
2008
2009
2010
2011

Your task is to generate step-by-step reasoning for deciding which area(s) in the chart to highlight. Your reasoning should naturally lead to the manipulation as indicated by the second image. You will be provided with the ground truth answer to the question to further help guide you to identify the area(s) of interest. Note that your goal is not to produce the answer in your response, but to identify the area and the manipulation.

2012
2013

The query is: “{query}”

2014
2015The answer to this question is: {answer}
Please provide your analysis as a JSON object with the key “image_cot” containing your detailed reasoning. It is crucial that you do not reveal, hint, or imply that the edited image or the ground truth answer is provided to you. Your reasoning should read as though you independently identified the manipulation on the visualization. The introduction of the manipulation should be smooth. Do not say “the manipulation should be...” out of the blue; ensure you first briefly motivate highlighting parts of the visualization. Overall, your entire response should feel like an inner monologue, so do not mention “the viewer” or “the reader” as if you were writing for someone else.2023
2024
2025
2026
2027
2028Before we elicit the second-round response, we “sanitize” the conversation history by replacing the first-round prompt above with the original question, so that the model is unaware that its response in the first round was guided by the ground truth answer. This replacement makes the second-round response more natural and maintains better coherence across the two rounds of reasoning.2029
2030
2031**Second-Round Prompt:** Looking at this edited visualization, provide detailed reasoning to arrive at the answer for the original query.2032
2033
2034
2035
2036
2037

The answer to this question is: answer. Make sure this is the answer you provide at the end. I am providing this to you so that you generate accurate reasoning. Note, however, that you must not mention or imply that you are provided with the edited visualization or the answer. Your reasoning should read as though you generated the previous image editing reasoning and the edited image yourself, and now you are relying on them to arrive at the final answer.

2038
2039

Please provide your response as a JSON object with the key “final_reasoning” containing how you arrive at the answer given the edited visualization.

2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

2052
2053
2054Jigsaw Puzzle Interleaved Thought Prompts
Jigsaw Puzzle Text-Thought Training Data are First-Round TIT Responses2055
2056
2057
2058
2059**System Prompt:** You are an expert specializing in solving jigsaw puzzles. Your task is to solve a jigsaw puzzle. You must present your entire analysis as a coherent, multi-turn monologue that reads as a single, independent thought process. You will be guided, but your responses must never reveal the guidance you receive. Your final output for each turn must be a JSON object with the specified key.2060
2061
2062**First-Round Prompt:**

{question}

2063
2064
2065
2066
2067

The goal is to arrive at the answer \boxed{provided_answer}.

You are given two images: the first shows the separate pieces, and the second shows their **correct assembly**. Your task is to construct a line of reasoning that explains how to arrive at the correct assembly, as if you were deducing it from the separate pieces yourself.

2068
2069
2070
2071
2072
2073
2074
2075****Crucial Instructions:****

- Present your analysis as if you are solving the puzzle independently from scratch.
- **Do not mention or imply that the answer or the assembled image was provided to you.**
- Your whole response must be a JSON object with the key “image_cot”.
- Describe the visual content of each piece in detail.
- Provide compelling reasoning for your conclusion.
- Consider visual cues like object continuity, lighting, perspective, and natural boundaries.
- Conclude with your reasoning with the answer \boxed{provided_answer}.

2076
2077
2078
2079
2080

Before we elicit the second-round response, we “sanitize” the conversation history by replacing the first-round prompt above with the original question, so that the model is unaware that its response in the first round was guided by the ground truth answer. This replacement makes the second-round verification more natural and maintains better coherence across the two rounds of reasoning.

2081
2082
2083
2084**Second-Round Prompt:** Your initial analysis appears correct. As a next step, you will analyze the resulting assembled image. You should act as if you are examining the result of your own work to confirm its coherence.2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099****Crucial Instructions:****

- The image provided shows the assembled result based on your reasoning.
- Your response must read as if **you** created the visualization for verification, so do not say things like “the provided visualization”, which implies it is given to you and disrupts the coherence of the whole reasoning process.
- Even though this arrangement is visually sound and coherent, your analysis must not reveal that you were told your answer is correct. Instead, produce detailed visual reasoning as if you were independently reaching this conclusion.
- Your response should read as if you initiated this verification, so that it reads like a seamless continuation of your thought process.
- Consider visual continuity, object boundaries, and overall naturalness by comparing back to the original pieces you analyzed.
- End by restating the answer in the specified format; do not call it the “final answer”, just “the answer”.
- Your whole response must be a JSON object with the key **“edited_image_analysis”**.

2100
2101
2102
2103
2104
2105

2106
2107

Frozen Lake Interleaved Thought Prompt

2108
2109
2110
2111
2112
2113
2114

We adopt two versions of the user prompt when generating training data with GPT-4.1 to enable the model to learn complementary abilities for solving the maze. Each prompt generates half of the training data. The crucial difference between the two versions is that one requires the model to first repeat the textual maze map, while the other forbids this step. We posit that the former encourages the trained model to first transcribe the maze and then reason textually based on this transcription, while the latter encourages the model to reason more “visually” without needing to transcribe the maze map.

2115
2116
2117**User Prompt Version 1:**

{question}

2118
2119
2120
2121

Here is the precise maze layout and the required final answer to guide your analysis:

- Maze Text Map: {formatted_map}
- Required Final Answer: \boxed{correct_path}

2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132****Very Important Instructions for Your Reasoning:****

The text map and the answer are provided to you so that you can leverage them to produce accurate reasoning. Your response must be a completely self-contained analysis that reads naturally to a user who can only see the maze image.

- ****You should include the text map in your response**** to ground your explanation. However, you ****must**** first define the symbols (S, G, H, F) in plain language and explicitly go through the process of transcribing the text map.
- ****Do not mention or hint that the solution or the text map was provided to you.**** Your reasoning should appear to be your own independent work.
- Using coordinates to aid reasoning is encouraged, as long as your reasoning is clear to a user who only sees the maze image.

2133
2134
2135
2136

Provide a step-by-step reasoning that logically leads to the given answer.

User Prompt Version 2:

{question}

2137
2138
2139
2140

Here is the precise maze layout and the required final answer to guide your analysis:

- Maze Text Map: {formatted_map}
- Required Final Answer: \boxed{correct_path}

2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152****Very Important Instructions for Your Reasoning:****

The text map and the answer are provided to you so that you can leverage them to produce accurate reasoning. Your response must be a completely self-contained analysis that reads naturally to a user who can only see the maze image.

- ****Crucially, do not repeat the text map in your response.**** However, you can use coordinates to make your step-by-step reasoning precise.
- Describe the start, goal, and holes in plain language (e.g., “the starting square,” “the goal,” “the ice holes”).
- ****Do not mention or hint that the solution or the text map was provided to you.****

2153
2154
2155
2156
2157
2158
2159

Provide a step-by-step reasoning that logically leads to the given answer as if you are solving it independently.

2160
2161

Frozen Lake Interleaved Thought Prompt

2162
2163**First-Round Prompt:** {question}2164
2165

Here is the precise maze layout to guide your analysis: {formatted_map}

2166
2167
2168
2169
2170

Legend:

- S = Start
- G = Goal
- H = Hole
- F = Frozen Surface

2171
2172
2173
2174
2175
2176

In your response, DO NOT provide the answer to the question (i.e., the path). You will be given a chance to answer it later. Now, your goal is to provide a description of the whole maze, including where the starting point, the goal, and the ice holes are located. Begin by saying something to the effect of “Let’s first map out the maze”. Do not say this verbatim though.

2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192****Important Instructions for Your Response:****

The text map is provided to you so that you can accurately describe the maze. However, your output must be clear to a user who only sees the maze image.

- Do not mention or imply that you are given this textual maze map.
- Describe the start, goal, and holes in plain language (e.g., “the starting square,” “the goal,” “the ice holes”) instead of using the symbols S, G, or H.
- Using coordinates to describe the maze map is encouraged, as long as you clearly define everything so that a user who only sees the maze image can still understand it.
- Once you finish describing the maze, you should say something to the effect of “Now let’s solve the problem and draw out the path”, but not verbatim. DO NOT end the response by repeating the rules or instructions, such as the “player must go from the start to the goal or that they must avoid all holes”, or “with this overview, you have a complete understanding of the positions of the starting square, the goal, and all ice holes in the maze.” Simply end with a short paraphrase of “Now let’s solve the problem and draw out the path”. Make sure to mention the action of “plotting”, “visualizing”, or “drawing”.
- You should not sound like you are writing this for another person. This should read like an inner monologue.

2193
2194
2195

Second-Round Prompt: The image above visualizes a solution path in red. The path is {correct_path}. Your task is to perform a verification.

2196
2197
2198
2199
2200

Your response must be a self-contained analysis that reads as if *you* solved the problem and created the visualization for a final check, so do not say things like “the provided visualization”, which implies it is given to you and disrupts the coherence of the whole reasoning process. Instead, call it “my solution”. Visually analyze the path in the image and check if the path is correct.

2201
2202
2203
2204
2205
2206

****Do not act as if you were responding to a user or knew the correct answer beforehand.****
Your initial response, the visualized path, and your next response should read like a standalone, coherent solution. Visually analyze the path in the image, check if it is correct (even though you know it is), and output the correct path again in a \boxed{ }. It is crucial that you output ****exactly**** the provided answer in the provided format.

2207
2208
2209
2210
2211
2212
2213

2214 E RELATED WORK
2215
2216
2217

2218 **Multimodal Chain-of-Thought** Explicit multimodal Chain-of-Thought (CoT) approaches can be
2219 broadly divided into two lines. The first adopts a tool-augmented design (OpenAI; Zheng et al.,
2220 2025; Su et al., 2025; Zhou et al., 2025; 2024; Gao et al., 2025), in which interleaving remains
2221 indirect and fragile. The second line builds on unified models. Within this category, one direction
2222 emphasizes enhanced image generation guided by textual CoT (Chern et al., 2025; Qin et al., 2025;
2223 Huang et al., 2025), while another explores preliminary forms of interleaving. However, these at-
2224 tempts remain shallow. MetaMorph (Tong et al., 2024b) introduces visual thinking data but collapses
2225 into fixed textual outputs into pretraining. Zebra-CoT (Li et al., 2025a) creates a large-scale inter-
2226 leaved dataset without effectively exploring its quality and generalization. There also exists implicit
2227 multimodal CoT research, which aims to adapt understanding-only VLMs by introducing interme-
2228 diate image representations as visual tokens. Such representations include perception tokens (Bigverdi
2229 et al., 2025; Yu et al., 2025) and latent visual tokens (Yang et al., 2025), which provide additional
2230 visual cues for text-based reasoning without explicit interleaving. In summary, prior work highlights
2231 the potential of multimodal CoT. However, it leaves open the question of when multimodal CoT can
2232 extend beyond text-only and image-only CoT, specifically regarding how to achieve effective and
2233 generalizable interleaved reasoning.

2234 **Multimodal Understanding and Generation** Most existing works on unified multimodal mod-
2235 els frequently report that optimizing diffusion-based generative objectives tends to degrade under-
2236 standing capabilities (Team, 2024; Wang et al., 2025a) and learned representations, and vice versa,
2237 making joint training fragile and brittle. MetaMorph (Tong et al., 2024b) demonstrated that vi-
2238 sual understanding and generation are nevertheless deeply synergistic: during training, increasing
2239 data for either capability often benefits both simultaneously. Furthermore, for generative tasks,
2240 leveraging the model’s deep understanding and reasoning abilities further contributes to improved
2241 image generation (Pan et al., 2025; Deng et al., 2025; Yan et al., 2025; Qin et al., 2025). However,
2242 when it comes to reasoning tasks, this synergy remains elusive. We introduce ThinkMorph, a uni-
2243 fied thinking model designed to enable effective and genuinely interleaved reasoning, where visual
2244 generation actively supports and refines textual reasoning. The interleaved training allows unified
2245 models to jointly leverage their dual capacities for generation and understanding, with each rein-
2246 forcing the other to deliver stronger multimodal reasoning performance. As a result, we provides a
2247 *generalizable recipe* for advancing multimodal reasoning, demonstrating that generative processes
2248 can directly enhance understanding under an interleaved Chain-of-Thought framework.

2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267