



Do Vision and Text Cues Exhibit Evidential Coupling?

UFO: A Benchmark for Compositional Multimodal Reasoning in Unified Models

Anonymous Authors¹

Abstract

Unified Foundation Models (UFMs), which support interleaved multimodal generation and understanding, have been proposed as a promising paradigm for reasoning about dynamic world states, yet it remains unclear whether the visual content they generate functions as grounded evidence for subsequent reasoning or merely as auxiliary output. Existing benchmarks largely evaluate generation and understanding as separate capabilities and do not test their functional dependence during reasoning. We introduce **UFO**, a benchmark designed to evaluate whether UFMs generate and use image and text cues as evidence for compositional multimodal reasoning. UFO spans three cue types, state determination, state reconstruction, and state augmentation, which correspond to progressively smaller transformations of the underlying world state. Our analysis reveals a significant modality gap, as models often achieve high prediction accuracy even when the generated visual cues exert limited influence on their decisions, indicating weakened evidential coupling and a reliance on textual shortcuts rather than robust cross modal grounding.

1. Introduction

Unified Foundation Models (UFMs) increasingly integrate multimodal understanding and generation within a single system, driven by the hypothesis that these capabilities are mutually reinforcing. The cornerstone of this integration is the concept of *coupling*. By producing outputs for multiple modalities and tasks in a shared latent space, UFMs are incentivized to maintain cross-modal consistency at inference time. From a world-model perspective, coupling is crucial for state-transition reasoning. A model must interpret the

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



Figure 1. The UFO reasoning framework. Unlike existing benchmarks that evaluate understanding and generation in isolation, UFO formulates reasoning as a **compositional multimodal process**. Given a question and input images, the model must first generate intermediate *textual and visual cues* representing a future state, and then answer the question by conditioning on these cues. This design enforces **evidential coupling**, requiring intermediate multimodal cues to function as grounded evidence rather than allowing single-step shortcut predictions.

current state and construct a plausible future state, supported by complementary evidence across modalities.

However, the impact of coupling on UFM performance remains largely underexplored, in part because existing benchmarks seldom evaluate coupling directly. In particular, current protocols fall short in three respects. **First**, they largely isolate tasks, treating generation and understanding as separate capabilities rather than as steps in a single inference procedure. **Second**, evaluation is often effectively unimodal. Understanding is probed via text, while generation is assessed via images, without verifying that the two modalities provide aligned, complementary, and grounded evidence. **Third**, most benchmarks focus on one-shot predictions, obscuring how evidence is accumulated and used across intermediate steps. Taken together, these gaps suggest that evaluating UFMs as world models requires benchmarks that jointly assess understanding and generation within explicit state transitions.

Table 1. **Comparison with Existing Multimodal Benchmarks.** We compare UFO with prior benchmarks along seven properties. ¹**Two-step** uses a two-stage procedure in which models first generate intermediate cues and then answer. ²**Inter. eval.** evaluates intermediate steps or cues rather than only the final prediction. ³**Multi. metrics** reports multidimensional metrics, beyond a single scalar score, for cue validity. ⁴**Hybrid eval.** supports both computational metrics and LLM-assisted evaluation for question answering. ⁵**Human labels** includes human annotations for judging or validating intermediate cues. ⁶**Contamination** checks for leakage by evaluating without the input image(s). ⁷**Filter** screens out examples in which the question can be answered without relying on multimodal cues. UFO is *the only benchmark* that satisfies all criteria and spans 10 task types.

Benchmark	Venue	Two-step	Inter. eval.	Multi. metrics	Hybrid eval.	Human labels	Contamination	Filtering	MCQ & OQ	#Types
ReasonPix2Pix (Jin et al., 2024)	ArXiv'24	✗	✗	✗	✗	✗	✗	✗	✗	1
ReasonEdit (Huang et al., 2024)	CVPR'24	✗	✗	✗	✗	✗	✗	✗	✗	1
EditWorld (Yang et al., 2024)	MM'25	✗	✗	✗	✗	✗	✗	✗	✗	7
Reason50K (He et al., 2025)	ArXiv'25	✗	✗	✗	✗	✗	✗	✗	✗	4
KRIS-Bench (Wu et al., 2025b)	NeurIPS'25	✗	✗	✓	✓	✓	✗	✗	✗	7
RISEBench (Zhao et al., 2025)	NeurIPS'25	✗	✗	✓	✓	✓	✗	✗	✗	4
WorldGenBench (Zhang et al., 2025a)	ArXiv'25	✗	✗	✓	✓	✗	✗	✗	✗	2
Unified-Bench (Yan et al., 2025)	ArXiv'25	✗	✗	✗	✗	✗	✗	✗	✗	1
MetaQuery (Pan et al., 2025b)	ArXiv'25	✓	✗	✗	✗	✗	✗	✗	✗	-
UFO (Ours)	-	✓	✓	✓	✓	✓	✓	✓	✓	10

To address these limitations, we formulate **compositional multimodal reasoning** for UFM as a structured inference process over state transitions. Given the current state, a model first generates intermediate multimodal cues that specify the next state, and then answers a question defined over that future state by conditioning on the generated cues.

This formulation enables the study of **evidential coupling**, which is defined as the extent to which the intermediate cues provide grounded, state-consistent evidence that is jointly sufficient for answering questions about the future states. This approach unifies understanding and generation within a single inference trace, spans modalities by producing and consuming visual and textual cues together, and makes the intermediate evidence explicit rather than leaving it hidden inside a single-step prediction.

As summarized in Table 1, prior benchmarks rarely provide the ingredients needed to study evidential coupling. We therefore introduce **UFO (Unified FOundation)**, a benchmark for evaluating multimodal reasoning under explicit evidential coupling.

UFO is motivated by a world-state perspective. We organize evaluation into three state-centric settings, all formulated with respect to *future* states. **State Determination** infers the future state from current observations. **State Reconstruction** treats the future state as partially specified and requires recovering missing or implicit components. **State Augmentation** treats the future state as insufficiently informative and requires enriching it with additional state information. Across all settings, we probe evidential coupling by validating the generated multimodal cues and measuring the degree to which they jointly support question answering about future states. Correct answers require consistent evidence (i.e., causal contribution) from both textual and visual cues.

To make coupling interpretable and testable, we adopt a **two-step** protocol that explicitly separates cue generation from answer prediction. We then conduct **intermediate evalua-**

tion to assess whether the generated cues are consistent with the underlying state implied by the inputs. Because coupling can fail in multiple ways, we report **multidimensional metrics** that disentangle *relevance*, *consistency*, *causal utility*, *specificity*, and *compactness*. At scale, **hybrid evaluation** combines lightweight automatic checks with LLM-assisted judging, while **human annotations** are used to calibrate the judges and verify fine-grained cue correctness. To guard against spurious gains, **contamination checks** evaluate models without input images to detect potential information leakage. Finally, **filtering** removes examples that can be solved without relying on multimodal cues.

Our contributions are summarised as follows:

- 1 We introduce UFO, a state-centric benchmark for UFM designed to evaluate two-step reasoning across state determination, reconstruction, and augmentation, with explicit measures of evidential coupling based on cue consistency and causal contribution.
- 2 Our experiments indicate that most UFM fall short of being fully unified, as evidential coupling does not consistently yield performance improvements across state determination, reconstruction, and augmentation tasks in the twelve models evaluated.
- 3 We also observe that although the required information gain decreases from state determination through reconstruction to augmentation, this increased ease of evidential coupling does not consistently lead to improved UFM performance.

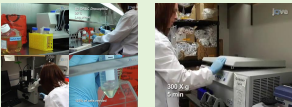
2. Related Works

Unified Foundation Models. The rapid progress of MLLMs (Bai et al., 2025; Dubey et al., 2024; Li et al., 2024a; Team et al., 2024; Yang et al., 2025a; Liu et al., 2024; 2025; Yang et al., 2025b; Zhang et al., 2025b; Yang et al.,

110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164

State Determination


Chemical



What specific laboratory instrument is required to execute the predicted next step?

A: A vortex mixer is required.
B: A hemocytometer (cell-counting chamber) is required.
C: An ultracentrifuge is required.
D: A centrifuge is required


Hybridisation



Identify the specific hybrid animal that results from breeding a male of the species shown in the second image with a female of the species shown in the first image. Based on the provided text, what was the name of the first cloned example of this hybrid to ever be born, and on what specific date did this event occur?

A: Idaho Gem; May 4, 2003.
B: Idaho Gem; May 5, 2003 (a hinny).
C: Idaho Gem; May 5, 2002.
D: Idaho Jewel; May 5, 2003


Multi-table



Among the departments currently listed in the Management table, calculate the 'Budget Allocation per Employee'. Identify the department with the lowest value for this metric. Return the names of the heads managing this specific department.

A: Dudley Hart, Franklin Langham,
B: K. J. Choi, Franklin Langham.
C: Jeff Maggert, Billy Mayfair,
D: Billy Mayfair

Multi-view




What is positioned next to the counter?

A: Towel rack,
B: Freestanding bathtub.
C: Bidet,
D: Toilet

State Reconstruction

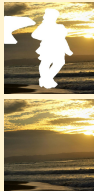
Exo-to-Ego



Analyze the interaction to determine whether the subject is depositing the item onto a stationary surface or fixture, adding contents to it, or transferring physical custody to another individual.

A: Wash the dishes in the sink.
B: Pour water into a cup.
C: Put the potted plant on the shelf.
D: Place the stainless bowl on the countertop


Inpainting







Upon restoring the scene, is the horizon line in the occluded center formed by open water or a continuous dark landmass?

A: A gap of open water between two separate headlands
B: Open water (from linear interpolation across the occluded center)
C: Continuous dark landmass.
D: Open water, because the dark band at the edges is just cloud shadow on the sea

Jigsaw

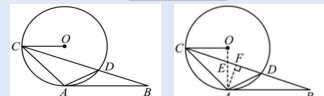


Which option correctly fills the missing area to complete the central skull's right horn?

A:  B: 
C:  D: 

State Augmentation

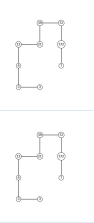
Geometry



In the figure, $\odot O$ is tangent to line AB at A and $OC \parallel AB$. Line BC intersects $\odot O$ at C and D . Let $E = OA \cap BC$, and let F be the foot of the perpendicular from A to BC . Using the radius of $\odot O$ and $\tan \angle ABC$ as indicated, find the length of AD .

A: $(12/5)/10$, B: $(12/6)/10$,
C: $(8/5)/5$, D: $(12/5)/5$

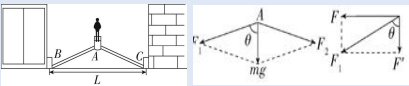
Logical



After the operation Replace the question-mark text inside the bottom-right circle with the number 240. Consider the sequence of numbers connected by lines. Let s define a function $f(n)$ as the sum of the digits of n . Starting with 3, apply $f(n)$ repeatedly until you reach a single-digit number. Then, multiply that single-digit number by 2 raised to the power of the number of times $f(n)$ was applied. Repeat this process for each number in the sequence. What is the sum of all the results?

A: 48, B: 46,
C: 42, D: 36

Physics



A student uses a simple hinged two-rod device to push a wardrobe: two identical rigid rods are pinned together at A and rotate freely; their other ends press against the wardrobe at B and a corner support at C . The student stands on A so the device bears his weight. Using the dimensions and given values in the diagram determine the horizontal thrust exerted by the wardrobe on the device at B ?

A: $mg/2L$, B: $[mg \tan(2h/L)]/2$,
C: $mgL/2H$, D: $mg(L^2+4h^2)/4h$

Figure 2. The UFO Benchmark for *Compositional Multimodal Reasoning*. UFO evaluates Unified Foundation Models (UFMs) by requiring them to generate *intervenable multimodal cues* as intermediate reasoning steps. The benchmark unifies three state-centric regimes: *State Determination* (inferring latent task-relevant variables), *State Reconstruction* (recovering information from partial observations), and *State Augmentation* (synthesizing auxiliary structural cues). Crucially, UFO moves beyond surface-level alignment to measure *evidential coupling*—quantifying the causal extent to which these intermediate cues are state-consistent and indispensable for the final prediction.

2025c) has driven the development of *Unified Foundation Models* (UFMs) that integrate multimodal understanding with generation via interleaved text-image interfaces. Early attempts such as Liquid (Wu et al., 2024b), Emu3 (Wang et al., 2024b), and Chameleon (Team, 2024) adopted unified tokenization (Kingma & Welling, 2013; Esser et al., 2021), but exposed an inherent *representation conflict*: tokens optimized for high-fidelity synthesis often fail to preserve the decision-critical semantics required for reasoning. Recent work addresses this tension either by strengthening unified interfaces (Tang et al., 2025; Qu et al., 2025) or, more commonly, by decoupling generation and understanding pathways (e.g., Janus (Ma et al., 2024; Chen et al., 2025b; Wu et al., 2024a), Show-o (Xie et al., 2025), and diffusion-based designs (Pan et al., 2025a; Wu et al., 2025a; Chen et al., 2025a)). These advances raise a central evaluation question: do interleaved intermediate cues encode *task-relevant, intervenable state* that can be monitored and reused across perception, generation, and reasoning, or are they merely plausible artifacts without causal utility?

Multimodal Reasoning Benchmarks. Multimodal evaluation has expanded from text-only reasoning (Cobbe et al., 2021; Hendrycks et al., 2021; Srivastava et al., 2022; Jin et al., 2023) to challenging domains including math/science (Lu et al., 2024; Wang et al., 2024a; Li et al., 2024c) and coding (Li et al., 2024b; Chen et al., 2024). However, a persistent validity threat is *modality redundancy*, where models exploit textual shortcuts to reach correct answers without substantive visual processing (Zhang et al., 2024; Yue et al., 2024b; Wang et al., 2025c). While process-oriented benchmarks like ROVER (Liang et al., 2025) and Uni-MMMU (Zou et al., 2025) record intermediate artifacts, they remain fundamentally *observational*: without controlled interventions, they cannot distinguish whether an intermediate cue encodes a **decision-critical state variable** or is a post-hoc hallucination. UFO addresses this limitation by organizing evaluation around state determination, reconstruction, and augmentation, and by introducing cue interventions to test cross-modal state consistency and counterfactual influence.

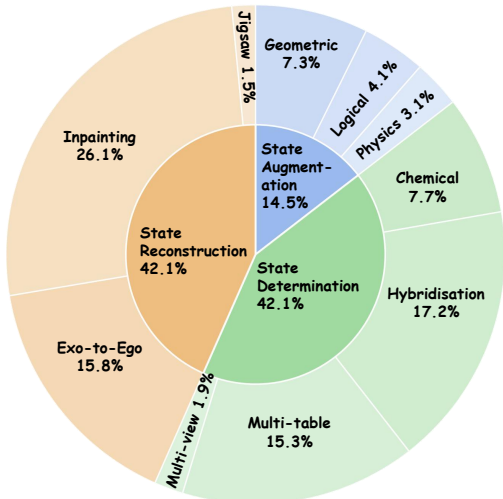


Figure 3. Composition of UFO.

Gap and Our Positioning. Taken together, fragmented UFM architectures and observational benchmarks create a blind spot: standard endpoint and similarity metrics cannot diagnose whether intermediate multimodal cues constitute *grounded evidence*—that is, whether they are *state-consistent* and *causally influential*. UFO bridges this blind spot with a **state-centric**, cue-based protocol that keeps the underlying problem structure fixed while intervening on the modality and content of evidence cues. By quantifying cross-modal state consistency and measuring prediction sensitivity under controlled interventions, our framework complements traditional leaderboards with a direct diagnostic of unified, state-consistent reasoning.

3. Benchmark

3.1. Overview of UFO

UFO is a **Unified FOundation** benchmark for **two-step compositional multimodal reasoning**. Each instance defines a *current* state through the input images and asks a question about a *future* state. A model first generates intermediate *textual and visual cues* that describe this future state. It then answers the question using the generated cues. UFO contains 3,936 curated questions spanning three types of state transition. These transition types differ in how the future state is constrained by the inputs, enabling us to examine evidential coupling across state determination, reconstruction, and augmentation, corresponding to decreasing levels of information gain. Across all three, the intermediate cues are treated as an explicit description of the future state, and evaluation checks (i) whether the cues are state-consistent and (ii) whether the final answer is actually justified by those cues.

State Determination covers settings in which the future

Table 2. Key statistics of UFO.

Statistic	Number	Percentage
Total questions	3,936	
- Multiple-choice questions	1,943	49.4%
- Free-form questions	1,993	50.6%
- Questions with answers	3,936	100.0%
Image in the question	3,936	100.0%
Problems with single images	1,182	30.0%
Problems with multiple images	2,754	70.0%
Maximum question length	756	
Maximum answer length	70	
Average question length	44.61	
Average answer length	2.90	

state is fully specified by the conditioning inputs, including the images and the question. All information required to answer the question is explicitly available, and the correct future state follows deterministically from this conditioning. The task therefore reduces to executing the implied state transition, rather than inferring missing content or introducing additional assumptions. This category includes tasks such as *Hybridisation*, where the offspring is uniquely determined by the observed parents; *Chemical*, which requires predicting the next experimental step from the current configuration; *Multi-table*, which involves deriving the relevant target table from multiple provided tables; and *Multi-view*, which requires integrating consistent information across different views of the same scene. These tasks provide a controlled baseline for evidential coupling, as intermediate cues should faithfully encode a uniquely determined future state grounded in the inputs.

State Reconstruction arises when the future state is only partially specified by the conditioning inputs. Although the images and the question constrain the state, some required information is absent due to occlusion, corruption, or limited viewpoint. The task is to recover the missing components of the future state while remaining strictly consistent with the available evidence. This category includes *Inpainting*, *Exo-to-Ego*, and *Jigsaw*, where successful reasoning depends on completing unseen regions, reconciling viewpoints, or assembling fragmented observations under strong contextual or geometric constraints. In this setting, evidential coupling tests whether intermediate cues recover information genuinely supported by the inputs, rather than introducing plausible but unsupported completions.

State Augmentation addresses cases in which the future state is correctly specified by the conditioning inputs, but answering the question requires making implicit constraints explicit. All entities and relations are observable, yet the raw state description is insufficient for verification, abstraction, or counterfactual reasoning. This category includes *Geometric*, *Logical*, and *Physics* tasks, which require ex-

Table 3. **Main results.** Comparison of different models’ performance on UFO across state determination, reconstruction, and augmentation multiple-choice question tasks. The bold font indicates the best performance among the direct, textual, visual, and joint schedules.

Model	# Params	State Determination				State Reconstruction				State Augmentation				Average			
		Direct	Textual	Visual	Joint	Direct	Textual	Visual	Joint	Direct	Textual	Visual	Joint	Direct	Textual	Visual	Joint
<i>Proprietary Unified MLLMs</i>																	
Gemini-3.0-pro _{withReasoning}	-	67.55	68.23	58.84	64.76	82.44	79.41	84.13	85.31	52.32	47.60	46.23	45.42	67.65	63.90	65.31	66.13
GPT-5	-	45.91	46.77	48.91	51.37	52.11	51.31	49.12	53.29	41.27	46.38	48.74	50.86	46.50	48.46	49.12	52.12
<i>Open-source Unified MLLMs</i>																	
UniPic1	1.5B	25.91	26.40	26.64	25.79	37.12	34.02	34.54	31.83	25.97	25.97	25.26	24.91	30.74	29.61	29.84	28.26
Ovis-U1	2.4 + 1B	22.56	22.80	19.72	19.82	44.11	30.00	39.00	39.22	27.72	30.88	26.67	31.23	32.59	26.99	29.03	29.84
OmniGen2	3B	34.03	33.33	34.30	33.92	51.52	37.41	50.36	61.18	37.19	35.44	34.74	35.44	42.02	35.39	41.28	45.88
Janus-Pro	1B	26.29	25.92	26.64	27.13	32.71	35.43	35.47	37.39	23.16	23.16	25.97	27.02	28.59	29.61	30.34	31.53
Bagel	7B MoT	33.28	33.10	35.02	35.20	52.86	52.83	55.33	50.11	38.95	34.74	37.89	35.79	42.54	41.83	44.18	41.70
EMU3	8.5B	25.18	25.06	25.06	26.16	32.32	25.84	23.24	21.16	25.26	23.15	21.40	22.11	28.26	25.12	23.74	23.41
Omni-R1	7B	25.67	24.90	24.94	25.06	28.33	29.07	28.45	29.02	24.56	20.42	22.10	23.50	26.65	26.04	26.03	26.53
UniCoT	7B	34.79	35.16	35.04	37.10	43.40	40.92	43.61	39.31	37.54	39.65	38.60	37.19	38.90	38.30	39.25	38.07
Janus-Pro	7B	28.39	28.28	27.94	29.64	37.80	40.07	38.37	36.87	20.00	20.35	21.76	22.46	31.21	32.19	31.52	31.70
UniWorld-V1	7B + 12B	32.48	32.97	26.64	35.28	50.91	51.08	47.72	48.48	32.98	35.44	30.18	33.68	40.49	41.12	36.23	40.73
UniPic2-Metaquery	9B	31.42	31.01	27.48	37.74	50.70	50.62	46.06	47.65	31.23	30.88	30.18	31.93	39.69	39.43	35.87	41.15

explicit relations, conditions, or quantities that are not directly expressed in the observations. The task is not to recover missing information, but to introduce auxiliary descriptions that enable principled reasoning.

Taken together, these settings differ in the information gain required to characterise the future state, spanning determination, reconstruction, and augmentation, while uniformly requiring both correct cue generation and answer prediction to assess evidential coupling. The key statistics of UFO are summarised in Table 2 and Figure 3.

3.2. Data Curation

Overall, UFO is curated through a two-stage process comprising image collection and question–answer generation, followed by LLM-assisted and human verification. Each instance includes a question paired with one or more images, together with a verified correct answer and corresponding ground-truth cues.

Stage I: Image Collection and Question–Answer Generation Based on the world state transition taxonomy, we identify three categories of tasks: state determination, state reconstruction, and state augmentation. We then survey publicly available datasets (Grauman et al., 2024; Yue et al., 2024a; Lin et al., 2014; Bhattad et al., 2025; Xu et al., 2025), wikipedia and select those that naturally support these transition types to construct UFO instances. As data availability varies across tasks, curation procedures differ slightly. We therefore describe the most complex case, Hybridisation, as a representative example, with the remaining tasks following the same general procedure and task-specific adaptations when information is missing or partially specified. In the Hybridisation task, hybridisation is defined as a compositional process in which two parent concepts combine to form a hybrid outcome. This notion extends beyond biological entities to include objects, sports, and cultural artifacts, such as a spork formed from a fork and a spoon. We use GPT-

5.1 to generate candidate hybridisation cases, which are cross-checked using Gemini 3 Pro. For each valid case, we retrieve the corresponding parent and offspring entries from Wikipedia, collecting images of the parents and the hybrid outcome, together with textual descriptions of the offspring. These offspring images and names serve as multimodal cues. Based on this material, we construct two-stage questions using GPT-5.1, first requiring identification of the hybrid outcome and then reasoning about it using the extracted text. To prevent shortcut reasoning, explicit references to parent and offspring names are masked using pronouns.

Stage II: LM-assisted and Human Verification. Ensuring question validity is a core design principle of UFO. Accordingly, we employ both human evaluation and LLM-assisted evaluation. Human evaluators verify that each question is properly constructed, with grounded evidential cues and a correct question–answer mapping, while LLM-assisted evaluation is used to detect and mitigate shortcut reasoning arising from legacy dataset biases. To eliminate legacy shortcuts, we regenerate QA pairs from raw media using diverse models (GEMINI-3-PRO, GPT-5.2), retaining only candidates with cross-model consensus. Crucially, we enforce *reasoning necessity* through a dual-layer adversarial stress test. First, to ensure *cue necessity*, we discard instances solvable by external validators (e.g., QWEN2.5-VL, GPT-5.1) via direct QA without intermediate steps. Second, to preclude **data contamination**, we implement a **blind-modality audit**: any instance solvable without visual input is flagged as parametric memorization and strictly excised. This guarantees that UFO evaluates grounded perception rather than latent knowledge or textual bias.

3.3. Comparison with Existing Benchmarks

Table 1 highlights a recurring pattern in prior multimodal benchmarks: they improve *individual* aspects of interleaved reasoning (e.g., exposing interleaving or enforcing visual

Table 4. **Main results.** Comparison of different models’ performance on UFO across state determination, reconstruction, and augmentation open question tasks. The bold font indicates the best performance among the direct, textual, visual, and joint schedules.

Model	# Params	State Determination				State Reconstruction				State Augmentation				Average			
		Direct	Textual	Visual	Joint	Direct	Textual	Visual	Joint	Direct	Textual	Visual	Joint	Direct	Textual	Visual	Joint
Proprietary Unified MLLMs																	
Gemini-3.0-pro _{withReasoning}	-	62.47	62.23	60.23	64.77	79.50	76.41	80.13	82.31	45.22	44.60	44.13	46.11	63.69	64.45	64.52	66.42
GPT-5	-	45.35	45.95	41.41	41.25	43.86	45.96	44.12	47.21	36.38	36.38	34.11	38.13	38.13	39.14	38.12	40.33
Open-source Unified MLLMs																	
UniPic1	1.5B	0.96	1.44	0.72	1.68	7.36	5.39	6.96	6.39	4.18	3.14	3.83	4.22	3.41	3.90	4.00	
Ovis-U1	2.4 + 1B	17.03	18.34	20.14	18.23	35.04	29.59	30.45	28.90	14.29	14.64	15.33	24.51	22.73	23.66	22.48	
OmniGen2	3B	10.67	9.35	8.99	8.87	29.20	32.32	27.32	30.51	14.01	11.15	14.63	11.15	19.26	19.66	17.82	18.67
Janus-Pro	1B	5.40	6.04	6.42	4.34	25.78	27.43	23.02	23.65	11.50	11.50	12.54	8.01	15.20	16.19	14.56	13.32
Bagel	7B MoT	8.87	8.87	8.87	7.79	36.59	31.19	37.02	29.67	16.73	14.98	15.68	13.59	22.13	19.52	22.17	18.20
EMU3	8.5B	3.72	3.24	3.24	3.48	22.09	18.36	20.19	18.39	5.57	4.88	6.62	4.53	12.02	10.09	11.14	10.15
Omni-R1	7B	9.32	8.35	8.41	9.80	28.72	30.92	26.31	27.64	4.83	8.29	8.05	6.54	17.16	18.22	16.19	17.13
UniCoT	7B	6.95	6.84	6.48	6.71	33.22	30.26	33.53	30.42	13.94	14.98	15.33	12.89	19.45	18.26	19.59	17.98
Janus-Pro	7B	5.16	5.40	4.79	5.16	29.24	26.71	26.77	25.45	9.06	11.85	10.80	10.45	16.26	15.65	15.27	14.80
UniWorld-V1	7B + 12B	26.14	26.14	25.64	27.28	34.55	30.13	31.13	29.76	14.29	13.94	14.64	12.20	28.11	26.13	26.46	26.19
UniPic2-Metaquery	9B	10.02	10.26	9.30	9.18	32.82	31.97	29.44	31.80	11.50	11.85	12.54	11.50	20.21	19.99	18.58	19.41

dependence), yet evaluation remains largely *observational*. As a result, intermediate cues can be fluent and well-aligned while remaining non-essential to the final decision. UFO reframes this by shifting the objective from “*getting the answer right*” to “*using the right state for the right reason*”. This causal perspective drives three key differentiators:

From Visual Dependence to Process Verification. Benchmarks that emphasize visual dependence ensure images matter in principle, but they rarely test *how* evidence is used along the way. UFO makes intermediate cues first-class objects via Process Evaluation: by enforcing cue necessity during curation and evaluating intermediate steps rather than just endpoints, success implies genuine state derivation rather than shortcut heuristics or post-hoc narration. **From Subject Collections to Regime-Level Diagnosis.** Standard subject-based groupings (e.g., Math, Physics) aggregate heterogeneous skills, thereby obscuring *why* models fail. UFO instead organizes tasks by **state operation**: determination, reconstruction, and augmentation. This regime-centric design turns a leaderboard into a diagnostic instrument, effectively localizing failures to distinct mechanisms (e.g., misextracted state vs. failed recovery) that topic-level scores cannot disentangle on their own.

From Holistic Accuracy to Multi-Source Cue Disentanglement. Standard benchmarks typically evaluate reasoning as a monolithic process, often masking the provenance of the solution. UFO disrupts this by structurally decomposing performance into four distinct inference pathways: *text-cued*, *image-cued*, and *joint-cued* reasoning, benchmarked against a *direct answer* baseline. This granular factorization allows us to isolate the specific contribution of each modality. Instead of merely checking the final output, we further verify whether success stems from genuine cross-modal synergy (where Joint > Unimodal) or reliance on single-modality priors, ultimately offering a precise audit of the model’s reasoning dependencies.

4. Experiment

4.1. Experimental Setup

Models. We evaluate twelve representative UFM in a zero-shot setting, including two proprietary models (Gemini-3.0-pro (Deepmind) and GPT-5.2 (OpenAI)) and ten open-source models. Based on their design paradigms and generation mechanisms, we group the open-source UFM into three architectural categories:

- **Pure autoregressive unified MLLMs.** These models employ a single autoregressive transformer trained with a unified next-token objective to support both multimodal understanding and image generation under one interface. A representative example is Emu3 (Wang et al., 2024b).¹
- **Hybrid unified models with a dedicated image generator.** These models combine an autoregressive reasoning backbone with a dedicated image generator, often diffusion- or DiT-based, to enhance visual fidelity while preserving strong instruction following and controllability. This group includes Ovis-U1 (Wang et al., 2025a), UniPic2-Metaquery (Wei et al., 2025), and OmniGen2 (Wu et al., 2025a).
- **Autoregressive unified models with architectural specialization.** These models remain autoregressive but introduce explicit specialization between understanding and generation, such as decoupled visual encodings, expert routing, or task-specific masking, while maintaining a unified interaction interface. This group includes Janus-Pro (Chen et al., 2025b), UniPic1 (Wang et al., 2025b), Bagel (Deng et al., 2025), Omni-R1 (Cheng et al., 2026), and UniCoT.

¹Some autoregressive unified models further introduce architectural specialization (e.g., decoupled visual encodings or expert routing). For clarity, we group such variants separately below.

Metrics: We use classification accuracy as the primary evaluation metric for all main experiments. Accuracy is appropriate in our setting since all benchmarks are formulated as single-answer tasks with unambiguous ground-truth.

4.2. Implementation Details

All experiments are conducted on NVIDIA H100 GPUs (80GB) using bfloat16 precision. To ensure fair comparison across models and pruning strategies, we fix the inference batch size and decoding hyperparameters for all settings, including temperature (0.2), top- k (1), diffusion timesteps (100), and text-guided scale (7.5). All remaining hyperparameters and implementation details strictly follow the official configurations released by the respective models.

4.3. Main Results

State-Centric Stratification and the Generative Gap. Consistent with the state-centric regimes defined in our framework, we observe a distinct performance hierarchy across Tables 3 and 4: *State Reconstruction* > *State Augmentation* > *State Determination*. This stratification empirically validates the theoretical difficulty implied by the state transitions: *State Reconstruction*, which involves recovering information from partial observations, benefits from tighter constraints on the hypothesis space, allowing models like OmniGen2 to reach 61.18% accuracy (MCQ). Conversely, *State Determination* proves most challenging, as it requires inferring latent future variables from current observations with high entropy. Crucially, evaluating these transitions reveals a severe *discriminative-generative gap*. While models maintain serviceable baselines in multiple-choice settings, performance collapses in open-ended generation (e.g., UniPic1 drops from 25.91% to 0.96% in Determination). This suggests that despite the promise of unified modeling, current UFM often rely on surface-level discrimination rather than constructing the robust, grounded world models required for explicit state generation.

The Challenge of Evidential Coupling. The central hypothesis of UFMs, that multimodal capabilities are mutually reinforcing, is challenged by our findings on *evidential coupling*. While intermediate cues (textual or visual) individually outperform direct inference in specific regimes, their combination does not consistently yield the expected additive gains. In grounded tasks like *State Reconstruction*, visual cues effectively serve as complementary evidence (e.g., OmniGen2 improves from 51.52% Direct to 61.18% Joint). However, in the more abstract *State Determination* and *Augmentation* regimes, the Joint schedule frequently trails the strongest single-modality baseline (e.g., UniCoT: 39.65% Textual vs. 37.19% Joint). These results indicate that simply producing outputs in a shared latent space does not guarantee effective coupling; instead, heterogeneous

Model	# Params	State Determination		State Reconstruction		State Augmentation	
		Textual	Visual	Textual	Visual	Textual	Visual
<i>Proprietary Unified MLLMs</i>							
Gemini-3.0-pro	-	41.42	59.17	33.52	71.51	22.55	38.49
GPT-5	-	44.96	47.13	24.91	72.41	51.90	12.45
<i>Open-source Unified MLLMs</i>							
UniPic1	1.5B	17.31	4.72	11.41	3.22	7.70	5.31
Ovis-U1	2.4 + 1B	12.35	6.31	9.50	4.43	12.10	2.30
OmniGen2	3B	19.13	4.21	17.5	5.22	11.1	5.19
Janus-Pro	1B	22.13	2.34	9.31	4.10	10.11	0.88
Bagel	7B MoT	21.30	17.41	19.31	13.55	22.47	11.41
EMU3	8.5B	12.35	5.71	10.39	7.21	14.12	4.66
Omni-R1	7B	12.80	8.08	14.52	6.90	14.13	3.84
UniCoT	7B	20.93	8.99	15.32	11.11	15.50	7.41
Janus-Pro	7B	17.99	13.45	14.46	9.04	10.63	3.11
UniWorld-V1	7B + 12B	13.13	9.74	9.73	7.41	16.22	4.12
UniPic2-Metaquery	9B	17.91	13.99	13.76	4.12	13.30	7.63

Table 5. Process-level cue accuracy.

cues can introduce interference or compete for attention. Consequently, achieving true evidential coupling, where intermediate cues function as jointly sufficient, grounded evidence, remains a non-trivial obstacle for current unified architectures.

4.4. Process Evaluation

To explain the limited gains from evidential coupling, Table 5 evaluates the quality of intermediate cues, revealing a severe imbalance between textual and visual fidelity. Models consistently articulate state transitions better than they instantiate them visually; for instance, the 1B Janus-Pro model achieves 22.13% textual accuracy versus only 2.34% for visual synthesis in State Determination. This disparity provides a structural explanation for the modality interference observed earlier, as low-quality visual cues act as high-entropy noise that disrupts rather than supports the inference process. Furthermore, the generally low absolute performance highlights a critical bottleneck in formulating UFMs as world models. While models can partially articulate future states in text, their ability to instantiate these predictions into a consistent visual reality remains nascent. Even for proprietary models like Gemini-3.0-pro, textual accuracy peaks at around 41%, indicating that the internal simulation of future states is far from perfect. These findings confirm that achieving effective evidential coupling requires more than shared latent spaces. It demands a fundamental improvement in the fidelity of intermediate generation to ensure that self-generated cues serve as reliable, high-quality supports for downstream compositional reasoning.

5. VLM as Judge

To establish a scalable and reliable evaluation pipeline, we construct a human-annotated validation set using intermediate cues generated by GEMINI-3-PRO and GPT-5.2. We employ this dataset to rigorously benchmark candidate VLM judges and scoring protocols, ensuring our automated metrics align with human consensus.

Judge Model	Visual-cue		Text-cue		Avg.	
	Acc. \uparrow	κ \uparrow	Acc. \uparrow	κ \uparrow	Acc. \uparrow	κ \uparrow
<i>Proprietary MLLMs</i>						
GPT-4o-mini	63.3	16.5	41.9	22.4	52.6	19.5
GPT-4o	66.5	23.6	33.8	31.4	50.2	27.5
GPT-5.1	61.6	23.4	37.9	28.3	49.8	25.9
Gemini-2.5-Flash	63.0	16.7	37.7	23.3	50.4	20.0
Gemini-2.5-Pro	64.4	22.1	38.0	22.8	51.2	22.5
Gemini-3-Flash	68.6	25.1	57.3	22.7	63.0	23.9
<i>Open-source MLLMs</i>						
Gemma-3-12B	56.6	10.9	48.8	20.0	52.7	15.5
Gemma-3-27B	65.8	18.6	62.2	18.2	64.0	18.4
Qwen2.5-VL-32B	50.5	13.8	69.7	18.7	60.1	16.3
Qwen2.5-VL-72B	55.1	15.6	44.1	31.0	49.6	23.3
Qwen3-VL-8B	63.1	20.2	45.5	17.8	54.3	19.0
Qwen3-VL-32B	69.1	27.2	46.4	26.6	57.8	26.9
Qwen3-VL-30B-A3B	59.6	14.9	53.6	21.8	56.6	18.4
Qwen3-VL-235B-A22B	60.6	19.7	47.2	28.1	53.9	23.9

Table 6. Alignment between candidate VLM judges and human labels under visual-cue and text-cue judging settings, categorized by proprietary and open-source models.

Judge Model	Methods	Visual-cue		Text-cue		Avg.	
		Acc. \uparrow	κ \uparrow	Acc. \uparrow	κ \uparrow	Acc. \uparrow	κ \uparrow
Qwen3-VL-32B	Binary	69.1	27.2	46.4	26.6	57.8	26.9
	Tournament	59.2	19.2	39.2	23.0	49.2	21.1
	Majority Voting	59.4	18.6	46.5	31.5	53.0	25.1
	Chain-of-Thought	57.5	20.0	49.1	25.1	53.3	22.6
	Confidence Weighted Voting	59.0	18.5	47.3	24.2	53.2	21.4
Gemini-3-Flash	Binary	68.6	25.1	57.3	22.7	63.0	23.9
	Tournament	50.5	15.2	49.9	24.1	50.2	19.7
	Majority Voting	47.6	11.7	59.0	29.1	53.3	19.7
	Chain-of-Thought	72.8	28.9	62.0	31.2	67.4	30.1
	Confidence Weighted Voting	49.5	13.6	58.3	24.4	53.9	19.0

Table 7. Alignment between candidate VLM judge methods and human labels under visual-cue and text-cue judging settings.

5.1. Meta-Evaluation of Judges

We identify the most capable judge models by benchmarking alignment with human annotations in Table 6. Among proprietary models, GEMINI-3-FLASH demonstrates the highest consistency, achieving a Kappa (κ) of 25.1 in visual evaluation. Notably, in the open-source landscape, QWEN3-VL-32B emerges as superior, outperforming significantly larger counterparts with a peak visual κ of 27.2. This indicates that parameter count does not strictly correlate with judging capability, which is likely due to better instruction-following in specific checkpoints.

Building on these selections, Table 7 further investigates optimal judging methodologies. We observe a modality-dependent preference: visual cues are robustly evaluated using Binary scoring, whereas textual cues, involving subtle semantic nuances, benefit significantly from Chain-of-Thought (CoT) reasoning. For instance, applying CoT to Qwen3-VL-32B improves textual alignment stability compared to binary scoring. Thus, we adopt QWEN3-VL-32B using a hybrid protocol (Binary for visual, CoT for text) as our primary open-source evaluation strategy.

Model	Params	Relevance	Faithfulness	Causal Util.	Specificity	Compactness	Average
<i>Proprietary Unified MLLMs</i>							
Gemini-3.0-pro	-	3.47	3.43	2.94	3.10	4.12	3.41
GPT-5	-	3.55	3.31	2.40	2.88	3.77	3.19
<i>Open-source Unified MLLMs</i>							
UniPic1	1.5B	1.48	2.35	1.45	1.43	1.75	1.69
Ovis-U1	2.4 + 1B	0.98	1.34	0.55	1.30	1.12	1.06
OmniGen2	3B + 4B	1.50	2.42	1.45	2.13	1.31	1.76
Janus-Pro	1B	1.32	1.12	0.93	1.36	1.04	1.15
Bagel	7B MoT	2.41	2.58	2.19	2.47	2.19	2.38
EMU3	8.5B	1.79	1.92	1.35	2.65	1.79	1.90
Omni-R1	7B	1.54	1.43	1.71	2.18	2.13	1.80
UniCoT	7B	1.32	2.41	1.39	1.77	2.11	1.80
Janus-Pro	7B	1.92	2.03	1.31	1.64	1.70	1.72
UniWorld-V1	7B + 12B	2.03	2.00	1.34	2.04	1.11	1.70
UniPic2-Metaquery	9B	2.01	2.16	1.95	2.53	2.31	2.19

Table 8. Evidence-oriented evaluation of visual cue generated.

5.2. Multi-dimensional Cue Analysis

To diagnose the breakdown of evidential coupling, we evaluate cue quality across five dimensions: *Relevance*, *Faithfulness*, *Causal Utility*, *Specificity*, and *Compactness* (rubrics in Appendix A.2.2). The results in Table 8 reveal a critical *semantic gap*: models consistently score higher on surface-level *Relevance* than on reasoning-centric *Causal Utility*. This suggests current UFMs function primarily as *associative retrievers* rather than grounded *world simulators*: producing generic scenes that are thematically related but lack the specific "decision hinge" required to resolve state uncertainty. Furthermore, universally low *Specificity* scores indicate that generated evidence is typically ambiguous or illustrative rather than decisive. Consequently, these vague visual outputs act as semantic noise. Instead of reinforcing the inference trace through grounded coupling, they dilute the model's attention, providing an explanation for the modality interference observed in our main results.

6. Conclusion

This work examines whether Unified Foundation Models effectively couple generation and understanding when reasoning over world states. We introduce UFO, a state centric benchmark for two-step compositional multimodal reasoning across state determination, reconstruction, and augmentation. Each instance includes question answering pairs with ground truth multimodal cues, allowing direct evaluation of whether intermediate cues are state consistent and causally relevant. Our results show that evidential coupling varies across models and tasks and does not consistently improve performance. Using human labeled cues and a vision language model as a judge, we further find that textual and visual contributions are often misaligned and can even degrade accuracy. By separating cue generation from answering, UFO exposes limitations of evaluation protocols that collapse reasoning into a single prediction and provides a diagnostic framework emphasising state consistency and causal contribution beyond aggregate accuracy metrics. We believe that UFO provides the community with a principled and comprehensive framework for exploring and evaluating Unified Foundation Models.

7. Impact Statement

Unified Foundation Models are increasingly expected to reason about hypothetical, future, or counterfactual states in domains where intermediate decisions matter. However, prevailing evaluation protocols often reduce reasoning to a single prediction, making it difficult to distinguish genuine evidence-based inference from shortcut behavior. By separating cue generation from answering and explicitly validating the causal role of intermediate multimodal evidence, UFO offers a principled means of diagnosing this distinction.

The primary impact of UFO lies in its ability to reframe how multimodal reasoning systems are evaluated rather than in optimizing any particular model architecture. By emphasizing state consistency, intervenability, and causal contribution of intermediate cues, the benchmark encourages the development of models whose reasoning processes are more transparent and auditable. We expect UFO to serve as a diagnostic tool for the community, enabling more fine-grained analysis of multimodal reasoning behaviors that are otherwise obscured by aggregate accuracy metrics.

References

- Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Bhattach, A., Preechakul, K., and Efros, A. A. Visual jenga: Discovering object dependencies via counterfactual inpainting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=yMXn86pzWx>.
- Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.
- Chen, N., Zhang, Y., Xu, J., Ren, K., and Yang, Y. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., and Ruan, C. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.
- Cheng, D., Li, Y., Ma, Z., Cai, H., Hu, Y., Wang, W., Nie, L., and Li, W. Omni-r1: Towards the unified generative paradigm for multimodal reasoning, 2026. URL <https://arxiv.org/abs/2601.09536>.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Deepmind, G. A new era of intelligence with gemini 3. <https://blog.google/products-and-platforms/products/gemini/gemini-3/>.
- Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv e-prints*, 2024.
- Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.
- Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.-J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Hareesh, S., Huang, J., Islam, M. M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K. J., Liu, J.-W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S. K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanov, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S. C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Damen, D., Engel, J., Farinella, G. M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C., Newcombe, R., Park, H. S., Rehg, J. M., Sato, Y., Savva, M., Shi, J., Shou, M. Z., and Wray, M. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.
- He, Q., Chen, X., Wang, C., Pan, Y., Hu, X., Gan, Z., Wang, Y., Wang, C., Li, X., and Zhang, J. Reasoning to edit: Hypothetical instruction-based image editing with visual reasoning. *arXiv preprint arXiv:2507.01908*, 2025.

- 495 Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart,
496 S., Tang, E., Song, D., and Steinhardt, J. Measuring math-
497 ematical problem solving with the math dataset. *arXiv*
498 *preprint arXiv:2103.03874*, 2021.
- 499 Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y.,
500 Zhou, J., Dong, C., Huang, R., Zhang, R., et al. Smartedit:
501 Exploring complex instruction-based image editing with
502 multimodal large language models. In *Proceedings of the*
503 *IEEE/CVF Conference on Computer Vision and Pattern*
504 *Recognition*, pp. 8362–8371, 2024.
- 506 Jin, Y., Ling, P., Dong, X., Zhang, P., Wang, J., and Lin,
507 D. Reasonpix2pix: instruction reasoning dataset for ad-
508 vanced image editing. *arXiv preprint arXiv:2405.11190*,
509 2024.
- 511 Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Zhiheng,
512 L., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan,
513 M., et al. Cladder: Assessing causal reasoning in lan-
514 guage models. In *Thirty-seventh conference on neural*
515 *information processing systems*, 2023.
- 516 Kingma, D. P. and Welling, M. Auto-encoding variational
517 bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- 519 Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang,
520 H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-
521 onevision: Easy visual task transfer. *arXiv preprint*
522 *arXiv:2408.03326*, 2024a.
- 524 Li, K., Tian, Y., Hu, Q., Luo, Z., and Ma, J. Mmcode:
525 Evaluating multi-modal code large language models with
526 visually rich programming problems. *arXiv preprint*
527 *arXiv:2404.09486*, 2024b.
- 528 Li, Z., Yang, X., Choi, K., Zhu, W., Hsieh, R., Kim, H., Lim,
529 J. H., Ji, S., Lee, B., Yan, X., et al. Mmsci: A multimodal
530 multi-discipline dataset for phd-level scientific compre-
531 hension. In *AI for Accelerated Materials Design-Vienna*
532 *2024*, 2024c.
- 534 Liang, Y., Chow, W., Li, F., Ma, Z., Wang, X., Mao, J.,
535 Chen, J., Gu, J., Wang, Y., and Huang, F. Rover: Bench-
536 marking reciprocal cross-modal reasoning for omnimodal
537 generation, 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2511.01163)
538 [2511.01163](https://arxiv.org/abs/2511.01163).
- 539 Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S.,
540 Yu, W., Wang, S., Ge, Y., et al. Uniworld: High-resolution
541 semantic encoders for unified visual understanding and
542 generation. *arXiv preprint arXiv:2506.03147*, 2025.
- 544 Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P.,
545 Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft
546 coco: Common objects in context. In *European Confer-*
547 *ence on Computer Vision*, 2014. URL [https://api.](https://api.semanticscholar.org/CorpusID:14113767)
548 [semanticscholar.org/CorpusID:14113767](https://api.semanticscholar.org/CorpusID:14113767).
- 549 Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao,
C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3
technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Liu, H., He, J., Jin, Y., Zheng, D., Dong, Y., Zhang, F.,
Huang, Z., He, Y., Li, Y., Chen, W., et al. Shotbench:
Expert-level cinematic understanding in vision-language
models. In *NeurIPS*, 2025.
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajjishirzi, H.,
Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Math-
vista: Evaluating mathematical reasoning of foundation
models in visual contexts. In *International Conference*
on Learning Representations (ICLR), 2024.
- Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan,
Z., Xie, Z., Zhang, H., Yu, X., Zhao, L., Wang, Y., Liu,
J., and Ruan, C. Janusflow: Harmonizing autoregression
and rectified flow for unified multimodal understanding
and generation, 2024.
- OpenAI. Introducing gpt-5.2. [https://openai.com/](https://openai.com/index/introducing-gpt-5-2/)
[index/introducing-gpt-5-2/](https://openai.com/index/introducing-gpt-5-2/).
- Pan, X., Shukla, S. N., Singh, A., Zhao, Z., Mishra, S. K.,
Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., Hou, J.,
and Xie, S. Transfer between modalities with metaqueries.
arXiv preprint arXiv:2504.06256, 2025a.
- Pan, X., Shukla, S. N., Singh, A., Zhao, Z., Mishra, S. K.,
Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.
Transfer between modalities with metaqueries. *arXiv*
preprint arXiv:2504.06256, 2025b.
- Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu,
C., Tan, Z., and Li, H. Uni-cot: Towards unified chain-
of-thought reasoning across text and vision, 2025. URL
<https://arxiv.org/abs/2508.05606>.
- Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y.,
Ye, H., Du, D. K., Yuan, Z., and Wu, X. Tokenflow:
Unified image tokenizer for multimodal understanding
and generation. In *CVPR*, 2025.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid,
A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A.,
Garriga-Alonso, A., et al. Beyond the imitation game:
Quantifying and extrapolating the capabilities of language
models. *arXiv preprint arXiv:2206.04615*, 2022.
- Tang, H., Xie, C., Bao, X., Weng, T., Li, P., Zheng, Y., and
Wang, L. Unilip: Adapting clip for unified multimodal
understanding, generation and editing. *arXiv preprint*
arXiv:2507.23278, 2025.
- Team, C. Chameleon: Mixed-modal early-fusion foundation
models. *arXiv preprint arXiv:2405.09818*, 2024.

- 550 Team, Q. et al. Qwen2 technical report. *arXiv preprint*
551 *arXiv:2407.10671*, 2024.
- 552 Wang, G.-H., Zhao, S., Zhang, X., Cao, L., Zhan, P., Duan,
553 L., Lu, S., Fu, M., Chen, X., Zhao, J., Li, Y., and Chen,
554 Q.-G. Ovis-u1 technical report, 2025a. URL <https://arxiv.org/abs/2506.23044>.
- 555 Wang, K., Pan, J., Shi, W., Lu, Z., Zhan, M., and Li, H. Mea-
556 suring multimodal mathematical reasoning with math-
557 vision dataset, 2024a.
- 558 Wang, P., Peng, Y., Gan, Y., Hu, L., Xie, T., Wang, X., Wei,
559 Y., Tang, C., Zhu, B., Li, C., Wei, H., Li, E., Song, X., Liu,
560 Y., and Zhou, Y. Skywork unipic: Unified autoregressive
561 modeling for visual understanding and generation, 2025b.
562 URL <https://arxiv.org/abs/2508.03320>.
- 563 Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang,
564 J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al. Emu3:
565 Next-token prediction is all you need. *arXiv preprint*
566 *arXiv:2409.18869*, 2024b.
- 567 Wang, X., Liu, J., Huang, C., Yu, X., Wang, Z., Sun, X.,
568 Wu, J., Yuille, A., Barsoum, E., and Liu, Z. Xmod-
569 bench: Benchmarking cross-modal capabilities and con-
570 sistency in omni-language models, 2025c. URL <https://arxiv.org/abs/2510.15148>.
- 571 Wei, H., Xu, B., Liu, H., Wu, C., Liu, J., Peng, Y., Wang,
572 P., Liu, Z., He, J., Xietian, Y., Tang, C., Wang, Z., Wei,
573 Y., Hu, L., Jiang, B., Li, W., He, Y., Liu, Y., Song, X., Li,
574 E., and Zhou, Y. Skywork unipic 2.0: Building kontext
575 model with online rl for unified multimodal model, 2025.
576 URL <https://arxiv.org/abs/2509.04548>.
- 577 Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu,
578 W., Xie, Z., Yu, X., Ruan, C., et al. Janus: Decoupling
579 visual encoding for unified multimodal understanding and
580 generation. *arXiv preprint arXiv:2410.13848*, 2024a.
- 581 Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y.,
582 Li, W., Jiang, X., Liu, Y., Zhou, J., et al. Omnigen2:
583 Exploration to advanced multimodal generation. *arXiv*
584 *preprint arXiv:2506.18871*, 2025a.
- 585 Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai,
586 S., and Bai, X. Liquid: Language models are scalable
587 and unified multi-modal generators. *IJCV*, 2024b.
- 588 Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu,
589 W., Schiele, B., Yang, M.-H., and Yang, X. Kris-bench:
590 Benchmarking next-level intelligent image editing mod-
591 els. *arXiv preprint arXiv:2505.16707*, 2025b.
- 592 Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q.,
593 Gu, Y., Chen, Z., Yang, Z., and Shou, M. Z. Show-o: One
594 single transformer to unify multimodal understanding and
595 generation. In *ICLR*, 2025.
- 596 Xu, Y., Wu, Y., Yu, J., Yan, Z., Jiang, T., He, Y., Zhao, Q.,
597 Chen, K., Qiao, Y., Wang, L., Okumura, M., and Wang, Y.
598 Expvid: A benchmark for experiment video understand-
599 ing reasoning, 2025. URL <https://arxiv.org/abs/2510.11606>.
- 600 Yan, Z., Lin, K., Li, Z., Ye, J., Han, H., Wang, Z., Liu, H.,
601 Lin, B., Li, H., Xu, X., et al. Can understanding and
602 generation truly benefit together—or just coexist? *arXiv*
603 *preprint arXiv:2509.09666*, 2025.
- 604 Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B.,
Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical
report. *arXiv preprint arXiv:2505.09388*, 2025a.
- Yang, L., Zeng, B., Liu, J., Li, H., Xu, M., Zhang, W.,
and Yan, S. Editworld: Simulating world dynamics
for instruction-following image editing. *arXiv preprint*
arXiv:2405.14785, 2024.
- Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang,
Y., Li, B., Qin, C., Lu, S., Li, X., and Bing, L. Longvt:
Incentivizing "thinking with long videos" via native tool
calling. *arXiv preprint arXiv:2511.20785*, 2025b.
- Yang, Z., Yuan, Y., Jiang, X., An, B., and Pang, W.
Inex: Hallucination mitigation via introspection and
cross-modal multi-agent collaboration, 2025c. URL
<https://arxiv.org/abs/2512.02981>.
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G.,
Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B.,
Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y.,
Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A
massive multi-discipline multimodal understanding and
reasoning benchmark for expert agi. In *Proceedings of*
CVPR, 2024a.
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G.,
Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A
massive multi-discipline multimodal understanding and
reasoning benchmark for expert agi. In *Proceedings of the*
IEEE/CVF Conference on Computer Vision and Pattern
Recognition, pp. 9556–9567, 2024b.
- Zhang, D., Jiang, C., Xu, R., Chen, B., Jin, Z., Lu, Y.,
Zhang, J., Yong, L., Luo, J., and Luo, S. Worldgen-
bench: A world-knowledge-integrated benchmark for
reasoning-driven text-to-image generation. *arXiv preprint*
arXiv:2505.01490, 2025a.
- Zhang, K., Wu, K., Yang, Z., Li, B., Hu, K., Wang, B.,
Liu, Z., Li, X., and Bing, L. Openmmreasoner: Pushing
the frontiers for multimodal reasoning with an open and
general recipe, 2025b. URL <https://arxiv.org/abs/2511.16334>.

605 Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P.,
606 Zhou, A., Lu, P., Chang, K.-W., Gao, P., et al. Mathverse:
607 Does your multi-modal llm truly see the diagrams in
608 visual math problems? *arXiv preprint arXiv:2403.14624*,
609 2024.

610 Zhao, X., Zhang, P., Tang, K., Li, H., Zhang, Z., Zhai, G.,
611 Yan, J., Yang, H., Yang, X., and Duan, H. Envisioning
612 beyond the pixels: Benchmarking reasoning-informed
613 visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

615 Zou, K., Huang, Z., Dong, Y., Tian, S., Zheng, D., Liu, H.,
616 He, J., Liu, B., Qiao, Y., and Liu, Z. Uni-mmmu: A
617 massive multi-discipline multimodal unified benchmark.
618 *arXiv preprint arXiv:2510.13759*, 2025.

619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659

Overview of the Appendix

A. Implementation Details

A.1. Model Architectures

Emu3 (Wang et al., 2024b) Emu3 employs a pure autoregressive architecture trained with a unified next-token prediction objective. By interleaving discrete visual codes (via a visual quantizer) with text tokens, it performs both understanding and generation within a single shared parameter space, eliminating the need for separate encoders or diffusion models.

Ovis-U1 (Wang et al., 2025a) Ovis-U1 utilizes a hybrid architecture that structurally aligns visual embeddings with language representations. It integrates an LLM with a dedicated image generator through a probabilistic mapping mechanism, leveraging the reasoning backbone for semantic guidance while decoding visual outputs via a specialized generator.

OmniGen2 (Wu et al., 2025a) OmniGen2 combines an autoregressive reasoning backbone with a diffusion-based generation module. It focuses on optimizing the interface between instruction following and visual synthesis by decoding the LLM’s latent representations into structural control signals, thereby balancing semantic controllability with generative fidelity.

UniPic2-Metaquery (Wei et al., 2025) UniPic2-Metaquery introduces a “Metaquery” mechanism within a hybrid framework to handle complex multimodal instructions. It explicitly separates reasoning from pixel-level synthesis, allowing the LLM to formulate structured queries that plan content layout before delegating rendering to a dedicated generator.

Janus-Pro (Chen et al., 2025b) Janus-Pro resolves the conflict between perception and generation through decoupled visual pathways. While sharing a single autoregressive transformer, it uses a SigLIP encoder for high-level semantic understanding and a separate VQ-tokenizer for fine-grained generation, ensuring optimal performance for both tasks without interference.

UniPic1 (Wang et al., 2025b) UniPic1 serves as a foundational autoregressive unified model, processing concatenated sequences of textual and visual tokens via a single transformer. It employs a standard VQ-VAE for visual tokenization, prioritizing a seamless unified interaction interface where understanding and generation are treated as equivalent sequence modeling tasks.

Bagel (Deng et al., 2025) Bagel addresses task interference through architectural specialization, such as Mixture-of-Experts (MoE) or task-specific routing. By dynamically allocating parameters for different modalities, it manages the gradient conflicts between understanding and generation losses, preserving LLM reasoning depth while enabling robust visual synthesis.

Omni-R1 (Cheng et al., 2026) Omni-R1 enhances the reasoning capabilities of unified models by integrating reinforcement learning strategies. It focuses on a “plan-then-generate” approach, ensuring that generated visual content is logically consistent with the multi-step deductions derived by the autoregressive backbone.

UniCoT (Qin et al., 2025) UniCoT integrates Chain-of-Thought (CoT) reasoning into the generative process. It trains the model to produce intermediate textual rationales before synthesizing visual tokens, thereby grounding image generation in explicit logical deductions and improving performance on multi-hop reasoning tasks.

UniWorld-V1 (Lin et al., 2025) UniWorld-V1 emphasizes world modeling by treating video or image sequences as state transitions. It employs a unified architecture trained on large-scale predictive tasks, learning to simulate future physical states based on current observations to serve as a predictive engine for dynamic environments.

A.2. Prompt

A.2.1. PROMPTS FOR RESPONSE GENERATION

Textual Cue Generation Prompt

My specific question is: {Question}

Your task is to generate a key textual cue based on the provided reference information. This textual cue should:

Identify key entities, attributes, and relationships pertinent to the query. Focus on critical visual features (e.g., structure, spatial layout, states) that drive reasoning. Ensure the response is concise and informative, limited to 1-2 sentences.

The generated text cue should serve as an abstract, language-only substitute for a visual cue, helping to answer the question above.

Visual Cue Generation Prompt

My specific question is: {Question}
Your task is to generate an image based on the provided reference images. This generated image should serve as a visual cue to help answer the question above.

A.2.2. PROMPTS FOR EVALUATION GENERATION

Multi-dim Cue Evaluation Prompt

Role. You are a strict evaluator of a model-generated Visual CoT (visual cue image) for a Visual QA task. Your goal is to judge whether the generated cue provides *usable visual evidence* to support the ground-truth answer, beyond answer correctness.

Inputs (ONLY these):

- Question: Q
- Ground-truth Answer: A_{gt}
- Generated Visual CoT image: I_{gen}

What to evaluate. Score the evidential quality of I_{gen} under five dimensions: (1) Evidence Relevance, (2) Faithfulness / Logical Consistency, (3) Causal Utility, (4) Localization / Specificity, (5) Non-redundancy / Compactness. Each dimension is scored as an **integer from 0 to 5**.

Hard rules.

- Judge evidence, not aesthetics.
- Only reward information that is *visually verifiable* in I_{gen} .
- If a factor is implied but not visually checkable, treat it as missing or weak.
- Penalize hallucinated or distracting content that could alter the decision.

Step 1: Identify decision-critical factors. From Q and A_{gt} , extract the minimal set of visual factors required to justify the answer (e.g., object identity, attribute, count, spatial relation, readable text). Mark each factor as **critical** (answer would flip if wrong) or **supporting**.

Step 2: Verify factors in I_{gen} . For each factor, assign one status: *verified* (clearly visible), *weak* (ambiguous/partial), *missing* (not shown), or *contradicted* (opposite evidence shown).

Step 3: Identify hallucinations or distractors. List any visually salient content that is not required by Q but may mislead the answer.

Output format (STRICT JSON ONLY).

```
{
  "decision_critical_factors": [
    {"factor": "...", "criticality": "critical|supporting",
     "status": "verified|weak|missing|contradicted", "why": "..."}
  ],
  "scores": {
    "relevance": 0,
    "faithfulness": 0,
    "causal_utility": 0,
    "specificity": 0,
    "compactness": 0
  },
  "failure_modes": [
    "missing_critical_evidence",
    "ambiguous_evidence",
    "contradiction_to_answer",
    "misleading_hallucination",
    "generic_scene"
  ],
  "notes": "Brief justification based only on visible evidence."
}
```

Multi-dim Cue Evaluation Scoring Rubrics Prompt

1. **Evidence Relevance.**

What: Coverage of **decision-critical visual factors** required to justify A_{gt} (only visually verifiable evidence counts).

Scoring (0–5):

- **0:** No decision-critical evidence.
- **1:** Thematic similarity only; no factor verifiable.
- **2:** Few factors weakly supported; most missing.
- **3:** Several factors verified, but at least one **critical** factor missing or ambiguous.
- **4:** Most critical factors clearly verified; minor gaps only.
- **5:** All or nearly all critical factors clearly and directly verified.

Rule: If any critical factor is missing or contradicted, relevance ≤ 3 .

2. **Faithfulness / Logical Consistency.**

What: Logical compatibility with Q and A_{gt} ; absence of contradictions or **misleading hallucinations**.

Scoring (0–5):

- **0:** Contradicts A_{gt} or strongly supports an alternative answer.
- **1:** Major inconsistency or hallucination likely flipping the decision.
- **2:** Inconsistency affecting a critical factor.
- **3:** Minor hallucinations, not affecting the decision.
- **4:** Fully consistent; only trivial extra details.
- **5:** Fully consistent; no competing or misleading content.

3. **Causal Utility (counterfactual proxy).**

What: Expected reduction of uncertainty relative to using Q alone; presence of a clear **decision hinge**.

Scoring (0–5):

- **0:** Misleading or harmful.
- **1:** No added evidence beyond the question.
- **2:** Weak confirmation; answer largely guessable.
- **3:** Meaningful support; uncertainty reduced.
- **4:** Clear decisive evidence resolving the ambiguity.
- **5:** Uniquely decisive evidence; cue is essential for confident resolution.

Rule: If no critical factor is clearly verifiable, causal utility ≤ 2 .

4. **Localization / Specificity.**

What: Precision and inspectability of evidence (not a generic illustration).

Scoring (0–5):

- **0:** Generic scene; queried evidence not checkable.
- **1:** Evidence implied only.
- **2:** Evidence present but poorly visible or ambiguous.
- **3:** Decisive evidence visible, partial ambiguity remains.
- **4:** Clear and specific; low ambiguity.
- **5:** Unmistakable, sharply localized decisive evidence.

5. **Non-redundancy / Compactness.**

What: Signal-to-noise ratio; focus on necessary evidence without competing distractors.

Scoring (0–5):

- **0:** Dominated by distractors or misleading content.
- **1:** Heavy clutter; evidence hard to isolate.
- **2:** Noticeable distractors competing with evidence.
- **3:** Some redundancy; evidence still salient.
- **4:** Clean composition; minimal distraction.
- **5:** Maximally compact; only decision-critical evidence shown.

A.3. Licenses

Table A.1 summarizes the licensing information of all models evaluated in this work. We rely exclusively on publicly released research models with official project pages or model cards that specify their usage terms. Several systems, including Ovis-U1, OmniGen2, Janus-Pro, Bagel, EMU3, and UniCoT, are distributed under the Apache-2.0 license, enabling broad academic use of their implementations. Other models, such as UniPic1, UniPic2-Metaquery, and UniWorld-V1, provide dedicated license files or usage statements within their repositories, which we follow as specified by the original authors. For Omni-R1, we adhere to the research-only usage terms described in its official model card. Across all cases, models are used solely for non-commercial academic evaluation, and no proprietary, closed-access, or restricted commercial resources are included.

Table A.1. License information for models evaluated in this paper.

Model	Official URL	License / Terms
UniPic1	Project Page	MIT License
Ovis-U1	Project Page	Apache-2.0 License
OmniGen2	Project Page	Apache-2.0 License
Janus-Pro	Project Page	MIT License
Bagel	Project Page	Apache-2.0 License
EMU3	Project Page	Apache-2.0 License
Omni-R1	Model Card	MIT License
UniCoT	Project Page	Apache-2.0 License
UniWorld-V1	Project Page	MIT License
UniPic2-Metaquery	Project Page	MIT License

B. More Experiments

B.1. Detailed Experimental Analysis

Disentangling answer accuracy from reasoning behavior. Tables B.1–B.4 report results for multi-choice questions under different inference protocols. Across models, direct question answering often yields competitive average accuracy, with several models differing by less than two points in overall performance. However, once intermediate cues are explicitly introduced, substantially larger task-level differences emerge. For example, on state reconstruction tasks such as inpainting, several open-source models improve from below 30 accuracy in the text-only setting (Table B.1) to above 40 when visual or joint cues are provided (Tables B.2, B.3). This contrast illustrates that similar answer accuracy under direct QA can mask qualitatively different reliance on intermediate evidence.

State-centric variation across tasks. A closer examination by task category shows that cue effectiveness aligns with the underlying state transition required by each task. For state determination tasks, such as chemistry and hybridisation, textual cues already provide a strong signal. For instance, UniWorld-V1 achieves 52.63 accuracy on chemistry with textual cues (Table B.1), while visual cues alone yield a lower score of 32.24 (Table B.2), indicating that abstract state variables are more readily conveyed through text. In contrast, for state reconstruction tasks such as jigsaw and inpainting, visual grounding becomes essential. Bagel improves from 41.30 on inpainting with text-only cues to 42.61 with visual cues (Tables B.1, B.2), and further stabilizes under joint cues (Table B.3), reflecting the necessity of recovering missing structure before answering.

Joint cues as evidence integration rather than score aggregation. Joint cue performance is not a trivial combination of text-only and visual-only results. Across multiple models, joint cues either yield modest but consistent gains or reduce variance across tasks. For example, UniWorld-V1 improves its average multi-choice accuracy from 26.67 under direct QA (Table B.4) to 28.33 with joint cues (Table B.3), despite similar averages under other protocols. At the same time, models such as Bagel show comparable average accuracy across protocols but exhibit pronounced task-level shifts once cues are introduced. These patterns indicate differences in how models integrate multimodal evidence, which remain largely invisible under direct answering alone.

Open-ended questions and robustness of cue reliance. Tables B.5–B.8 extend this analysis to open-ended questions. As expected, absolute accuracy is lower than in the multi-choice setting. Nevertheless, relative trends across inference protocols persist. For models that benefit from intermediate cues in the multi-choice setting, the transition to open-ended evaluation leads to smaller performance changes. For instance, OmniGen2 maintains an average accuracy of 18.03 under open-ended joint cues (Table B.7), compared to 20.00 under multi-choice joint cues (Table B.3), corresponding to a modest drop. In contrast, under direct QA, the same model decreases from 25.00 in the multi-choice setting (Table B.4) to 19.67 in the open-ended setting (Table B.8), indicating greater sensitivity to answer format when intermediate evidence is not explicitly leveraged.

Implications for benchmark design. Taken together, these results show that UFO enables a more discriminative evaluation of multimodal reasoning by making the role of intermediate evidence measurable. By comparing direct answering with

text-only, visual-only, and joint cue settings across both multi-choice and open-ended questions, the benchmark reveals systematic differences in how models construct, integrate, and rely on multimodal cues. These differences are not reliably reflected by aggregate accuracy alone, but become evident once inference is decomposed around explicit state transitions. As a result, UFO complements existing benchmarks by providing diagnostic insight into reasoning behavior that remains obscured under conventional evaluation protocols.

Table B.1. Main Result. Multi-choice question with text cue only. Note that for the Inpainting task, results are presented as two sub-metrics (Sub-task A + Sub-task B) to reflect its internal subsets.

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	28.29	25.08	27.0	26.65	27.39+30.00	43.23	20.00	27.78	25.93	21.67
Ovis-U1	2.4 + 1B	31.58	23.55	21.62	18.26	30.00	61.29	26.67	34.03	32.10	21.67
OmniGen2	3B + 4B	51.32	37.45	13.51	23.65	34.12+12.50	66.77	23.33	45.14	28.40	21.67
Janus-Pro	1B	30.26	30.12	10.81	21.86	26.96+28.33	49.35	16.67	23.61	22.22	23.33
Bagel	7B MoT	48.68	38.17	18.92	23.05	34.78+48.33	70.00	26.67	40.97	33.33	21.67
EMU3	8.5B	32.89	27.42	16.22	20.36	29.13+36.67	12.90	23.33	27.08	19.75	18.33
Omni-R1	7B	30.26	23.08	24.32	24.15	31.30 + 30.00	26.13	36.67	22.22	11.35	28.33
UniCoT	7B	51.97	39.13	27.03	24.85	30.43+40.00	49.03	26.67	43.06	49.38	18.33
Janus-Pro	7B	48.03	31.27	16.22	17.96	30.43+31.67	55.48	26.67	21.53	18.52	20.00
UniWorld-V1	7B + 12B	52.63	38.13	24.32	20.36	38.70+43.33	70.32	6.67	40.97	30.86	28.33
UniPic2-Metaquery	9B	52.63	33.33	16.88	20.66	37.83+41.67	69.03	30.00	36.11	29.63	20.00

Table B.2. Main Result. Multi-choice question in visual cue only

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic	1.5B	32.24	26.09	29.73	24.25	26.09+28.33	47.10	20.00	23.61	29.63	23.33
Ovis-U1	2.4 + 1B	37.50	17.76	18.92	13.47	27.83+45.00	40.32	23.33	34.03	25.93	10.00
OmniGen2	3B + 4B	51.32	41.79	21.62	21.26	34.71+50.00	62.26	20.00	41.67	35.80	16.67
Janus-Pro	1B	28.95	32.43	8.11	22.46	27.83+31.67	44.19	30.00	29.17	23.46	21.67
Bagel	7B MoT	53.29	39.77	29.73	23.05	42.61+58.33	62.26	23.33	44.44	37.04	23.33
EMU3	8.5B	32.24	27.76	27.03	19.16	29.13 +30.00	12.90	23.33	22.92	14.81	26.67
Omni-R1	7B	30.26	23.08	18.92	24.85	27.39+31.67	24.84	36.67	25.69	12.35	26.67
UniCoT	7B	51.32	40.13	21.62	24.55	37.83+43.33	49.35	20.00	47.22	37.04	20.00
Janus-Pro	7B	46.05	31.66	10.81	18.26	29.57+25.00	56.77	43.33	26.39	20.99	11.67
UniWorld-V1	7B + 12B	32.24	26.09	29.73	24.25	37.39+40.00	65.16	10.00	34.03	30.86	20.00
UniPic2-Metaquery	9B	56.58	22.22	17.31	20.06	36.96+43.33	55.48	30.00	37.50	23.46	21.67

Table B.3. Main Result. Multi-choice question in text cue and visual cue. Note that for the Inpainting task, results are presented as two sub-metrics (Sub-task A + Sub-task B) to reflect its internal subsets.

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	28.29	23.75	27.03	26.35	24.78+26.67	42.58	16.67	27.78	23.46	20.00
Ovis-U1	2.4 + 1B	34.21	17.37	2.70	17.37	31.30+33.33	52.26	13.33	36.11	34.57	15.00
OmniGen2	3B + 4B	48.03	39.00	11.43	25.45	32.94+75.00	66.13	20.00	43.06	33.33	20.00
Janus-Pro	1B	34.21	30.12	13.51	22.75	30.87+35.00	44.84	23.33	28.47	28.40	21.67
Bagel	7B MoT	40.79	40.93	27.03	28.44	35.22+50.00	60.97	23.33	41.67	35.80	21.67
EMU3	8.5B	31.58	28.09	29.73	21.56	21.74+26.67	14.84	23.33	22.92	20.99	21.67
Omni-R1	7B	24.34	23.75	18.92	27.25	27.83+31.67	25.81	40.00	22.22	27.13	21.67
UniCoT	7B	57.89	37.46	27.03	28.44	31.30+40.00	44.52	23.33	41.67	44.44	16.67
Janus-Pro	7B	47.37	31.66	16.22	21.26	30.00+31.67	47.10	26.67	24.31	20.99	20.00
UniWorld-V1	7B + 12B	53.95	36.45	24.32	26.95	38.26+43.33	63.23	10.00	37.50	30.86	28.33
UniPic2-Metaquery	9B	51.32	44.44	17.22	27.84	40.87+41.67	60.00	23.33	36.11	32.10	21.67

Table B.4. **Main Result.** Multi-choice question in direct qaNote that for the Inpainting task, results are presented as two sub-metrics (Sub-task A + Sub-task B) to reflect its internal subsets.

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	30.26	28.76	21.62	21.86	26.52+31.67	50.32	20.00	22.92	30.86	26.67
Ovis-U1	2.4 + 1B	31.58	23.55	16.22	18.26	28.70+38.33	61.61	13.33	33.33	29.63	11.67
OmniGen2	3B + 4B	50.66	41.70	27.03	20.36	34.12+50.00	65.81	18.87	43.06	35.80	25.00
Janus-Pro	1B	30.26	30.12	18.92	21.86	22.61+23.33	49.35	20.00	29.17	13.58	21.67
Bagel	7B MoT	48.68	39.00	16.22	23.05	34.35+48.33	70.32	26.67	47.22	34.57	25.00
EMU3	8.5B	32.89	27.42	18.92	20.36	29.13+ 33.33	34.19	20.00	28.47	20.99	23.33
Omni-R1	7B	30.26	23.08	32.43	25.15	30.87+28.33	26.13	36.67	22.22	23.46	31.67
UniCoT	7B	51.97	39.13	18.92	24.85	33.04+45.00	49.35	23.33	44.44	37.04	21.67
Janus-Pro	7B	48.03	31.66	15.41	17.96	28.26+26.67	56.13	23.33	23.61	23.46	6.67
UniWorld-V1	7B + 12B	52.63	38.13	13.51	20.36	36.09+43.33	70.97	10.00	35.42	33.33	26.67
UniPic2-Metaquery	9B	53.95	33.33	15.10	21.26	38.26+41.67	69.35	26.67	35.42	30.86	21.67

Table B.5. **Main Result.** open question with textual cue. Note that for the Inpainting task, results are presented as two sub-metrics (Sub-task A + Sub-task B) to reflect its internal subsets.

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	3.95	0.99	5.41	0.29	13.48+ 5.0	0.97	10.0	3.47	3.66	1.64
Ovis-U1	2.4 + 1B	15.13	22.19	21.62	16.03	44.35+33.33	17.74	23.33	12.50	9.76	26.23
OmniGen2	3B + 4B	14.47	12.58	13.51	3.79	50.0+40.0	15.16	16.67	9.03	10.98	16.39
Janus-Pro	1B	8.55	7.42	18.92	2.33	34.35+40.0	10.0	16.67	11.81	9.76	13.11
Bagel	7B MoT	16.45	10.6	8.11	4.08	41.30+38.33	19.68	6.67	14.58	9.76	22.95
EMU3	8.5B	3.29	2.98	16.22	2.04	32.17+20.0	8.71	20.0	2.78	3.66	11.48
Omni-R1	7B	10.26	7.55	9.46	8.1	35.22+38.33	21.94	10.0	5.0	7.32	17.38
UniCoT	7B	13.82	6.29	16.22	3.21	40.0+41.67	13.87	6.67	14.58	8.54	24.59
Janus-Pro	7B	10.53	4.64	21.62	2.04	35.65+34.35	13.87	16.67	9.03	7.32	24.59
UniWorld-V1	7B + 12B	17.76	38.13	16.22	20.36	40.0+ 36.67	18.71	13.33	13.19	12.20	18.03
UniPic2-Metaquery	9B	23.03	11.11	16.22	3.21	40.0+38.33	21.94	13.33	9.03	7.32	24.59

Table B.6. **Main Result.** open question with visual cue

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	0.66	1.32	2.7	0	10.87+10.0	1.29	6.67	3.47	2.44	6.56
Ovis-U1	2.4 + 1B	13.82	24.50	32.43	17.78	40.43+40.0	15.48	13.33	11.81	9.76	21.31
OmniGen2	3B + 4B	11.18	12.91	10.81	4.37	47.45+28.33	15.48	23.33	13.89	13.41	18.03
Janus-Pro	1B	7.24	10.12	13.51	2.04	33.04+28.33	10.32	33.33	11.11	9.76	19.67
Bagel	7B MoT	13.82	10.93	18.92	3.79	46.96+50.0	18.71	13.33	10.42	10.98	34.43
EMU3	8.5B	4.61	2.32	27.03	0.87	27.39+26.67	8.06	26.67	6.94	2.44	11.48
Omni-R1	7B	11.58	8.21	13.24	6.65	30.43+33.33	17.74	6.67	7.36	3.41	15.9
UniCoT	7B	12.50	6.62	10.81	3.21	42.61+46.67	15.48	10.0	14.58	12.20	21.31
Janus-Pro	7B	9.21	4.30	13.51	2.33	36.96+33.04	14.84	16.67	6.94	6.10	26.23
UniWorld-V1	7B + 12B	17.76	37.12	18.92	19.76	37.39+41.67	17.1	13.33	11.81	10.98	26.23
UniPic2-Metaquery	9B	15.13	11.11	18.92	4.08	36.67+37.39	17.74	13.33	11.11	6.10	24.59

Table B.7. **Main Result.** open question with jointly cue. Note that for the Inpainting task, results are presented as two sub-metrics (Sub-task A + Sub-task B) to reflect its internal subsets.

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	5.26	1.32	0.0	0.58	15.65+6.67	0.97	6.67	3.47	2.44	4.92
Ovis-U1	2.4 + 1B	12.50	21.19	32.43	16.62	44.35+33.33	16.77	13.33	13.19	8.54	29.51
OmniGen2	3B + 4B	13.82	12.58	13.51	2.92	47.84+36.67	15.48	13.33	10.42	7.32	18.03
Janus-Pro	1B	5.26	4.71	21.62	1.75	31.3+33.33	9.68		6.94	6.10	13.11
Bagel	7B MoT	12.50	10.93	10.81	2.62	40.0+35.0	19.03	16.67	13.19	9.76	19.67
EMU3	8.5B	5.92	1.66	29.73	1.17	31.74+20.0	9.03	20.0	2.08	2.44	13.11
Omni-R1	7B	16.32	8.58	14.05	7.52	36.09+31.67	19.68	13.33	2.83	6.10	15.9
UniCoT	7B	11.84	7.28	13.51	3.21	40.87+40.0	15.81	6.67	13.19	6.10	21.31
Janus-Pro	7B	9.21	3.64	18.92	3.21	35.65+31.3	13.87	16.67	8.33	3.66	24.59
UniWorld-V1	7B + 12B	13.82	35.79	16.22	26.95	40.43+36.67	17.42	13.33	10.42	9.76	19.67
UniPic2-Metaquery	9B	16.45	11.11	16.22	3.50	40.87+40.0	19.35	10.0	9.03	9.76	19.67

Table B.8. **Main Result.** open question in direct answer without any cue. Note that for the Inpainting task, results are presented as two sub-metrics (Sub-task A + Sub-task B) to reflect its internal subsets.

Model	# Params	State Determination				State Reconstruction			State Augmentation		
		Chemistry	Multi-table	Multi-view	Hybridisation	Inpainting	Exo-to-Ego	Jigsaw	Geometric	Logical	Physics
<i>Open-source Unified MLLMs</i>											
UniPic1	1.5B	1.32	1.32	2.7	0.29	11.74+11.67	0	6.67	3.47	4.88	4.92
Ovis-U1	2.4 + 1B	15.79	20.53	16.22	14.58	47.39+46.67	17.10	10.0	13.89	9.76	21.31
OmniGen2	3B + 4B	17.11	14.24	8.11	4.95	43.33+35.0	16.45	10.87	13.33	10.98	19.67
Janus-Pro	1B	8.55	6.64	13.51	2.04	38.33+31.67	11.29	33.33	11.11	7.32	18.03
Bagel	7B MoT	16.45	10.6	10.81	3.79	44.78+48.33	20.97	10.0	12.5	10.98	34.43
EMU3	8.5B	3.29	2.98	18.92	2.92	38.26+28.33	6.45	16.67	2.08	3.66	16.39
Omni-R1	7B	10.92	6.89	12.16	10.44	32.17+36.67	19.68	6.67	2.83	2.68	12.46
UniCoT	7B	14.47	6.62	10.81	3.50	42.17+48.33	13.23	6.67	12.50	8.54	24.59
Janus-Pro	7B	10.53	4.64	13.51	2.33	43.04+36.96	13.23	23.33	6.25	7.32	18.03
UniWorld-V1	7B + 12B	17.76	38.13	16.22	20.36	44.35+45.0	19.03	13.33	11.11	9.76	27.87
UniPic2-Metaquery	9B	20.39	11.11	16.22	3.79	40.87+38.33	23.87	13.33	8.33	6.1	26.23

B.2. Example of Data

Figures C.1–C.10 collectively reveal a systematic gap in how current multimodal benchmarks assess reasoning. Across all ten cases, the dominant failure mode is not a lack of perceptual recognition or linguistic competence, but the inability to construct *intermediate evidence* that is both state-consistent and causally actionable.

In process-driven domains such as chemistry and physics (Figs. C.1, C.10), correct answers are only reachable when latent transition variables or causal trajectories are explicitly encoded in the cue. Models that bypass this step can still appear competent under one-shot evaluation, yet fail immediately once the reasoning is decomposed. Similarly, in symbolic and relational settings, including geometry, logical graphs, and multi-table reasoning (Figs. C.3, C.7, C.8), the decisive factor is whether the cue exposes hidden structure, enabling subsequent operations to be performed on a well-defined intermediate state rather than on memorized patterns.

Vision-centric tasks further stress the necessity of evidential coupling. Hybridisation, multi-view, and exo-to-ego reasoning (Figs. C.4, C.9, C.2) demonstrate that visually plausible generations are insufficient if the produced cues fail to preserve discriminative attributes aligned with the textual constraint. Likewise, reconstruction-based tasks such as inpainting and jigsaw reasoning (Figs. C.5, C.6) expose models that recover visually coherent content but omit the specific spatial or compositional evidence required for answering.

Taken together, these cases show that conventional answer-only accuracy systematically overestimates reasoning ability by conflating *shortcut inference* with evidence-based reasoning. UFO directly targets this blind spot by enforcing a two-step protocol and validating the causal contribution of intermediate multimodal cues. As a result, the benchmark does not merely ask whether a model can answer correctly, but whether it can *justify that answer through grounded, intervenable evidence*. This property is essential for treating UFM as reliable world models rather than black-box predictors.

Chemical reasoning question in UFO



Input Images

GT Image cue

Question: If the predicted next step shown in the protocol is omitted, what specific reaction property can no longer be calculated?

- A: The reaction order with respect to the limiting reagent,
- B: The average perfusion/flow rate through the system (mL/min),
- C: The activation energy (E_a) via an Arrhenius analysis",
- D: The reaction rate or kinetics.

Answer: D

Ground Truth Text Cue: Record completion time

Reasoning with Textual cue

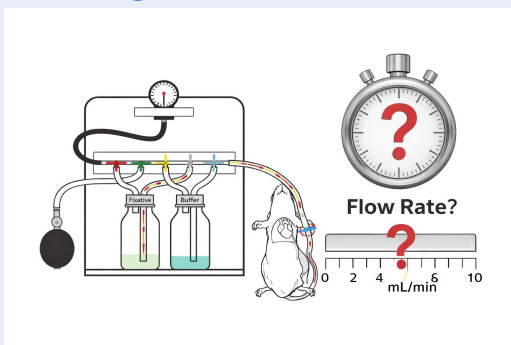
text_cue_generated_gpt:

If the protocol omits the predicted "next step" (the subsequent elementary reaction/transition state), you cannot calculate the reaction rate/kinetics—specifically the activation energy (ΔG^\ddagger) and the associated rate constant—because the energy barrier between current and next states is undefined.

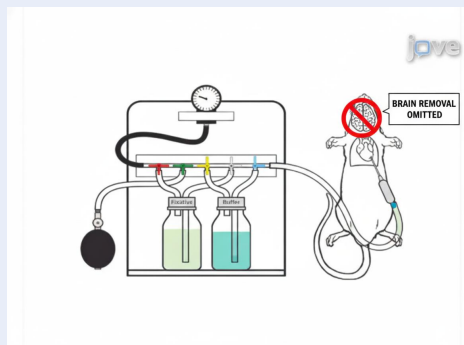
text_cue_generated_gemini:

Based on the provided images, the key textual cue is: The visual evidence features a digital timer and a data sheet with specific fields for recording the start and end times of the perfusion steps, highlighting the precise monitoring of the procedure's duration. This temporal measurement is essential for calculating the flow rate of the fluids administered during the reaction.,

Reasoning with Visual cue



GPT Generated Visual Cue



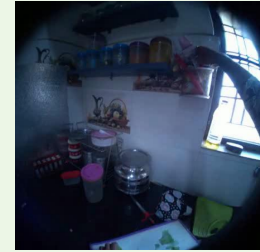
Gemini Generated Visual Cue

Figure C.1. Case 1: Chemical reasoning (state determination). The question is answerable only when the intermediate cue correctly encodes the task-relevant reaction attribute, illustrating how cue quality governs downstream answerability under the two-step protocol.

Exo-to-Ego reasoning question in UFO



Input Images



GT Image cue

Question: Analyze the physical points of contact in the scene to identify whether the task is being executed by the left hand, the right hand, or the simultaneous use of both hands.

A: Placing cups on the rack,

B: Cutting vegetables on a board,

C: Put away the peeler in the utensil basket,

D: Pouring water into a cup

Answer: C

Ground Truth Text Cue: The scene shows a vegetable-chopping task that requires two-handed coordination: the left arm (identified by a pink watch strap) stabilizes the vegetable on the cutting board while the right hand grips the knife or reaches for items. Both hands are actively engaged, making the task inherently bimanual rather than one-handed.

Reasoning with Textual cue

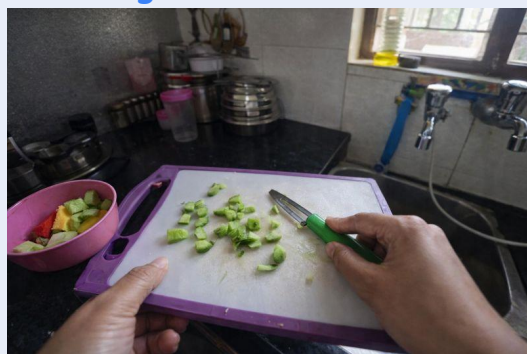
text_cue_generated_gpt:

Determine which hand(s) directly touch and manipulate the target object: note contact points (fingers/palm) and whether one hand alone grips/presses while the other is off-object. If both hands simultaneously hold/steady opposite sides or perform coordinated actions, classify as both hands.

text_cue_generated_gemini:

The left hand stabilizes the vegetable on the cutting board, while the right hand grips the knife to slice it.

Reasoning with Visual cue



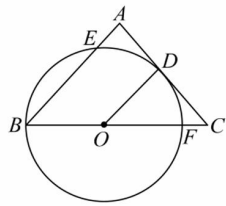
GPT Generated Visual Cue



Gemini Generated Visual Cue

Figure C.2. **Case 2: Exo-to-ego reasoning (viewpoint-conditioned action).** The instance requires grounding the decision in physical contact and hand-object interaction, where subtle visual evidence must be carried into the cue to support a correct classification.

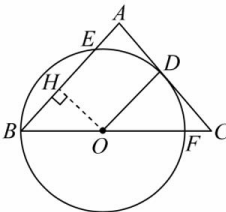
Geometry reasoning question in UFO



Input Images

Question: In $\triangle ABC$ with $\angle A = 90^\circ$, points O and D lie on BC and AC respectively and satisfy $CD \cdot BC = AC \cdot OC$. With center O and radius OD , draw a circle passing through B , intersecting AB at E and BC at F . Using the given segment lengths marked for AE and CF , find the inradius r of $\triangle ABC$.

A: 10 B:8 C:12 D:9



GT Image cue

Answer: D

Ground Truth Text Cue: Draw $OH \perp AB$ through point O , and draw the vertical foot as H

Reasoning with Textual cue

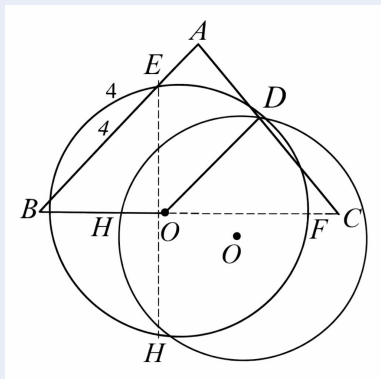
text_cue_generated_gpt:

Right $\triangle ABC$ with $\angle A=90^\circ$ ($AB \perp AC$); O on hypotenuse BC , D on AC with condition $CD \cdot BC = AC \cdot OC$. Circle centered at O with radius OD passes through B , meets AB at E and BC at F ; drop $OH \perp AB$. Given $AE=4$ and $CF=5$, find inradius r .

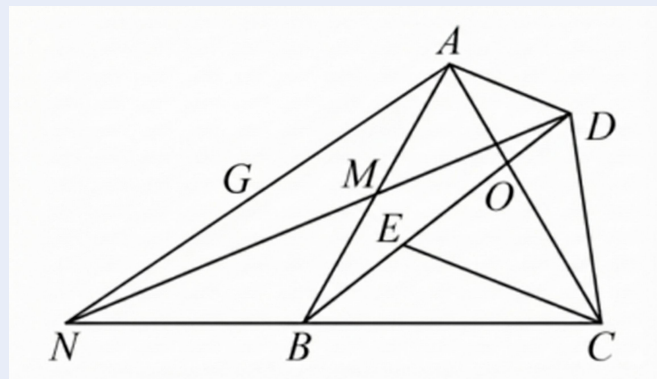
text_cue_generated_gemini:

The image depicts a right-angled triangle ABC with a circle centered on the hypotenuse BC that passes through vertex B and is tangent to leg AC at point D . Key geometric features include the construction of $OH \perp AB$, which forms a rectangle with the radius OD and bisects the chord on AB , thereby connecting the given segment lengths $AE=4$ and $CF=5$ to the circle's radius.

Reasoning with Visual cue



GPT Generated Visual Cue



Gemini Generated Visual Cue

Figure C.3. Case 3: Geometry reasoning (constructive invariants). The cue must specify the missing auxiliary construction (e.g., a perpendicular drop) that makes the solution trace executable, turning implicit geometric structure into explicit intermediate evidence.

Hybridisation reasoning question in UFO



Input Images



GT Image cue

Question: Based on the visual characteristics of the two feline varieties depicted in the images, specifically the spotted coat of the first and the distinct ear structure of the second, what hybrid breed results from their cross? regarding the breed identified in the first part, what is the stated ultimate aim of its development according to the text?

- A: The ultimate aim is to produce a cat that looks similar to a serval by introducing recent serval blood and then breeding back to domestic lines.",
- B: The ultimate aim is to produce a cat with a leopard-like spotted coat that is smaller and more manageable than a Bengal, with no focus on serval resemblance.,
- C: The ultimate aim is to standardize an extreme Oriental-type head and ear set on a spotted pattern, prioritizing show-ring silhouette over resemblance to any particular wild species.,
- D: The ultimate aim is to produce a cat that looks similar to a serval, without using any recent wild cat blood.

Answer: D

Ground Truth Text Cue: Serengeti

Reasoning with Textual cue

text_cue_generated_gpt:

The first cat has a golden, leopard-like rosetted/spotty coat (Bengal), and the second shows prominent outward-curved ears (American Curl); crossing them yields a Bengal Curl hybrid. The Bengal' s stated ultimate development aim is to create a domestic cat with an exotic "wild" spotted look but a gentle pet temperament

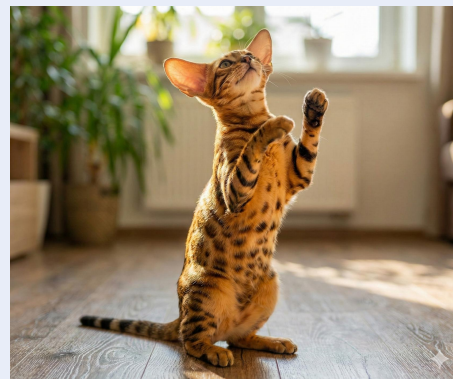
text_cue_generated_gemini:

The images depict a Bengal cat, identified by its leopard-like rosetted coat, and an Oriental Shorthair, distinguished by its large, triangular ears. Crossing these two breeds produces the Serengeti cat, a hybrid developed with the ultimate aim of resembling the wild Serval without incorporating wild bloodlines.

Reasoning with Visual cue



GPT Generated Visual Cue



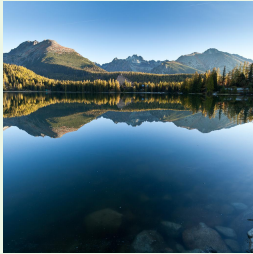
Gemini Generated Visual Cue

Figure C.4. Case 4: Hybridisation reasoning (attribute composition). The question couples fine-grained visual attributes with a text-described goal, so the intermediate cue must preserve discriminative appearance cues while aligning them with the referenced description.

Inpainting reasoning question in UFO



Input Images



GT Image cue

Question: After restoring the occluded center section to reveal the full structure on the shoreline, what 2D geometric shape best describes the front profile of this building?

- A: Triangle,
- B: Parallelogram,
- C: Trapezoid,
- D: Diamond

Answer: A

Ground Truth Text Cue: The large building located at the center of the shoreline, nestled in the trees.

Reasoning with Textual cue

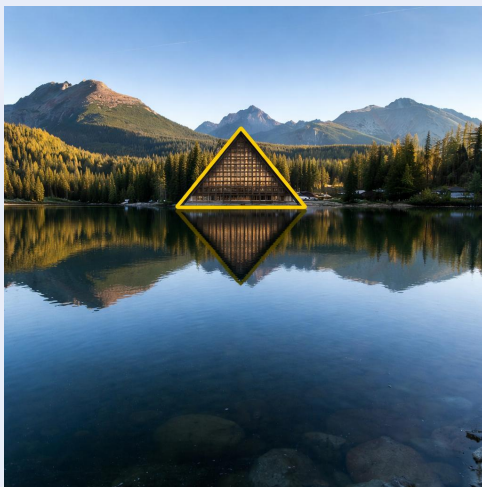
text_cue_generated_gpt:

A shoreline building appears symmetrical; when the occluded middle is restored, its front face forms a broad “A” silhouette with two slanted sides meeting at a peak, like a triangular gable. The overall front profile is best described as a triangle.

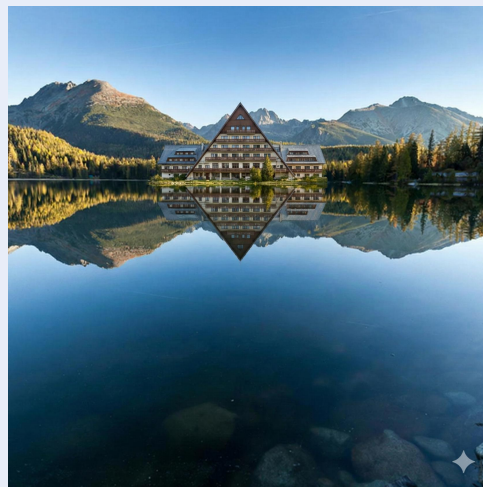
text_cue_generated_gemini:

The building on the shoreline features a steeply pitched roof with sides that slope upward to converge at a sharp central apex, creating a distinctive triangular profile that is clearly mirrored in the lake's reflection.

Reasoning with Visual cue



GPT Generated Visual Cue



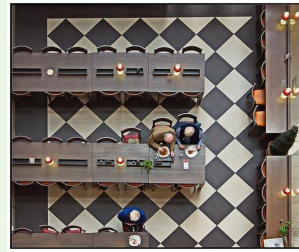
Gemini Generated Visual Cue

Figure C.5. Case 5: Inpainting-based reasoning (reconstruction). Correct answering depends on recovering the occluded structure and summarizing its salient geometric profile in the cue, revealing whether reconstruction contributes causally to the final decision.

Jigsaw reasoning question in UFO

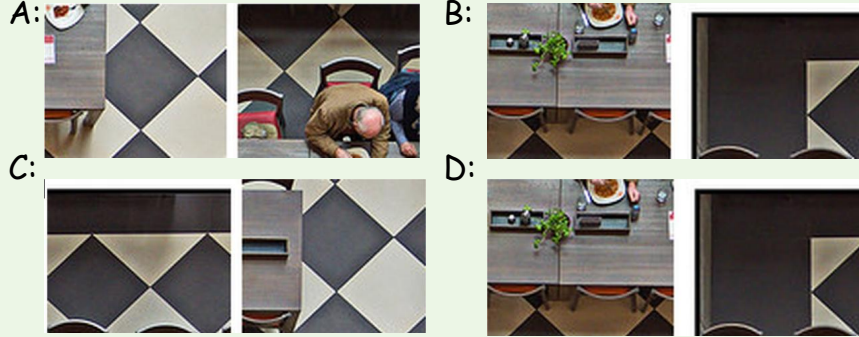


Input Images



GT Image cue

Question: Which option correctly fills the missing areas, determining from the reference whether the concealed table section contains a plant, a person, or neither?



Answer: C

Ground Truth Text Cue: The large building located at the center of the shoreline, nestled in the trees.

Reasoning with Textual cue

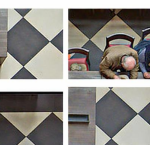
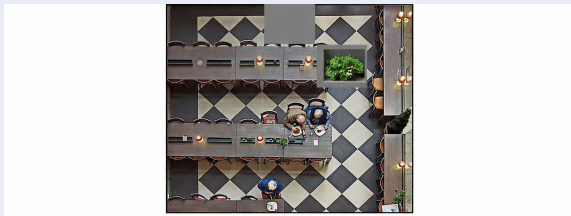
text_cue_generated_gpt:

The building on the shoreline features a steeply pitched roof with sides that slope upward to converge at a sharp central apex, creating a distinctive triangular profile that is clearly mirrored in the lake's reflection.

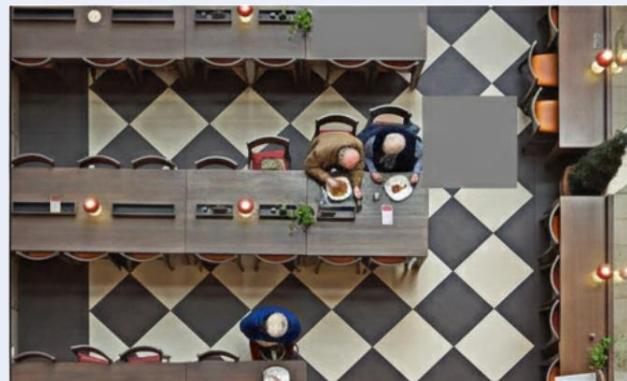
text_cue_generated_gemini:

Based on the visual evidence in the reference image, here is the textual cue: Observe the vertical column of tables in the center of the layout; since the visible tables in the rows below each feature a small plant, the concealed table section in the top row should follow this pattern and also contain a plant.

Reasoning with Visual cue



GPT Generated Visual Cue

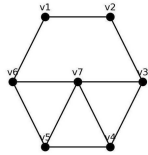


Gemini Generated Visual Cue

Figure C.6. Case 6: Jigsaw reasoning (compositional completion). The task requires selecting the correct missing patch by matching local visual patterns, where the cue must encode the specific spatial correspondence needed for the completion.

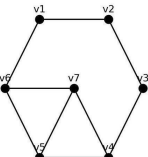
Logical reasoning question in UFO

Graph Representation of Vertices and Edges



Input Images

Graph Representation of Vertices and Edges



GT Image cue

Question: After the operation Remove the single line segment connecting the central node v7 to the right mid-level node v3. Calculate the ratio of the sum of degrees of all vertices with a degree strictly greater than 2, to the total number of edges in the graph.

- A: 10/9,
- B: 3/2,
- C: 4/3,
- D: 5/3

Answer: C

Ground Truth Text Cue: Remove the single line segment connecting the central node v7 to the right mid-level node v3.

Reasoning with Textual cue

text_cue_generated_gpt:

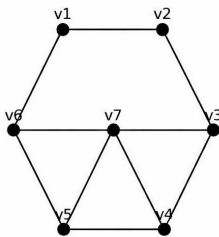
Graph has 7 labeled vertices with a central node v7 connected to all six outer nodes; additionally, outer nodes form a ring with edges v1-v2-v3-v4-v5-v6-v1. Remove edge v7-v3; then degrees are v7=5, v3=2, others=3, with total edges 11.

text_cue_generated_gemini:

The graph consists of a hexagonal cycle (v1-v2-v3-v4-v5-v6-v1) with a central vertex v7 connected to vertices v3, v4, v5, and v6. This topology results in a total of 10 edges, where v7 initially has a degree of 4, vertices v3 through v6 have a degree of 3, and vertices v1 and v2 have a degree of 2.

Reasoning with Visual cue

After removing the single line segment connecting v7 to v3, calculate the ratio of the sum of degrees of all vertices with a degree strictly greater than 2 to the total number of edges in the graph.



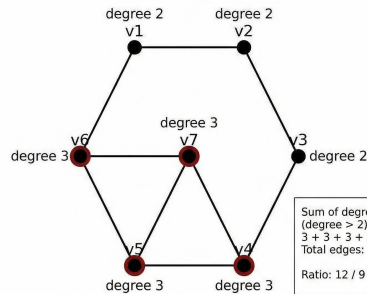
Sum of degrees of vertices with degree > 2:
deg(v1) + deg(v2) = 3 + 3 = 6

Total number of edges in the graph:

$$\frac{6}{8} = \frac{3}{4}$$

GPT Generated Visual Cue

Graph Representation of Vertices and Edges



Sum of degrees (degree > 2):
3 + 3 + 3 + 3 = 12
Total edges: 9
Ratio: 12 / 9 = 4/3

Gemini Generated Visual Cue

Figure C.7. Case 7: Logical graph reasoning (state transition via edit). The cue formalizes the intervention (edge removal) and its structural consequences, enabling the answer step to depend on updated degrees/edges rather than a shortcut guess.

MultiTable reasoning question in UFO

Table 1: People

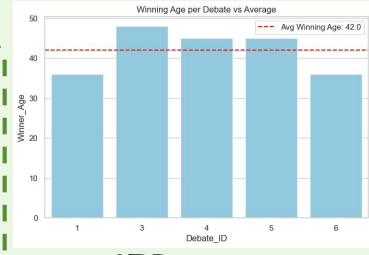
Person_ID	Name	Gender	Prof	Age
1	Charles Van Wyck	Male	Debate	36
2	George Briggs	Male	Debate	42
3	Thomas J. Barr	Male	Debate	48
4	William B. Maclay	Male	Debate	36
5	Horace F. Clark	Male	Debate	45

Table 2: Debate

Debate_ID	Date	Winner	Num_of_Audience
1	October 21, 2011	Negative	362
2	December 17, 2014	Affirmative	134
3	August 3, 2015	Affirmative	90
4	October 27, 2015	Negative	259
5	January 12, 2017	Negative	313
6	January 21, 2017	Affirmative	159

Table 3: Debate People

Debate_ID	Affirmative	Negative	If_Affirmative_Win
1	1	10	F
5	2	8	F
3	4	7	T
6	5	6	T
4	5	8	F



Input Images

GT Image cue

Question: Compute the average age of the winning speaker for all debates with recorded outcomes. Identify the dates of the debates where the Negative side won and the Negative speaker's age was strictly greater than this calculated average winning age.

- A: January 12, 2017 and October 27, 2015
 - B: August 3, 2015 and January 12, 2017,
 - C: October 27, 2015,
 - D: December 17, 2014 and October 27, 2015
- Answer: A

Ground Truth Text Cue: Merge tables to associate speaker ages, calculate the average winner's age (42.0), and retrieve the dates for debates where the Negative winner was strictly older than that average.

Reasoning with Textual cue

text_cue_generated_gpt:

Locate debates with a recorded winner; for each, use the winning side's speaker age to compute the overall average winning age. Then filter rows where winner = "Negative" and Negative speaker age > that average, and report the corresponding debate dates.

text_cue_generated_gemini:

Cross-reference the Debate People and People tables to determine the ages of winning speakers based on the `If_Affirmative_Win` column ('T' for Affirmative, 'F' for Negative) and calculate the average. Then, identify dates from the Debate table for debates where the Negative side won and the speaker's age is strictly greater than this average.

Reasoning with Visual cue

Average Age of Winning Speaker: 38.67

Dates of debates where the Negative side won, and the Negative speaker's age was strictly greater than 38.67:

Date	Speaker as Negative	Age
Oct 21, 2011	Charles Van Wyck	36
Jan 12, 2017	George Briggs	42

GPT Generated Visual Cue

Table 4: Winning Speakers, Ages, and Target Debate Analysis

Debate_ID	Date	Winning_Side	Winning_Speaker_Name	Winning_Speaker_Age	Negative_Win_& Age_Above_Average
1	October 21, 2011	Negative	Charles Van Wyck	36	
5	January 12, 2017	Negative	Horace F. Clark	45	Yes (45 > 42)
3	August 3, 2015	Affirmative	Thomas J. Barr	48	
6	January 21, 2017	Affirmative	William B. Maclay	36	
4	October 27, 2015	Negative	Horace F. Clark	45	Yes (45 > 42)

Average age of all winning speakers: 42 years. Target debates (Negative win, age > 42) are highlighted in the last column.

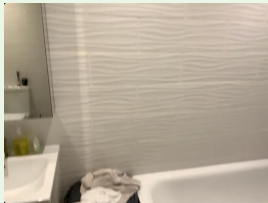
Gemini Generated Visual Cue

Figure C.8. Case 8: Multi-table reasoning (relational composition). The cue must specify the join keys and the derived statistic before filtering, making the evidence chain explicit and auditable across intermediate operations.

Multi-view reasoning question in UFO



Input Images



GT Image cue

Question: The second image shows the result of panning the camera to the right (clockwise) from the starting view. What large plumbing fixture is revealed in this new view?

- A: Shower Stall,
- B: Toilet,
- C: Bathtub,
- D: Vanity with integrated sink basin

Answer: C

Ground Truth Text Cue: Identification relies on matching the distinctive horizontal wavy texture pattern on the white wall tiles to orient the views.

Reasoning with Textual cue

text_cue_generated_gpt:

After panning the camera to the right from the initial view, a large white bathtub is revealed along the bathroom wall, occupying a prominent portion of the new frame.

text_cue_generated_gemini:

Based on the provided image, here is the textual cue: The image displays a bathroom interior featuring a large mirror on the right wall that clearly reflects a white toilet and toilet roll holder located on the opposite side of the room. A white radiator is also visible, mounted on the textured tiled wall to the left of the mirror.,

Reasoning with Visual cue



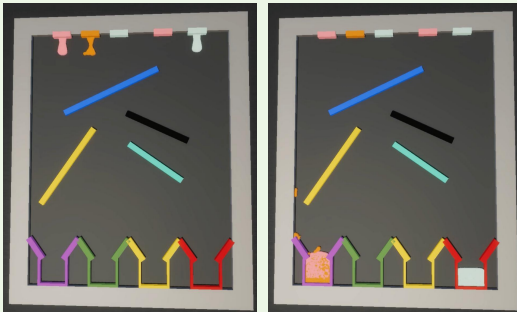
GPT Generated Visual Cue



Gemini Generated Visual Cue

Figure C.9. Case 9: Multi-view reasoning (view alignment). The instance tests whether the cue captures the correspondence between views via shared textures or landmarks, supporting a correct identification after viewpoint change.

Physics reasoning question in UFO



Input Images

GT Image cue

Question: Which container catches the liquid falling from the orange emitter?

- A: second container from the left (the purple-green adjacent pocket),
 - B: yellow container,
 - C: center green container,
 - D: purple container
- Answer: D

Ground Truth Text Cue: The orange object rolls down the blue and yellow ramps to land in the purple container.

Reasoning with Textual cue

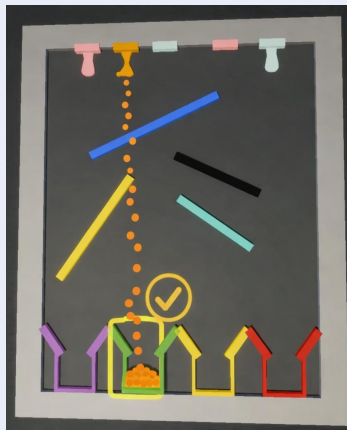
text_cue_generated_gpt:

The liquid stream from the orange emitter falls directly into the container positioned immediately below it; identify the container aligned under the emitter's spout/outlet and intersecting the falling liquid path, not adjacent containers offset to the side.

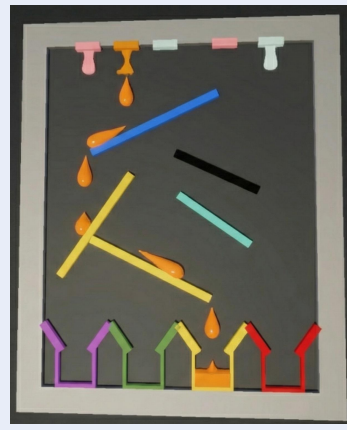
text_cue_generated_gemini:

The orange emitter drops liquid onto a blue bar angled downwards to the left, which directs the flow onto a similarly angled yellow bar below. This path guides the liquid into the leftmost purple container.

Reasoning with Visual cue



GPT Generated Visual Cue



Gemini Generated Visual Cue

Figure C.10. Case 10: Physics reasoning (causal trajectory). The cue must encode the causal path of the liquid/object through intermediate ramps, so the final answer depends on a faithful physical trace rather than surface-level saliency.

1650 **C. Limitations and Future Work**

1651 While UFO is designed to expose whether intermediate multimodal cues are causally utilized during reasoning, it does not
1652 aim to fully characterize all forms of multimodal intelligence. Our benchmark focuses on question-driven state transitions
1653 with explicitly defined future states, a formulation that enables controlled evaluation of evidential coupling but does not
1654 encompass open-ended reasoning scenarios in which the target state is implicit, evolving, or negotiated through interaction.
1655 Extending UFO to such settings would require mechanisms for tracking, validating, and attributing evidence across longer
1656 horizons, which we leave as an important direction for future work.
1657

1658 UFO currently targets vision–language reasoning, assuming that intermediate evidence can be expressed as textual and
1659 visual cues. In many real-world systems, however, state transitions unfold over richer and temporally continuous modalities,
1660 including audio, video, depth, or proprioceptive signals. Generalizing our evaluation protocol to these domains would
1661 require rethinking both the representation of intermediate cues and the notion of evidential sufficiency under temporal
1662 aggregation. We view such extensions as complementary to the current benchmark and defer their exploration to future
1663 studies.
1664

1665 Finally, while UFO makes the reliance on intermediate evidence measurable, it does not prescribe how such reliance
1666 should be induced during training. Our empirical results show that some models generate cues that are nominally state-
1667 aligned, yet only weakly condition on them when producing the final answer. This gap suggests that existing training
1668 objectives insufficiently couple evidence generation and evidence utilization. Future work may investigate learning strategies
1669 that explicitly encourage causal dependence on intermediate cues, such as intervention-aware supervision or consistency
1670 constraints across reasoning steps.
1671

1672 **D. Ethical Considerations**

1673 UFO is an evaluation benchmark and does not introduce new generative capabilities or deployment mechanisms. As such,
1674 it does not directly amplify risks associated with content generation, misinformation, or model misuse. Nevertheless,
1675 by promoting evaluation protocols that emphasize interpretable and intervenable intermediate representations, UFO may
1676 indirectly support the development of more controllable and accountable multimodal systems.
1677

1678 At the same time, evidential coupling alone does not guarantee correctness, fairness, or robustness. A model may rely
1679 consistently on its intermediate cues while those cues reflect biased or incomplete interpretations of the input. We therefore
1680 view UFO as complementary to existing efforts in safety, bias mitigation, and robustness evaluation. Ensuring that
1681 intermediate evidence is not only causally effective but also socially and ethically grounded remains an open challenge
1682 beyond the scope of this work.
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704