ONTOLOGY-HARNESSING MULTI-AGENT SYSTEM FOR ABSTRACTIVE VISUAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal large language models (MLLMs) have made rapid progress in visual understanding, yet abstractive visual representations remain challenging because their interpretation depends on recovering entities, relations, and structural constraints beyond visual appearance alone. In this paper, we propose the Ontology-Harnessing Multi-Agent system (OHM), a training-free framework that decomposes abstractive visual understanding into perception, validation, and task-specific reasoning around an explicit symbolic graph. An ontology provides shared semantic and structural knowledge for intermediate validation and disambiguation, while gradually incorporating reliable observations during inference to support subsequent reasoning. Extensive experiments across multiple tasks and model backbones demonstrate consistent improvements over direct inference, while further revealing the complementary roles of reliable structure recovery and effective downstream reasoning.

1 INTRODUCTION

Recent advances in multi-modal large language models (MLLMs) have demonstrated strong visual understanding and reasoning capabilities across diverse visual content, from natural images to structured representations such as tables, code snippets, and mathematical expressions (Liu et al., 2023; Dai et al., 2023; Bai et al., 2025b; Lu et al., 2024).

Among these forms of visual content, abstractive visual representations have received comparatively limited attention (Zhang et al., 2025; 2026b). As illustrated in Figure 1, such representations encode entities, relations, and higher-level structure through visual abstractions rather than directly depicting real-world scenes. Unlike natural images or highly regular formats such as tables, their semantics depend on interpreting both individual visual elements and the abstract relations among them (Zhang et al., 2025).

Understanding such abstractive visual representations is challenging because it requires not only recognizing individual visual elements but also recovering the relational structure that connects them. This challenge is therefore not solely a matter of improving visual encoding. Instead of mapping an image directly to an answer, a more explicit process can first recover entities and relations, validate the resulting structure, and then perform task-specific reasoning such as counting, diagnosis, or completion (Yi et al., 2018; Gupta & Kembhavi, 2023). This decomposition naturally motivates a multi-agent workflow, in which specialized agents collaborate across different stages of understanding (Guo et al., 2024; Zhang et al., 2026a).

Such decomposition alone, however, leaves the agents without a shared model of the structural regularities that make an interpretation plausible. A useful intuition comes from how unfamiliar abstract representations are understood: individual observations are interpreted against a tentative conceptual model that can be refined as evidence accumulates (Danesi, 1990; Arnheim, 2023). We instantiate this shared model as an ontology, which captures entity concepts, relation patterns, and structural constraints (Gruber, 1993; Noy & McGuinness, 2001). It provides common semantic and structural context for validating recovered representations and resolving ambiguous interpretations, while incorporating reliable observations over time.

To this end, we propose the **O**ntology-**H**arnessing **M**ulti-Agent system (OHM), a training-free framework that reconstructs an explicit graph and maintains a lightweight ontology as shared seman-

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

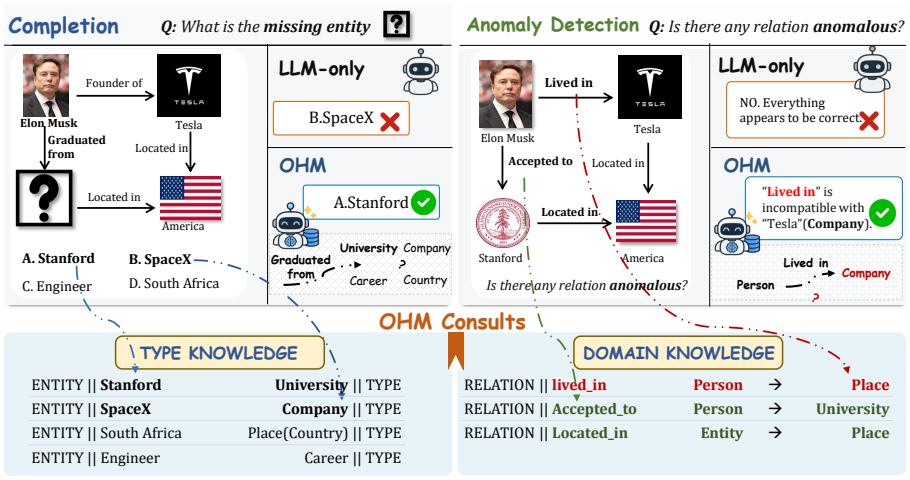


Figure 1: Illustrative examples of ontology-harnessed reasoning in OHM. In knowledge graph (KG) completion (left), type knowledge helps identify the missing entity. In anomaly detection (right), relation-domain knowledge reveals that *Lived in* is incompatible with Tesla as a company. The ontology provides supporting evidence rather than a hard decision rule.

tic and structural knowledge. The recovered graph provides an explicit state between perception and task-specific reasoning, while the ontology supplies type and structural evidence for validation and disambiguation. As illustrated in Figure 1, type knowledge can help resolve an ambiguous missing entity, while relation-domain knowledge can expose an anomalous edge.

Our contributions are summarized as follows:

- We formulate abstractive visual understanding as perception, validation, and task-specific reasoning, and develop a training-free multi-agent workflow around an explicit graph state.
- We introduce an ontology-harnessing mechanism that captures structural constraints as shared evidence for graph validation and task-specific reasoning.
- We evaluate OHM across multiple tasks and model backbones, demonstrating consistent improvements and revealing the complementary roles of reliable structure recovery and downstream reasoning.

2 Related Work

We discuss three lines of work most relevant to OHM.

Structured Visual Understanding Structured visual understanding has been studied across diagrams, charts, and synthetic scenes. AI2D focuses on diagram parsing and question answering (Kembhavi et al., 2016), ChartQA on visual and logical reasoning over charts (Masry et al., 2022), and CLEVR and GQA on compositional object–relation reasoning (Johnson et al., 2017; Hudson & Manning, 2019). More closely related, M3STR introduces abstractive visual understanding over rendered multi-modal KGs through counting, detection, and completion tasks (Zhang et al., 2025), while STAR improves such reasoning through task-specific training (Zhang et al., 2026b). In contrast, OHM targets training-free inference by explicitly recovering a symbolic graph before validation and task-specific reasoning.

Ontology and Neuro-Symbolic Reasoning Neuro-symbolic methods make intermediate structure explicit between perception and reasoning. NS-VQA constructs symbolic scene representations for program execution (Yi et al., 2018), while later approaches combine neural perception with logic rules or modular visual programs (Eiter et al., 2025; Gupta & Kembhavi, 2023); PixelCraft (Zhang et al., 2026a) further extends modular reasoning to structured images through multi-agent coordination. Ontology engineering similarly provides structured representations of concepts, relations,

and constraints (Gruber, 1993; Noy & McGuinness, 2001; The Gene Ontology Consortium, 2000). Building on these lines, OHM uses an explicit entity–relation graph as the shared state and ontology-derived knowledge as semantic and structural evidence for validation and disambiguation.

Multi-Agent Systems for Structured Knowledge LLM-based multi-agent systems coordinate specialized roles through messages, workflows, or shared artifacts (Guo et al., 2024), with representative systems including AutoGen (Wu et al., 2024), MetaGPT (Hong et al., 2024), CAMEL (Li et al., 2023), and multi-agent debate (Liang et al., 2023). Related systems have also applied LLMs to structured knowledge construction, including KARMA for KG enrichment (Lu et al., 2025) and ontology-guided extraction for schema-constrained outputs (van Cauter & Yakovets, 2024). Unlike these approaches, OHM coordinates task-specific agents through a shared graph state recovered from visual inputs, using it for validation and downstream reasoning rather than persistent knowledge-base construction.

3 METHODOLOGY

In this section, we present the methodology of OHM, a training-free framework for abstractive visual understanding that decomposes inference into perception, validation, and task-specific reasoning under the coordination of an orchestration agent. We first describe the overall multi-agent workflow and its main components. Then, we describe the construction and refinement of the ontology and how it supports validation and task-specific reasoning.

3.1 ONTOLOGY-HARNESSING MULTI-AGENT SYSTEM

At a high level, OHM comprises five components: an orchestration agent, a perception module, a validation module, a reasoning module, and an ontology. Once the orchestration agent selects the execution path, the perception module recovers entities and relations from the visual input. The validation module then examines the resulting symbolic state with support from the ontology, before the reasoning module produces the task-specific answer.

Specifically, let I denote the input visual representation, q the task-specific question, and $\mathcal{O}_t$ the current ontology. The orchestration agent maps q to a task family

$$\tau = f_{\text{orch}}(q), \qquad \tau \in \mathcal{T} = \{\text{count}, \text{detect}, \text{complete}\}, \tag{1}$$

and selects the corresponding execution path. The subsequent workflow can be summarized as

$$\hat{G} = f_{\text{perc}}(I, h_\tau) \qquad \text{(Perception)} \tag{2}$$

$$(\tilde{G}, c) = f_{\text{val}}(\hat{G}, \mathcal{O}_t) \qquad \text{(Validation)} \tag{3}$$

$$a = f_\tau(\tilde{G}, q, c) \qquad \text{(Reasoning)} \tag{4}$$

where $\hat{G}$ is the graph recovered from the visual input, $\tilde{G}$ is its validated state, c denotes semantic and structural compatibility evidence, h_τ denotes task-specific guidance, and a is the final prediction. This explicit graph state separates visual perception from task-specific reasoning and provides a shared interface for intermediate validation and refinement.

Perception The perception stage converts the visual input I into an explicit symbolic graph $\hat{G} = (\hat{V}, \hat{E})$, where $\hat{V}$ denotes the recognized entities and $\hat{E}$ the extracted relations. The entity recognition agent first produces

$$\hat{V} = f_{\text{ent}}(I, h_\tau), \tag{5}$$

where h_τ denotes task-specific guidance. Conditioned on the image and the recognized entities, the relation extraction agent then produces

$$\hat{E} = f_{\text{rel}}(I, \hat{V}, h_\tau), \qquad e = (v_i, r, v_j), \tag{6}$$

where r is a normalized relation label and v_i and v_j denote the ordered head and tail entities. Lightweight post-processing normalizes labels and endpoint references and removes duplicate records. The resulting graph provides an explicit interface between visual perception and subsequent reasoning, while keeping extraction errors accessible to the validation stage.

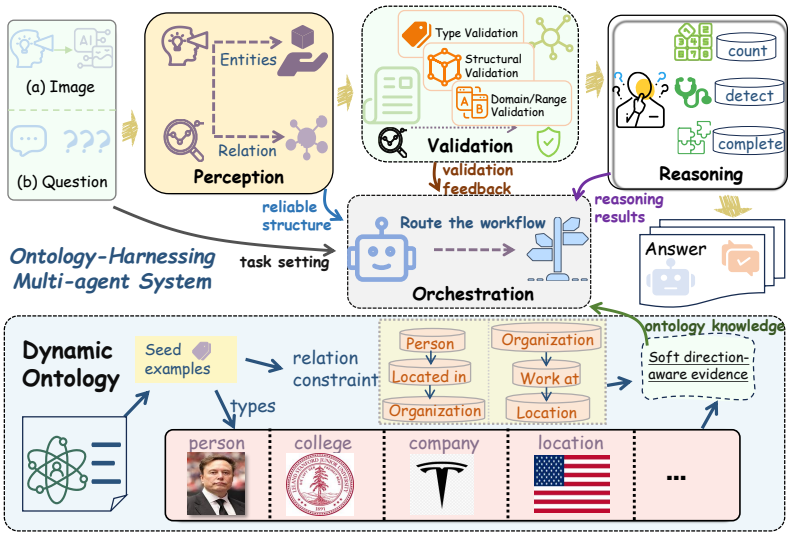


Figure 2: Workflow of OHM. Given a structured visual representation and a task-specific question, the orchestration agent routes the input to perception agents that recover entities and relations. The validation module examines the recovered graph with support from the ontology, after which task-specific reasoning agents produce the final answer.

Validation The validation stage examines the extracted graph before task-specific reasoning. Given the current ontology $\mathcal{O}_t$, the validation module computes

$$(\tilde{G}, c) = f_{\text{val}}(\hat{G}, \mathcal{O}_t), \tag{7}$$

where $\tilde{G}$ denotes the validated graph and c collects semantic and structural compatibility evidence. The module checks graph integrity, entity–relation compatibility, relation domain and range, and structural constraints such as cardinality and one-to-one relations. It also detects common extraction errors such as orphan endpoints and isolated entities. Potential inconsistencies are treated as soft signals: they may trigger state revision or re-extraction, but do not automatically invalidate an observation. The resulting evidence is retained for subsequent reasoning.

Reasoning The reasoning stage produces the task-specific answer from the validated graph. Given $\tilde{G}$, the question q, and the validation evidence c, the corresponding reasoning agent computes

$$a = f_{\tau}(\tilde{G}, q, c), \tag{8}$$

where a denotes the final prediction. For counting, the system directly returns the number of entities or relations in $\tilde{G}$. For detection, the reasoning agent determines whether an anomaly is present in the queried entities, relations, or the graph as a whole. For completion, it selects the missing entity or relation from the provided candidates based on the local graph structure and compatibility evidence in c. Although the decision procedures differ across tasks, they operate on the same validated symbolic state, separating visual structure recovery from task-specific reasoning.

Ontology The ontology $\mathcal{O}_t$ captures entity concepts, relation patterns, and lightweight structural constraints. It provides semantic and structural compatibility evidence for intermediate validation and task-specific reasoning. Initialized from a small set of seed examples, it can be conservatively refined as reliable observations accumulate during inference.

3.2 Ontology Construction and Refinement

We represent the ontology at step t as

$$\mathcal{O}_t = (\mathcal{C}_t, \mathcal{R}_t, \mathcal{A}_t)\,, \tag{9}$$

where $\mathcal{C}_t$, $\mathcal{R}_t$, and $\mathcal{A}_t$ denote entity concepts, relation patterns, and structural constraints, respectively. The initial ontology $\mathcal{O}_0$ is constructed from a small set of seed examples, providing coarse semantic categories and structural regularities without assuming complete domain coverage.

As new visual inputs are processed, reliable intermediate states provide additional evidence for ontology refinement. Given the validated graph $\tilde{G}_t$, the ontology is updated as

$$\mathcal{O}_{t+1} = f_{\mathrm{upd}}(\mathcal{O}_t, \mathcal{S}_t), \tag{10}$$

where $\mathcal{S}_t$ contains observations supported by consistent extraction and compatible local structure. Ambiguous or structurally conflicting observations are excluded from updates, while new concepts and relation patterns are introduced only when supported by accumulated evidence.

Ontology refinement uses only visible inputs and validated intermediate representations. Ground-truth answers are never accessed, and masked entities or relations in completion tasks are excluded from update evidence. Additional implementation details are provided in Appendix C.

3.3 Ontology-Guided Reasoning

The ontology serves as auxiliary evidence during task-specific reasoning rather than as a hard constraint. Given the validated graph $\tilde{G}$, question q, and ontology-derived compatibility evidence c, the reasoning agent operates primarily on the recovered structure while consulting c when semantic or structural ambiguity arises.

Counting Counting is performed directly on the validated graph:

$$\hat{n}_V = |\tilde{V}|, \qquad \hat{n}_E = |\tilde{E}|. \tag{11}$$

No additional ontology reasoning is applied after validation.

Detection For anomaly detection, the reasoning agent combines the local graph context with compatibility evidence from entity types, relation domains and ranges, and structural constraints. These signals expose potential inconsistencies and support the final binary decision.

Completion For completion, let M denote the masked entity or relation and $\mathcal{A}_q$ the candidate set. Each candidate $y \in \mathcal{A}_q$ is evaluated from its local graph context and compatibility evidence:

$$\hat{y} = \arg\max_{y \in \mathcal{A}_q} f_{\mathrm{comp}}\big(y, \mathrm{ctx}(M, \tilde{G}), c\big). \tag{12}$$

For missing entities, relation direction provides expected endpoint types; for missing relations, the endpoint concepts provide compatibility evidence over candidate predicates.

4 Experiments

In this section, we evaluate OHM through system-level comparisons and controlled studies of its main design choices.

4.1 Experimental Setup

Benchmarks We evaluate OHM on two benchmarks covering different forms of structured visual understanding. M3STR (Zhang et al., 2025) includes counting, anomaly detection, entity completion, and relation completion over rendered KGs, with 4,500 counting images, 3,500 detection images, and 5,600 images for each completion task. AI2D (Kembhavi et al., 2016) contains 3,088 multiple-choice questions over 814 scientific diagrams, providing a complementary setting with different visual structures. Details of the benchmarks are provided in Appendix B.

Models and Metrics We evaluate Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), InternVL3.5 (Wang et al., 2025), and MiniCPM-V 4.5 (Yu et al., 2025), together with the hosted DeepSeek-Flash (DeepSeek-AI, 2026), Gemini-3.1-Flash (Google, 2026), and GPT-4o-mini (OpenAI, 2024). Same-model configurations use the same backbone for Direct inference and for both

Table 1: Main results on M3STR across different model configurations. Green shading highlights the strongest results in each metric column, with darker shades indicating higher performance. Boldface denotes the best reported result within each backbone block.

Method	Perception	Reasoning	Count-E	Count-R	Detect-E	Detect-R	Detect-Mix	Complete-E	Complete-R	Complete-Avg
Direct	Qwen2.5-VL-7B	–	44.12	33.00	15.98	14.05	24.24	82.14	86.11	84.13
OHM	Qwen2.5-VL-7B	Qwen2.5-VL-7B	**72.20**	**60.64**	25.49	28.52	66.29	78.82	86.99	82.91
OHM	Qwen2.5-VL-7B	DeepSeek-Flash	71.51	60.08	**56.40**	**28.58**	**68.84**	**91.76**	**91.55**	**91.66**
Direct	Qwen3-VL-4B	–	49.93	53.70	22.12	13.05	24.53	82.46	85.29	83.88
OHM	Qwen3-VL-4B	Qwen3-VL-4B	**82.06**	**74.32**	52.85	29.63	69.27	85.06	88.18	86.62
OHM	Qwen3-VL-4B	DeepSeek-Flash	82.04	74.25	**57.33**	**33.20**	**71.77**	**94.12**	**91.32**	**92.72**
Direct	Qwen3-VL-8B	–	66.44	50.67	52.29	28.83	67.54	87.33	87.40	87.37
OHM	Qwen3-VL-8B	Qwen3-VL-8B	84.54	76.51	57.42	30.33	70.42	85.90	87.67	86.79
OHM	Qwen3-VL-8B	DeepSeek-Flash	**84.69**	**76.77**	**59.13**	**33.15**	**72.50**	**93.96**	**91.85**	**92.91**
Direct	Qwen3-VL-30B	–	60.42	61.22	40.28	29.87	65.81	85.50	84.64	85.07
OHM	Qwen3-VL-30B	Qwen3-VL-30B	**82.89**	78.58	53.91	29.99	68.11	90.04	90.54	90.29
OHM	Qwen3-VL-30B	DeepSeek-Flash	82.69	**78.85**	**58.80**	**33.59**	**72.46**	**94.32**	**91.48**	**92.90**
Direct	InternVL3.5-8B	–	21.55	43.24	51.97	26.99	60.86	82.58	86.13	84.36
OHM	InternVL3.5-8B	InternVL3.5-8B	70.46	63.97	55.08	29.72	68.76	86.62	86.83	86.72
OHM	InternVL3.5-8B	DeepSeek-Flash	**70.62**	**64.19**	**57.81**	**31.38**	**71.00**	**93.60**	**90.63**	**92.12**
Direct	MiniCPM-V-4.5	–	47.21	20.19	48.00	24.74	53.00	84.21	81.23	82.72
OHM	MiniCPM-V-4.5	MiniCPM-V-4.5	**70.50**	**59.06**	48.61	30.61	68.90	84.82	78.43	81.63
OHM	MiniCPM-V-4.5	DeepSeek-Flash	70.41	58.77	**57.24**	**32.31**	**71.32**	**93.10**	**91.21**	**92.15**
Direct	DeepSeek-Flash	–	57.92	64.08	51.98	29.89	68.28	90.65	89.27	89.96
OHM	DeepSeek-Flash	DeepSeek-Flash	**89.60**	**85.32**	**58.93**	**35.46**	**73.50**	**94.95**	**92.27**	**93.61**
Direct	Gemini-3.1-Flash	–	75.08	**80.75**	**57.27**	**30.95**	**68.12**	94.95	92.75	93.85
OHM	Gemini-3.1-Flash	Gemini-3.1-Flash	**78.87**	78.56	53.10	30.59	64.23	**95.61**	**94.42**	**95.01**
Direct	GPT-4o-mini	–	48.19	40.03	53.51	24.77	65.60	**89.86**	81.97	85.92
OHM	GPT-4o-mini	GPT-4o-mini	**81.35**	**69.10**	**59.92**	**27.92**	**67.79**	87.87	**84.34**	**86.11**

perception and reasoning within OHM, whereas heterogeneous configurations pair a perception backbone with a different reasoning model. We report exact accuracy for counting and answer accuracy for completion and AI2D. Anomaly detection is evaluated by positive-class F1 due to class imbalance. Additional implementation details are provided in Appendix C.

4.2 Main Results on Abstractive Visual Reasoning

Table 1 summarizes the main results. OHM delivers its most consistent gains on counting, where explicit graph reconstruction substantially improves entity and relation enumeration across almost all backbones. For example, Qwen2.5-VL-7B improves from 44.12% to 72.20% on entity counting and from 33.00% to 60.64% on relation counting, while DeepSeek-Flash improves from 57.92% to 89.60% and from 64.08% to 85.32%, respectively. Unlike direct inference, which must jointly interpret the image and produce the final count, OHM first externalizes the visual content into an explicit set of entities and relations. Counting can therefore be performed over a recovered symbolic state rather than inferred implicitly from the image. This advantage is particularly pronounced when direct perception is unreliable, suggesting that explicit structure recovery reduces the burden of performing visual extraction and task reasoning in a single step.

Anomaly detection also benefits from the structured pipeline, but exhibits greater variation across models and anomaly types. Same-model OHM substantially improves mixed-anomaly detection for several backbones, including Qwen2.5-VL-7B from 24.24% to 66.29% and Qwen3-VL-4B from 24.53% to 69.27%. The gains are less uniform for individual entity or relation anomalies, and some strong Direct configurations show limited or negative changes. This difference is consistent with the nature of the task: anomaly detection requires not only recovering the graph, but also judging whether a locally plausible entity or relation is globally compatible with its context. The explicit graph and validation signals provide additional evidence for this decision, but their usefulness depends on the quality of both the recovered structure and the downstream interpretation.

Completion shows an even stronger dependence on reasoning capability. Same-model OHM improves several configurations, but yields limited or negative gains for others, indicating that explicit structure alone is insufficient when the reasoner cannot effectively exploit it. In contrast, replacing the same-model reasoner with DeepSeek-Flash consistently strengthens completion across multiple perception backbones. With Qwen2.5-VL-7B perception, completion accuracy increases

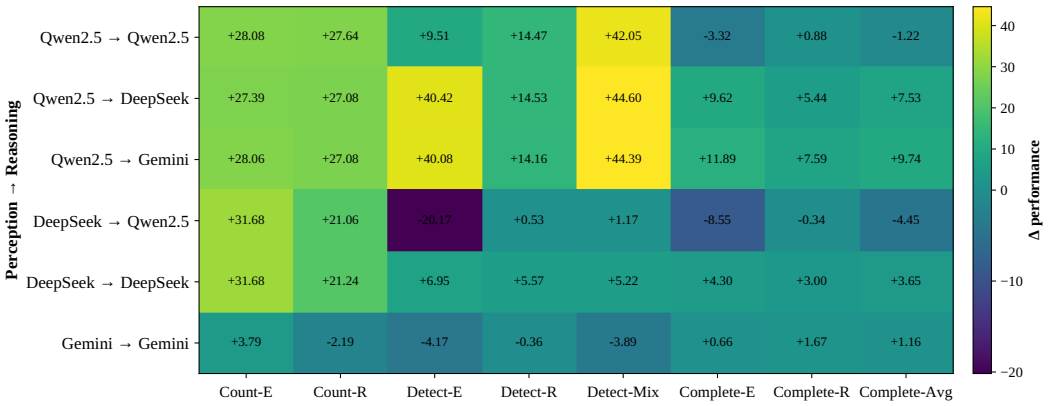


Figure 3: Performance differences of different perception–reasoning configurations relative to their corresponding Direct baselines. Each row denotes the perception and reasoning models used in OHM, and each cell reports the score difference between OHM and Direct.

from 82.91% to 91.66%, while MiniCPM-V-4.5 improves from 81.63% to 92.15%. Notably, these changes occur while the perception backbone remains fixed, indicating that the recovered graph provides a reusable interface whose value depends on the reasoning model operating over it. Taken together, the results suggest that OHM benefits from two complementary factors: reliable structure recovery is essential for constructing a useful symbolic state, while sufficient reasoning capacity determines how effectively that state can be translated into the final prediction.

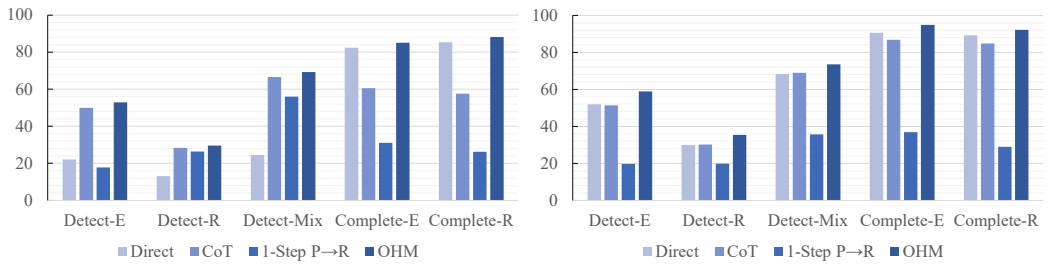
4.3 Perception–Reasoning Complementarity Across Backbones

Figure 3 further disentangles the roles of perception and reasoning by varying the two components independently. Under fixed Qwen2.5-VL-7B perception, counting gains remain nearly unchanged when the reasoner is replaced by DeepSeek or Gemini, with entity and relation counting consistently improving by about 27–28 points. In contrast, reasoning-intensive tasks respond strongly to the choice of reasoner. Entity-anomaly detection improves by more than 40 points with either DeepSeek or Gemini reasoning, compared with less than 10 points in the same-model configuration, while average completion changes from a slight degradation to gains of 7.53 and 9.74 points. Since the recovered graph is unchanged across these configurations, these differences isolate the ability of the reasoning model to interpret and exploit the same symbolic state.

The reverse pairing provides complementary evidence. With DeepSeek perception, replacing the DeepSeek reasoner with Qwen2.5 preserves almost identical counting gains, yet substantially weakens anomaly detection and completion; entity-anomaly detection even falls below the corresponding Direct baseline. This separation suggests that perception primarily determines the fidelity of the symbolic state, whereas downstream reasoning determines how effectively that state is converted into task-specific decisions. The interaction is nevertheless not monotonic with model strength. Gemini, whose Direct baseline is already strong, exhibits smaller or occasionally negative changes under same-model OHM, indicating reduced headroom and showing that a structured pipeline does not uniformly benefit every configuration. Overall, the results point to a complementary rather than substitutive relationship between perception and reasoning: reliable reconstruction provides the necessary structural evidence, while capable reasoning is required to make effective use of it.

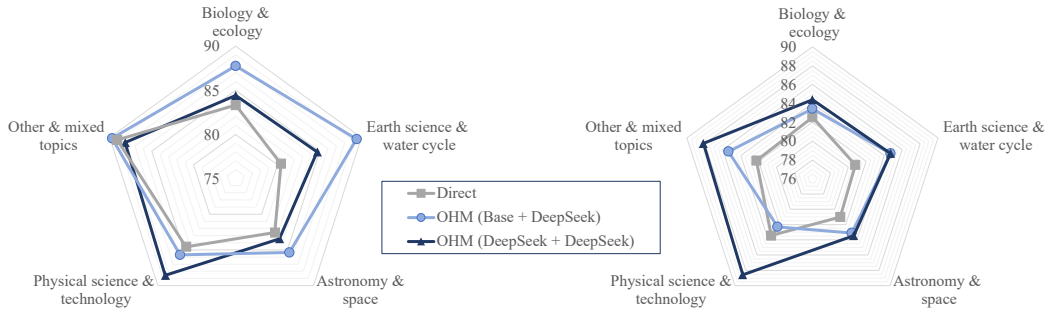
4.4 Controlled Comparison with Simplified Pipelines

Figure 4 examines whether the gains of OHM can be reproduced by stronger prompting or a simpler agent decomposition. CoT improves several anomaly-detection metrics with Qwen3-VL-4B, suggesting that additional reasoning steps can help expose local inconsistencies that are easily missed by direct answering. However, these gains do not extend consistently to completion and become much less pronounced with DeepSeek-Flash. This indicates that prompting the same model to reason more explicitly does not by itself provide a stable solution across tasks or backbones.



(a) Qwen3-VL-4B **(b)** DeepSeek-Flash

Figure 4: Controlled comparison of OHM against Direct inference, chain-of-thought prompting (CoT), and a simplified one-step perception-to-reasoning pipeline (1-Step P→R) on M3STR. Results are shown with **(a)** Qwen3-VL-4B and **(b)** DeepSeek-Flash.



(a) Qwen2.5-VL-7B **(b)** Qwen3-VL-4B

Figure 5: Topic-wise accuracy on AI2D. Each panel compares Direct inference with OHM using the panel-specific backbone for perception and DeepSeek-Flash for reasoning (Base + DeepSeek), together with a DeepSeek-Flash perception–reasoning configuration. **(a)** Qwen2.5-VL-7B as the base backbone. **(b)** Qwen3-VL-4B as the base backbone.

The 1-Step P→R baseline provides a stronger structural comparison: it separates perception from reasoning, but directly passes the extracted content to the reasoner without validation and ontology-guided mechanisms. Its performance remains unstable, and in several cases completion degrades substantially despite the perception–reasoning split. In contrast, full OHM consistently improves over Direct across all detection and completion metrics under both backbones. The comparison therefore suggests that the gains of OHM do not arise merely from longer reasoning traces or from introducing multiple stages. Rather, organizing the recovered content as a graph, validating the intermediate state, and incorporating structural knowledge provide a more reliable interface between perception and reasoning. This advantage becomes particularly important for anomaly detection and completion, where the decision depends not only on extracting visual content but also on assessing relations and constraints among the recovered elements.

4.5 Transfer to Scientific Diagrams

Figure 5 examines whether the structured perception–reasoning pipeline transfers beyond the rendered KGs to scientific diagrams. Across both base backbones, OHM with DeepSeek-Flash reasoning improves over Direct inference on most topic groups, including biology and ecology, Earth science and water cycle, and astronomy and space. The gains are therefore not concentrated in a single semantic category, suggesting that the benefit of separating visual structure recovery from downstream reasoning extends across diverse diagram content rather than relying on a particular knowledge domain.

The comparison also reinforces the perception–reasoning interaction observed on M3STR. Replacing the base perception model with DeepSeek-Flash further changes the topic-wise profile and yields

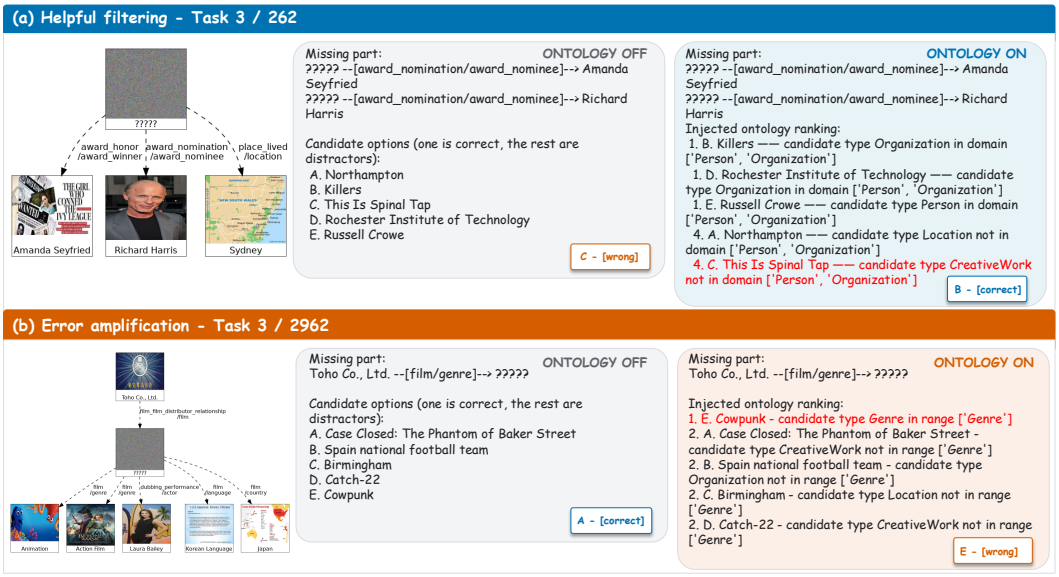


Figure 6: Paired ontology-off/on examples on M3STR with Qwen2.5-VL-7B as backbone. **(a)** Reliable reconstruction allows type compatibility to suppress implausible candidates and support the correct prediction. **(b)** Erroneous reconstruction produces misleading compatibility evidence and amplifies the upstream error.

particularly strong performance on several categories, while no single configuration dominates every topic. This variation indicates that transfer performance depends on both the quality of the recovered visual structure and the reasoning model operating over it. More broadly, the results show that the multi-stage design of OHM remains useful when the visual representation changes substantially, supporting its applicability beyond KG images.

4.6 Case Study: When Does Ontology Evidence Help?

Figure 6 provides two paired examples explaining the model-dependent gains observed above. In Figure 6(a), the recovered structure is reliable, allowing type compatibility to suppress implausible candidates and guide the reasoner toward the correct answer. The ontology does not uniquely determine the prediction, but narrows the plausible interpretations while leaving the final decision to the reasoning model. Figure 6(b) shows the opposite case. An upstream reconstruction error incorrectly associates the masked film with a `film/genre` relation, causing compatibility evidence to favor an incorrect Genre candidate and overturn an initially correct prediction.

These cases reveal an interaction between perception and ontology-guided reasoning. Ontology evidence is most useful when the recovered graph provides a reliable structural premise. Accurate reconstruction not only improves the graph itself but also makes downstream compatibility evidence more trustworthy, whereas perception errors can be reinforced by the same mechanism. This helps explain why the gains of OHM depend jointly on perception quality and downstream reasoning.

5 Conclusion

In this paper, we presented OHM, a training-free multi-agent framework for abstractive visual understanding that organizes perception, validation, and task-specific reasoning around an explicit symbolic graph. By separating structure recovery from downstream reasoning and using lightweight ontology knowledge as auxiliary evidence, OHM provides a more reliable interface for structured visual reasoning. Our results further reveal the complementarity between perception and reasoning, suggesting that robust structure recovery and effective exploitation of the recovered state are both essential to the success of multi-agent visual understanding.

AI USE STATEMENT

Generative AI tools were used only to support manuscript preparation, primarily for language refinement, grammatical correction, and improving clarity of presentation. All AI-assisted text was subsequently reviewed and revised by the authors. The authors are fully responsible for the technical content, experimental results, interpretations, and conclusions presented in this work.

ETHICS STATEMENT

This study uses publicly available datasets and does not involve human-subject experiments, private user data, or newly collected sensitive information. Our experiments focus on structured visual reasoning over benchmark data and are conducted solely for research purposes. We are not aware of additional ethical concerns specific to the methodology or experimental setting, and the authors take responsibility for the appropriate use and reporting of the results.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we include the source code and relevant processed data in the supplementary materials. The appendix documents the main experimental settings, model configurations, evaluation protocols, and implementation details. Together, these materials provide the information and resources needed to reproduce the principal results reported in this paper.

REFERENCES

Rudolf Arnheim. *Visual thinking*. Univ of California Press, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Marcel Danesi. Thinking is seeing: Visual metaphors and the nature of abstract thought. *Semiotica*, 80(3-4):221–238, 1990. doi: 10.1515/semi.1990.80.3-4.221.

DeepSeek-AI. DeepSeek-V4.1-Flash. DeepSeek API Documentation, 2026. URL https://api-docs.deepseek.com/updates/. Accessed September 2026.

Thomas Eiter, Jan Hadl, Nelson Higuera Ruiz, Lukas Lange, Johannes Oetsch, Bileam Scheuvens, and Jannik Strötgen. Explainable zero-shot visual question answering via logic-based reasoning. In *Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning*, volume 284 of *Proceedings of Machine Learning Research*, pp. 977–991, 2025. URL https://proceedings.mlr.press/v284/eiter25a.html.

Google. Gemini api models. Google AI for Developers, 2026. URL https://ai.google.dev/gemini-api/docs/models. Accessed September 2026.

Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. doi: 10.1006/knac.1993.1008.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8048–8057, 2024. doi: 10.24963/ijcai.2024/890. URL https://www.ijcai.org/proceedings/2024/890.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019. doi: 10.1109/CVPR.2019.00686.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*. OpenReview.net, 2024.

Yuxing Lu, Wei Wu, Xukai Zhao, Rui Peng, and Jinzhuo Wang. KARMA: Leveraging multi-agent LLMs for automated knowledge graph enrichment. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1887. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/517f9b9c227b9dd51dba4560f37165ed-Abstract-Conference.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177.

Natalya F. Noy and Deborah L. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, 2001.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence, July 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Zeno van Cauter and Nikolay Yakovets. Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models*, pp. 75–84, 2024. doi: 10.18653/v1/2024.kallm-1.8.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=uAjxFFing2.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *CoRR*, abs/2509.18154, 2025. doi: 10.48550/arXiv.2509.18154. URL https://arxiv.org/abs/2509.18154.

Shuoshuo Zhang, Zijian Li, Yizhen Zhang, Jingjing Fu, Lei Song, Jiang Bian, Jun Zhang, Yujiu Yang, and Rui Wang. Pixelcraft: A multi-agent system for high-fidelity visual reasoning on structured images. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=HtpjSCs3g5.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Min Zhang, Wen Zhang, and Huajun Chen. Abstractive visual understanding of multi-modal structured knowledge: A new perspective for mllm evaluation, 2025.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Wen Zhang, and Huajun Chen. Structured and abstractive reasoning on multi-modal relational knowledge images. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15518–15540, 2026b. URL https://aclanthology.org/2026.findings-acl.761/.

Table 2: Representative ontology entries for the *Person* and *Location* concepts. Each concept is described by a semantic definition, positive examples, boundary rules, and negative examples, which together provide lightweight semantic evidence for validation and reasoning.

	Person	**Location**
Semantic definition	named, uniquely identifiable human individuals, including real or historical people but not fictional characters or generic profession labels	geographic locations, including administrative regions (countries, cities), natural features, and time zones
Positive examples	`Tom Hanks`, `Madonna`, `J.K. Rowling`	`New York City`, `United States of America`, `Eastern Time Zone`
Boundary rules	a concrete individual is Person, while the role they hold is Profession; a person is not classified as their work or organization	administrative regions and natural features are Location, while buildings and institutions are Organization; a time zone is treated as a Location subtype
Negative examples	`Actor-GB`, `Film Director`	`Harvard University`, `Taj Mahal`

A Ontology Representation and Examples

The ontology in OHM is represented as a lightweight collection of entity concepts, relation patterns, and structural constraints. Rather than attempting to encode an exhaustive taxonomy or a complete domain model, it maintains only the conceptual information needed to support intermediate validation and task-specific reasoning. Entity concepts describe the semantic categories to which recognized entities may belong, while relation patterns characterize plausible ordered endpoint types and structural constraints capture regularities such as cardinality or relation-specific restrictions.

Each entity concept is represented by several complementary fields. A semantic definition specifies the intended scope of the concept, while positive examples provide representative instances that ground the definition in concrete entities. Boundary rules distinguish closely related concepts and clarify ambiguous cases that cannot be resolved from a category name alone. Negative examples further identify common confusions and provide explicit counterexamples during type assessment. Together, these fields form a compact concept description that can be used as soft evidence when validating extracted entities and relations.

Table 2 shows two representative entries, *Person* and *Location*. For example, the *Person* concept includes named human individuals while excluding profession labels such as `Actor-GB` and `Film Director`. Its boundary rules further distinguish a person from the profession, organization, or work associated with that individual. Similarly, the *Location* concept covers geographic regions and related spatial categories, while distinguishing them from institutions and other non-geographic entities. These distinctions are particularly useful when the visual evidence alone admits multiple plausible interpretations.

The ontology entries act as compatibility evidence rather than hard classification rules. Observations that conflict with existing concept descriptions are retained for further inspection rather than automatically discarded. During ontology refinement, reliable intermediate evidence may refine existing concepts or introduce additional structure when recurring observations are not adequately captured by the current ontology. This representation therefore supports type and relation reasoning while remaining adaptable to previously unseen concepts and relation patterns.

B Benchmark Details

We provide additional details of the two benchmarks used in our experiments. They cover complementary forms of structured visual understanding, ranging from rendered KGs to scientific diagrams. Table 3 summarizes their main characteristics.

Table 3: Overview of the benchmarks used in our experiments. M3STR provides broad task coverage over multi-modal knowledge-graph images, while AI2D provides a complementary setting with scientific diagrams and diverse visual layouts.

Benchmark	Visual Input	Tasks	Scale
M3STR (Zhang et al., 2025)	Multi-modal KG images	Counting, Detection, Completion	19,200 images
AI2D (Kembhavi et al., 2016)	Scientific diagrams	Multiple-choice QA	3,088 questions / 814 diagrams

M3STR M3STR (Zhang et al., 2025) evaluates abstractive visual understanding over rendered multi-modal KGs. It contains four task families: counting, anomaly detection, entity completion, and relation completion, with counting further covering both entities and relations. The released benchmark contains 4,500 counting images, 3,500 anomaly-detection images, and 5,600 images for each completion task. We use the official annotations throughout evaluation and the provided candidate sets for completion.

Before inference, we perform a fixed input-validity check and exclude images that cannot be reliably loaded or processed. This removes 83 instances in total, leaving 4,482 counting images, 3,481 detection images, 5,581 entity-completion images, and 5,573 relation-completion images. These exclusions are determined before evaluation and are therefore independent of model predictions.

AI2D AI2D (Kembhavi et al., 2016) contains 3,088 multiple-choice questions over 814 elementary-science diagrams. Compared with the rendered KGs in M3STR, AI2D exhibits substantially different visual organization, including arrows, labels, spatial groupings, and diagram-specific graphical conventions. Each question is paired with four answer choices. We preserve the original answer-choice order and use the official references for evaluation. We include AI2D as a complementary setting for examining structured visual understanding beyond rendered KGs.

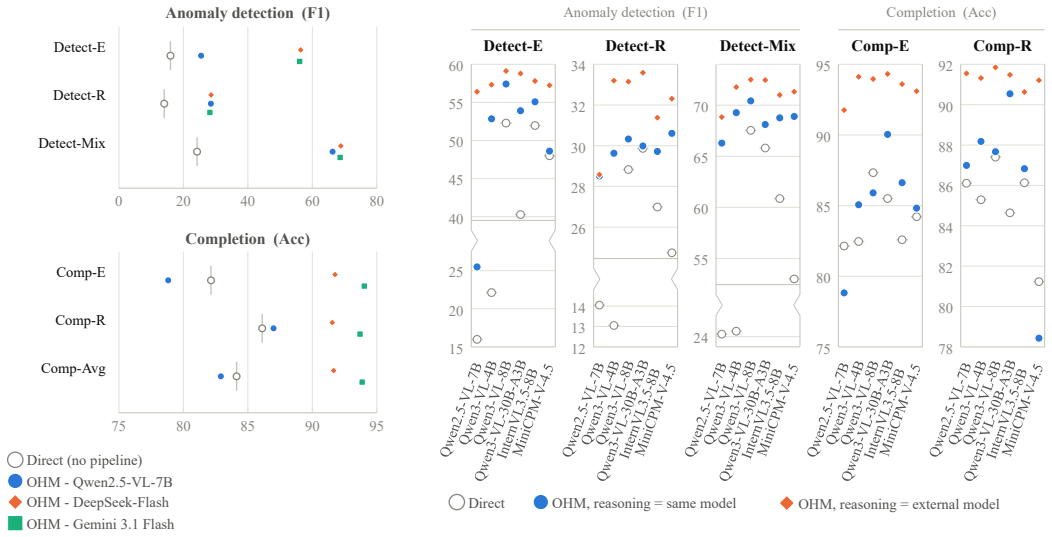
C Implementation and Controlled Evaluation Details

System Implementation OHM is implemented as a modular inference pipeline in which perception and reasoning models can be configured independently. Local models are served through vLLM, while hosted models are accessed through OpenAI-compatible interfaces. The orchestration logic is deterministic: each input is assigned to the corresponding counting, detection, or completion pipeline without an additional planning call. Model outputs are requested in structured form and normalized before being passed between stages. We use deterministic decoding throughout evaluation, with model-specific context and output limits chosen to accommodate their serving interfaces.

The perception stage first extracts entities and then recovers directed relations conditioned on the recognized entity set. The resulting graph is normalized by resolving minor label variations, duplicate entities, duplicate triples, and endpoint references. These operations are intentionally lightweight and do not introduce information absent from the visual input. Reasoning then operates on this normalized symbolic state. Counting directly returns the number of recovered entities or relations. For anomaly detection, entity- and relation-level modules assess the consistency of the corresponding graph elements, after which a deterministic aggregation rule produces the binary prediction. For completion, the reasoner receives the masked structure, its local graph context, and the original candidate set, and selects the most plausible candidate.

Ontology Support When enabled, the ontology provides auxiliary semantic and structural evidence rather than replacing visual perception or determining the final answer. Entity concepts and relation domain–range patterns are used to identify potentially inconsistent structures and to provide compatibility evidence during completion. For a masked entity, the direction of its incident relations determines the expected endpoint types; for a masked relation, the concepts of its two endpoints provide evidence about compatible predicates. This information is exposed to the reasoning model as a non-binding hint. If the ontology provides no discriminative evidence, no preference is introduced.

For M3STR, the initial ontology is derived from DB15K, outside the target evaluation set. Target answers are never used for initialization or refinement. Online refinement uses only validated entities and relations recovered from processed examples; masked content, candidate answers, generated



(a) reasoning-model comparison **(b)** cross-backbone comparison

Figure 7: Detailed M3STR results from two complementary views. **(a)** Fixed Qwen2.5-VL-7B perception with different reasoning models. **(b)** Direct, same-model OHM, and external-reasoner OHM configurations across perception backbones.

answers, and deliberately corrupted anomaly instances are excluded from updates. Updates affect only subsequent examples and never revise previous predictions.

Controlled Pipeline Baselines For the controlled comparison in Figure 4, Direct uses ordinary image-to-answer inference, CoT augments the same model with a chain-of-thought instruction, and 1-Step P→R performs a single perception pass followed directly by reasoning without intermediate validation or ontology-guided evidence. All variants use the same evaluation instances, backbone, candidate order, and decoding settings.

Implementation details not required to interpret the experiments, including prompt templates, serialization formats, serving configurations, and the exact update procedures, are provided with the source code in the supplementary materials.

D Additional Backbone Analysis

Figure 7 complements the relative improvements reported in Figure 3 with absolute performance across representative perception–reasoning configurations. Figure 7(a) fixes Qwen2.5-VL-7B perception and varies the reasoning model, showing that stronger reasoners mainly improve anomaly detection and completion while leaving the recovered visual structure unchanged. Figure 7(b) provides the corresponding comparison across perception backbones, where external reasoning generally yields stronger performance on reasoning-intensive tasks. Together, these results provide an absolute-score view of the perception–reasoning interaction analyzed in Section 4.3.