RefGround: Learning Reference-Dependent Visual Reasoning for Large Multimodal Models

Anonymous ACL submission

Abstract

Humans effortlessly leverage reference contexts to solve novel visual problems, yet Large Multimodal Models (LMMs) persistently default to memorized statistical patterns when inference-time guidance conflicts with their pre-trained priors. We term this challenge Reference-Dependent Visual Reasoning and introduce RefGround, a framework that systematically addresses this limitation. We first develop a Reference-Grounded Data Synthesis Engine that constructs strictly controlled task instances through closed-loop verification, yielding both diverse training data and a rigorous evaluation benchmark, RefGround-Bench. Evaluation on RefGround-Bench reveals that state-of-the-art models including GPT-5.2 and Claude-Opus achieve below 25% accuracy on reference-conditioned tasks, with grounding performance falling under 2.5%. To bridge this gap, we propose a curriculum-based reinforcement learning strategy with novel Noisy Contrast and Clue Verification rewards that enforce genuine visual dependency rather than prior exploitation. Experiments demonstrate that RefGround achieves 39.25% accuracy with an 8B backbone, substantially outperforming proprietary models, while exhibiting strong zero-shot transfer to MuirBench and BLINK benchmarks.

1 Introduction

Humans possess a remarkable ability to solve novel visual problems by dynamically integrating reference contexts with perceptual analysis. When confronted with unfamiliar objects or rules, we naturally extract relevant information from provided guidance and apply it to accomplish tasks that would be impossible through prior knowledge alone.

Large Multimodal Models have achieved impressive performance across diverse visual understanding benchmarks (Ying et al., 2024; Fu et al., 2023; Zou et al., 2025; Wang et al., 2024; Lu et al., 2024). However, these successes largely reflect scenarios where correct answers align with patterns memorized during pre-training. A fundamentally different picture emerges when models encounter tasks requiring strict adherence to inference-time references. As illustrated in Figure 1, existing LMMs (Zhang et al., 2024; Wu et al., 2024; Tong et al., 2024; Laurençon et al., 2024; Liu et al., 2023; OpenAI, 2024) exhibit a systematic failure: rather than grounding predictions in provided references, they default to pre-training priors and salience bias, selecting statistically frequent objects over explicitly specified targets. This failure stems from web-scale pre-training (Li et al., 2023; Alayrac et al., 2022) that encodes strong priors about object frequencies and semantic associations. When inference-time guidance conflicts with these memorized patterns, models implicitly favor parametric memory over explicit visual reference, leading to hallucinations that undermine reasoning.

We term this challenge Reference-Dependent Visual Reasoning, formally defined as tasks where correct solutions are strictly contingent on explicit guidance provided at inference time, rendering general pre-trained knowledge insufficient. Under this formulation, models cannot rely on internal statistical priors but must dynamically anchor their reasoning to specific references, whether visual patterns, textual definitions, or logical rules. Systematically investigating this challenge requires both rigorous benchmarks and sufficient training data, yet suitable instances that genuinely depend on provided guidance are scarce in existing datasets, and manual annotation is prohibitively expensive.

To address this bottleneck, we develop a Reference-Grounded Data Synthesis Engine that constructs task instances under strict visual and geometric constraints through closed-loop verification, ensuring all annotations derive from verified visual evidence rather than generative

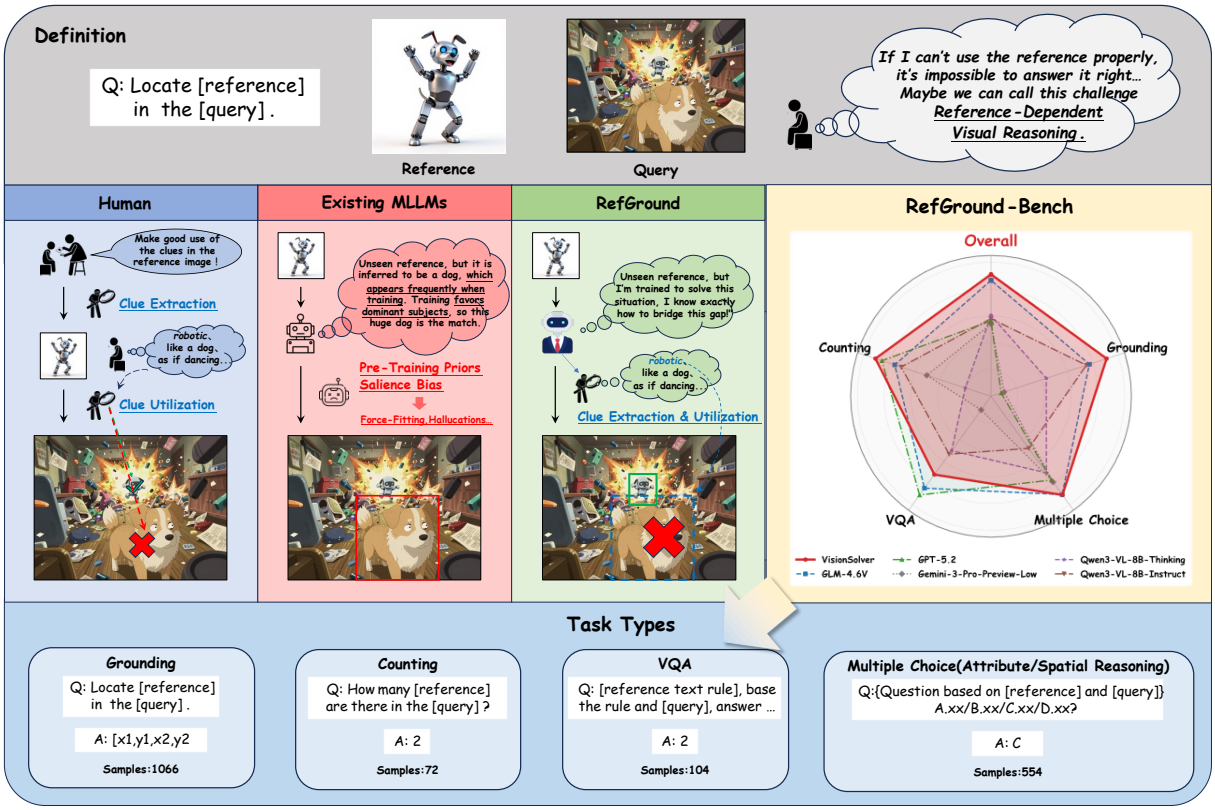

Figure 1: **Overview of the Reference-Dependent Visual Reasoning challenge and the RefGround framework.** The top panel illustrates the core definition: locating or reasoning about a specific *reference* within a complex *query* image. The middle panel contrasts reasoning processes: while humans and our proposed **RefGround** successfully bridge the gap between the reference and query (correctly finding the robot), existing MLLMs often fail at such tasks due to salience bias and pre-training priors. The right and bottom panels introduce **RefGround-Bench**, a comprehensive benchmark comprising four task types (Grounding, Counting, VQA, and Multiple Choice) to evaluate this capability across various models.

prompts. This approach enables scalable generation of contamination-resistant data serving dual purposes: diverse training instances and a comprehensive evaluation benchmark, which we term **RefGround-Bench**.

RefGround-Bench reveals striking deficiencies across current systems. Leading proprietary models including GPT-5.2 (OpenAI, 2025) and Claude-Opus (Anthropic, 2025) achieve only approximately 24% overall accuracy, with grounding performance falling below 2.5%. Even the largest open-source models struggle substantially, confirming that parameter scale does not resolve this fundamental limitation.

Building on this foundation, we introduce RefGround, a framework that enables MLLMs to master Reference-Dependent Visual Reasoning through curriculum-based reinforcement learning. As shown in Figure 1, RefGround learns to bridge the gap between reference clues and visual targets through a multi-task training framework with custom rewards designed specifically for this challenge. Our strategy employs Group Relative Policy Optimization (Shao et al., 2024) with two novel mechanisms. The Noisy Contrast Reward penalizes models that produce correct answers despite corrupted reference inputs, directly targeting reliance on parametric priors. The Clue Verification Reward ensures that generated reasoning faithfully reflects visual analysis by validating that textual descriptions alone suffice to solve tasks. Together, these rewards enforce genuine visual dependency rather than superficial pattern matching.

Extensive experiments demonstrate that RefGround achieves 39.25% accuracy on our benchmark using an 8B parameter backbone, substantially outperforming both proprietary giants and much larger open-source models. Notably, our approach exhibits strong zero-shot transferability to challenging external benchmarks including MuirBench (Wang et al., 2025) and BLINK (Fu et al., 2024), achieving a 12.4% absolute improvement

over the base model on multi-image reasoning tasks.

Our contributions are as follows. First, we formally define Reference-Dependent Visual Reasoning and develop a scalable data synthesis engine. Using this engine, we construct a large-scale training data and **RefGround-Bench**, a rigorous benchmark for evaluating reference-dependent capabilities. Second, we propose a curriculum-based reinforcement learning strategy with novel reward mechanisms that enforce genuine visual dependency. Third, we demonstrate that RefGround achieves state-of-the-art performance while exhibiting robust transfer to unseen domains, establishing reference-dependent reasoning as a learnable capability.

2 Related Work

Reinforcement Learning for Multimodal Reasoning. Post-training alignment for Large Multimodal Models has evolved from Supervised Fine-Tuning to Reinforcement Learning techniques (Christiano et al., 2017; Bai et al., 2022; Rafailov et al., 2023; Schulman et al., 2017) that enhance instruction following and reasoning capabilities. Recently, Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Yao et al., 2025) has emerged as a memory-efficient alternative to critic-based methods, demonstrating success in complex reasoning tasks across both textual (DeepSeek-AI, 2025) and multimodal domains (Huang et al., 2025). Building upon these advances, RefGround adapts GRPO with tailored reward mechanisms that specifically target the reference-dependent reasoning challenge, enforcing strict visual dependency through contrastive evaluation and self-verification loops.

Visual Prompting and Reference-Based Benchmarks. Recent studies have employed visual prompting techniques and specialized benchmarks (Fu et al., 2024; Wang et al., 2025) to assess how well LMMs utilize contextual information. BLINK (Fu et al., 2024) reveals fundamental challenges in core visual perception, while Muir-Bench (Wang et al., 2025) evaluates multi-image reasoning under diverse instructions. Concurrently, synthetic data engines such as UNO (Wu et al., 2025) and BAGEL (Deng et al., 2025) leverage generative models to produce training data for improving image generation quality. RefGround leverages generative engine for a different purpose: constructing strictly controlled reasoning tasks where synthesis metadata provides precise supervision signals for reference-dependent visual reasoning.

3 Method

In this section, we present the RefGround framework. We begin by formally defining Reference-Dependent Visual Reasoning and establishing the problem scope. We then introduce our Reference-Grounded Data Synthesis Engine that constructs strictly controlled task instances for both evaluation and training. Finally, we describe our curriculum-based reinforcement learning strategy with novel reward mechanisms designed to enforce genuine visual dependency.

3.1 Problem Formulation

We formally define Reference-Dependent Visual Reasoning as a class of visual tasks where correct solutions are strictly contingent on explicit guidance provided at inference time, rendering general pre-trained knowledge insufficient.

Definition. Let $\mathcal{R}$ denote a reference context, which may take the form of visual exemplars $\mathcal{R}_v$ (e.g., cropped object images) or textual specifications $\mathcal{R}_t$ (e.g., attribute descriptions, rule definitions). Let $\mathcal{I}_q$ denote a query image and $\mathcal{Q}$ denote a task query. A visual reasoning task is *reference-dependent* if and only if:

$$P(y^*|\mathcal{I}_q, \mathcal{Q}, \mathcal{R}) \gg P(y^*|\mathcal{I}_q, \mathcal{Q}) \qquad (1)$$

where y^* represents the correct answer. This formulation captures the essential constraint that the reference $\mathcal{R}$ provides information necessary for solving the task that cannot be inferred from the query image and task description alone.

Task Types. Under this formulation, we identify four primary task categories that span the spectrum of reference-dependent reasoning: *Reference-Based Grounding*, *Reference-Based Counting*, *Rule-Based Visual Question Answering*, and *Reference-Conditioned Multiple Choice*. Notably, we utilize the Multiple Choice format specifically for **Spatial Reasoning** and **Attribute Identification** tasks. This design choice mitigates the ambiguity inherent in evaluating open-ended descriptions of spatial layouts or fine-grained attributes, ensuring robust automatic evaluation. Detailed definitions and formulations for each task are provided in Appendix A.

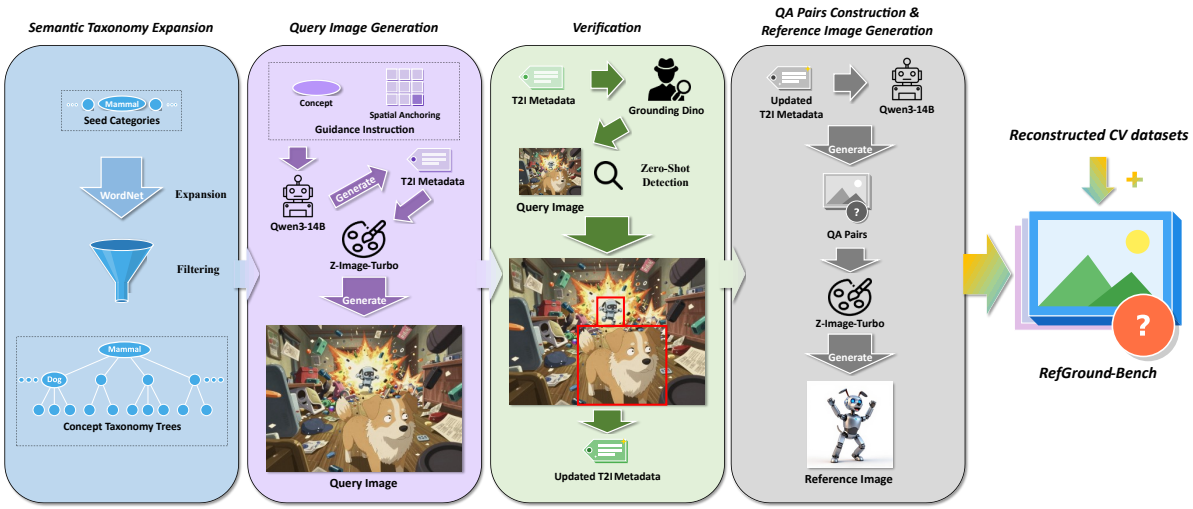

Figure 2: **The data construction engine of RefGround-Bench.** The process consists of four stages: (1) **Semantic Taxonomy Expansion**, where seed categories are expanded and filtered via WordNet to construct diverse concept taxonomy trees; (2) **Query Image Generation**, which utilizes Qwen3-14B for layout planning and Z-Image-Turbo for synthesizing complex query images based on spatial anchoring; (3) **Verification**, employing Grounding Dino to perform zero-shot detection for quality control and annotation refinement; and (4) **VQA Construction & Reference Image Generation**, where the corresponding reference images are synthesized and QA pairs are generated. The final RefGround-Bench is composed of these high-quality synthetic data combined with reconstructed public CV datasets. More details are in Appendix C, D, E, F, G

Distinction from Prior-Dependent Tasks. Reference-dependent tasks are fundamentally distinct from conventional visual recognition. In standard VQA or grounding, $P(y^*|\mathcal{I}_q, \mathcal{Q})$ alone often yields high accuracy due to language biases. Our formulation explicitly targets settings where $P(y^*|\mathcal{I}_q, \mathcal{Q}) \approx$ uniform, making $\mathcal{R}$ causally necessary. For example, in reference-dependent grounding, the target may be a specific industrial component or tabletop game piece defined entirely by the provided exemplar, not by any category in the training distribution. This distinction enables rigorous evaluation of whether models follow explicit guidance or exploit parametric memory—a capability essential for deployment in biological classification, industrial quality control, and scientific taxonomy.

3.2 Reference-Grounded Data Synthesis Engine

Constructing reference-dependent tasks at scale presents fundamental challenges. Suitable instances must satisfy stringent criteria: unambiguous targets, clean references, and queries where solutions genuinely depend on provided guidance rather than exploiting statistical shortcuts. Existing datasets rarely meet these requirements, and manual annotation proves prohibitively expensive. To address this bottleneck, we develop a fully automated synthesis engine operating through four sequential stages.

Semantic Taxonomy Expansion. To prevent overfitting to common object categories, we implement an adaptive taxonomy mining strategy grounded in WordNet (Miller, 1995). Starting from curated seed categories spanning artifacts, biological entities, and environmental objects, we recursively traverse the hypernym hierarchy (max depth=7) to extract long-tail concepts. We apply filtering heuristics to prune abstract terms and microscopic entities, retaining only visually distinctive, macroscopic objects. Category-specific branching constraints (max siblings=10 for animal/marine categories, 20 for others) prevent domination by dense biological subtrees, ensuring balanced coverage across diverse visual domains. This process yields thousands of concepts exhibiting semantic diversity far exceeding standard detection benchmarks.

Structured Scene Generation. Leveraging the mined taxonomy, we employ Qwen3-14B as a scene composition agent to generate structured metadata for text-to-image synthesis via Z-Image-Turbo (1344×768 resolution, 8 inference steps, guidance scale=0.0, bfloat16 precision). The agent

receives a mandatory spatial region for target placement, counteracting center bias. We prompt for cluttered environments where targets fit naturally but remain inconspicuous, and introduce long-tail concepts as soft guidance through illustrative examples rather than rigid constraints. The output is a structured specification containing target objects, background themes, distractor elements, and detailed generation prompts.

Closed-Loop Verification. A critical challenge in synthetic data is misalignment between generation prompts and actual visual content. Using Grounding DINO (Liu et al., 2024) as an open-vocabulary detector (box confidence=0.55, text similarity=0.50, max bbox area ratio=0.50, reference score≥0.65), we extract verified bounding boxes for all scene elements, treating generated images as the sole source of truth. Abstract prompt descriptions are overridden with detected labels, and geometric analysis computes precise spatial relationships. This closed-loop verification eliminates hallucination loops where models might exploit prompt artifacts.

Task Construction. Based on verified metadata, we synthesize reference-dependent visual tasks spanning grounding, counting, spatial reasoning, and attribute identification. For visual reference tasks, we generate isolated reference images by extracting target descriptors and rendering objects against neutral backgrounds. This workflow produces complete task tuples $\langle \mathcal{R}, \mathcal{I}_q, \mathcal{Q}, y^* \rangle$ with verified ground truth. The synthesized data serves dual purposes: partitioned into training and evaluation splits, it provides both diverse optimization signals and rigorous benchmark assessment.

3.3 Curriculum-Based RL with Visual Dependency Rewards

To enable MLLMs to master reference-dependent visual reasoning, we employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with a progressive curriculum and novel reward mechanisms designed to enforce genuine visual grounding.

Structured Output Format. We structure model outputs into an explicit chain-of-thought format with three components: `<clue>` for visual analysis of reference images, `<thinking>` for step-by-step reasoning connecting references to queries, and `<answer>` for final task outputs. This structure enables fine-grained reward assignment and interpretable reasoning chains.

Curriculum Stratification. We organize training into progressive stages based on sample difficulty. For each training instance, we execute multiple inference rollouts using the base model and compute the pass rate. Stage 1 comprises samples with moderate pass rates, targeting knowledge consolidation on achievable tasks. Stage 2 advances to samples with low pass rates, focusing on complex reasoning that challenges current capabilities. This curriculum prevents early-stage collapse on difficult instances while ensuring continued improvement on challenging cases.

Composite Reward Function. Our optimization objective combines four reward signals:

$$R_{\text{total}} = R_{\text{fmt}} + R_{\text{acc}} + R_{\text{noise}} + R_{\text{clue}} \quad (2)$$

The format reward R_{fmt} ensures structural compliance with the required output schema. The accuracy reward R_{acc} provides task-specific feedback based on answer correctness, computed via IoU for grounding tasks and exact match for classification.

Noisy Contrast Reward. A critical failure mode is blind correctness, where models produce correct answers by exploiting parametric priors while ignoring visual inputs. To penalize this behavior, we introduce contrastive evaluation against corrupted references. For each sample, we generate a degraded counterpart by applying severe augmentations that destroy reference image. We then compare performance on clean inputs (S_{clean}) versus corrupted inputs (S_{noise}):

$$R_{\text{noise}} = S_{\text{clean}} \cdot \mathbb{I}(S_{\text{clean}} > \theta \cdot S_{\text{noise}}) \quad (3)$$

where θ is a margin threshold. This mechanism assigns zero reward when models succeed despite corrupted references, forcing the policy to genuinely depend on valid visual cues.

Clue Verification Reward. To ensure generated reasoning faithfully reflects visual analysis rather than producing generic text, we implement a self-verification loop. We extract the textual content from generated `<clue>` tags and construct a new prompt where the visual reference is replaced entirely by this extracted description. The model is re-queried using only the textual clue and query image. If the clue accurately captures discriminative visual features, the model should solve the task

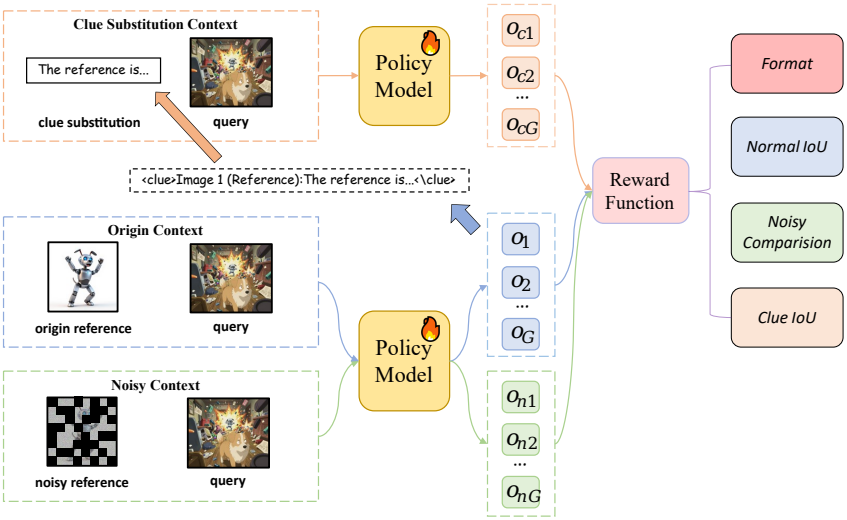

Figure 3: Training Framework

using text alone:

$$R_{\text{clue}} = \text{Acc}_{\text{verify}} \cdot \mathbb{I}(\text{Acc}_{\text{verify}} \geq \tau) \quad (4)$$

This reward compels explicit verbalization of visual features into actionable descriptions, bridging perception and reasoning while validating that generated clues contain genuine task-relevant information. Together, R_{noise} ensures predictions causally depend on visual inputs, while R_{clue} ensures reasoning chains faithfully reflect perceived content.

4 Experiments

4.1 Experimental Setup

Dataset Construction. We construct our dataset through two complementary sources. First, we reconstruct existing computer vision datasets including LVIS (Gupta et al., 2019) and domain-specific collections from Roboflow Universe spanning medical imaging, tabletop games, industrial products, and scientific taxonomy. These real-world sources contribute 1,236 samples (**68.8%** of the evaluation benchmark). Details are in Appendix B and Appendix H. Second, we employ our Reference-Grounded Data Synthesis Engine to generate 560 filtered synthetic instances (31.2%). The combined dataset is partitioned into training and evaluation splits with an 8:2 ratio, yielding approximately 1,800 evaluation samples spanning four task categories, formally establishing **RefGround-Bench**.

Training Configuration. We utilize Qwen3-VL-8B-Instruct as our backbone policy model. To stratify training difficulty, we execute $N = 8$ inference rollouts for each training sample using the base model and compute the pass rate k. Stage 1 comprises samples where $4 \leq k \leq 6$, targeting knowledge consolidation, while Stage 2 comprises samples where $1 \leq k \leq 3$, targeting complex reasoning. Samples with extreme pass rates ($k = 0$ or $k \geq 7$) are filtered out to ensure meaningful learning signals. We employ Group Relative Policy Optimization with a learning rate of 1×10^{-6}, KL coefficient of 0.01, and global batch size of 8. For the Noisy Contrast Reward, we apply Gaussian noise with standard deviation of 100 and patch blackening with probability 0.4. The margin threshold is set to $\theta = 0.9$. All experiments are conducted on 8 NVIDIA A6000 GPUs.

Category	Count
Data Sources	
Reconstructed Existing Dataset	1,236
Generated Synthetic Data	2,435
Selected Synthetic Data(Used)	560
Total Benchmark Size	**1,796**
Task Types	
Grounding	1,066
Counting	72
VQA	104
Multiple Choice	554

Table 1: Statistics of **RefGround-Bench**. The benchmark integrates reconstructed real-world data with filtered synthetic instances. For quality control in the synthetic subset, we specifically select samples where the base model achieves a pass rate of $1 \leq k \leq 4$ across 8 inference rollouts, ensuring moderate difficulty and validity.

Evaluation Protocol. We evaluate models across four task types with task-specific metrics. For

Model	Grounding	Counting	VQA	Multiple Choice	Overall
Proprietary Models					
Claude-Opus-4-5-20251101-Thinking	1.88	**36.11**	21.15	66.43	24.28
Gemini-3-Pro-Preview-Low	2.44	13.89	3.85	68.41	23.33
GPT-5.2	1.88	23.61	**27.88**	66.79	24.28
Open-source Model					
GLM-4.6V	18.20	20.83	25.96	78.34	37.31
Qwen3-VL-235B-A22B-Instruct	20.83	27.78	21.15	61.19	33.57
Qwen3-VL-235B-A22B-Thinking	21.95	33.33	**27.88**	70.58	37.75
Qwen3-VL-32B-Instruct	**22.05**	22.22	20.19	65.52	35.36
Qwen3-VL-32B-Thinking	19.42	27.78	19.23	72.92	36.25
Qwen3-VL-Plus	20.54	23.61	23.08	67.15	35.19
Qwen3-VL-8B-Thinking	10.23	5.56	15.38	61.01	26.00
Qwen3-VL-4B-Instruct	17.54	11.11	8.65	51.99	27.39
Qwen3-VL-4B-Thinking	10.51	4.17	13.46	65.88	27.51
Qwen3-VL-8B-Instruct(Base)	17.64	19.44	16.35	41.52	25.00
RefGround-Stage1	18.39	20.83	16.35	78.70	36.97
RefGround-Stage2	21.39	25.00	22.12	**78.70**	**39.25**

Table 2: Accuracy comparison by task type. We report the performance of various models without modifying their original identifiers. **Bold** and underline indicate the best and second-best results in each category. Prompt details are in Appendix K

grounding tasks, we compute Intersection over Union (IoU) between predicted and ground-truth bounding boxes, considering predictions correct when IoU exceeds 0.5. For other tasks, we employ exact match on extracted answers. We report accuracy percentages for each task category and overall weighted accuracy across the full evaluation set.

Baselines. We compare against proprietary models (Claude-Opus-4-5-Thinking, Gemini-3-Pro-Preview, GPT-5.2) and open-source models including the Qwen3-VL family (4B–235B, Instruct and Thinking variants) and GLM-4.6V.

4.2 Evaluation Results

Benchmarking Reference-Dependent Reasoning. Table 2 presents evaluation results across RefGround-Bench, which serves as a *diagnostic benchmark* designed to isolate and measure the specific capability of reference-dependent reasoning. The results reveal substantial limitations in current foundation models on reference-dependent tasks. Despite massive scale, proprietary models including GPT-5.2 and Claude-Opus achieve only 24.28% overall accuracy. The 235B-parameter Qwen3-VL-Thinking variant reaches 37.75%, while even the best-performing baseline GLM-4.6V attains 37.31%. These results validate our hypothesis that existing models struggle with strict reference-guided reasoning when visual contexts conflict with parametric priors. We note that the absolute accuracy levels reflect the inherent difficulty of tasks where pre-trained knowledge is deliberately insufficient.

Performance Analysis by Task Type. Examining task-specific performance reveals distinct patterns. In Grounding tasks, proprietary models perform particularly poorly ($< 2.5\%$), suggesting strong reliance on statistical priors rather than reference-based localization. RefGround-Stage2 achieves 21.39% on Grounding, comparable to the much larger Qwen3-VL-32B-Instruct (22.05%) and substantially outperforming the base model (17.64%). In Multiple Choice tasks, our approach achieves 78.70% accuracy, representing a significant improvement over the base model (41.52%) and surpassing all baselines including GLM-4.6V (78.34%). For VQA tasks, our curriculum strategy drives improvement from 16.35% to 22.12%, enabling our 8B model to outperform Qwen3-VL-235B-Instruct (21.15%). Detailed per-domain results are provided in Appendix I.

Overall Performance. RefGround-Stage2 achieves state-of-the-art accuracy of 39.25% on RefGround-Bench, a **+14.25%** absolute improvement over the base model (25.00%) and **+14.97%** over the best proprietary model (24.28%). The progressive improvement from Stage 1 (36.97%) to Stage 2 (39.25%) validates the curriculum

strategy. While the absolute accuracy reflects the extreme difficulty of the benchmark—where proprietary models achieve under 2.5% on grounding—our 8B model matching or exceeding 235B-parameter models confirms that targeted reference-dependency training is more effective than parameter scaling.

Model	MuirBench	BLINK
Qwen3-VL-8B-Thinking	**76.8**	68.7
Qwen3-VL-8B-Instruct(Base)	64.4	69.1
RefGround-Stage1	74.1	**69.5**
RefGround-Stage2	**76.8**	69.3

Table 3: Accuracy comparison on MuirBench and BLINK benchmarks. We report the performance of various models based on official technical reports.

4.3 Transferability to Unseen Benchmarks

To verify generalizability, we evaluate on **MuirBench** and **BLINK**, two challenging multi-image benchmarks. As shown in Table 3, RefGround demonstrates exceptional transferability. On MuirBench, it achieves a **+12.4%** improvement over the base model, matching specialized "Thinking" models by effectively anchoring reasoning in visual references. Meanwhile, on BLINK, performance remains stable ($\sim$69%), indicating that our RL strategy avoids the common "alignment tax" and preserves fundamental perceptual capabilities while enhancing complex reasoning.

Model / Configuration	Stage1 Score	Δ Improvement
Base (w/o R_{noise} & R_{clue})	35.74	-
w/o R_{clue} (Noisy only)	36.75	+1.01
w/o R_{noise} (Clue only)	35.18	-0.56
RefGround-Stage1 (Full)	**36.97**	**+1.23**

Table 4: Ablation study of the proposed R_{noise} and R_{clue} rewards on Stage 1 performance.

4.4 Ablation Study

We analyze the contribution of each reward component in Table 4.

- **Impact of Noisy Contrast (R_{noise}):** This is the primary performance driver (+1.01%), confirming that penalizing correct answers under corrupted contexts is essential to break prior reliance.
- **Synergy with Clue Reward (R_{clue}):** While R_{clue} alone causes regression (-0.56%) by encouraging "hallucinated reasoning," it complements R_{noise} effectively. The combination ensures reasoning is both structured (via Clue) and visually grounded (via Noise), yielding the best performance.

4.5 Qualitative Analysis

We provide a detailed case study in Appendix J, comparing GPT-5.2 and RefGround on a Chess Grounding task. The comparison reveals that GPT-5.2 reduces the reference to a coarse category ("a black piece") and defaults to prior-driven counting, while RefGround's structured `<clue>`→`<thinking>`→`<answer>` format enforces fine-grained feature extraction from the reference before reasoning over the query, enabling precise localization.

5 Conclusion

In this paper, we addressed the challenge of **Reference-Dependent Visual Reasoning**, identifying the critical conflict between inference-time visual guidance and memorized statistical priors. We proposed REFGROUND, a comprehensive framework that incorporates a *Reference-Grounded Data Synthesis Engine* for constructing contamination-free tasks and a curriculum-based Reinforcement Learning strategy utilizing *Noisy Contrast* and *Clue Verification* rewards to enforce visual dependency. Experiments demonstrate that REFGROUND enables a model with a modest parameter count to substantially outperform significantly larger state-of-the-art baselines, validating that explicit training for visual causality yields robust and generalizable reasoning capabilities.

Limitations

While RefGround demonstrates significant improvements in reference-dependent reasoning, several limitations remain, pointing towards important directions for future research.

Scarcity of High-Quality Real-World Data. Constructing a "clean" reference-dependent benchmark from real-world sources remains inherently difficult. Even when leveraging massive repositories like LVIS and Roboflow, filtering for samples that simultaneously satisfy high-quality criteria—unambiguous targets, clean references, and disjoint query contexts—yields a surprisingly low

acceptance rate. The natural distribution of real-world images is often too noisy or lacks the specific long-tail categories required to rigorously test fine-grained reference following, limiting the scale of "natural" training data we can acquire. Consequently, this scarcity disproportionately restricts the construction of tasks demanding higher data fidelity, resulting in a limited volume of available samples for categories such as VQA and Counting.

Dependency on Verification Pipelines. Our synthetic engine relies on a "generate-then-verify" paradigm to circumvent the limitations of current Text-to-Image (T2I) models. Since existing generative models struggle to strictly follow complex spatial and attribute binding instructions, we cannot trust the generated images to perfectly match the metadata. Consequently, we are forced to employ an auxiliary open-vocabulary detector (Grounding DINO) to extract ground-truth metadata post-hoc. This "detour" increases computational overhead and caps the complexity of our synthetic tasks by the performance ceiling of the detector itself.

The Modality Gap. Fundamentally, our approach optimizes the reasoning policy but does not alter the underlying model architecture. Current LMMs suffer from the well-documented "Modality Gap", where visual details are compressed or lost during the encoding phase before interacting with the language component. While RefGround aligns the model's behavior to respect visual signals, it cannot recover fine-grained features (e.g., subtle texture differences in similar objects) that may have been discarded by the visual encoder (e.g., CLIP, SigLIP) due to this architectural bottleneck. Several short-term mitigations could partially address this gap: (1) unfreezing the visual encoder for co-training alongside the RL policy, allowing gradient signals from reference-dependent tasks to refine low-level feature extraction; (2) leveraging high-resolution dynamic cropping strategies native to architectures like Qwen-VL to preserve fine-grained details; and (3) employing multi-scale feature fusion to retain both global context and local texture information. We leave exploration of these complementary strategies to future work, as our primary contribution establishes the reasoning-level alignment that serves as a necessary foundation for any architectural enhancement.

Future Outlook. Looking ahead, we envision two primary pathways to address these challenges:

- **Scalable Synthesis via Next-Gen T2I:** As T2I models evolve towards precise instruction following and better spatial control, we aim to streamline our engine, eliminating the need for post-hoc verification. This will allow for the direct, scalable generation of highly complex, fine-grained datasets that are currently impossible to construct.

- **Architecture-Agnostic Adaptation:** We plan to extend RefGround to emerging native multimodal architectures. By integrating our reference-dependent objectives into models with unified tokenization (reducing the modality gap), we hope to unlock even deeper synergies between visual perception and logical reasoning.

6 Ethics Statements

This work presents RefGround, a framework and benchmark designed to enhance the reliability of Large Multimodal Models (LMMs). The RefGround-Bench dataset is constructed using a combination of public datasets (LVIS, Roboflow) and synthetic data generated via a strictly controlled engine. We strictly adhere to the licenses of all source datasets and have implemented rigorous filtering protocols to ensure the synthetic data contains no personally identifiable information (PII) or offensive content. Furthermore, our research explicitly aims to mitigate model hallucinations and statistical biases by enforcing genuine visual dependency, thereby contributing to the development of safer and more interpretable AI systems.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Anthropic. 2025. Anthropic model report: Claude opus 4.5. https://www.anthropic.com/transparency/model-report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,

Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. 2025. Emerging properties in unified multimodal pretraining. *CoRR*, abs/2505.14683.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: multimodal large language models can see but not perceive. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, Lecture Notes in Computer Science, pages 148–166. Springer.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *CoRR*, abs/2408.12637.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, Lecture Notes in Computer Science, pages 38–55. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

OpenAI. 2025. Update to gpt-5 system card: Gpt-5.2. `https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. 2025. Less-to-more generalization: Unlocking more controllability by in-context generation. *CoRR*, abs/2504.02160.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *CoRR*, abs/2412.10302.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaxing Huang. 2025. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *CoRR*, abs/2505.16673.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, and 3 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 57116–57198. PMLR / OpenReview.net.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, and 4 others. 2024. MM1.5: methods, analysis & insights from multimodal LLM fine-tuning. *CoRR*, abs/2409.20566.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

A Detailed Task Definitions

We provide comprehensive definitions for the four task categories constituting the RefGround-Bench:

Reference-Based Grounding. Given a visual reference $\mathcal{R}_v$ depicting a target object, the model must localize instances of the same object category within the query image $\mathcal{I}_q$. The challenge lies in distinguishing the target from visually similar distractors based solely on features extracted from the reference, rather than relying on semantic category priors.

Reference-Based Counting. Extending grounding to enumeration, this task requires counting all instances matching the visual reference within the query scene. Success demands both accurate instance discrimination and exhaustive spatial coverage.

Rule-Based Visual Question Answering. Given textual specifications $\mathcal{R}_t$ defining classification criteria, domain-specific rules, or attribute mappings, the model must apply these rules to analyze the query image. This tests the capacity to ground abstract textual definitions in concrete visual evidence.

Reference-Conditioned Multiple Choice. This category is specifically designed to evaluate **Spatial Reasoning** and **Attribute Identification**. Evaluating these capabilities through open-ended VQA poses significant challenges, as correct reasoning can often be expressed in varied, unstructured natural language that is difficult to parse automatically. By constraining the output space to a multiple-choice format, we isolate the model's ability to utilize reference information from the complexities of open-ended generation, ensuring that the evaluation strictly measures reasoning accuracy.

B Reconstructing Existing Datasets for Reference-Dependent Reasoning

In this section, we provide detailed protocols on how we reconstructed existing datasets to evaluate reference-dependent reasoning, specifically focusing on the one-shot localization setup for LVIS and context-dependent benchmarks from Roboflow Universe.

B.1 Reconstructing LVIS for One-Shot Localization

To establish a robust baseline for visual reference following, we leverage the Large Vocabulary Instance Segmentation (LVIS) dataset. With over 1,200 categories, LVIS provides a natural long-tail distribution that challenges models beyond common objects. However, standard supervised training on LVIS allows models to memorize category-specific features rather than learning to verify them against a context. To transform this into a test of reference-dependent reasoning, we reconstruct the dataset into a **One-Shot Object Localization** benchmark, applying a rigorous "Hard Mining" strategy to filter out trivial instances.

Disjoint Reference-Query Pairing. For each target category, we construct pairs consisting of a *Reference Image* and a *Query Image*. To prevent the model from exploiting low-level artifacts or performing simple pattern matching, we strictly enforce that the reference crop and the query scene originate from distinct source images ($I_{ref} \neq I_{query}$). The reference image is cropped to the object's bounding box to serve as a focused visual definition, while the query image remains the full scene, requiring the model to locate the corresponding object based solely on the provided visual cue.

Difficulty-Aware Sampling. LMMs often rely on pre-trained saliency maps to detect large, central objects, ignoring the visual prompt. To counteract this, we implement a multi-stage filtering pipeline to mine "Hard" samples:

- **Scale Constraint:** We sort instances by their object-to-image area ratio, prioritizing samples where the target occupies a minimal portion of the scene (specifically, area ratio < 0.35).
- **Background Interference:** We quantify the distinctiveness of the object against its environment by computing the histogram divergence between the object crop and the full scene. We select samples with high interference (score > 1500), ensuring the object does not trivially "pop out" from the background.

Semantic Consistency Verification. To ensure that the difficulty stems from reasoning complexity rather than data noise or occlusion, we employ a CLIP-based filtration step. We calculate the semantic similarity between the reference crop and the detected target in the query image. We retain only those pairs where the visual similarity exceeds a strict threshold (e.g., > 0.55). This ensures that while the task is challenging due to scale and clutter, the reference remains a valid and semantically consistent prototype for the target object.

B.2 Constructing Context-Dependent Benchmarks from Long-Tail Domains

To strictly evaluate the model's ability to reason from provided references rather than parametric memory, we turn to the **Roboflow Universe**, a repository characterized by its rich collection of long-tail, domain-specific, and often user-defined computer vision datasets. Unlike standard benchmarks (e.g., COCO or ImageNet) where categories are semantically distinct and widely memorized by LMMs, these datasets often involve fine-grained classifications or specialized tasks that represent a "cold start" scenario for general-purpose models.

We meticulously select subsets where the task is *impossible* to solve without explicit contextual guidance. Based on the modality of the necessary guidance, we categorize these tasks into two distinct types of reference-dependent reasoning:

Visual Reference: Fine-Grained Instance Matching. For tasks requiring precise visual identification of specific objects, we employ an automated pipeline (as implemented in our data processing script) to reconstruct standard detection datasets into a one-shot format. Specifically, we extract a cropped region of a target category from a support image to serve as the *visual reference*, and require the model to locate the corresponding object within a query image. We selected datasets where semantic labels alone are insufficient, such as:

- **Tabletop Games:** Locating specific tiles in *Mahjong* or identifying cards in custom *Poker* decks. Here, the model cannot rely on the generic concept of a "tile" or "card" but must match the specific pixel-level patterns provided in the reference crop.
- **Industrial Quality Control:** Distinguishing between visually similar product variants, such as identifying a specific flavor of *chocolate* based on surface texture or distinct packaging details defined only in the reference image.

This format forces the model to perform feature mapping between the reference crop and the query scene, explicitly suppressing the bias to output high-frequency classes from its pre-training distribution.

Textual Reference: Rule-Based Logical Grounding. While visual references address "what to look for," textual references address "how to judge." We curate datasets where the definition of categories relies on expert knowledge or arbitrary rules that may contradict or refine general common sense. In these scenarios, the *reference* is provided as a textual description or rule set. Examples include:

- **Scientific Taxonomy:** Identifying stages of *Mitosis* in microscopic cells or classifying *Cloud* types. The model must map visual features (e.g., "chromosomes pulling apart") to the definitions provided in the text, rather than hallucinating based on generic cell imagery.
- **Positional Rules:** Mapping players in *Rugby* to their roles (e.g., "Scrum-half") based on formation rules described in the prompt, rather than visual appearance alone.
- **Symbolic Interpretation:** Translating *Sign Language* gestures where the mapping between hand pose and meaning is strictly defined by the provided dictionary context.

By reorganizing these long-tail datasets into this reference-query format, we ensure that RefGround is trained to treat the provided context—whether visual pixels or textual rules—as the primary source of truth, thereby mitigating hallucination in unseen domains.

C Generative AI-driven Synthetic Data Engine Details

While the reconstruction of existing computer vision datasets serves as a necessary initial step, relying solely on existing sources presents distinct limitations. Standard datasets are finite, often permeate the pre-training corpora of modern LMMs (leading to data leakage), and rarely yield samples of appropriate difficulty that are conducive to the effective construction of reasoning tasks. To address these challenges and support sustainable scaling, we introduce a fully automated **Synthetic Data Engine**. This engine not only **reduces the barrier to** data creation by operating at **low** marginal cost but also provides a **scalable** stream of **clean**, contamination-free samples. The engine operates in four sequential stages and is designed for **high** efficiency on consumer-grade hardware.

C.1 Semantic Taxonomy Expansion

- **Filtering Heuristics:** We apply specific filters to prune abstract concepts (e.g., “measurement”, “process”) and biological micro-entities (e.g., “cells”, “tissues”), focusing solely on visually distinct, macroscopic objects.
- **Balanced Sampling:** To counteract the natural dominance of certain classes (e.g., animals), we enforce dynamic sibling limits during tree traversal, ensuring a long-tail distribution of concepts that prevents visual repetition and mode collapse during the generation process.

C.2 Spatial-Aware Scene Generation

1. **Spatial Anchoring:** The LLM is assigned a mandatory spatial region (e.g., a 3×3 grid location like *Bottom-Right*) for the target object.
2. **Contextual Density:** The prompt engineering explicitly requests “cluttered, **visually complex**, or dynamic” settings where the target object fits naturally but is not centrally conspicuous. This prevents trivial solutions and forces the vision model to perform active searching.
3. **Soft Concept Guidance:** We introduce the mined concepts as illustrative examples (within *i.e.* clauses) rather than absolute requirements. This implicit prompting strategy effectively diversifies the semantic distribution beyond common categories while preserving the model’s creative freedom. By avoiding the forced inclusion of obscure terms that might degrade scene coherence, we ensure a balance between semantic diversity and visual realism.
4. **Prompt Output:** The system outputs a JSON structure containing the target object, background theme, surrounding distractor objects, and a detailed T2I prompt.

C.3 Closed-Loop Verification: Constraint Enforcement

- **Semantic Realignment:** The initial descriptions from the **Generative** LLM are often ab-

stract (e.g., "a vessel of ancient murmurs"). Grounding DINO filters these through a detection process. We force the final QA module to use the **verified detected label** (e.g., "cracked clay pot") rather than the original prompt. While Grounding DINO acts as a proxy metric rather than an **absolute ground truth**, this process eliminates "hallucination loops" and significantly enhances factual alignment compared to unverified generation.

- **Geometric Constraints:** We reject vague spatial descriptions. Instead, we compute precise geometric metadata regarding angle and size ratios. This metadata serves as a **strict constraint**. The final LLM is allowed to phrase the question naturally, but it is strictly forbidden from altering the facts provided by these hard constraints.

- **Human-in-the-Loop Validation:** To further guarantee data fidelity, we perform random manual spot-checks on the filtered dataset. This qualitative verification confirms that the visual content strictly aligns with the generated reasoning tasks, serving as a final safeguard against subtle generation artifacts.

D Taxonomy Construction Details

To ensure semantic diversity and avoid concept stagnation, we curated a set of seed categories and applied strict negative filtering to remove non-visual or microscopic entities.

D.1 Seed Categories

The taxonomy expansion initiates from the manually curated seed mapping in WordNet, ensuring a coverage of distinct, tangible objects across biological and artificial domains.

Super-Category	Seed Concepts
Animals	placental, bird, insect...
Vehicles	car, vessel, aircraft...
Architecture	building, shelter, shop
Interior & Household	furniture, seat, home_appliance...
Tools & Devices	electronic_device, hand_tool...
Clothing	garment, headgear, footwear
Food	edible_fruit, vegetable...

Table 5: Curated Seed Categories for Taxonomy Mining.

D.2 Filtering Blocklist

We implement a blocklist mechanism to prune nodes that are semantically abstract, biological internal structures, or otherwise unsuitable for visual object detection. The following keywords trigger the exclusion of a synset:

> *genus, family, order, group, class, tissue, organ, cell, muscle, vein, artery, nerve, part, portion, fragment, section, piece, substance, matter, fluid, liquid, gas, flesh, meat, remains, carcass, body, limb, excrement, waste, secretion, abstract, abstraction, relation, measure, unit, quantity, phenomenon, process, state.*

E Prompt Templates

We employ distinct system prompts for the Query Image Generation and the QA Generation.

E.1 Query Image Generation

The following prompt instructs the LLM to hallucinate a scene based on the mined concepts and spatial constraints.

Query Image Generation System Prompt

System: You are a Creative Visual Director. Your task is to design a unique, vivid scene for a Text-to-Image dataset based on specific constraints.
INSTRUCTIONS: 1. **INPUTS:** You will be given suggested concepts and a mandatory spatial location. 2. **OBJECT:** Select a target object based on the suggestions (or a creative expansion of them). 3. **SPATIAL CONSTRAINT:** The target object MUST be located in the specified region of the image. 4. **COMPLEXITY:** The scene must be cluttered, atmospheric, or dynamic. The object should fit naturally but might be hard to spot immediately. 5. **FORMAT:** Return JSON only.
User Instruction Template: Random Seed ID: {unique_id} Task: Invent a completely random, unique scene and a target object. 1. **Target Object Suggestions (e.g.)**: {suggestion_str}. 2. **Mandatory Location**: The target object must appear in the **{spatial_loc}** of the frame. Output the scene description specifically for a Text-to-Image model. Make it hard to spot the object.

System Prompt Configuration (JSON Object)


```
{
    "id": 4,
    "data": {
        "setting_theme": "Postwar underground sanctuary with overgrown flora",
        "target_object": "fallout shelter door handle",
        "spatial_location": "Middle-Left",
        "surrounding_objects": [
            "moss-covered concrete walls",
            "entangled ivy",
            "dripping water from cracks",
            "rustic wooden chair with torn cushion",
            "a broken lamp casting long shadows",
            "a pile of moldy books",
            "a map of the subway system with faded ink"
        ],
        "t2i_prompt": "A postwar underground sanctuary, overgrown with lush greenery and moss-
             covered concrete walls, with entangled ivy and dripping water from cracks. In the
             middle-left, a rusted fallout shelter door handle is partially hidden by a cluster
              of vibrant ferns and creeping vines. A rustic wooden chair with a torn cushion
             sits nearby."
    }
}
```


Figure 4: Original Metadata

E.2 QA Generation

This prompt governs the logic for determining whether to use support images and how to construct the final reasoning tasks. We use the concept of few-shot learning to make model easy to comprehend our aim.

Task Construction System Prompt

System: You are an AI Expert creating a Few-Shot Learning Visual Question Answering dataset. Your task is to generate questions about an image based on metadata. Use `[support]` in the question and generate a T2I prompt.
STRICT RULE FOR T2I PROMPTS (Support Images): - The prompt MUST describe **ONLY the object itself** (color, texture, shape). - **FORBIDDEN**: Do NOT describe the environment, location, furniture, or background.

F Metadata Examples

F.1 Original T2I Metadata

F.2 Updated T2I Metadata

Formatted Data Entry (ID: 4)


```
{
    "id": 4,
    "data": {
        ...
    },
    "grounding_dino_result": {
        "fallout shelter door handle": [
            {
                "bbox": [549.62, 273.07, 603.36, 388.04 ],
                "score": 0.6571,
                "label_matched": "fallout shelter door handle"
            }
        ],
        ...
    }
}
```


Figure 5: Updated Metadata

G Filtering and Verification Metrics

G.1 Detection Thresholds

In the Closed-Loop Verification stage, we utilize Grounding DINO with specific confidence thresholds to balance recall and precision.

Parameter	Value
Box Confidence Threshold	0.55
Text Similarity Threshold	0.50
Maximum BBox Area Ratio	0.50
Reference Object Score Threshold	0.65

Table 6: Hyperparameters for Grounding DINO Verification.

G.2 Spatial Relation Definitions

Spatial relationships are rigorously defined based on the angle θ (in degrees) between the centroid of the target box (C_t) and the reference box (C_r). The angle is calculated using $\text{atan2}(dy, dx)$, normalized to $[0, 360)$.

$$\text{Relation} = \begin{cases} \text{Right} & \text{if } 337.5^\circ \leq \theta < 22.5^\circ \\ \text{Bottom-Right} & \text{if } 22.5^\circ \leq \theta < 67.5^\circ \\ \text{Bottom} & \text{if } 67.5^\circ \leq \theta < 112.5^\circ \\ \text{Bottom-Left} & \text{if } 112.5^\circ \leq \theta < 157.5^\circ \\ \text{Left} & \text{if } 157.5^\circ \leq \theta < 202.5^\circ \\ \text{Top-Left} & \text{if } 202.5^\circ \leq \theta < 247.5^\circ \\ \text{Top} & \text{if } 247.5^\circ \leq \theta < 292.5^\circ \\ \text{Top-Right} & \text{if } 292.5^\circ \leq \theta < 337.5^\circ \end{cases} \quad (5)$$

H Subset Examples(QA is simplified)

H.1 Chess

Task: Reference-based Object Grounding and Counting The Chess dataset consists of images of chess pieces collected from publicly available sources. Each task provides a reference image showing a specific chess piece and a query image containing multiple pieces. The model is required to either locate the object in the query image that matches the reference (grounding) or count the number of pieces that match the reference (counting). This evaluates fine-grained object recognition, spatial localization, and counting abilities.

Question: Locate the object in the query image that is most similar to the object shown in the reference image.

Answer: [1638, 975, 1808, 1170].

Question: Count the objects in the query image that are most similar to the object shown in the reference image.

Reference image

Query image

Answer: 2.

H.2 Chocolate

Task: Reference-based Fine-grained Object Grounding The Chocolate dataset contains images of various chocolate products from publicly available sources. Each task provides a reference image of a chocolate and a query image containing multiple chocolates. The model is required to locate the chocolate in the query image that matches the reference, testing fine-grained visual similarity recognition.

Reference image

Query image

Question: Locate the chocolate in the query image that is most similar to the object shown in the reference image.

Answer: [1182, 545, 1336, 790].

H.3 Cloud

Task: Rule-based Cloud Pattern Classification (VQA) The Cloud dataset contains images of different cloud formations collected from publicly available sources. Instead of a reference image, textual definitions describe the four cloud patterns: Sugar, Gravel, Flowers, and Fish. The model is required to classify the query image according to these rules.

Question: Distinguish the cloud pattern in the image based on the textual description of Sugar, Gravel, Flowers, and Fish.

Answer: Gravel.

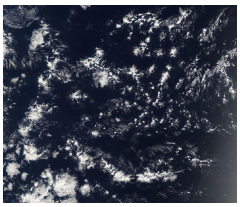

Figure 6: Query image

H.4 Currency

Task: Text-referenced Grounding and Counting The Currency dataset uses images of coins with textual descriptions specifying the target coin (e.g., size, color, figure on coin). The model is required to either locate the coin in the query image matching the description (grounding) or count how many instances exist (counting).

Figure 7: Query image

Question: Locate the Dime in the query image. It is the smallest silver-colored coin, thin and light. It features a profile of Franklin D. Roosevelt on the obverse.

Answer: [259, 339, 290, 378].

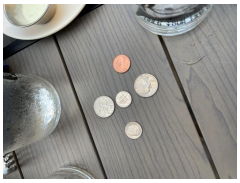

Figure 8: Query image

Question: Count the Dimes in the image. It is the smallest silver-colored coin, thin and light. It features a profile of Franklin D. Roosevelt on the obverse.

Answer: 2.

H.5 Figure

Task: Rule-based Hand Gesture Recognition (VQA) The Figure dataset contains images of hand gestures representing numbers from 0 to 13. Reference is textual rules defining hand configurations for each number. The model is required to identify the number shown in the query image based on the rule definitions.

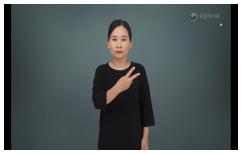

Figure 9: Query image

Question: Identify the number represented by the hand gesture in the image.

Answer: 0.

H.6 Football

Task: Role-based Player Grounding The Football dataset contains images of American football formations. Reference is textual definition of player roles and spatial relationships (e.g., Quarterback is behind the Center). The model is required to locate the player of interest in the query image.

Figure 10: Query image

Question: Locate the Quarterback in the football snap formation.

Answer: [781, 280, 803, 343].

H.7 Lung

Task: Rule-based Chest X-ray Diagnosis (VQA) The Lung dataset contains chest X-ray images with textual definitions of common conditions. The model is required to identify the condition present in the query image based on these rules.

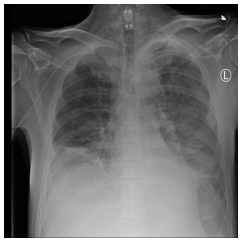

Figure 11: Query image

Question: Identify the condition present from the list: Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, Pneumothorax.

Answer: Atelectasis.

H.8 Mahjong

Task: Reference-based Fine-grained Object Grounding The Mahjong dataset contains images of Mahjong tiles. Reference image shows a single tile, query image has multiple tiles. The model is required to locate the tile most similar to the reference.

Reference image

Query image

Question: Locate the object in the query image that is most similar to the object shown in the reference image.

Answer: [849, 304, 980, 440].

H.9 Mitosis

Task: Rule-based Mitotic Cell Grounding The Mitosis dataset contains microscopic images of cells. Reference is textual description of mitotic cells. The model is required to locate mitotic cells in the query image based on these features.

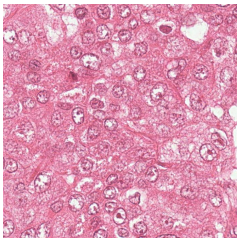

Figure 12: Query image

Question: Locate the mitotic cell in the image. Mitotic cells are characterized by condensed, hyperchromatic (darkly stained) chromatin, irregular shapes with hairy or spiky projections, and the absence of a clear nuclear envelope. They appear denser and more chaotic than the surrounding interphase cells.

Answer: [182, 195, 244, 235].

H.10 Detailed Dataset-Level Analysis

In this section, we provide a fine-grained evaluation across specific data sources. Table 7 presents accuracy across individual datasets. Synthetic datasets yield substantially higher performance, particularly for thinking variants, indicating that models benefit from structured and pattern-consistent scenarios.

In contrast, real-world and long-tail datasets such as LVIS-Hard, Figure, and Football remain difficult across all models. Even the strongest baselines achieve below 11% accuracy on LVIS-Hard, reflecting persistent failures in fine-grained reference grounding under cluttered conditions. Interestingly, smaller models occasionally outperform larger ones on specific datasets, suggesting that scale alone does not guarantee robust reference utilization.

I Evaluations on Domains

J Qualitative Comparison: GPT-5.2 vs. RefGround

We present a case study on a Chess Grounding task to illustrate the difference between prior-dominated reasoning and reference-dependent reasoning. Both models receive the same inputs: a reference image depicting a Black Rook and a query image of a cluttered chessboard. The task is to locate the object matching the reference.

Reference image

Query image

GPT-5.2 (Incorrect)

Output: "The reference image shows a black chess piece. Looking at the query image, I can identify the black pieces on the board..."
Answer: 5 ✗

Analysis: GPT-5.2 coarsely recognizes the reference as "a black chess piece" without extracting discriminative features. It then counts *all* black pieces on the board rather than locating the specific rook, demonstrating a reliance on categorical priors over fine-grained reference matching.

RefGround (Correct)

<clue>: "The reference depicts a rook—a tall piece with a sturdy cylindrical base and a flat top featuring a distinctive rectangular notch (crenellation). The surface is smooth and dark-colored."
<thinking>: "Scanning the query chessboard for pieces matching this specific design. I can clearly see one piece with the characteristic crenellated top and cylindrical body, situated near the right side of the board."
Answer: 1 ✓

This comparison highlights a key failure mode: GPT-5.2 reduces the reference to a coarse category ("black piece") and defaults to counting by prior, while RefGround's structured <clue>→<thinking>→<answer> format enforces fine-grained feature extraction from the reference before reasoning over the query, enabling precise localization.

Table 7: Accuracy across individual datasets (%). Synthetic scores are averaged over three sub-collections.

Model	Chess	Choc.	Cloud	Curr.	Fig.	Foot.	Lung	Mahj.	Mito.	LVIS	Synth.
Claude-Opus-4-5-20251101-Thinking	36.07	0.00	25.00	20.00	7.14	0.00	18.75	1.79	0.00	2.07	65.71
Gemini-3-Pro-Preview-Low	16.39	6.25	10.00	0.00	0.00	0.00	0.00	5.36	0.00	1.52	67.86
GLM-4.6V	60.66	41.25	**40.00**	40.00	3.57	**13.33**	17.19	30.95	10.00	9.66	77.86
GPT-5.2	19.67	3.75	25.00	35.00	3.57	0.00	**29.69**	0.60	0.00	1.79	66.07
Qwen3-VL-235B-A22B-Instruct	60.66	**65.00**	22.50	**50.00**	**10.71**	6.67	20.31	38.69	30.00	8.97	60.89
Qwen3-VL-235B-A22B-Thinking	**67.21**	61.25	30.00	45.00	7.14	**13.33**	26.56	40.48	**45.00**	10.21	70.18
Qwen3-VL-32B-Instruct	63.93	58.75	27.50	30.00	7.14	6.67	15.62	**44.64**	30.00	9.79	65.36
Qwen3-VL-32B-Thinking	57.38	63.75	20.00	**50.00**	3.57	3.33	18.75	42.86	35.00	6.90	72.14
Qwen3-VL-Plus	57.38	61.25	25.00	**50.00**	7.14	6.67	21.88	38.10	30.00	8.97	66.96
Qwen3-VL-8B-Thinking	14.75	21.25	30.00	15.00	0.00	0.00	6.25	22.62	10.00	6.07	60.36
Qwen3-VL-4B-Instruct	34.43	36.25	10.00	20.00	**10.71**	6.67	7.81	33.33	15.00	10.48	51.61
Qwen3-VL-4B-Thinking	6.56	30.00	20.00	10.00	3.57	0.00	9.38	19.64	10.00	6.34	65.71
Qwen3-VL-8B-Instruct(Base)	49.18	48.75	17.50	25.00	3.57	3.33	15.62	29.17	25.00	9.79	41.25
RefGround-Stage1	45.90	55.00	20.00	25.00	10.71	3.33	12.50	30.95	25.00	9.38	78.75
RefGround-Stage2	57.38	53.75	22.50	20.00	14.29	3.33	21.88	36.31	15.00	**12.28**	**78.93**

K Inference and Evaluation Settings

To ensure a fair and reproducible evaluation, we detail the specific prompting strategies and inference configurations used for our model and the baseline models.

K.1 RefGround Inference Setup

For our proposed **RefGround**, we maintain strict consistency between training and inference. We utilize the exact same System Prompt used during the RL training phase. This prompt enforces a structured `<clue>` → `<thinking>` → `<answer>` format, compelling the model to explicitly analyze every image in the input stream (both reference and query) before deducing the answer.

The full System Prompt is provided below:

RefGround System Prompt

You are a visual AI assistant. Reference information is always provided to guide your task. The input will strictly follow ONE of these two scenarios:
1. Text Reference: "reference_text" is provided, along with exactly one image (the "query_image").
2. Image Reference: No text is provided; instead, one or more "reference_images" are provided followed by the "query_image".
Crucial: The "query_image" is ALWAYS the last image provided. Any preceding images are "reference_images".
Your goal is to perform the requested visual task using the following strict XML structure:
1. <clue> Tag: Visual Analysis
MANDATORY REQUIREMENT: You MUST generate a separate <clue> tag for EVERY image in the input stream, strictly in the order they appear.
- Check the total number of input images (N). You MUST output exactly N <clue> tags.
- Inside each tag, start by explicitly stating the image role: "Image [i] (Reference)" or "Image [N] (Query)".
- **Warning:** Do NOT skip the reference images. Do NOT only describe the query image.
- Content Structure: - If Image i is a Reference: Describe the visual pattern, key object, or defining attribute that serves as the "Source of Truth". - If Image i is the Query (Last Image): Describe the visual features relevant to the target task.
2. <thinking> Tag: Reasoning Process
Detail the step-by-step logic connecting your <clue> analysis to the solution.
- Step 1: Synthesize the "Source of Truth" from the reference <clue> tags (or reference text).
- Step 2: Apply this criteria to the query_image.
- Step 3: Derive the final answer.
3. <answer> Tag: Final Output
Output ONLY the final result using the specific format for the task type:
- Grounding: {"bbox_2d": [x1,y1,x2,y2]}
- Counting: An Arabic numeral (e.g., 5).
- Multiple Choice / QA: The exact option, object name, or key phrase only.
Format Examples:
- (Text Reference - 1 Image total): <clue> Image 1 (Query): ...analysis... </clue> <thinking> ... </thinking> <answer> ... </answer>
- (2 References + 1 Query - 3 Images total): <clue> Image 1 (Reference): ...pattern... </clue> <clue> Image 2 (Reference): ...pattern... </clue> <clue> Image 3 (Query): ...analysis... </clue> <thinking> ... </thinking> <answer> ... </answer>

K.2 Baseline Models Setup

For baseline models (e.g., Qwen-Thinking series, Claude-Thinking), we do not impose the complex `<clue>` structure, as they were not fine-tuned with this specific granular requirement. Instead, to allow them to leverage their inherent reasoning capabilities, we only require a standard Chain-of-Thought format <think></think><answer></answer>. This ensures that baselines are evaluated on their ability to reason naturally without being penalized for failing to follow an unseen formatting rule.

K.3 Special Handling for Gemini Models

We explicitly utilize the `Gemini-3-Pro-Preview-Low` variant for our evaluations. During our preliminary experiments with the standard `Gemini-3-Pro-Preview` model, we observed a persistent failure mode in reference-dependent tasks: the model frequently entered recursive self-verification loops within its reasoning chain. This behavior resulted in excessive token generation without converging to a valid output, often exceeding context limits or timing out.

To address this, we adopted the `Low` reasoning effort variant and constrained the prompt to strictly prioritize the final answer. This configuration successfully mitigated the infinite loop issue, allowing the model to produce valid, evaluatable responses.