KnowDR-REC: Auditing Knowledge-Conditioned Visual Grounding in Referring Expression Comprehension

Guanghao Jin[1,2,3,4†], **Jingpei Wu**[4†], **Tianpei Guo**[2], **Yiyi Niu**[5], **Weidong Zhou**[1,2], **Linyi Yang**[3*], **Guoyang Liu**[1,2*]
[1]Shenzhen Research Institute of Shandong University, Shenzhen 518000, China
[2]School of Integrated Circuits, Shandong University, Jinan 250199, China
[3]Southern University of Science and Technology, Shenzhen, China
[4]LMU Munich, Munich, Germany
[5]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
G.Jin@campus.lmu.de, jingpei.wu@lmu.de, 202200201042@mail.sdu.edu.cn
yiyi.niu@tum.de, wdzhou@sdu.edu.cn, yangly6@sustech.edu.cn, gyliu@sdu.edu.cn
† Equal contribution
* Corresponding author

Abstract

While Multimodal Large Language Models (MLLMs) have demonstrated the capacity for multi-modal reasoning, current Referring Expression Comprehension (REC) benchmarks lag behind, predominantly relying on intra-image cues and neglecting the integration of external world knowledge, which significantly impedes the evolution of REC towards real-world applications. This limitation obscures a model's true capability to conduct textual reasoning (entity resolution), resolve spatial location (visual grounding), and verify reference validity (hallucination rejection). To address this, we introduce KnowDR-REC, a targeted audit benchmark comprising 1,042 positive triplets derived from real-world knowledge, along with rigorously matched negative samples. Unlike traditional datasets, we implement a controllable counterfactual evaluation mechanism that subjects textual expressions to single-factor perturbations (entity, relation, or time) to test sensitivity to fine-grained factual changes. Extensive evaluation of 18 state-of-the-art MLLMs exposes a critical "binding hallucination," revealing that current high performance is often built on fragile visual shortcuts rather than true understanding. KnowDR-REC thus serves as a pivotal diagnostic instrument, steering future research toward the genuine integration of perception and reasoning.

1 Introduction

Multimodal large language models (MLLMs) have achieved remarkable progress in vision–language understanding tasks (Liu et al., 2023; Zang et al., 2025). In particular, in referring expression comprehension (REC) (Qiao et al., 2020), representative models such as CogVLM have reached near-saturated performance on the RefCOCO family of benchmarks (Yu et al., 2016; Wang et al., 2024a). Existing evaluation metrics therefore appear to suggest that these models have acquired fine-grained visual grounding capabilities. However, we argue that high accuracy on standard benchmarks may mask fundamental limitations in the underlying reasoning mechanisms. Most prior work primarily exploits intra-image visual cues, such as color attributes or spatial relations, to localize target objects (Wu et al., 2020), while largely overlooking the fact that human referring expressions frequently depend on external knowledge that spans temporal, factual, and relational dimensions.

Consider the query, "Who is the flag bearer of the US delegation at the 2024 Paris Olympics?" Correctly resolving such a request within the REC framework requires more than visual pattern matching; it requires a reasoning process naturally decomposed into three stages: Who (identifying the correct individual), Where (visually grounding the resolved entity) and Whether (verifying existence to abstain when the entity is absent or the description contradicts reality. However, current evaluation protocols overwhelmingly emphasize the Where stage, providing little independent scrutiny of the Who and Whether components. Consequently, it remains unclear if strong model performance reflects genuine knowledge-conditioned reasoning or merely the exploitation of visual shortcuts, such as selecting the most salient object (Liu et al., 2019). To systematically evaluate this gap, we in-

Findings of the Association for Computational Linguistics: ACL 2026, pages 38607–38629
July 2-7, 2026

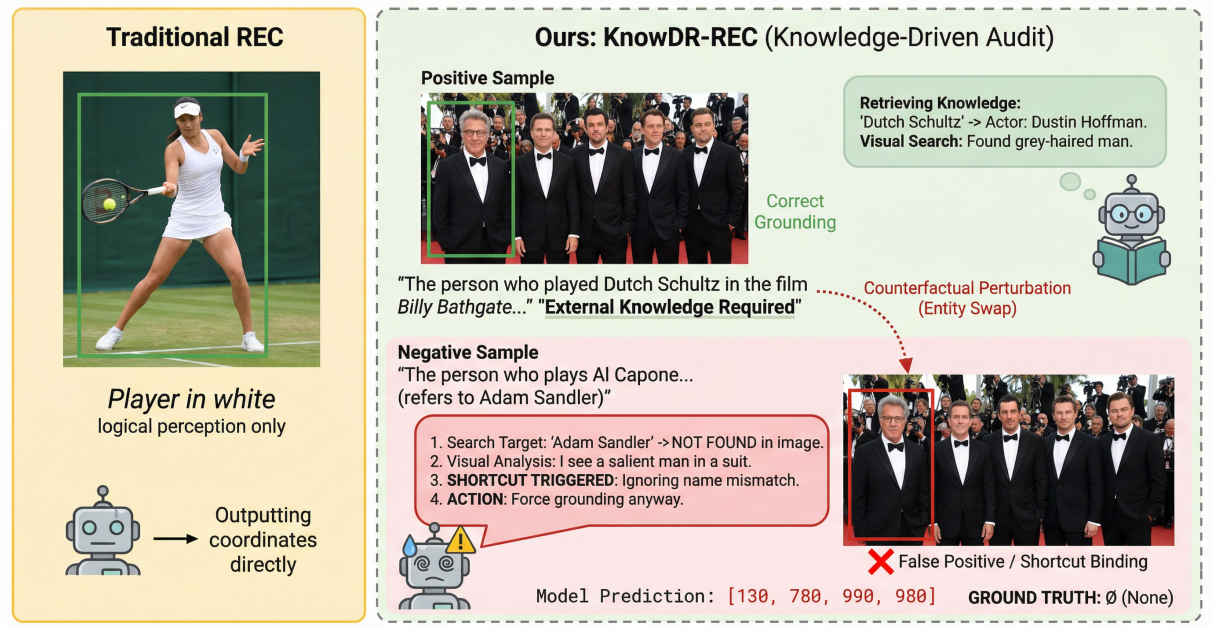

Figure 1: KnowDR-REC is a novel benchmark that aims to integrate external world knowledge into the textual expressions used in referring expression comprehension (REC).

troduce KnowDR-REC(Figure 1). KnowDR-REC is not merely a harder dataset, but a diagnostic audit framework explicitly designed to disentangle knowledge reasoning from visual grounding. Unlike benchmarks that increase difficulty by lengthening descriptions or amplifying perceptual complexity (Liu et al., 2024c; Xu et al., 2024), KnowDR-REC is constructed around structured factual knowledge derived from multi-hop question answering sources. Each referring expression is grounded in an explicit semantic representation, enabling controlled analysis of how models respond to changes in underlying factual logic.

Central to KnowDR-REC is a counterfactual audit protocol. For a given image and a factually correct referring expression, we extract its underlying semantic tuple (s, r, o, t), corresponding to subject, relation, object, and time. We then generate a paired counterfactual expression by perturbing a single factor—such as altering the year or substituting the award recipient—while preserving grammatical fluency and surface plausibility. This paired design allows us to directly test a key hypothesis: when the semantic logic of the input is reversed, does the model's localization behavior change accordingly, or does it remain invariant?

We evaluate 18 representative MLLMs (Bai et al., 2023; Comanici et al., 2025) under this protocol and observe substantial robustness deficiencies on knowledge-intensive REC tasks. Under an instruction-guided binary decision setting, most models exhibit pronounced difficulty in abstaining from factually invalid queries, resulting in a marked increase in error rates. More critically, even when the referring expression is deliberately modified into a counterfactual form that contradicts the original factual semantics, many models continue to predict the same bounding box as for the corresponding positive sample, often with high confidence. This semantic–localization mismatch—where textual meaning changes but spatial predictions remain invariant—suggests that models may bypass textual reasoning altogether and instead rely on spurious shortcut associations between visual features and entity types. Such shortcut binding behavior reveals a fundamental decoupling between knowledge reasoning and visual grounding in current MLLMs, rather than a true integration of the two modalities.

In summary, this work makes the following contributions:

(1) We introduce KnowDR-REC, a diagnostic benchmark designed to rigorously evaluate knowledge-conditioned visual grounding capabilities.

(2) We propose a counterfactual negative sample generation method for REC based on semantic tuple perturbation.

(3) Our evaluation of 18 MLLMs reveals a critical

reasoning–grounding decoupling. We identify "shortcut binding" as a dominant failure mode, where models rely on visual saliency while ignoring contradictions in textual semantics.

2 Related Work

2.1 Referring Expression Comprehension

Referring Expression Comprehension (REC) is a core task in multimodal understanding that detects the visual target based on natural language. Classic benchmarks such as the RefCOCO series rely on short spatial expressions and have been nearly saturated—models like CogVLM (Wang et al., 2024a) achieve strong results even in zero-shot settings, with 92.44% on RefCOCO, 88.55% on RefCOCO+, and 90.67% on RefCOCOg (Acc@0.5), revealing a performance ceiling and limited room for further progress.

To advance the task, recent datasets increase complexity: HC-RefLoCo (Wei et al., 2024) uses long human-centric expressions, FineCops-Ref (Liu et al., 2024c) focuses on compositional reasoning, ReSeDis (Huang et al., 2025) introduces cross-dataset and open-world settings, and MC-Bench (Xu et al., 2024) extends REC to the multi-image domain.

Despite progress, existing benchmarks lack hard negatives, temporal cues, and knowledge-based references—e.g., referring to people via historical or biographical facts—limiting real-world applicability. This calls for more realistic, knowledge-intensive REC datasets and evaluation protocols.

2.2 Knowledge-intensive Understanding

Integrating external knowledge is essential for reasoning in multimodal tasks, as many real-world questions require factual, temporal, or relational information beyond what is visible in an image. In Natural Language Processing (NLP), benchmarks like HotpotQA (Yang et al., 2018) and MetaQA (Puerto et al., 2021) test multi-hop reasoning across documents or knowledge graphs. In vision, datasets like OVEN (Hu et al., 2023), WIKIPerson(Sun et al., 2022), and KVQA (Shah et al., 2019) require linking visual entities to external knowledge bases, enabling more accurate and context-aware understanding.

In contrast, existing Referring Expression Comprehension (REC) datasets rely mostly on commonsense or intra-image cues. KB-Ref (Wang et al., 2020) incorporates basic commonsense knowledge, but lacks integration of rich factual or temporal knowledge. This exposes a key gap in REC evaluation—namely, the inability to test knowledge-intensive reasoning in multimodal grounding tasks.

2.3 Multimodal Model Evaluation and Hallucination

Hallucination has become central to evaluating robustness in MLLMs, particularly vision-language models (VLMs) (Liu et al., 2024b; Bai et al., 2024). A special case is "unsolvable problem detection (UPD)," where models should ideally abstain due to the absence or mismatch of visual content. Prior studies systematically explored single-object hallucination (Li et al., 2023), multi-object hallucination (Chen et al., 2024), abstention in multimodal QA scenarios (Miyai et al., 2024) . However, current REC benchmarks, despite constructing negative samples (Liu et al., 2024c), fail to provide an explanation for the causes of hallucination. To bridge this gap, our benchmark introduces controlled negative samples through temporal knowledge graph manipulations and employs two evaluation settings, enabling a systematic analysis of hallucination triggers within REC tasks.

3 Benchmark

To overcome the reliance on intra-image cues in standard benchmarks, KnowDR-REC leverages real-world multi-person images where resolution requires external factual, relational, and temporal knowledge. We focus on persons as the primary carriers of rich relational and temporal facts, and construct paired counterfactual negatives. Accordingly, our diagnostic protocol is designed to rigorously audit the alignment between textual reasoning and visual grounding.

3.1 Task Formulation

KnowDR-REC formulates knowledge-conditioned referring expression comprehension as a text-driven instance reasoning and grounding task under strong intra-class ambiguity. Given an image I containing multiple instances of the same object category $\mathcal{P} = \{p_1, \ldots, p_n\}$ and a natural-language expression x whose resolution cannot be determined from intra-image evidence or dataset-level statistical biases alone, the model is required to perform explicit reasoning beyond visual cues, identify the unique instance in $\mathcal{P}$ that satisfies the external knowledge constraints implied by x, and localize it in the image. If no instance in I is consistent with

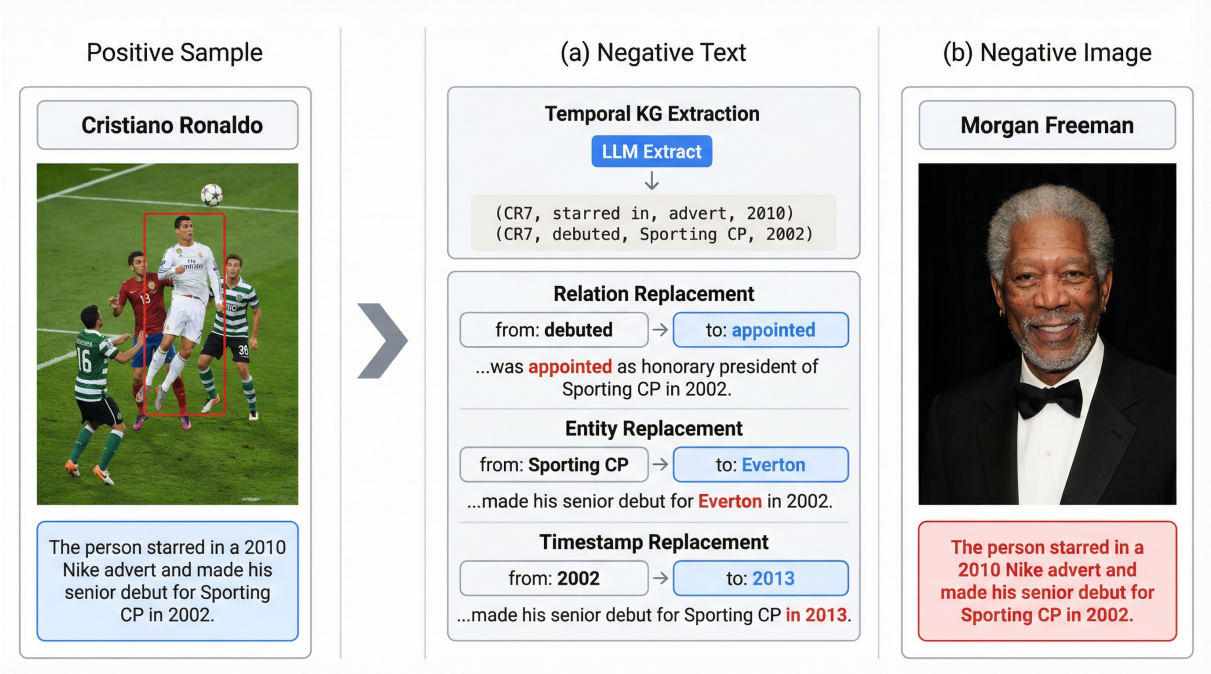

Figure 2: Negative sample construction in KnowDR-REC. Starting from an image-text pair, we extract a temporal knowledge graph from the textual expression. Fine-grained negative samples are introduced by perturbing factual triples within the graph. Additionally, coarse-grained negatives are obtained by pairing the original text with semantically unrelated images using image-text re-ranking.

the reasoning outcome, the model should abstain from grounding.

Formally, the task is defined as

$$f(I, x) \rightarrow \begin{cases} b \in \mathbb{R}^4, & \exists p_i \in \mathcal{P} \text{ s.t. } p_i \sim x, \\ \varnothing, & \text{otherwise.} \end{cases}$$

where b denotes the bounding box of the selected instance p_i, and $p_i \sim x$ indicates that instance p_i is consistent with the knowledge-dependent constraints expressed in x.

3.2 Dataset Construction

KnowDR-REC consists of a positive set $\mathcal{D}^+ = \{(I, x, b^*)\}$, where I is the image, x is the referring expression, and b^* is the ground-truth bounding box; and a negative set $\mathcal{D}^- = \{(I, x)\}$ comprising (i) minimally edited counterfactual negative expressions and (ii) paired contrast negatives formed by mismatching images and expressions. Statistics of our dataset are presented in Appendix 6.2.

Expression sources and rewriting. To obtain natural, diverse, and knowledge-intensive referring expressions, we curate person-centric questions from four widely used multi-hop QA datasets—ComplexWebQuestions (Talmor and Berant, 2018), HotpotQA (Yang et al., 2018), KQA Pro (Cao et al., 2020), and MetaQA (Puerto et al., 2021). We manually filter candidate questions to ensure each target person is unambiguous and uniquely determined by the textual constraints. We then use GPT-4o (Hurst et al., 2024) to rewrite questions into declarative REC-style expressions, improving fluency and linguistic variety while preserving the original factual constraints. This pipeline yields 2,537 referring expressions forming the expression corpus of KnowDR-REC.

Image retrieval and filtering. For each target person, we retrieve candidate images from Wikipedia (Bridge, 2001). To avoid trivial single-portrait grounding and to introduce realistic distractors, we retain only visually rich images that contain multiple persons, span diverse categories, and meet a minimum resolution requirement (e.g., $> 300 \times 300$). We use YOLOv8 (Sohan et al., 2024) as a pre-filter to verify multi-person criteria and image quality, rather than as an automatic annotator.

Positive triplets. We annotate instance-level bounding boxes for a subset of image–expression pairs where the referred person is present and uniquely identifiable. To improve reliability, we adopt two independent annotation routes: one team localizes the target using publicly available appear-

ance descriptions, while another team uses GPT-4o to generate descriptive cues and then annotates accordingly. We keep only samples where both routes consistently identify the same instance. In addition, we manually filter triplets whose referent can be resolved from intra-image cues (e.g., color, clothing, position) or generic commonsense priors (e.g., height, gender), ensuring that correct grounding requires external knowledge. This results in 1,042 verified triplets.

Counterfactual negative expressions. To construct fine-grained negatives, we parse textual expressions into temporal knowledge tuples (s, r, o, t), where $s, o \in \mathcal{E}$ represent the subject and object entities, $r \in \mathcal{R}$ denotes the relation, and $t \in \mathcal{T}$ indicates the timestamp. We employ GPT-4o to perturb a single factor within the tuple and rewrite the expression, thereby generating counterfactuals that preserve the original surface form (e.g., length and syntax) to avoid trivial biases (see Figure 2). All samples are manually verified to ensure they introduce a definitive semantic contradiction with the image while retaining linguistic fluency.

Negative image–text pairs. We additionally construct coarse-grained mismatched negatives to serve as a clean baseline. Detailed comparative analysis between these coarse-grained negatives and the fine-grained counterfactuals is provided in Appendix 6.3 .

3.3 Evaluation Protocol

We evaluate models with a unified stage-wise audit protocol that measures textual reasoning, visual grounding, and refusal on negative samples, and further quantifies the coupling between textual reasoning and visual localization.

Textual reasoning. For each positive sample, we extract the predicted entity from the model's chain-of-thought response. We report textual reasoning accuracy (Acc_{Text}) using alias-based matching, where a prediction is deemed correct if it matches any entry in a predefined alias table (covering alternative spellings, transliterations, and abbreviations). This metric isolates the model's ability to resolve entities based on textual expression rather than intra-image visual cues.

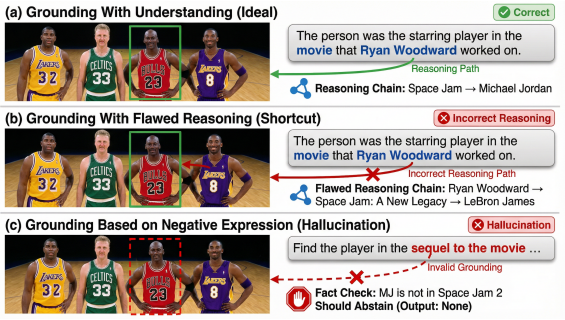

Figure 3: Visual grounding behaviors observed in REC tasks: (a) correct grounding based on correct understanding, (b) correct grounding based on incorrect understanding, and (c) "correct" grounding under negative descriptions.

Visual grounding. We measure grounding accuracy at multiple IoU thresholds $\tau \in \{50, 75, 90\}$:

$$Acc_\tau = \frac{|\{(I, x, b^*) \in \mathcal{D}^+ : \text{IoU}(b, b^*) \geq \tau\%\}|}{|\mathcal{D}^+|},$$

where $\text{IoU}(b, b) = \frac{\text{area}(b \cap b)}{\text{area}(b \cup b^*)}$ measures the spatial overlap between the predicted and ground-truth bounding boxes.

Refusal on negatives. We use two complementary negative protocols. In the *Basic Setting*, the model is evaluated using the same prompt as for positive samples, directly outputting a bounding box without explicit refusal instructions; we report the Error rate:

$$\text{Error rate} = \frac{|\{(I, x) \in \mathcal{D}^- : f(I, x) \neq \varnothing\}|}{|\mathcal{D}^-|}.$$

Regarding the *Instruction-guided Setting*, prior research highlights that the reward mechanism in RLVR can inadvertently induce sycophancy, where models prioritize maximizing rewards by forcing responses rather than admitting ignorance when facing uncertainty (Kalai et al., 2025). To mitigate this tendency, the prompt in this setting explicitly instructs the model to first make a binary existence decision (yes/no) prior to grounding. We report the balanced accuracy over an equal number of positive and negative samples:

$$\text{BalAcc} = \frac{1}{2}\left(\frac{\text{TP}}{|\mathcal{D}^+|} + \frac{\text{TN}}{|\mathcal{D}^-|}\right),$$

where TP denotes true positives (correctly identified as present) and TN denotes true negatives (correctly identified as absent).

4 Experiments and Results

In this section, we conduct a comprehensive evaluation of 18 state-of-the-art MLLMs on KnowDR-REC. Beyond standard performance metrics, we aim to disentangle the model's capability for *visual grounding* from its ability to perform *reasoning based on textual input*. Furthermore, we assess the models' rejection capabilities on negative samples under two distinct evaluation settings.

Our analysis unveils a critical paradox: models often achieve high grounding precision yet frequently fail to resolve underlying entities or identify factual contradictions (Figure 3). This phenomenon suggests that on our proposed benchmark, models predominantly rely on spurious visual shortcuts rather than genuine multimodal reasoning. This raises serious concerns regarding their reliability in reference tasks that require complex textual understanding and knowledge beyond the visible image.

4.1 Experimental Setup

We evaluate all models in a zero-shot setting. Our selection encompasses three categories of representative models: closed-source general MLLMs, including Gemini 3 Pro/Flash and Grok-2 Vision; open-source general MLLMs, such as the Qwen-VL series and SPHINX; and open-source specialized MLLMs designed for visual grounding, such as VLM-R1 and Ferret.

To systematically audit the three stages of knowledge-driven Referring Expression Comprehension (REC), we employ a multi-dimensional metric system. Specifically, for visual grounding, we report standard accuracy at different IoU thresholds (0.5, 0.75, and 0.9) to measure localization precision. For textual reasoning, we utilize Chain-of-Thought (CoT) prompting and calculate reasoning accuracy (Acc_{Text}) through specific character dictionaries and alias matching. To assess rejection capability on negative samples, we measure the Error Rate in the *basic setting* and Balanced Accuracy (BalAcc) in the *instruction-guided setting*. All experiments are conducted on NVIDIA RTX 4090 GPUs.

4.2 Main Results: The Paradox of High Grounding and Low Reasoning

As shown in Table 1, strong baselines such as Qwen-VL-Max and Qwen2.5-VL-72B demonstrate impressive localization capabilities, achieving $Acc_{0.5}$ scores of 72.2% and 71.0%, respectively. Even smaller models like Qwen2.5-VL-3B exhibit competitive performance (60.6%). We additionally report the mean localization accuracy mAcc averaged over IoU thresholds $\{0.5, 0.75, 0.9\}$ to better reflect precision degradation at stricter overlap criteria. The majority of evaluated models surpass 50% accuracy on $Acc_{0.5}$ in the zero-shot setting, but their mAcc values are consistently lower, indicating that coarse localization success does not necessarily translate into high-precision grounding. Superficially, these results seem to imply that current MLLMs have mastered fine-grained visual grounding.

However, in-depth analysis reveals a severe disconnect in model capabilities. First, textual reasoning significantly lags behind grounding performance. For instance, while Qwen-VL-Plus achieves a grounding accuracy of 64.6%, its reasoning accuracy is only 39.8%, implying it often localizes the "correct" object without knowing "who" the object is. Second, performance in the "existence verification" stage is even more concerning. In the evaluation of negative sample error rates, most models suffer catastrophic failures. In the *basic setting* for negative samples, open-source models like Qwen2.5-VL and VLM-R1 exhibit error rates approaching 100%, blindly predicting bounding boxes for factually invalid descriptions. In the *instruction-guided setting*, model accuracy approximates random guessing, failing to distinguish whether the referred entity actually exists in the image. This paradox of "High Grounding, Low Reasoning, and Low Rejection" indicates that relying solely on standard metrics ($Acc_{0.5}$) to judge visual grounding may mask fundamental flaws in the models' logic. Additional analyses on one-shot prompting and Gemini box-formatting behavior are provided in Appendix 6.8.4 and Appendix 6.8.6, respectively.

4.3 Diagnostic Analysis: Unveiling Shortcut Binding

To explain the paradox observed above—where models correctly "ground" entities they fail to identify—we employ a counterfactual audit mechanism. Our analysis demonstrates that this perceived grounding success is primarily driven by *shortcut binding* rather than deep semantic understanding.

Methods	Size	Positive Sample							Negative Sample	
		$\text{Acc}_{0.5}$	$\text{Acc}_{0.75}$	$\text{Acc}_{0.9}$	mAcc	Acc_{Text}	A_{50}^{match}	A_{50}^{mismatch}	Error Rate	BalAcc
Closed-Source Generalist MLLMs										
Gemini 2.5 Flash(Comanici et al., 2025)	–	44.1	16.3	2.7	21.0	84.7	46.2	32.2	79.3	82.0
Gemini 2.5 Pro(Comanici et al., 2025)	–	21.9	4.6	0.8	9.1	89.1	23.5	9.4	97.1	79.4
Gemini 3 Flash(DeepMind, 2025a)	–	53.5	32.4	11.2	32.4	98.2	53.9	41.5	71.5	70.8
Gemini 3 Pro(DeepMind, 2025b)	–	43.8	27.5	11.8	27.7	91.5	40.2	56.5	74.2	76.4
Grok-2 Vision(Xie et al., 2025)	–	56.7	16.6	8.4	27.2	71.0	61.4	45.0	83.5	78.5
Qwen-VL-Plus(Bai et al., 2023)	–	64.6	54.9	24.6	48.0	39.8	83.7	51.9	72.5	51.7
Qwen-VL-Max(Bai et al., 2023)	–	72.2	65.6	47.2	61.7	61.4	80.4	59.0	70.7	57.4
Open-Source Generalist MLLMs										
Qwen2.5-VL(Bai et al., 2025)	3B	60.6	54.5	38.8	51.3	34.6	86.4	46.9	100.0	52.3
Qwen2.5-VL(Bai et al., 2025)	7B	59.1	48.4	20.4	42.6	40.8	82.0	43.3	98.7	51.1
Qwen2.5-VL(Bai et al., 2025)	32B	57.6	45.9	28.3	43.9	43.2	71.6	43.8	81.3	55.5
Qwen2.5-VL(Bai et al., 2025)	72B	71.0	63.8	45.9	60.2	61.6	68.0	56.6	81.2	54.8
SPHINX-Tiny(Lin et al., 2023)	1.1B	8.3	2.1	1.6	4.0	10.4	21.7	6.7	100.0	55.1
SPHINX-v2-1k(Liu et al., 2024a)	13B	38.1	14.3	3.6	18.7	32.5	42.8	35.7	93.1	61.0
Open-Source Specialist MLLMs										
VLM-R1(Shen et al., 2025)	3B	63.7	61.5	40.5	55.2	36.8	78.1	55.9	100.0	56.0
GroundingGPT(Li et al., 2024)	7B	49.6	41.4	16.4	35.8	16.1	37.3	52.3	100.0	47.0
Ferret(You et al., 2023)	7B	35.7	23.6	12.7	24.0	23.7	50.8	31.4	88.1	51.5
Ferret(You et al., 2023)	13B	44.5	31.4	16.6	30.8	24.9	57.3	39.8	94.3	52.4
REESEEK(Jiang et al., 2025)	7B	46.2	40.7	35.1	40.7	–	–	–	78.5	–

Table 1: Comparison of baselines on KnowDR-Bench. In the positive sample, we summarize the model's performance in textual reasoning and visual grounding, and analyzed the impact of the success of textual reasoning on visual grounding. In the negative sample, we evaluated the performance under two settings.

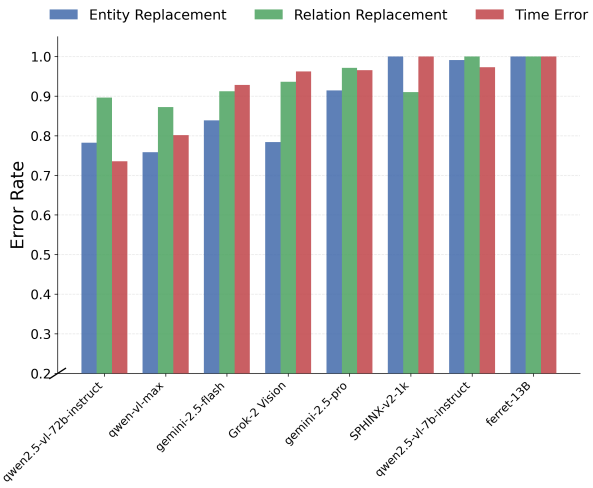

Figure 4: A comparison of error rates across multiple models under different types of negative samples.

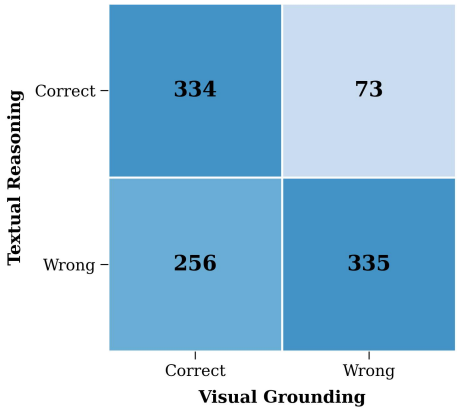

Figure 5: Confusion matrix analysis for Qwen2.5-VL-7B.

4.3.1 Counterfactual Insensitivity and Shortcut Binding

To rigorously investigate model sensitivity to fine-grained semantic changes, we audited localization behavior by directly comparing positive sample triplets with their paired counterfactual negative samples. The core premise is straightforward: if a model truly comprehends the referring expression, a semantic inversion of the prompt (e.g., changing the timeline or relation) should necessitate a shift in visual attention or a refusal to predict. However, the quantitative evidence in Table 2 reveals a pervasive failure mode we term "Shortcut Binding."

We observe an alarmingly high degree of prediction invariance across leading models. For instance, Qwen-VL-Plus exhibits a consistency score (Neg-Pos) of 63.8%, and Qwen2.5-VL-72B reaches 67.2%. This implies that for nearly two-thirds of the test cases, even when the input text is modified to factually contradict the image content, the models stubbornly predict the exact same bounding box as they did for the correct query. This behavior is

Methods	Size	Acc@0.5 (%)		
		Pos-GT	Neg-GT	Neg-Pos
Closed-Source Generalist MLLMs				
Gemini 2.5 Flash	–	44.1	1.0	1.8
Gemini 2.5 Pro	–	21.9	9.1	11.5
Gemini 3 Flash	–	53.5	5.2	7.8
Gemini 3 Pro	–	43.8	24.5	28.9
Grok-2 Vision	–	56.7	4.3	6.5
Qwen-VL-Plus	–	64.6	59.6	63.8
Qwen-VL-Max	–	72.2	56.9	61.5
Open-Source Generalist MLLMs				
Qwen2.5-VL	3B	60.6	52.1	56.8
Qwen2.5-VL	7B	59.1	25.3	29.5
Qwen2.5-VL	32B	57.6	46.5	50.4
Qwen2.5-VL	72B	71.0	61.9	67.2
SPHINX-Tiny	1.1B	8.3	6.2	7.0
SPHINX-v2-1k	13B	38.1	56.2	48.5
Open-Source Specialist MLLMs				
VLM-R1	3B	63.7	49.4	54.2
GroundingGPT	7B	49.6	49.6	53.0
Ferret	7B	35.7	32.3	34.8
Ferret	13B	44.5	46.6	47.5
REESEEK	7B	46.2	27.1	31.2

Table 2: Analysis of reasoning-grounding consistency. Pos-GT: Standard grounding accuracy on positive samples. Neg-GT: Accuracy calculated by comparing predictions under negative expressions against the original ground truth. Neg-Pos: Consistency accuracy calculated by measuring the IoU overlap between the predicted boxes of positive and negative expressions.

not merely an error in judgment but a fundamental bypass of the textual reasoning stage. The models appear to treat the textual query as a generic trigger for object detection rather than a set of semantic constraints. Consequently, the high scores on standard benchmarks ($Acc_{0.5}$) are largely illusory: models are not grounding the specific entity described by the complex text, but are mechanically binding the input to the most visually salient or prototypical object in the scene, rendering them insensitive to the actual logical content of the expression. To rule out the concern that this effect is driven by positives that were already failed in the original task, we further performed a conditioned counterfactual audit restricted to samples that were correctly localized under the original positive expression. The same qualitative pattern persists: strong models still tend to preserve highly overlapping boxes after the semantic constraints are flipped, confirming that the observed invariance is not merely noise from already-incorrect positives. Additional examples and implementation details are summarized in Appendix 6.8.3.

4.3.2 Fine-grained Sensitivity Analysis across Perturbation Types

We further decompose the negative samples into three perturbation types—entity, relation, and time—to pinpoint the triggers of hallucination. As shown in Figure 4, models generally lack the granularity to verify specific factual constraints. While generalist models like Qwen-VL-Max show marginal robustness, most architectures, especially specialist grounding models like Ferret, exhibit error rates nearing 100% on relation and time perturbations. Notably, abstract semantic constraints (time/relation) consistently induce higher error rates than concrete entity attributes. This suggests that while models may detect visual mismatches (e.g., gender), they are largely "blind" to non-visual contradictions, defaulting to grounding based solely on object presence. We additionally stratified performance by target saliency, using target size and spatial centrality as coarse proxies. The resulting trend is consistent with the shortcut-binding hypothesis: models are substantially more likely to output the correct box when the target is large or near the image center, even when textual reasoning is incorrect. This explains why shortcut binding can sometimes appear successful on positive samples: salient targets make visually driven guessing more likely to coincide with the annotated ground truth. A concise summary is provided in Appendix 6.8.5.

4.3.3 Decoupling Reasoning from Grounding

To statistically verify this reliance, as shown in Table 1, we report A_{50}^{match} and A_{50}^{mismatch}, which represent the grounding accuracy conditioned on whether the textual reasoning was correct or incorrect, respectively. Ideally, if a model fails to identify the correct target entity (reasoning error), its subsequent localization should be akin to a random guess, resulting in low grounding accuracy. However, the metric A_{50}^{mismatch} in Table 1 reveals a severe decoupling.

Take Qwen-VL-Max as a representative example: it achieves an A_{50}^{mismatch} of 59.0%. This statistic indicates that in cases where the model explicitly retrieves the wrong entity name or fails to reason about the identity, there is still nearly a 60% probability that it will hit the correct bounding box. This phenomenon is counter-intuitive and points to a "Lucky Guessing" mechanism driven by dataset bias or visual saliency rather than logical derivation.

To dissect this further, we visualize the confusion matrix for Qwen2.5-VL-7B (Figure 5). A significant density of samples clusters in the "Wrong Reasoning but Correct Grounding" quadrant. These are instances where the model's Chain-of-Thought reveals a complete hallucination or misidentification of the target person, yet the final bounding box output is technically correct. This provides strong visual and statistical evidence that the reasoning and grounding modules are functionally decoupled. The model effectively "guesses right for the wrong reasons," confirming that current state-of-the-art performance is inflated by spurious correlations between visual features and entity types, independent of the complex knowledge reasoning required by the prompt.

5 Conclusion

In this paper, we introduce KnowDR-REC, a diagnostic benchmark designed to scrutinize the knowledge-conditioned reasoning capabilities of MLLMs within the context of Referring Expression Comprehension. Unlike traditional benchmarks that focus primarily on visual localization driven by intra-image cues, KnowDR-REC requires models to resolve the target identity based on external knowledge , localize its spatial position, and make a rejection decision regarding invalid queries .

Our comprehensive evaluation of 18 state-of-the-art MLLMs reveals a critical paradox: while models achieve high performance on standard grounding metrics, they exhibit severe deficiencies in textual reasoning and hallucination rejection. Through counterfactual auditing, we diagnose this phenomenon as "shortcut binding," where models bypass semantic reasoning and instead rely solely on visual features for reasoning. These findings suggest that the high grounding accuracy in current benchmarks often masks a fundamental decoupling of perception and reasoning. KnowDR-REC encourages a shift from surface-level pattern matching toward knowledge-driven multimodal understanding.

Limitations

Despite the diagnostic value of KnowDR-REC, several limitations characterize our current work. First, our benchmark focuses exclusively on person-centric entities. While humans are the most knowledge-dense subjects for examining external references, this design choice limits our ability to evaluate knowledge-conditioned grounding for inanimate objects, landmarks, or biological species. Second, the dataset is currently English-only, and the lack of multilingual queries restricts the evaluation of cross-cultural knowledge transfer capabilities. Finally, our assessment of textual reasoning relies on explicit Chain-of-Thought generation. While we employ robust alias-based matching, this proxy may not perfectly capture a model's internal belief state (Turpin et al., 2023), and it is possible that some models possess implicit knowledge that is not fully verbalized in the generated rationale, potentially affecting the precision of our reasoning-grounding decoupling analysis.

Acknowledgments

This work was supported in part by the Shenzhen Fundamental Research Program (No. JCYJ20250604124702003), the Guangdong Basic and Applied Basic Research Foundation (Nos. 2025A1515011826 and 2026A1515010768), the National Natural Science Foundation of China (Nos. 62401342 and 62271291), and the Natural Science Foundation of Shandong Province (Nos. ZR2024QF092, ZR2024LZH007, and ZR2025ZD24).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Astoria-Megler Bridge. 2001. Wikipedia, the free encyclopedia. *San Francisco (CA): Wikimedia Foundation*.

Borui Cai, Yong Xiang, Longxiang Gao, He Zhang, Yunfeng Li, and Jianxin Li. 2022. Temporal knowledge graph completion: A survey. *arXiv preprint arXiv:2201.08236*.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2020. Kqa pro: A dataset with explicit compositional programs for complex question answering over knowledge base. *arXiv preprint arXiv:2007.03875*.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Siyang Dai, Jun Liu, and Ngai-Man Cheung. 2024. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16985–16995.

Google DeepMind. 2025a. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Model card, published Dec 2025.

Google DeepMind. 2025b. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Model card, published Nov 2025.

Chenyang Gao, Biao Yang, Hao Wang, Mingkun Yang, Wenwen Yu, Yuliang Liu, and Xiang Bai. 2023. Textrec: A dataset for referring expression comprehension with reading comprehension. In *International Conference on Document Analysis and Recognition*, pages 402–420. Springer.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

Ziling Huang, Yidan Zhang, and Shin'ichi Satoh. 2025. Resedis: A dataset for referring-based object search across large-scale image collections. *arXiv preprint arXiv:2506.15180*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Qin Liu, and Lei Zhang. 2025. Referring to any person. *arXiv preprint arXiv:2503.08507*.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Shuhei Kurita, Naoki Katsura, and Eri Onami. 2023. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15214–15224.

Seongmin Lee, Jaewook Shin, Youngjin Ahn, Seokin Seo, Ohjoon Kwon, and Kee-Eung Kim. 2024. Zero-shot multi-hop question answering via monte-carlo tree search with large language models. *arXiv preprint arXiv:2409.19382*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yuandi Li, Hui Ji, Fei Yu, Lechao Cheng, and Nan Che. 2025. Temporal multi-modal knowledge graph generation for link prediction. *Neural Networks*, 185:107108.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, and 1 others. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, and 1 others. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, and 1 others. 2024a. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. 2024c. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. 2024. Unsolvable problem detection: Robust understanding evaluation for large multimodal models. *arXiv preprint arXiv:2403.20331*.

U Penchalaiah and Siva Kumar Vg. 2018. Design of high-speed and energy-efficient parallel prefix kogge stone adder. In *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*, pages 1–7. IEEE.

Haritz Puerto, Gözde Gül Şahin, and Iryna Gurevych. 2021. Metaqa: Combining expert agents for multi-skill question answering. *arXiv preprint arXiv:2112.01922*.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *Preprint*, arXiv:2504.07615.

Mupparaju Sohan, Thotakura Sai Ram, and Ch Venkata Rami Reddy. 2024. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 529–545. Springer.

Wenxiang Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2022. Visual named entity linking: A new dataset and a baseline. *arXiv preprint arXiv:2211.04872*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Peng Wang, Dongyang Liu, Hui Li, and Qi Wu. 2020. Give me something to eat: Referring expression comprehension with commonsense knowledge. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 28–36.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024a. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

Xiaochen Wang, Junqing He, Liang Chen, Reza Haf Zhe Yang, Yiru Wang, Xiangdi Meng, Kunhao Pan, and Zhifang Sui. 2024b. Sg-fsm: A self-guiding zero-shot prompting paradigm for multi-hop question answering based on finite state machine. *arXiv preprint arXiv:2410.17021*.

Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. 2024. A large-scale human-centric benchmark for referring expression comprehension in the lmm era. *Advances in Neural Information Processing Systems*, 37:69566–69587.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2024. Cofca: A step-wise counterfactual multi-hop qa benchmark. *arXiv preprint arXiv:2402.11924*.

Liuyue Xie, George Z Wei, Avik Kuthiala, Ce Zheng, Ananya Bal, Mosam Dabhi, Liting Wen, Taru Rustagi, Ethan Lai, Sushil Khyalia, and 1 others. 2025. Maverix: Multimodal audio-visual evaluation reasoning index. *arXiv preprint arXiv:2503.21699*.

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2024. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. *arXiv preprint arXiv:2410.12332*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Keunwoo Peter Yu. 2023. Constructing temporal dynamic knowledge graphs from interactive text-based games. *arXiv preprint arXiv:2311.01928*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer.

Zhihan Yu and Ruifan Li. 2024. Revisiting counterfactual problems in referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13438–13448.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2025. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2):825–843.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *arXiv preprint arXiv:2402.14116*.

6 Appendix

6.1 Code Availability

The code for KnowDR-REC is available at https://github.com/LetItBe12345/KnowDR-REC.

6.2 Data Statistic

The statistical overview of our dataset is shown in Figure 7.

6.3 Comparative Analysis of Negative Sample Categories

To provide a fine-grained understanding of the grounding capabilities of MLLMs, we categorize the negative samples into two distinct types: *Counterfactual Negative Samples* and *Negative Image-Text Pairs*. The former aggregates samples containing subtle textual hallucinations (i.e., entity replacement, relation replacement, and time errors), while the latter represents coarse-grained image-text mismatches.

As illustrated in Figure 6, we observe a significant performance disparity between these two categories. Across the 12 evaluated models, the average error rate on Counterfactual Negative Samples reaches 86.9%, which is substantially higher than the 61.2% error rate observed on Negative Image-Text Pairs.

As illustrated in Figure 6, we observe a significant performance disparity between these two categories. Across the 12 evaluated models, the average error rate on Counterfactual Negative Samples reaches 86.9%, which is substantially higher than the 61.2% error rate observed on Negative Image-Text Pairs.

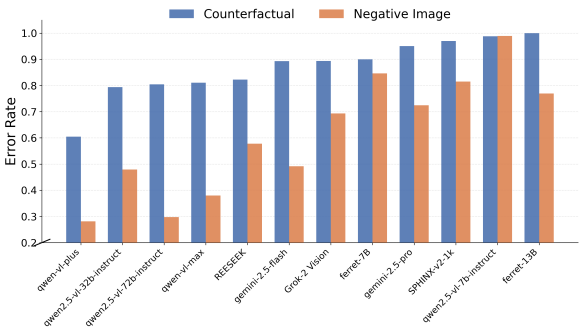

Figure 6: Performance comparison of 12 MLLMs on two categories of negative samples. The blue bars represent the average error rate on Counterfactual Negative Samples (fine-grained), while the orange bars represent the error rate on Negative Image-Text Pairs (coarse-grained).

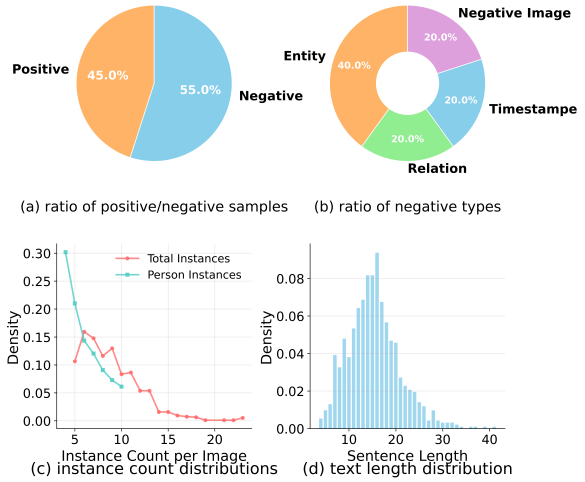

Figure 7: Data Statistics.

This discrepancy suggests that while current MLLMs possess reasonable capabilities in distinguishing completely irrelevant images (coarse-grained alignment), they struggle severely with fine-grained visual-semantic verification. When the text is partially aligned with the image but contains subtle factual manipulations—such as changed entities or relations—most models fail to reject the false positive. Notably, while models like `qwen-vl-plus` demonstrate relatively better robustness (0.604 error rate on counterfactuals), several models (e.g., `ferret-13B`, `SPHINX-v2-1k`) exhibit near-saturation error rates (≈ 1.0) on counterfactual samples, highlighting a critical bottleneck in the precise grounding capabilities of state-of-the-art MLLMs.

6.4 Future Work

Although this study proposes a novel paradigm for evaluating knowledge-driven referring expression understanding, there remain some non-negligible limitations. Our currently constructed benchmark dataset primarily focuses on "person" entities. This design choice is justifiable—persons often serve as key carriers and conveyors of complex world knowledge and frequently appear in real-world applications. However, this focus inevitably limits the evaluation coverage, thereby constraining the generalizability and applicability of the proposed paradigm to some extent.

Specifically, while person entities are indeed representative in terms of knowledge expression and understanding, they do not encompass all entity types that possess knowledge value and pose reasoning challenges in the real world. For instance, landmark buildings with historical or cultural significance, artworks, natural landscapes, scientific inventions, and other such entities are also important targets for knowledge-driven reference. These types of entities often exhibit characteristics and challenges in knowledge representation, contextual understanding, and reasoning that differ significantly from those of person entities. Moreover, an overreliance on "person" as the primary research object may lead to performance degradation or capability bottlenecks when the system is applied to other entity types, thereby affecting its generalization ability and practical value in diverse and complex scenarios.

Therefore, future research could proceed along the following directions: (1) When constructing benchmark datasets, further diversify the types of entities included, applying the methodology to more fine-grained and richly categorized entities, such as historical events, cultural heritage sites, famous artworks, scientific achievements, or natural geographic units. (2) Design evaluation tasks and annotation schemes tailored to the differing demands of knowledge reasoning, contextual ambiguity, and expression modalities brought by various entity types, thereby enhancing the benchmark's comprehensiveness and level of challenge. (3) In the process of data expansion and methodological refinement, emphasize the coordinated use of multimodal information (e.g., images, text, structured knowledge) to further improve the paradigm's adaptability to complex real-world applications.

6.5 Foundation Dataset

6.5.1 ComplexWebQuestions

ComplexWebQuestions (Talmor and Berant, 2018) is a benchmark dataset for evaluating the capabilities of question answering systems under complex natural language queries. This dataset is built upon simple questions from WebQuestionsSP, which are structurally expanded into more complex SPARQL queries and then paraphrased into natural language questions via crowdsourcing. The final dataset combines these questions with Freebase entity answers to create high-quality multi-hop QA samples. A central feature of CWQ is that it requires models to perform compositional reasoning, such as integrating information under entity constraints, comparisons, and conjunctive conditions.

Although the original goal of ComplexWebQuestions was to train and evaluate multi-hop QA systems that integrate information retrieval and reading comprehension, assuming access to external snippets or knowledge graphs, the high-quality design of its questions has led to its widespread use in recent years for evaluating the multi-hop reasoning capabilities of large language models under zero-shot settings.

6.5.2 HotpotQA

HotpotQA (Yang et al., 2018) is a benchmark for evaluating multi-hop reasoning over natural language passages. It contains crowdsourced questions that require integrating facts from multiple Wikipedia articles. Each sample provides not only the answer but also sentence-level supporting facts, enabling supervision for both reasoning and explanation.

The dataset includes two question types—bridge and comparison—and is split into distractor and fullwiki settings. Its design encourages models to perform document-level inference with interpretable outputs. Though limited to Wikipedia and prone to annotation noise, HotpotQA remains a central resource for supervised and zero-shot multi-hop QA research.

6.5.3 MetaQA

MetaQA (Puerto et al., 2021) is a synthetic QA dataset focused on multi-hop reasoning over a movie-domain knowledge graph. It features templated questions spanning 1 to 3 hops and comes in three formats: vanilla (text), NTM (paraphrased), and audio (TTS).

Its structured design allows precise evaluation of reasoning depth, but the reliance on templates and a narrow domain restricts generalization. Despite this, MetaQA is widely used in early-stage QA experiments and KG reasoning studies, especially for weak supervision and transfer learning setups.

6.5.4 KQA Pro

KQA Pro (Cao et al., 2020) is a compositional QA dataset over Wikidata that combines natural language questions with executable KoPL and SPARQL programs. It covers a wide range of reasoning skills, including logical, numerical, and multi-hop operations.

Each question aligns with symbolic reasoning steps, supporting explainable and program-guided QA. Though limited to a subset of Wikidata and requiring familiarity with program supervision, KQA Pro provides a strong benchmark for training interpretable and modular QA models over structured knowledge.

6.6 Preliminaries

6.6.1 Multi-hop Question Answering

Multi-hop Question Answering (Multi-hop QA) refers to tasks where answering a question requires traversing two or more pieces of information and reasoning through a series of intermediate facts (hops) to arrive at the correct answer. Unlike single-hop QA, which often allows direct inference from a single text passage or a single triple, multi-hop QA imposes higher demands on long-range retrieval, information integration, and logical reasoning, while also naturally providing opportunities for explainability evaluation (e.g., requiring supporting sentences or explicit reasoning paths). Multi-hop QA also serves as a crucial benchmark for evaluating the reasoning limits of large language models (LLMs) under zero-shot conditions, and becomes particularly challenging in the closed-book setting. Recently, a growing body of work has begun to systematically explore the capabilities of LLMs under this setting. For instance, the FanOutQA (Zhu et al., 2024) significantly increases the depth of reasoning chains (with an average of 7 hops); under the constraint of no external retrieval, even GPT-4 achieves only around 47% loose accuracy and less than 10% strict accuracy. Moreover, methods such as MZQA (Lee et al., 2024) and SG-FSM (Wang et al., 2024b) attempt to mitigate error propagation across reasoning steps through structured search mechanisms like Monte Carlo Tree Search and finite-state machine prompting paradigms, yet their overall performance remains far below human level. This task also exposes the issue of pretraining data leakage in closed-book evaluation, as demonstrated by the CofCA benchmark (Wu et al., 2024), which constructs counterfactual samples to more faithfully reveal the knowledge gaps of models. These studies collectively suggest that multi-hop QA is not only a litmus test for reasoning capabilities but also highlights the limitations of LLMs in memory-based knowledge reasoning, thereby sparking renewed interest in integrating structured external knowledge sources such as knowledge graphs.

6.6.2 Temporal Knowledge Graph

In recent years, knowledge graphs have demonstrated powerful capabilities in representation and reasoning across a wide range of natural language processing tasks. To better capture the dynamic evolution of real-world relations, temporal knowledge graphs (TKGs) (Cai et al., 2022) have been proposed. TKGs extend traditional fact triples by incorporating temporal information, constructing structured quadruples in the form of ⟨subject, relation, object, timestamp⟩. Compared with static knowledge graphs, TKGs are more suitable for representing and reasoning over time-sensitive event knowledge, and have been widely applied in tasks such as event forecasting, knowledge completion, and question answering. To address the growing need for temporal modeling in information extraction from text, recent research has started to explore directly constructing TKGs from natural language text. On one hand, methods such as Temporal Discrete Graph Updater (Yu, 2023) attempt to dynamically update temporal graph structures from text-based event logs. On the other hand, the Temporal KG Generation Dataset provides the first benchmark dataset tailored to document-level extraction, supporting the extraction of structured temporal quadruples from event-centric text. Moreover, multimodal modeling has also emerged as a promising direction, with approaches like TMMKGG (Li et al., 2025) aiming to jointly encode images and text into temporally-grounded multimodal knowledge graphs. In this work, we leverage temporal knowledge graphs to extract structured elements from textual data, including entities, relations, and timestamps, and further perturb these extracted elements to construct fine-grained textual negative samples.

6.7 Comparison to Related Datasets

As summarized in Table 3, some concurrent works have proposed multi-image MLLM benchmarks for more general purposes, covering multiple fields and disciplines, including RefCOCO (Yu et al., 2016), RefCOCOg/+ (Mao et al., 2016), HC-RefLoCo (Wei et al., 2024), FineCops-Ref (Liu et al., 2024c), KB-Ref (Wang et al., 2020), ReSeDis (Huang et al., 2025), MC-Bench (Xu et al., 2024), C-RefCOCO/g/+ (Yu and Li, 2024), SK-VG (Penchalaiah and Vg, 2018), REC-8K (Dai et al., 2024), TextREC (Gao et al., 2023), RefEgo (Kurita et al., 2023). In this section, we compare our benchmark with these concurrent works, as shown in Table 3.

6.8 Experimental Details

6.8.1 Computing Infrastructure

All experiments were conducted on a server running Ubuntu 22.04, equipped with two NVIDIA RTX 4090 GPUs. To ensure that dependencies do not interfere with each other, we created an isolated Conda environment for each model.

6.8.2 Inference Settings

All models were evaluated in a closed-book setting, meaning they did not access any external knowledge sources or the internet during inference. We explicitly disabled all network connections to guarantee that the entire evaluation process was conducted offline.

6.8.3 Conditioned Counterfactual Audit

Following the rebuttal discussion, we additionally re-evaluated counterfactual negatives only on instances whose corresponding positive samples were correctly localized in the original setting. This conditioned protocol removes the confound that a model may have already failed before the semantic perturbation is applied. The same failure pattern remains: strong models often preserve highly overlapping predictions even after the factual constraint is flipped, supporting our claim that shortcut binding reflects semantic insensitivity rather than mere carryover from already-failed positives.

6.8.4 In-Context Learning Analysis

We further tested positive-only, negative-only, and mixed one-shot demonstrations to examine whether in-context learning can repair reasoning-grounding decoupling. We find that few-shot prompting mainly changes refusal behavior and output format biases. For example, a preceding positive demonstration often encourages the model to always output a box, while a negative demonstration increases abstention. However, these prompt-level shifts do not consistently fix the core mismatch between semantic reasoning and localization, so the zero-shot audit remains informative about the underlying failure mode.

6.8.5 Saliency Stratification

To analyze why shortcut binding can occasionally yield an apparently correct prediction, we stratified samples by two simple target-saliency factors: bounding-box size and distance to the image center. We observe that localization accuracy rises noticeably for larger and more central targets, even in cases where textual reasoning is unreliable. This pattern supports the interpretation that many successful predictions are driven by a "most salient person" heuristic rather than by faithful grounding of the knowledge-conditioned expression.

6.8.6 Gemini Box-Formatting Diagnosis

To better understand the relatively low grounding precision of Gemini models despite their strong textual reasoning scores, we manually inspected representative outputs. A recurring issue is coordinate formatting rather than entity resolution: Gemini models frequently produce oversized boxes or out-of-bound coordinates, which sharply lowers IoU-based grounding scores. This suggests that part of the performance drop is attributable to output formatting and localization precision, not solely to deficiencies in knowledge-conditioned reasoning.

6.8.7 Privacy and Policy Confounds

Because KnowDR-REC is person-centric, a natural concern is that some failures may arise from privacy or safety policies rather than reasoning limitations. Our manual inspection indicates that the dominant error is not refusal, but overconfident box generation on negative samples. Moreover, we sourced images from Wikipedia/Wikimedia Commons with preserved attribution and per-image licenses, and excluded API models with strong default identity refusals from the reported benchmark in order to reduce policy-related confounds.

6.9 Licenses

The benchmark dataset introduced in this paper consists of textual and visual data derived from publicly available resources, as shown in Table

Table 3: Comparative overview of 12 referring-expression comprehension datasets.

Datasets	Neg.	Inst.	Know.	Task Types
RefCOCO/g/+	✗	✓	✗	Classic single-image REC
HC-RefLoCo	✗	✓	✗	Long-form person REC
FineCops-Ref	✓	✓	✗	Fine-grained composite
KB-Ref	✗	✓	✓	Knowledge-based object REC
ReSeDis	✓	✓	✗	Retrieval + Localization
MC-Bench	✗	✓	✗	Multi-view localization
C-RefCOCO/g/+	✓	✓	✗	Counterfactual REC
SK-VG	✗	✓	✓	Story-knowledge grounding
REC-8K	✗	✓	✗	Expression counting
TextREC	✗	✓	✓	Text-aware REC with OCR
RefEgo	✓	✓	✗	Egocentric video REC
KnowDR-REC(Ours)	✓	✓	✓	Knowledge-driven person REC

Model	**PosCorrect Error Rate**	**Orig. Error Rate**	**PosCorrect Neg-Pos@0.5**	**Orig. Neg-Pos@0.5**
Gemini-2.5-Pro	84.3	97.1	25.4	11.5
Grok-2-Vision	88.5	83.5	32.5	6.5
Qwen-VL-Plus	97.3	72.5	72.4	63.8
Qwen2.5-VL-72B	92.5	81.2	77.7	67.2
Qwen2.5-VL-7B	98.7	98.7	25.0	29.5

Table 4: PosCorrect-conditioned counterfactual audit reproduced from the rebuttal. "PosCorrect" restricts the evaluation to cases where the paired positive sample was correctly localized.

14 Each resource retains its original license, and specific usage details are summarized below.

6.9.1 Licensing Strategy

All included records preserve the original licensing information specified in their respective metadata. The JSON structure of the benchmark, excluding original content from the source datasets, is released under the CC0 1.0 Universal license.

6.9.2 Special Note on MetaQA

Due to potential NoDerivs (ND) restrictions associated with the MetaQA dataset, this benchmark only includes the original, unmodified MetaQA textual content. Modifications are provided solely in the form of scripts and transformation prompts, which are separately licensed under the MIT License.

6.9.3 Image Attribution

Images from Wikimedia Commons include comprehensive attribution in an accompanying metadata file (`attribution.tsv`), specifying title, author, source URL, and per-image licenses. Users must follow these licenses when reusing images.

6.9.4 Usage Recommendations

When redistributing or adapting data from this benchmark:

- Clearly attribute original sources and adhere to specified licenses.
- Preserve Apache-2.0 NOTICE when applicable (ComplexWebQuestions subset).
- Redistribute adaptations of CC BY-SA 4.0 content (HotpotQA, KQA Pro) under the same license.

6.10 Prompt

6.10.1 Prompt for Rewriting

To enhance the naturalness and linguistic diversity of referring expressions, we use GPT-4o to rewrite questions from four QA datasets (HotpotQA, MetaQA, ComplexWebQuestions, and KQA Pro) into declarative forms suitable for referring expression comprehension (REC). The prompt instructs the model to preserve factual content and produce unambiguous, grounding-oriented expressions.

Prompt Template:

> Please rewrite the following question as a declarative sentence. The rewriting requirements are:
>
> 1. Convert the interrogative sentence into a declarative sentence beginning with **"The person..."**
> 2. Retain the complete meaning and all information from the original sentence

Model	Zero-shot	Positive-ICL	Mixed-ICL	Negative-ICL
gemini-2.5-pro	97.1	99.3	22.3	2.8
grok-2-vision-latest	83.5	98.1	55.7	5.1
qwen-vl-plus	72.5	97.4	61.2	8.6
qwen2.5-vl-72b-instruct	81.2	99.0	58.4	10.3
VLM-R1-instruct	100.0	100.0	72.1	22.4

Table 5: Basic-setting adversarial negatives under different one-shot ICL strategies, reproduced from the rebuttal. The metric is error rate (%), where lower is better.

Model	Size bin	Acc@0.5	Acc@0.75	Acc@0.9	mean IoU	median IoU
gemini-2.5-pro	Q1	3.3	0.4	0.0	0.108	0.009
gemini-2.5-pro	Q2	8.4	0.4	0.0	0.179	0.099
gemini-2.5-pro	Q3	19.8	2.9	0.0	0.319	0.328
gemini-2.5-pro	Q4	57.0	15.6	3.3	0.510	0.549
grok-2-vision-latest	Q1	21.3	2.5	1.5	0.266	0.270
grok-2-vision-latest	Q2	58.4	12.4	5.0	0.480	0.570
grok-2-vision-latest	Q3	70.0	20.7	10.5	0.555	0.651
grok-2-vision-latest	Q4	77.5	31.1	16.6	0.614	0.674
qwen-vl-plus	Q1	40.1	24.1	5.9	0.343	0.110
qwen-vl-plus	Q2	63.9	54.1	19.3	0.567	0.770
qwen-vl-plus	Q3	74.7	69.1	34.8	0.670	0.844
qwen-vl-plus	Q4	79.5	72.1	38.5	0.722	0.876
qwen2.5-vl-72b-instruct	Q1	50.6	41.4	16.9	0.455	0.580
qwen2.5-vl-72b-instruct	Q2	68.9	59.8	43.9	0.628	0.876
qwen2.5-vl-72b-instruct	Q3	77.2	71.2	55.4	0.715	0.918
qwen2.5-vl-72b-instruct	Q4	87.2	82.7	67.5	0.817	0.940
VLM-R1-instruct	Q1	35.0	32.0	12.0	0.320	0.180
VLM-R1-instruct	Q2	60.0	58.0	36.0	0.560	0.720
VLM-R1-instruct	Q3	75.0	72.0	50.0	0.680	0.860
VLM-R1-instruct	Q4	84.8	84.0	64.0	0.780	0.910

Table 6: Positive samples stratified by ground-truth target size, reproduced from the rebuttal. Q1–Q4 denote quartiles from the smallest to the largest targets.

3. Ensure the declarative sentence clearly refers to the person: `{person_name}`
4. Use correct grammar, clear semantics, fluent sentence structure, and rich vocabulary
5. Return only the rewritten declarative sentence without any additional explanation

Original question: `{question_text}`
Target person: `{person_name}`

6.10.2 Prompt for Annotation Assistance

To support the manual annotation process, we design prompts that guide a language model (e.g., GPT-4o) to generate visual descriptions of a specific person in a given image. These prompts are used in one of two parallel annotation paths, where the model generates detailed appearance descriptions to help locate the target individual.

Prompt Template:

You are tasked with describing a person in a cropped image. Please follow the guidelines below:

1. Describe unique characteristics that distinguish the person.
2. Provide details on overall appearance, including clothing, hairstyle, etc.
3. Mention any visible interaction with other people or objects.
4. Note any visible text on the person (e.g., logos or labels).
5. Identify actions the person is performing.
6. Specify the person's relative position in the scene (e.g., "second from the left").

Model	Position bin	Acc@0.5	Acc@0.75	Acc@0.9	mean IoU	median IoU
gemini-2.5-pro	center	29.8	9.5	2.5	0.340	0.327
gemini-2.5-pro	mid	15.0	3.1	0.0	0.246	0.203
gemini-2.5-pro	edge	21.4	1.9	0.0	0.248	0.198
grok-2-vision-latest	center	69.0	17.5	12.0	0.562	0.632
grok-2-vision-latest	mid	54.6	15.6	8.0	0.481	0.520
grok-2-vision-latest	edge	46.4	16.7	5.2	0.392	0.467
qwen-vl-plus	center	81.6	68.7	31.0	0.717	0.833
qwen-vl-plus	mid	64.7	54.8	22.1	0.579	0.780
qwen-vl-plus	edge	47.2	40.9	20.6	0.429	0.353
qwen2.5-vl-72b-instruct	center	82.8	78.2	63.1	0.778	0.934
qwen2.5-vl-72b-instruct	mid	74.7	62.2	40.7	0.668	0.873
qwen2.5-vl-72b-instruct	edge	55.6	50.9	33.8	0.515	0.771
VLM-R1-instruct	center	78.0	76.0	52.0	0.720	0.880
VLM-R1-instruct	mid	64.0	62.0	40.0	0.600	0.780
VLM-R1-instruct	edge	49.1	46.5	29.5	0.500	0.640

Table 7: Positive samples stratified by ground-truth target position, reproduced from the rebuttal.

Model	Size bin	BBox output %	mean IoU(pred, pos_GT)
gemini-2.5-pro	Q1	93.8	0.050
gemini-2.5-pro	Q2	96.8	0.085
gemini-2.5-pro	Q3	99.2	0.145
gemini-2.5-pro	Q4	99.2	0.303
grok-2-vision-latest	Q1	78.0	0.055
grok-2-vision-latest	Q2	84.0	0.095
grok-2-vision-latest	Q3	86.4	0.145
grok-2-vision-latest	Q4	87.5	0.215
qwen-vl-plus	Q1	47.7	0.322
qwen-vl-plus	Q2	55.2	0.577
qwen-vl-plus	Q3	60.8	0.606
qwen-vl-plus	Q4	43.3	0.663
qwen2.5-vl-72b-instruct	Q1	68.0	0.373
qwen2.5-vl-72b-instruct	Q2	65.6	0.568
qwen2.5-vl-72b-instruct	Q3	68.8	0.627
qwen2.5-vl-72b-instruct	Q4	65.8	0.778
VLM-R1-instruct	Q1	100.0	0.285
VLM-R1-instruct	Q2	100.0	0.408
VLM-R1-instruct	Q3	100.0	0.493
VLM-R1-instruct	Q4	100.0	0.621

Table 8: Counterfactual negatives stratified by ground-truth target size, reproduced from the rebuttal.

7. Include any relevant context that may aid identification.

Input:
`{Image}`
`{person_name}`

6.10.3 Prompt for Knowledge Graph Extraction

To enable fine-grained negative sample generation, we guide the model to extract a temporal knowledge graph from each referring expression. The output is a tuple (`subject, relation, object, time`).

Prompt Template:

You are tasked with extracting a temporal knowledge graph from a given referring expression in order to support fine-grained negative sample generation. The output should be a quadruple in the form (`subject, relation, object, time`).

Please follow the guidelines below:

1. Interpret the input as a declarative sentence containing one or more factual statements.

Model	Position bin	BBox output %	mean IoU(pred, pos_GT)
gemini-2.5-pro	center	99.4	0.209
gemini-2.5-pro	mid	95.8	0.108
gemini-2.5-pro	edge	96.2	0.115
grok-2-vision-latest	center	89.4	0.195
grok-2-vision-latest	mid	82.7	0.135
grok-2-vision-latest	edge	78.1	0.085
qwen-vl-plus	center	51.2	0.663
qwen-vl-plus	mid	53.0	0.568
qwen-vl-plus	edge	51.2	0.388
qwen2.5-vl-72b-instruct	center	71.8	0.741
qwen2.5-vl-72b-instruct	mid	64.3	0.593
qwen2.5-vl-72b-instruct	edge	65.0	0.384
VLM-R1-instruct	center	100.0	0.533
VLM-R1-instruct	mid	100.0	0.429
VLM-R1-instruct	edge	100.0	0.318

Table 9: Counterfactual negatives stratified by ground-truth target position, reproduced from the rebuttal.

Model	mean IoU	median IoU	Acc@0.5	Acc@0.75	Acc@0.9	missing %	invalid %	OOB %	full-image-like %
gemini-2.5-pro	0.278	0.245	22.1	4.8	0.8	0.0	0.0	14.0	0.0
gemini-2.5-flash	0.420	0.451	44.1	16.3	2.8	0.0	1.7	0.6	2.9
gemini-3-pro-preview	0.445	0.490	43.8	27.5	11.8	0.0	0.0	0.0	0.0
gemini-3-flash-preview	0.465	0.535	53.5	32.4	11.2	0.0	0.0	26.3	0.0

Table 10: Gemini-family grounding quality and box legality, reproduced from the rebuttal. Lower missing/invalid/out-of-bound (OOB) rates are better.

2. For each statement, identify the subject, relation, object, and time expression.
3. Return the extracted information as one or more structured quadruples in the format: `(subject, relation, object, time)`.
4. Ensure that the extracted entities are specific and semantically correct. Avoid abstract or overly general concepts.
5. Return only the list of quadruples, one per line, without any additional explanation.

Input:
`{Refering_Expression}`

6.10.4 Prompt for Negative Text Generation

We construct adversarial expressions by perturbing a single element in the extracted tuple (e.g., changing the object or time). The model is prompted to generate a fluent but factually incorrect version of the original sentence based on the corrupted tuple.

Each input may contain multiple factual tuples. Only one tuple should be selected and only one component (entity, relation, or time) should be perturbed. The resulting sentence must remain grammatically fluent but contain a factual error due to the corruption.

Entity Corruption

Prompt Template: Given a natural language sentence and a set of factual tuples it expresses, the task is to generate a grammatically fluent but factually incorrect version of the sentence by corrupting the **subject or object entity** in exactly one of the tuples. Each tuple is in the format `(subject, relation, object, time)`.

Only one tuple should be selected for corruption, and only the entity (either subject or object) should be modified. The remaining parts of all tuples should stay unchanged. Ensure the corrupted sentence remains fluent and natural, but contains a factual error caused by the entity change.

Ensure the response includes only the corrupted sentence, without any additional text, characters, or explanations.

Input sentence: *[Original Natural Language Sentence]*

Tuples: `[(subject_1, relation_1,`

Model	pred area (median)	pred area (p90)	oversized %	tiny %	median log-area gap	median log-aspect gap
gemini-2.5-pro	0.245	0.620	15.6	1.8	-0.025	0.242
gemini-2.5-flash	0.330	0.676	20.8	1.5	0.234	0.059
gemini-3-pro-preview	0.243	0.790	21.1	0.0	-0.069	0.215
gemini-3-flash-preview	0.377	1.165	42.1	5.3	-0.088	0.001

Table 11: Gemini-family box tightness and style mismatch, reproduced from the rebuttal.

Model	mean "**" tokens	max "**" tokens
gemini-2.5-pro	0.01	8
gemini-2.5-flash	0.10	24
gemini-3-pro-preview	0.00	0
gemini-3-flash-preview	0.00	0

Table 12: Gemini-family output-format noise indicators, reproduced from the rebuttal.

Model	< 0.1	0.1–0.3	0.3–0.5	≥ 0.5
gemini-2.5-pro	36.3	19.6	22.1	22.1
gemini-2.5-flash	23.3	13.1	19.5	44.1
gemini-3-pro-preview	25.0	15.4	15.8	43.8
gemini-3-flash-preview	30.7	10.5	5.3	53.5

Table 13: Gemini-family IoU distribution, reproduced from the rebuttal.

`object_1, time_1), (subject_2, relation_2, object_2, time_2), ...]`

Relation Corruption

Prompt Template: Given a natural language sentence and a set of factual tuples it expresses, the task is to generate a grammatically fluent but factually incorrect version of the sentence by corrupting the **relation** in exactly one of the tuples. Each tuple is in the format `(subject, relation, object, time)`.

Only one tuple should be selected for corruption, and only the relation should be changed to a plausible but incorrect alternative. The rest of the sentence and tuples should remain unmodified. The resulting sentence must be grammatically correct but contain a factual error due to the relation change.

Ensure the response includes only the corrupted sentence, without any additional text, characters, or explanations.

Input sentence: *[Original Natural Language Sentence]*

Tuples: `[(subject_1, relation_1, object_1, time_1), (subject_2, relation_2, object_2, time_2), ...]`

Time Corruption

Prompt Template: Given a natural language sentence and a set of factual tuples it expresses, the task is to generate a grammatically fluent but factually incorrect version of the sentence by corrupting the **time expression** in exactly one of the tuples. Each tuple is in the format `(subject, relation, object, time)`.

Select only one tuple to modify, and change only the time component to a different but plausible value. All other components and tuples should remain unchanged. The corrupted sentence should sound natural but contain a factual error due to the incorrect time.

Ensure the response includes only the corrupted sentence, without any additional text, characters, or explanations.

Input sentence: *[Original Natural Language Sentence]*

Tuples: `[(subject_1, relation_1, object_1, time_1), (subject_2, relation_2, object_2, time_2), ...]`

6.10.5 Prompt for Positive and Negative Sample Evaluation

Positive Samples For positive samples, we make minimal modifications to the default prompts of all Specialist MLLMs and certain Generalist MLLMs(Lin et al., 2023; Liu et al., 2024a) to ensure that predictions of the target objects are explicitly included in the output. For other Generalist MLLMs without predefined prompts, we adopt a unified prompt that yields consistently strong performance across models, as illustrated in Figure 8.

Negative Samples under the Basic Setting For negative samples under the basic setting, our goal is to evaluate whether the models are capable of appropriately refusing to answer in typical usage scenarios. For all Specialist MLLMs and certain Generalist MLLMs (Lin et al., 2023; Liu et al., 2024a) with predefined visual prompts, we directly reuse their default prompts. For Generalist MLLMs

Dataset	Original License	Modification Status
HotpotQA (Yang et al., 2018)	CC BY-SA 4.0	Modified, distributed under CC BY-SA 4.0
ComplexWebQuestions (Talmor and Berant, 2018)	Apache-2.0	Modified, Apache-2.0 (NOTICE preserved)
KQA Pro (Cao et al., 2020)	CC BY-SA 4.0	Modified, distributed under CC BY-SA 4.0
MetaQA (Puerto et al., 2021)	CC BY-ND 3.0	Original text only; modifications via scripts
Wikimedia Commons	Per-image licenses	Individual image attribution provided

Table 14: Licenses of the datasets used in the benchmark.

```
Given an image and a question, follow these reasoning steps to identify and localize the main person referenced in the question:

Step 1: Understand the question and extract key clues or facts about the person mentioned.

Step 2: Based on the extracted information, infer the identity of the person.
- Use commonsense and factual knowledge to identify who matches all the given attributes.
- IMPORTANT: Once identified, explicitly state the target person in the format:
  Target: [person's full name - NOT a description]
  → For example: Target: Emma Watson

Step 3: Locate the identified person in the image using facial recognition, context, or other visual cues.

Step 4: Determine the bounding box of the person in the image.
- Provide coordinates in the format [x1, y1, x2, y2], where (x1, y1) is the top-left and (x2, y2) is the bottom-right corner.
- Use only integers for the coordinates.

---

Now analyze the image:
Question: {insert question here}

Please follow the reasoning steps above and output your result at the end in this format:
Bounding Box: [x1, y1, x2, y2]
```

Figure 8: Prompt template for positive samples

without predefined prompts, we employ the prompt shown in Figure 9.

Negative Samples under the Instruction-guided Setting For the evaluation of negative samples under the instruction-guided setting, we adopt a unified prompt across all models, as shown in Figure 10.

6.11 Illustrative CoT-Based Evaluation Example

Our reasoning-stage evaluation requires the model to explicitly state the resolved target entity before outputting a box. Figure 8 shows the prompt template used for this stage, and Figure 11 provides a representative successful example in which the model identifies the target person and outputs a matching box. Together, these examples clarify how the reasoning output is exposed for alias-based matching in Acc_{Text}.

6.12 More Samples

In this section, we present additional examples as shown in Figures 11, 12, and 13.

```
"""Given an image and a referring description, locate the target and provide its bounding box coordinates.

Question: {question}

Please provide the bounding box coordinates in the following format:
Bounding Box: [x1, y1, x2, y2]

Where (x1,y1) is the top-left corner and (x2,y2) is the bottom-right corner.
"""
```

Figure 9: Prompt template for negative samples under the basic setting

```
"""Given an image and a description, please determine if the described content is present in the image.

Question: {question}

Please respond with only one word: "Yes" or "No"

If the described person, object, or content is present in the image, answer "Yes".
If the described person, object, or content is not present in the image, answer "No".
"""
```

Figure 10: Prompt template for negative samples under the instruction-guided setting

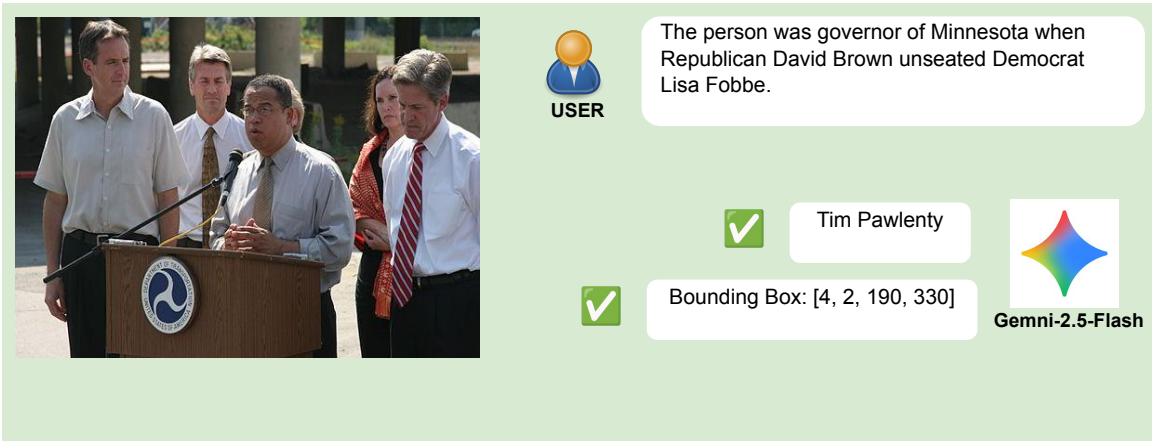

Figure 11: Successful cases under the positive sample

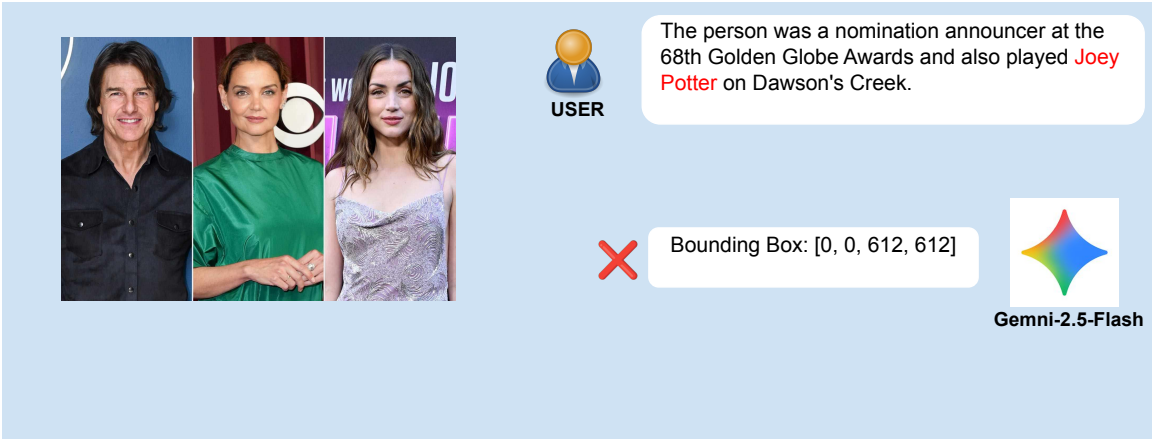

Figure 12: Failure cases under the negative sample setting (basic setting)

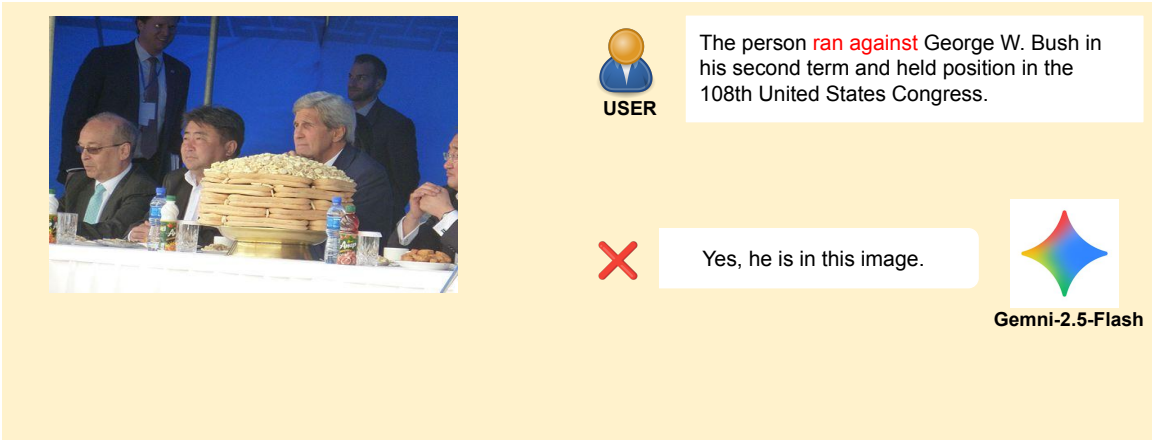

Figure 13: Failure cases under the negative sample setting (instruction-guided setting)