CLEVR-3D-DeRef: A Benchmark for Robust 3D Spatial Reasoning with De-Identified Referring Expressions

Mary Lynn Martin, Martha Palmer, Maria Leonor Pacheco
University of Colorado Boulder
{mary.martin, martha.palmer, maria.pacheco}@colorado.edu

Abstract

Vision-language models (VLMs) often struggle to interpret spatial referring expressions that require relational reasoning rather than reliance on surface-level cues. These models frequently identify referents through explicit visual attributes such as color or shape, rather than understanding spatial relationships (e.g., "to the left of the red cube"). To systematically analyze these limitations, we introduce CLEVR-3D-DeRef, a synthetic and extensible benchmark dataset modeled after CLEVR-Ref+, designed to evaluate spatial reasoning in multi-modal systems. CLEVR-3D-DeRef extends the original framework by incorporating depth information for 3D spatial reasoning, introducing de-identified context-dependent referring expressions that require relational inference to disambiguate referent objects, and expanding the range of spatial relations beyond the original four. We further extend our dataset by producing expressions with and without ordinal language and diversifying the language and structure of expressions while preserving meaning. CLEVR-3D-DeRef provides a controlled and reproducible resource for studying robust referring expression comprehension by enhancing the diversity of expressions while significantly reducing surface-level artifacts.

Keywords: vision-language models, robust spatial reasoning, de-identified referring expressions

1. Introduction

Understanding spatial relationships through language is a fundamental aspect of grounded visual reasoning. For vision-language models (VLMs), spatial understanding is essential for interpreting how objects relate within three-dimensional scenes and for reasoning about configurations that enable manipulation. Spatial reasoning bridges perception with action: to follow an instruction like "place the red cube between the two spheres," a model must first recognize objects, infer their relative positions, and reason about geometric constraints.

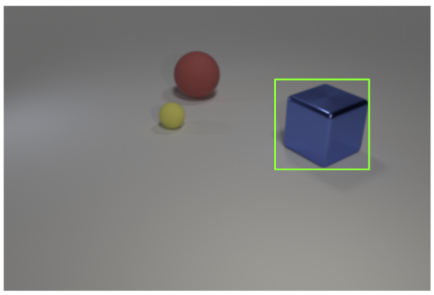
However, despite rapid advances in multimodal learning, current VLMs often rely on surface-level visual cues, such as color or shape, rather than performing genuine spatial reasoning (Zhang et al., 2025b,a; Qi et al., 2025). This issue is amplified by existing benchmarks, which often contain dataset biases that allow models to succeed by exploiting shallow correlations rather than understanding spatial configurations (Cirik et al., 2018; Kafle et al., 2019). For example, when given an instruction to identify "the cube to the left of the red sphere," a model may produce a correct solution simply because there is only one cube near the red sphere, which does not require actual relational inference. These shortcuts reveal a broader limitation; high benchmark performance does not necessarily imply spatial grounding, which is crucial for reasoning, planning, and physical interaction in robotic and embodied settings.

Several benchmarks have been proposed to evaluate spatial and relational reasoning in multimodal systems. For example, CLEVR-Ref+ focuses on referring expression comprehension, that is, the task of locating objects in images based on natural language descriptions (Liu et al., 2019). GQA evaluates compositional reasoning on real-world images in a visual question-answering setup (Ainslie et al., 2023), and NLVR2 tests comparative and logical reasoning by determining whether natural language statements accurately describe pairs of real images (Suhr et al., 2019). Among these, CLEVR-Ref+ remains a key controlled benchmark due to its transparent and reproducible programmatic referring expression generation pipeline, yet it is restricted to 2D scenes, a limited set of spatial relations, and attribute-dominant language that often allows models to bypass relational reasoning.

To address these limitations, we introduce CLEVR-3D-DeRef, a benchmark dataset designed to test *genuine* spatial reasoning in referring expression comprehension tasks. Our dataset builds on the controlled compositional structure of CLEVR-Ref+ while introducing the following extensions:

1. Incorporate depth information to produce 3D spatial scenes, enabling models to reason beyond 2D projections.
2. Expand the relation set beyond the original four relation types (i.e., left, right, front, and behind) to include proximity-based relations (i.e., between, nearest, and farthest) which require reasoning about relative distance.
3. Diversify the dataset by generating alternative formulations for all expressions that preserve meaning, object identity, and spatial relations while varying lexical and syntactic patterns.
4. Introduce underspecified context-dependent re-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9490–9503
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026



Unmasked: *The blue cube* that is to the right of *the yellow sphere* that is in front of the red cube.
Attribute-Masked: *The thing* that is to the right of *the object* that is in front of the red sphere.

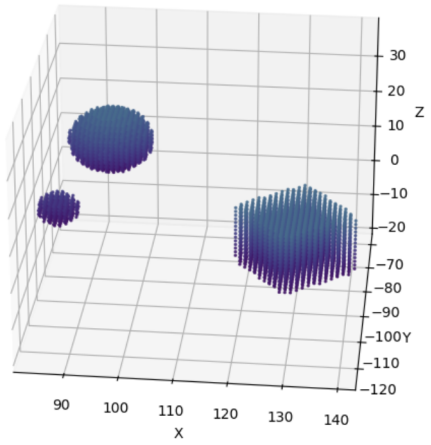


Figure 1: Example scene. The top image shows the 2D rendered image with a green bounding box indicating the referent object. The bottom image shows the corresponding 3D point cloud representation. Between the views are two referring expressions: The original (attribute-preserving) expression and its attribute-masked variant.

ferring expressions to evaluate model performance in cases where multiple objects satisfy the same surface-level description, requiring spatial reasoning to disambiguate the referent.

Together, these extensions establish CLEVR-3D-DeRef as a robust benchmark for 3D spatial reasoning in referring expression comprehension tasks, minimizing surface-level artifacts that can distort evaluation.

2. Related Work

Existing research on spatial grounding spans a wide spectrum, from synthetic diagnostic datasets designed to isolate compositional reasoning skills (Johnson et al., 2017; Liu et al., 2019; Ainslie et al., 2023; Chen et al., 2020b) to large-scale natural benchmarks assessing generalization in complex scenes (Yang et al., 2020; Zheng et al., 2022; Agarwal et al., 2025). While this diversity has advanced our understanding of multimodal reasoning, few benchmarks emphasize robust, bias-resistant evaluation, jointly capturing reasoning depth, object de-identification, and relational reasoning within a controlled and reproducible framework. Our work situates itself at this intersection, aiming to provide a dataset that isolates genuine spatial understanding by explicitly minimizing surface-level and dataset-specific artifacts.

Early diagnostic benchmarks such as CLEVR (Johnson et al., 2017) and CLEVR-Ref+ (Liu et al., 2019) introduced synthetic, programmatically generated scenes that enabled precise analysis of compositional reasoning and referring expression comprehension. Their controlled environments and functional-program annotations provided interpretability but were limited to 2D scenes, four spatial relations, and expressions riddled with object attributes that could be used to bypass spatial reasoning. Subsequent work has extended this paradigm to test spatial reasoning in modern VLMs. WSSR (Banerjee et al., 2021), VIPHY (Singh et al., 2023), VSR (Liu et al., 2023a), and SpatialVLM (SVLM) (Chen et al., 2024) integrate depth estimation or synthetic 3D cues to evaluate geometric understanding and relational grounding. While these approaches improve performance on specific spatial tasks, they reveal a persistent reliance on surface-level attributes and limited generalization to 3D or ambiguous spatial expressions. In general, synthetic and diagnostic datasets provide essential control and interpretability but remain static and nonembodied, offering little insight into how models use spatial understanding to guide physical reasoning or action.

A complementary research area focuses on the grounding of natural language in 3D environments. Datasets such as ReferIt3D (Achlioptas et al., 2020b) and NR3D extend referring expression comprehension to 3D point clouds derived from ScanNet (Dai et al., 2017), enabling reasoning about volumetric spatial relations (Achlioptas et al., 2020a). ScanRefer (Chen et al., 2020a) and SQA3D (Ma et al., 2023) further connect natural language with 3D scene scans to evaluate spatial localization and question answering. These benchmarks advance spatial understanding beyond 2D imagery, but they primarily test comprehension rather than spatial transformation or manipulation, and they offer limited linguistic variability or explicit probing of de-identified expressions. As a result, while they enable grounding in 3D structure, they do not address how models might use this information to generate spatial actions.

3. Methodology

We introduce a comprehensive spatial referring expression dataset designed to evaluate spatial reasoning in vision-language models and probe their reliance on surface-level artifacts for identifying referent objects. In this section, we first briefly introduce CLEVR-Ref+ (Sec. 3.1). Then, we describe our methodology to build CLEVR-3D-DeRef by converting 2D images into 3D point clouds (Sec. 3.2), adding new relation types (Sec. 3.3), diversifying referring expressions (Sec. 3.4) and de-identifying referent objects (Sec. 3.5).

3.1. CLEVR-Ref+ Overview

CLEVR-Ref+ is a synthetic benchmark dataset developed to evaluate referring expression comprehension and visual reasoning in multimodal systems (Liu et al., 2019). As an extension of the original CLEVR dataset (Johnson et al., 2017), it replaces question-answer pairs with natural language expressions that describe specific target objects within rendered 2D scenes. Each scene contains multiple objects that differ in shape, size, color, and material, arranged with carefully controlled spatial relationships. These expressions are generated automatically through a template-based process built on CLEVR's functional program structure. Each program specifies a sequence of operations, such as filtering by attributes or relating objects spatially, that are then translated into natural language using varied linguistic templates and synonyms. For example, a program might select a red sphere and then locate a cube to its right, resulting in the expression "the cube to the right of the red sphere". This generation process ensures that all expressions are semantically precise, compositional, and fully grounded in the visual scene.

We keep the original CLEVR-Ref+ generation framework because it gives us direct control over what each expression means and how it maps to the scene. Since the expressions are derived from functional programs, each one can be checked against the scene structure to confirm that it refers to the intended object and reflects the correct sequence of reasoning steps. This is especially important for our dataset, where we intentionally vary reasoning depth, add new relation types, and mask identifying attributes. Using a programmatic pipeline helps us make these changes in a controlled and reproducible way without introducing unintended errors. We use LLMs only later in the pipeline to create paraphrased variants of already validated expressions, so that we can add linguistic variety without changing the underlying meaning.

3.2. Creating 3D Point Clouds

To provide an accompanying 3D representation for each rendered scene, we integrate point cloud generation directly into the CLEVR image generation pipeline. During scene generation, a raycasting procedure in Blender samples the scene volume at a fixed spatial density (approximately 12 samples per meter) to determine which regions are occupied by objects. For each sampled location, rays are cast to detect mesh intersections, producing a 3D occupancy grid that is aligned with the 3D camera view. This data is used to create and store a 3D point cloud along with its corresponding 2D image (see Fig. 2). These point clouds capture the full 3D scene geometry, allowing models capable of processing 3D inputs to exploit spatial structure and depth information during reasoning.

3.3. Adding Relation Types

We extend the CLEVR referring-expression generation pipeline with two relation families beyond the original directional set (*left/right/front/behind*): (i) **between** relations that capture geometric betweenness with respect to two reference objects and (ii) **distance** relations that support *nearest* and *farthest* reasoning. These additions broaden the spatial vocabulary and enable the evaluation of models on geometric and metric reasoning, rather than solely on ordinal directions. The implementation is carried out in two stages.

Scene metadata augmentation. For each scene, we compute the pairwise 3D Euclidean distances between all objects and store (a) exact distances and (b) per-object rankings (nearest to farthest). For *between* detection, we first form candidate triplets from pairs of opposing directional relations (e.g., an object that is left of A and right of B, or in front of A and behind B). We then apply a geometric proximity test: an object is retained as *between* the pair if its position lies within an adaptive tolerance of the pair's midpoint. The tolerance is set from scene statistics (mean minus one standard deviation of inter-object distances), which scales the criterion to each scene's layout.

Template integration. The metadata required to determine distance and betweenness are written alongside existing CLEVR metadata at scene generation time. For each scene object, the metadata stores its 3D position, pairwise distances to all other objects, and the ordered ranking of those distances. In addition, it records valid object pairs that each object lies *between*, that is, instances where the object's position satisfies the geometric betweenness criterion relative to two reference objects.

During generation, the engine uses these metadata to determine which relation templates to instantiate. It queries stored positions, distance rankings, and between-pair records to identify valid *nearest*, *farthest*, and *between* configurations, ensuring that each generated expression corresponds to an actual spatial relationship present in the scene. These new templates are integrated with the existing directional ones, allowing us to define the composition of the expressions (e.g., *"the cube nearest to the object between X and Y"*). A final merge step consolidates the output from all template families and preserves consistent indexing for downstream evaluation. By grounding relation generation in explicit scene metadata, we extend coverage beyond directional reasoning to encompass distance-based relational reasoning.

3.4. Expression Diversification

To enhance the linguistic diversity and naturalness of referring expressions, we integrate a large language model (LLM)-based diversification module into the referring expression generation pipeline. While the base system produces structured, templated expressions designed for precision and compositional consistency, these outputs often lack the variability and natural phrasing found in human language. The diversification component introduces controlled linguistic variation by paraphrasing existing expressions through an LLM, producing multiple alternative phrasings that remain semantically faithful and spatially consistent with the original expression. For example, an expression such as *"the blue sphere that is right of the red cube"* may yield variants such as *"the blue sphere sitting to the right of the red cube"* or *"the sphere on the right side of the red cube."* This step enriches the dataset with stylistic and syntactic diversity without compromising referential clarity. The diversification pipeline operates in two main stages.

Prompt Construction and Execution. Each base expression is paired with contextual scene information (including object attributes and relevant spatial relations) to form a structured prompt. This prompt instructs the model to generate alternative formulations that preserve meaning, object identity, and spatial relations while varying lexical and syntactic patterns. We use OpenAI 3.5 turbo with specific generation parameters: a temperature of 0.8, one variation per expression, and a batch size of 10 expressions processed simultaneously.

Post-Processing and Filtering. The generated candidates are automatically evaluated to ensure semantic consistency with the original. Expressions are discarded if they introduce attribute mismatches, alter spatial relations, or contain malformed syntax. The remaining expressions form a verified and diversified set aligned with the original meaning. To systematically assess the output, we define several quantitative measures:

- **Meaning Preservation:** Assesses whether semantic content–object attributes (color, shape, material) and spatial relations–are preserved across original and diversified expressions. For each category, the evaluator measures the proportion of attributes or relations retained using set intersection, then averages across categories to obtain an overall meaning preservation score.
- **Linguistic Diversity:** Quantifies linguistic diversity by measuring variation in vocabulary, sentence length, and structure. The evaluator computes word diversity as the ratio of unique to total words, captures length variance as a measure of expression complexity, and assesses syntactic diversity using unique sentence starters. The word and structure diversity are averaged to form the overall diversity score.
- **Naturalness:** Assesses grammatical and stylistic fluency through the analysis of article usage, preposition diversity, conjunction balance, and relative clause variety. The evaluator measures the proportion of definite articles, the diversity of prepositions and conjunctions, and the frequency of relative clauses, averaging these indicators to produce an overall naturalness score.

To evaluate the fidelity of LLM-diversified expressions, we compute a composite *quality score* as a weighted average of meaning preservation (40%), linguistic diversity (30%), and naturalness (30%). This score is used for post-hoc validation rather than filtering. Since the diversification prompt (Appendix section B) tightly constrains the LLM to preserve all spatial relations and object references, the resulting rephrasings obtain high preservation scores (mean $= 0.917$). These results suggest that the pipeline maintains referential accuracy reliably.

After merging all expression types into a unified dataset, the diversification module generates configurable variants of selected expressions. The system outputs two files: (i) a **unified expressions file** with only original expressions and (ii) a **comprehensive file** containing both originals and diversified variants. By maintaining referential alignment between originals and variants, the augmented dataset enables evaluation of linguistic robustness rather than reinterpretation of scene semantics. Appendix section B details all prompts used for LLM diversification.

3.5. De-Identifying Referent Objects

To extend the expressivity and difficulty of spatial referring expression datasets, we implement a *de-*

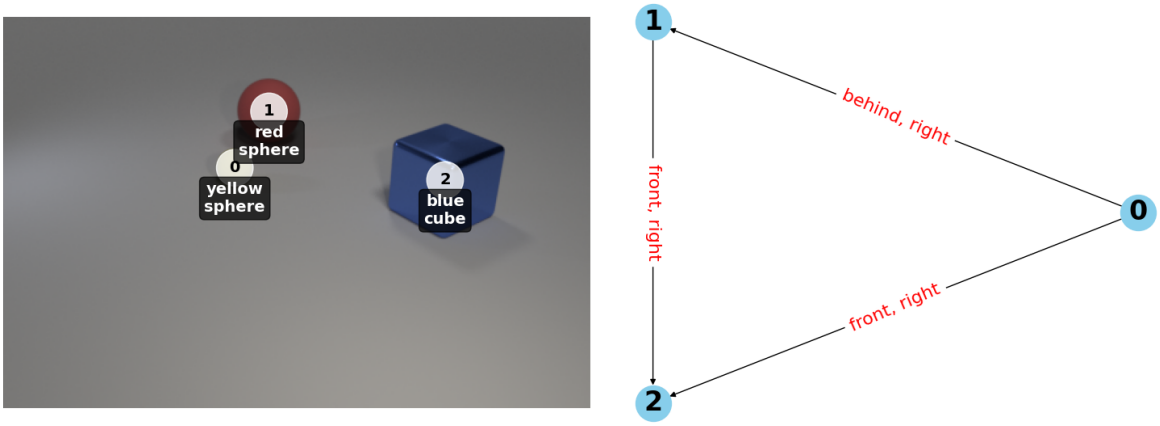


Figure 2: Scene graph for the 2D scene in figure 1. Scene graph structures are used to identify unique spatial configurations that allow de-identification of the referent in referring expressions.

identification mechanism that selectively removes explicit object descriptors when unique spatial relations are detected. This mechanism produces referring expressions that lose access to object attributes and can only be resolved via spatial and relational reasoning. It operates by analyzing the spatial scene graph for uniqueness patterns and dynamically masking object attributes (e.g., color, shape) in generated expressions (see Fig. 2).

Each scene is represented as a directed labeled graph

$$G = (V, E, L)$$

where each node $v_i \in V$ corresponds to an object in the scene and each directed edge $e_{ij} \in E$ denotes a spatial relation $r_{ij} \in L$ from object v_i to object v_j. For instance, $e_{ij} = (v_i, r_{ij}, v_j)$ may represent "object i is to the left of object j." Incoming and outgoing relations for a node v_i are defined as:

$$R_{\text{in}}(v_i) = \{(v_j, r_{ji}) \mid (v_j, r_{ji}, v_i) \in E\},$$
$$R_{\text{out}}(v_i) = \{(r_{ij}, v_j) \mid (v_i, r_{ij}, v_j) \in E\}.$$

The ambiguity mechanism performs **1-edge** and **2-edge** uniqueness analyses to identify relations or relation paths that occur uniquely in the scene.

For each node v_i, a relation $r_{ij} \in R_{\text{out}}(v_i)$ is considered unique if:

$$|\{v_k \mid (v_i, r_{ik}, v_k) \in E\}| = 1.$$

Analogously, for incoming relations:

$$|\{v_k \mid (v_k, r_{ki}, v_i) \in E\}| = 1.$$

If an object participates in a relation that no other object exhibits, that relation provides sufficient discriminative power to identify the object, enabling masking of intrinsic descriptors in the expression.

Two-edge paths are defined as sequences

$$p_{ijk} = (v_i, r_{ij}, v_j, r_{jk}, v_k)$$

where $(v_i, r_{ij}, v_j) \in E$ and $(v_j, r_{jk}, v_k) \in E$. A path p_{ijk} is unique for terminal node v_k if no other distinct path terminates at v_k with the same ordered pair of relation labels:

$$|\{(v_a, v_b) \mid (v_a, r_{ab}, v_b, r_{bk}, v_k) \in E^2, (r_{ab}, r_{bk}) = (r_{ij}, r_{jk})\}| = 1.$$

Two-edge uniqueness enables masking of both the referent and, when applicable, the intermediate object when their spatial configuration uniquely identifies them within the scene.

Given the uniqueness analysis, the masking process substitutes or removes explicit object descriptors in the generated referring expression $\mathcal{E}$. Let $\mathcal{E} = f(o_t, r, o_r)$ denote a template-based expression describing a **target object** o_t (the referent) in relation r to a **reference object** o_r.

If r (or a two-edge relation sequence involving r) is unique for o_t, then:

$$\mathcal{E}' = \text{mask}(f(o_t, r, o_r))$$

where the masking operator removes or replaces intrinsic descriptors of o_t:

$$\text{mask}(\text{"the blue cube"}) = \text{"the thing"}.$$

Masking rules are as follows:

- **Color descriptors** $(c) \rightarrow$ removed
- **Shape descriptors** $(s) \rightarrow$ replaced with "thing" or "object"
- **Material descriptors** $(m) \rightarrow$ removed

The resulting expression $\mathcal{E}'$ remains semantically valid and resolvable using spatial reasoning alone.

4. Resulting Dataset

The proposed dataset consists of 54,575 referring expressions paired with 1,000 synthetically generated 3D scenes, each containing between 3 and

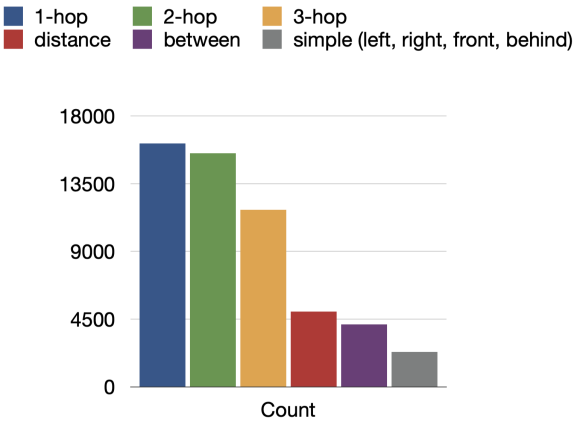


Figure 3: Multi-hop and zero-hop type distribution

10 objects. Every expression uniquely identifies a single object within its scene, yielding an average of approximately 54.6 expressions per scene.

The dataset integrates two complementary generation methods. The first is a template-based system, which produces 45,430 expressions (83.2%) through executable programs capturing 0-3-hop compositional reasoning. The second is a metadata-based system, which generates 4,140 "between" relations (7.6%) and 5,005 "distance" relations (9.2%) directly from precomputed geometric metadata. These metadata-derived expressions extend the spatial relation inventory to include proximity-based (*nearest/farthest*) and ternary (*between*) configurations, providing coverage beyond pairwise directional reasoning. Among distance-based expressions, 2,538 describe the *nearest* object and 2,467 describe the *farthest*, maintaining a balance between proximity extremes.

4.1. Reasoning Complexity and Relation Type Distribution

Expressions vary in complexity depending on the number of spatial reasoning steps, or "hops", required to identify the referent. This design supports the evaluation of the increasing relational depth and compositional structure.

Nearly 95% of all expressions require at least one spatial reasoning step, and almost 60% involve two or more chained relations, emphasizing the dataset's focus on multi-hop reasoning. Template-based expressions employ four canonical directional relations (*left*, *right*, *front*, *behind*), while metadata-based expressions add proximity (*nearest*, *farthest*) and ternary (*between*) structures. The resulting distribution of multi-hop and zero-hop relations per type can be observed in Fig. 3

4.2. Attribute Masking Distribution

To evaluate reasoning based purely on spatial information, the dataset incorporates a de-identification mechanism, replacing attribute descriptors (e.g., "red cube") with generic terms ("thing" or "object") when spatial relations alone uniquely identify the referent. This subset enables the targeted evaluation of geometric reasoning without reliance on appearance cues.

The masking distribution (Fig. 4) shows that the distance and between expressions often allow for unique spatial identification, while template-based ones typically require both spatial and attribute cues. Note that the template-based split contains all of the multi-hop relations, which explains its larger size. The 16.7% masked subset isolates attribute-independent spatial reasoning.

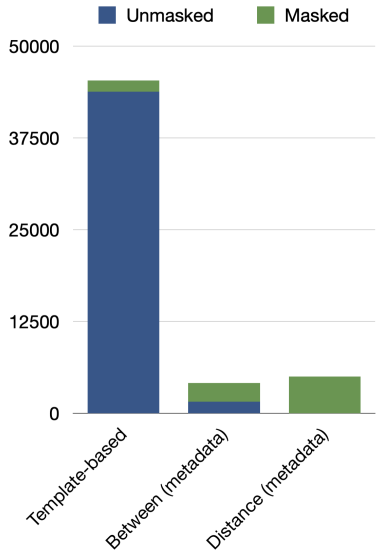


Figure 4: Attribute masking statistics

4.3. Expression Validation

All expressions were validated for textual and geometric consistency. Template-based expressions were synthesized directly from executable programs, ensuring perfect clause-to-relation correspondence. Metadata-based expressions followed identical referent-uniqueness criteria based on scene geometry. All 54,575 expressions passed validation with zero descriptor mismatches, zero relation mismatches, and zero invalid entries.

Across 1,000 scenes, expressions per image range from 26 to 50, with a median of 46. Rejection sampling during generation ensures balanced referent coverage and prevents overrepresentation of any single object or spatial configuration. This uniformity enables an unbiased evaluation of relation types, reasoning depths, and masking conditions.

5. Experimental Evaluation

To assess the spatial reasoning capabilities of current VLMs, we evaluated a range of architectures on the CLEVR-3D-DeRef dataset. Our goal was to examine how different modeling paradigms handle spatial referring expressions that require relational inference beyond surface-level cues. Each model was tested on its ability to correctly identify the spatially described referent within complex synthetic scenes. This evaluation provides insights into how architectural design and training objectives influence spatial grounding, compositional reasoning, and sensitivity to attribute-based shortcuts.

5.1. Models and Implementation Details

ReCLIP (Subramanian et al., 2022): ReCLIP was selected as a baseline for spatial referring expression comprehension due to its explicit alignment between visual and linguistic embeddings through contrastive fine-tuning. Its training objective optimizes for correspondence between a text description and a localized image region, making it well-suited for a bounding box selection task.

For evaluation, each input sample consists of the target image and an accompanying referring expression. ReCLIP receives a set of candidate bounding boxes, and the model's task is to assign the highest similarity score to the region that best matches the described referent. This approach aligns with ReCLIP's native zero-shot retrieval formulation, requiring minimal modifications. Appendix section A.1 details the prompts used for ReCLIP evaluation.

LLaVA-13b (Liu et al., 2023b): LLaVA-13b was used to evaluate how a multimodal instruction-tuned model performs spatial reasoning when presented with structured prompts. Unlike ReCLIP, LLaVA is a generative model that produces text responses conditioned jointly on an image and a natural language input. To adapt this to a bounding box selection task, each sample was framed as a multiple-choice question listing all candidate bounding boxes along with their coordinates and approximate dimensions. The model was asked to select the option that matched the described referent. This format was not directed by the model's architecture, but was introduced to make its generative output directly comparable to discrete selection tasks. In all setups, the goal was to identify the region that corresponds to the described object. Appendix section A.2 details the prompts used for LLaVA evaluation.

GPT-4o (OpenAI et al., 2024): A smaller-scale comparison was conducted using GPT-4o on a curated subset of 5,115 non-ordinal referring expressions drawn from the large-scale evaluation data. This subset excludes templates containing ordinal language (e.g., "The second sphere from the left") to focus evaluation on purely spatial and relational reasoning rather than order-based reference resolution. GPT-4o was selected instead of GPT-5 due to practical prompting constraints. GPT-5's current API interface proved problematic for a bounding-box selection task that required consistent output. GPT-4o allowed for reliable prompting in a standardized text format analogous to the LLaVA and ReCLIP setup. A smaller subset was chosen due to resource limitations. Appendix section A.3 details the prompts used for GPT-4o evaluation.

5.2. Results and Analysis

The evaluation compares ReCLIP, LLaVA-13b, and GPT-4o (on a smaller subset) on the CLEVR-3D-DeRef dataset. Each model was tasked with selecting the bounding box corresponding to the described referent. Accuracy was calculated as the proportion of correctly identified regions out of all evaluated samples.

5.2.1. Full-Scale Evaluation

ReCLIP consistently outperformed LLaVA-13b in all evaluation metrics in the entire set of 54,575 spatial referring expressions. Both models followed similar qualitative trends, performing best on proximity-based relations and struggling with complex multi-hop reasoning. However, ReCLIP achieved nearly double the overall accuracy, scoring 0.452 compared to 0.226 for LLaVA-13b. ReCLIP also maintained more stable performance across varying levels of relational complexity. These trends suggest that ReCLIP's embedding-based retrieval approach provides a consistent grounding signal across structural variations.

Performance by Template Type. Both models achieved their highest accuracy on distance relations (nearest, farthest), indicating that these spatial primitives are easier to ground. Accuracy decreased steadily with the depth of reasoning, reflecting the difficulty of chaining spatial clauses. ReCLIP consistently outperformed LLaVA-13b in all template families, with the largest margin in distance-based templates.

Performance by Spatial Relation. Directional relations ("left", "right", "front", "behind") proved more challenging for both models, yielding accuracies near 0.40 for ReCLIP and below 0.21 for LLaVA-13b. Moderate performance on "between" relations indicates that ternary spatial reasoning is somewhat achievable, though still constrained in accuracy and consistency.

Performance on Attribute-Masked Expressions. The CLEVR-3D-DeRef dataset includes an de-identification mechanism that replaces object-specific descriptors (e.g., "the red cube") with

Table 1: Model accuracy by template type.

Template Type	ReCLIP	LLaVA-13b
farthest	0.782	0.500
nearest	0.751	0.406
between	0.673	0.335
simple	0.584	0.328
one_hop	0.421	0.195
two_hop	0.352	0.181
three_hop	0.287	0.152

Table 2: Model accuracy by spatial relation type.

Spatial Relation	ReCLIP	LLaVA-13b
farthest	0.782	0.500
nearest	0.751	0.406
between	0.673	0.335
front	0.428	0.204
behind	0.412	0.165
right	0.401	0.207
left	0.398	0.181

generic terms such as “the thing” or “the object” when spatial relations alone uniquely identify the referent. This isolates spatial reasoning performance from the dependence on visual attributes.

Table 3: Performance on masked expressions.

Expression Type	ReCLIP	LLaVA-13b
Masked	0.212	0.396
Unmasked	0.523	0.191

As shown in Table 3, ReCLIP’s accuracy dropped from 0.523 to 0.212 on masked expressions, indicating heavy reliance on object descriptors for grounding. In contrast, LLaVA-13b’s accuracy increased from 0.191 to 0.396 when attributes were removed, suggesting improved reasoning when relying on spatial cues. This difference likely reflects both models’ reasoning strategies and the nature of the masked subset, which is dominated by distance (“nearest”, “farthest”) and “between” relations. These spatial expressions favor relational reasoning, rather than appearance-based matching, which could explain LLaVA’s relative improvement. Overall, LLaVA seems to benefit from a purer spatial reasoning signal, while ReCLIP performs best when attribute and spatial information are available in combination.

5.2.2. Non-Ordinal Subset Evaluation

A focused evaluation was conducted using a 4,931 sample subset of the CLEVR-3D-DeRef dataset consisting exclusively of non-ordinal referring expressions. Non-ordinal expressions include those that do not use rank-based modifiers such as “The second cube from the left”. By excluding ordinal references, the subset isolates spatial reasoning without the added complexity of positional ordering.

In this subset, GPT-4o achieved the highest overall precision (49. 8%), surpassing ReCLIP (39.1%) and LLaVA-13b (20.3%). Despite similar patterns of decline in accuracy with greater relational depth, GPT-4o remained the most consistent performer across all expression types.

Performance by Template Type. GPT-4o consistently outperformed both models across nearly all template types, particularly for between and multi-hop expressions. ReCLIP performed best on distance-based templates (“nearest”, “farthest”), surpassing GPT-4o in both categories, suggesting an advantage in spatial proximity reasoning. LLaVA-13b achieved lower but structurally consistent performance, mirroring the general pattern of degradation with increasing depth of reasoning.

Table 4: Template-wise accuracy.

Template	GPT-4o	ReCLIP	LLaVA-13b
between	0.585	0.133	0.272
farthest	0.235	0.329	0.127
nearest	0.227	0.279	0.131
1-hop (no ord.)	0.574	0.467	0.206
2-hop (no ord.)	0.544	0.508	0.221
3-hop (no ord.)	0.567	0.500	0.183

Performance by Spatial Relation. Across relation types, GPT-4o showed the strongest performance in binary spatial relations (front / behind, left / right), achieving over 55% accuracy, while ReCLIP had an advantage in distance-based reasoning. This is also demonstrated in Table 4 because the “between”, “nearest”, and “farthest” relations are isolated in individual templates.

Performance on Attribute-Masked Expressions. GPT-4o maintained the highest accuracy (34.4%) on attribute-masked expressions, suggesting its superiority in relying on spatial cues alone. Similarly to our findings in Section 5.2.1, the accuracy of ReCLIP decreased to 23.5%, demonstrating its dependence on explicit visual descriptors. LLaVA-13b achieved an accuracy of 19.6% on the masked portion of the data subset. LLaVA’s degradation in performance may be due to a change in the distribution of template types that were included in the

Table 5: Accuracy by spatial relation type.

Relation Type	GPT-4o	ReCLIP	LLaVA-13b
between	0.585	0.133	0.272
farthest	0.235	0.329	0.127
front/behind	0.553	0.509	0.191
left/right	0.564	0.482	0.214
nearest	0.227	0.279	0.131

data subset.

6. Limitations and Discussion

Although CLEVR-3D-DeRef extends CLEVR-Ref+ toward more robust 3D spatial reasoning, several limitations remain. Most notably, there is still a gap between the structure of the dataset and the models evaluated in this work. Each scene is paired with a 3D point cloud representation, but the current experiments only evaluate models that operate over 2D images or RGB-based inputs. As a result, the present study does not yet test whether access to explicit 3D structure improves referring expression comprehension, nor does it provide a direct comparison against models designed for 3D vision-language reasoning. The current evaluation therefore shows that the dataset is challenging for existing multimodal systems, but it does not fully measure the value of the 3D representation itself.

A second limitation is the breadth of the model comparison. While ReCLIP and LLaVA-13b were evaluated on the full dataset, GPT-4o was evaluated only on a smaller non-ordinal subset due to practical constraints. These results offer a useful comparison across model types, but they should not be interpreted as a complete account of state-of-the-art performance. A broader evaluation with additional contemporary models, especially systems built for 3D grounding or embodied reasoning, would better situate the benchmark.

There are also limitations in how some aspects of the methodology are assessed. The diversification pipeline relies on automatic measures of meaning preservation, linguistic diversity, and naturalness, which are useful for large-scale filtering and analysis but do not fully replace human judgment. In particular, naturalness and interpretability are difficult to capture with automatic metrics alone. For this reason, these measures should be understood as diagnostic signals rather than definitive indicators of expression quality. To improve reproducibility and transparency, we include detailed prompt templates, response formats, and post-processing steps in appendix sections A and B.

The benchmark itself is also still evolving. We are now extending the generation pipeline to support multi-hop expressions that combine *between*, *nearest*, and *farthest* with the original directional relations. These additions increase the compositional range of the dataset and better reflect the broader relational space we aim to study. Future work will also expand the validation and evaluation to these new expression types. We view the current version of CLEVR-3D-DeRef as a strong initial benchmark release, with additional extensions still underway.

7. Conclusions and Future Work

This paper introduced the CLEVR-3D-DeRef dataset, a synthetic and extensible benchmark designed to evaluate spatial reasoning in vision-language models. By incorporating depth information, expanding the set of spatial relations, diversifying linguistic expressions, and implementing a de-identification mechanism, the dataset enables systematic evaluation of spatial understanding beyond surface-level cues. Our experiments show that contemporary models, including ReCLIP, LLaVA-13b, and GPT-4o, exhibit sensitivity to spatial complexity and relational structure, with performance drops on multi-hop and attribute-masked expressions.

While CLEVR-3D-DeRef advances spatial referring expression comprehension in 3D, it currently focuses exclusively on object identification. As a next step, we plan to extend the benchmark to support action-oriented spatial reasoning through a paired scene rearrangement pipeline. This pipeline generates compositional instructions that both identify target objects and specify their intended spatial configuration (e.g., “Place the red cube to the left of the green sphere.”). This extension enables evaluation of both perception and manipulation reasoning within a unified framework. A critical step for future work will be to evaluate the transference of spatial reasoning in this synthetic dataset to a dataset comprised of real images.

8. Acknowledgements

This work was supported in part by computational resources provided by CU Research Computing at the University of Colorado Boulder, including the Blanca Condo Cluster.

9. Bibliographical References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020a. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision – ECCV 2020: 16th*

European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I, page 422–440, Berlin, Heidelberg. Springer-Verlag.

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. 2020b. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*.

Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, and Dong-Kyu Chae. 2025. MVTamperBench: Evaluating robustness of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18804–18828, Vienna, Austria. Association for Computational Linguistics.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. Weakly supervised relative spatial reasoning for visual question answering.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. 2020a. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, page 202–221, Berlin, Heidelberg. Springer-Verlag.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee Kenneth Wong, and Qi Wu. 2020b. Cops-ref: A new dataset and task on compositional referring expression comprehension. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10083–10092.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, Volume 2 - 2019.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4180–4189.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek

Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. 2025. Beyond semantics: Rediscovering spatial awareness in vision-language models.

Shikhar Singh, Ehsan Qasemi, and Muhao Chen. 2023. VIPHY: Probing "visible" physical commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7113–7128, Singapore. Association for Computational Linguistics.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. 2022. ReCLIP: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, Dublin, Ireland. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-structured referring expression reasoning in the wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9949–9958.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and

Wanglu Wanglu. 2025a. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11591–11609, Vienna, Austria. Association for Computational Linguistics.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025b. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*.

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Eric Wang. 2022. Vlmbench: a compositional benchmark for vision-and-language manipulation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

A. Model Prompting Details

This section describes the prompt formats and input representations used for each model in the referring expression comprehension evaluation. All models receive the same underlying referring expressions and candidate object bounding boxes; the formats differ to match each model's interface and to maximize each model's ability to demonstrate spatial reasoning.

A.1. ReCLIP (Zero-Shot Visual Grounding

Format

ReCLIP operates as a zero-shot visual grounding model and does not receive a natural language prompt. Instead, it takes a structured input consisting of:

- The raw referring expression text.
- The scene image.
- A set of candidate bounding boxes with integer identifiers.

```
{
  "image_id": 0,
  "file_name": "
     CLEVR_train_000000.png",
  "sentences": [{"raw": "<
     referring expression>"}],
  "anns": [
    {"id": 0, "bbox": [x, y, w, h
       ]},
    {"id": 1, "bbox": [x, y, w, h
       ]},
    ...
  ],
  "ann_id": [<
     ground_truth_object_id>]
}
```

Motivation

ReCLIP grounds referring expressions by computing CLIP-based similarity between the expression and image regions cropped to each candidate bounding box. No explicit prompt engineering is required or possible; the model directly scores each candidate region against the expression. This provides a baseline that reflects pure vision-language alignment without prompt-induced biases.

A.2. LLaVA-1.5-13b (Multi-Choice Selection)

Format

LLaVA receives the scene image and a multiple-choice text prompt that lists each candidate object's bounding box as a lettered option. The model is asked to select the option matching the referring expression.

```
I'm looking for: '<referring
   expression>'

I can see several objects in the
   image. Here are their
   positions:

Option A: An object at position (
   x0, y0) with dimensions w0 by
   h0
Option B: An object at position (
   x1, y1) with dimensions w1 by
   h1
Option C: An object at position (
   x2, y2) with dimensions w2 by
   h2
...

Which one matches what I'm
   looking for? Just tell me the
   letter.
```

The expected response here is a single capital letter (e.g., C).

Motivation

LLaVA is a generative vision-language model that produces free-form text. An open-ended prompt (e.g., “Where is X located?”) causes the model to generate verbose natural-language descriptions rather than selecting a specific object. The multiple-choice format constrains the output space to a single letter, enabling reliable automatic evaluation. Providing explicit bounding box coordinates as options tests whether the model can jointly reason about spatial language and numerical position information in the image.

A.3. GPT-4o (Visual Selection with Annotated Image)

Format

GPT-4o receives the scene image with bounding boxes drawn and labeled with letters (A, B, C, ...) directly on the image. A system prompt establishes the task context, and a user prompt asks which labeled object matches the referring expression.

Listing 1: System prompt

```
You are an assistant helping with
    a computer vision research
    study on spatial
reasoning. You will be shown a
    synthetic 3D scene image with
    objects labeled
A, B, C, etc. and asked which
    labeled object best matches a
    referring expression.
Respond with ONLY the single
    capital letter of the correct
    object - no explanation.
```

Listing 2: User prompt

```
Referring expression: "The cube
    that is left of the blue
    sphere"

The objects in the scene are
    labeled A through F.
Which single letter labels the
    object described by the
    expression above?
```

The image sent to the model has red bounding boxes with white letter labels drawn at the top-left corner of each object.

Motivation

GPT-4o supports high-resolution image inputs with native visual understanding. Rather than encoding bounding box coordinates as text (as with LLaVA), the labels are rendered directly on the image. This leverages GPT-4o’s strong visual perception to associate spatial language with visually marked regions, avoiding any confound from the model’s ability to parse numerical coordinates. The system prompt reduces refusals by framing the task as an academic research study, and requesting only a single letter minimizes response parsing ambiguity.

A.4. Summary of Differences

Table 6: Comparison of model input and output formats.

Aspect	ReCLIP	LLaVA-1.5-13B	GPT-4o
Input modality	Image + text + bboxes	Image + text prompt	Annotated image + text prompt
Prompt style	None (structured API)	Multiple-choice (text coords)	Visual labels on image
Candidate repr.	Bbox coordinates (JSON)	Bbox coordinates (text)	Drawn on image
Output	Bbox ID (automatic)	Single letter	Single letter
Coord. reasoning	Implicit (crop + CLIP)	Explicit (text)	Implicit (visual)

B. LLM Diversification Prompting

Template-based referring expression generation produces syntactically rigid outputs (e.g., “The red cube that is left of the blue sphere”). To increase linguistic variety while preserving referential accuracy, each template expression is passed through an LLM that generates natural, conversational paraphrases. The LLM receives the original expression together with a summary of the scene context and returns up to N variations per expression.

B.1. Scene Context Construction

Before prompting, a scene summary is constructed from the scene’s object list. Each object is described by its material, color, and shape. At most five objects are listed explicitly; any remaining objects are summarized with a count.

Format

```
Scene contains: <material> <color
    > <shape>, <material> <color>
    <shape>, ... and <k> other
    objects
```

Example

```
Scene contains: metal red cube,
    rubber blue sphere, metal
    green cylinder, rubber yellow
    cube, metal purple sphere and
    3 other objects
```

B.2. System Prompt

The system message is kept brief to avoid over-constraining the model's style:

```
You are an expert at creating
    natural language variations
    while preserving meaning.
```

B.3. User Prompt (Diversification Prompt)

The user prompt provides the scene context, original expression, task instructions, formatting requirements, and examples.

Prompt Template

```
You are an expert at creating
    natural, conversational
    variations of referring
expressions for visual scenes.

SCENE CONTEXT: <scene_summary>

ORIGINAL EXPRESSION: "<referring
    expression>"

TASK: Create <N> natural,
    conversational variations of
    this expression that:
1. Refer to the EXACT SAME object
    (s) as the original
2. Maintain the same spatial
    relationships and meaning
3. Use different sentence
    structures and word choices
4. Sound natural and
    conversational
5. Are grammatically correct

REQUIREMENTS:
- Keep all spatial relations
    accurate (left, right, front,
    behind, between,
   nearest, farthest)
- Preserve object attributes (
    color, shape, material) but
    vary how you
   describe them
- Use different sentence
    structures (active/passive,
    different clause orders)
- Vary word choice while keeping
    meaning identical
- Make expressions sound like
    natural human speech

OUTPUT FORMAT: Return only the
    variations, one per line,
    without numbering
or explanations.

EXAMPLES:
Original: "The red cube that is
    left of the blue sphere"
Variations:
- "That red cube sitting to the
    left of the blue sphere"
- "The cube on the left side of
    the blue sphere"
- "Red cube positioned left of
    the blue sphere"
- "The cube that's left of the
    blue sphere"

Now create variations for the
    given expression:
```

Expected Response

One variation should be returned per line, without numbering or bullet markers, for example:

```
That red cube sitting to the left
     of the blue sphere
The cube on the left side of the
    blue sphere
Red cube positioned left of the
    blue sphere
The cube that's left of the blue
    sphere
```

B.4. Post-Processing and Filtering

Post-Processing Filters

After the LLM returns its response, we apply the following filters before a variation is accepted:

1. **Minimum length** — Variations shorter than 10 characters are discarded.
2. **Identity check** — Variations that are identical to the original expression, ignoring case, are discarded.
3. **Artifact removal** — Lines beginning with `Original:` or `Variation` are discarded to remove residual prompt formatting.
4. **Truncation** — At most `max_variations` variations are retained for each expression.