Learning Unified Probabilistic Spatial Relation Representation from Visual Demonstrations

Paul Emil Hannuschka and Jianfeng Gao and Tamim Asfour

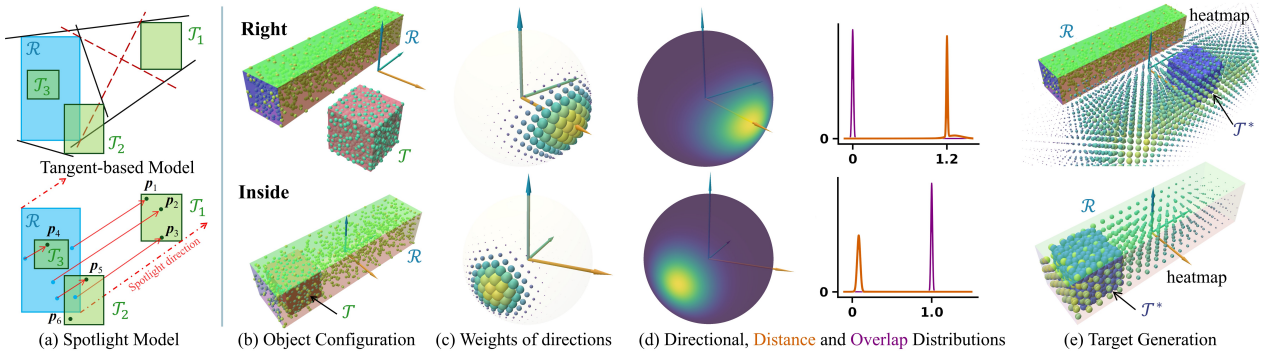

Fig. 1: Spotlight-based spatial relation representation. (a) Unlike tangent-based models, which fail when target objects ($\mathcal{T}_i, i = 1, 2, 3$) overlap the reference object ($\mathcal{R}$), our spotlight model resolves this by sweeping light across all angles to identify illuminated target points (e. g., $\boldsymbol{p}_1$–$\boldsymbol{p}_5$ are lit; $\boldsymbol{p}_6$ is unlit). (b)-(e) Examples of *right* and *inside* (more precisely, *inside front*) spatial relation representations learned from a single demonstration of object configuration. For each, the spotlight model computes directional weights (c), which are used to learn the directional distribution, alongside distance and overlap distributions (d). At generation (e), the directional likelihood heatmap (point size indicates directional suitability) is filtered by distance and overlap distributions to sample the target position ($\mathcal{T}^*$).

***Abstract*— The ability to interpret and reason about spatial relations is fundamental for robotic manipulation tasks. For instance, a robot must understand that "inside" requires different geometric constraints than "touching", and "closer" involves dynamic changes in distance relationships. Despite progress in modeling spatial relations, existing approaches face two critical limitations: they either oversimplify object geometry to points or bounding boxes, or they lack generative capabilities for synthesizing new spatial configurations. This paper introduces a novel generative and probabilistic model that jointly encodes object sizes, distances, and orientations within a unified representation, which captures distance-based, directional, and topological spatial relations while providing explicit uncertainty quantification. The model learns both static and dynamic semantic spatial relations from one or a few visual demonstrations and generalizes to novel contexts and configurations. We evaluate our approach across a set of spatial reasoning and robot manipulation tasks, demonstrating the model's robust performance with varied object shapes, sizes, and spatial arrangements. Videos and source code are available at https://sites.google.com/view/spatial-relations.**

I. INTRODUCTION

Consider a robot learning to organize a kitchen: it observes a human placing a cup "to the right of" a plate, then must generalize this spatial relation to place a bowl "to the right of" a different plate of a different size and shape. This seemingly simple task requires understanding how object geometry affects spatial relations – something that varies dramatically between a thin plate and a wide basket, or between placing objects "inside" versus "touching". Understanding and generating spatial relations is core to robotic manipulation, enabling tasks like object rearrangement [1], scene interpretation [2], task planning, and human-robot interaction [3]. Spatial relations serve as a bridge between subsymbolic geometric representations and symbolic semantic reasoning. This allows robots to interpret spatial object configurations, e. g., a tool being *inside* a drawer or a cup *to the right of* a plate, and execute metric displacements, such as moving an object $10\,\mathrm{cm}$ *more to the right*.

Following established taxonomies [4], [5], spatial relations are typically divided into three categories: 1) *Directional relations*, encoding relative orientation (e. g., left of, above); 2) *Distance-based relations*, quantifying proximity (e. g., close to, far from); and 3) *Topological relations*, characterizing connectivity and containment (e. g., inside, touching, between).

Accurate interpretation, discrimination, and generation of these relations are fundamental requirements for autonomous robotic manipulation. Despite extensive research, existing representations face two critical limitations. First, they typically lack generative capabilities and provide no mechanisms for quantitative comparison [6]. Second, they frequently oversimplify object geometry [3], [7], discarding essential properties like shape, size, and orientation that are essential for practical applications. Furthermore, while recent Vision-Language Models (VLMs) [1], [8]–[12] show promise in spatial reasoning, they remain highly data-hungry and struggle with the precise metric and subsymbolic generation required for 3D manipulation.

To address these limitations, we propose a probabilis-

The research leading to these results has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the SFB 1574 and the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. E-mails: paul.hannuschka@student.kit.edu, {jianfeng.gao, asfour}@kit.edu

tic representation of spatial relations learned from a few demonstrations. Our key insight is to model spatial relations using a "spotlight metaphor" (see Fig. 1.a-*bottom*): imagine the reference object as a rotating spotlight that illuminates the target object. By analyzing how much of the target is "illuminated" in each direction while considering geometric properties such as scale, orientation, relative distance, and overlap, we create a unified representation that simultaneously captures all three types of spatial relations.

The contributions of this paper are threefold: 1) A unified probabilistic spatial relation representation based on the spotlight model that captures distance-based, directional, and topological relations via three complementary distributions. Specifically, directional relations are modeled with Kernel Density Estimation (KDE) on the Riemannian manifold $\mathcal{S}^2$, distance relations with a Gaussian Mixture Model (GMM) in $\mathbb{R}_+$, and overlap relations with a univariate Gaussian in $[0, 1]$. 2) A systematic formalization of semantic spatial relations (e. g., *left*, *other side*, *closer*, *inside*) within this probabilistic framework. 3) A framework that learns spatial relation representations from only one or a few demonstrations, enabling both spatial reasoning and generative object manipulation. We validate our approach on generative manipulation tasks (robot pick-and-place) and discriminative reasoning tasks (spatial relation classification), showing superior performance compared to existing geometric models and competitive results with large-scale VLMs (e. g., GPT-5.2 [13] and Gemini-3-Pro [14]), while requiring only a few demonstrations for learning. It supports both qualitative semantic reasoning and quantitative geometric analysis and generation, thereby bridging symbolic and subsymbolic perspectives.

II. Related Works

Research on spatial relations spans both geometric models and data-driven approaches. We first review projective and metric models that emphasize geometric structure, followed by VLMs that focus on semantic reasoning.

A. Projective and Metric Spatial Relation

A central challenge in modeling projective and metric spatial relations lies in accurately capturing how object geometry, including size, shape, and orientation, affects the acceptance regions of spatial relations. This challenge has proven surprisingly difficult to address effectively. Many prior approaches oversimplify objects to single points [3], [7], [15] or bounding boxes [16]–[20], thereby discarding crucial geometric properties that are fundamental to human-like spatial cognition [21]. Bounding-box normalization methods [7] attempt to mitigate this issue but remain highly sensitive to object dimensions, often producing counterintuitive or infeasible spatial configurations for robotic manipulation.

To capture directional and metric relations more explicitly, fuzzy descriptor models have been combined with angular analysis [22], [23] or force histograms [24], [25]. Statistical models [26] extend these concepts by analyzing relative directions between arbitrary object points, though they primarily rely on central tendencies or dispersion measures rather than full distributional information. Despite these advances, existing approaches face several fundamental limitations. First, they fail to capture uncertainty or provide mechanisms for quantifying similarity between spatial relations, preventing generative reasoning [6]. Second, most metric models are restricted to cardinal direction relations in 2D domains, such as geographic data or image analysis, limiting their applicability to 3D robotics scenarios. Third, methods requiring calculation of external and internal tangent lines [16], [22] encounter difficulties when extended to 3D environments or overlapping configurations (see Fig. 1.a-*top*), where topological relations become essential. Building on these insights, our approach explicitly integrates object geometry within a unified probabilistic framework that addresses these limitations by supporting uncertainty quantification, similarity measurements, and generative reasoning.

B. Vision-Language Models for Spatial Reasoning

Recent vision-language models (VLMs) [1], [8]–[14] demonstrate promising spatial reasoning capabilities, yet they face limitations that restrict their practical applicability in dynamic robotic environments. While human infants acquire complex spatial relations such as containment from only a few demonstrations [27], VLMs require massive datasets and still fall short in generalization to novel scenarios [11], [28]. Furthermore, Current VLMs excel at semantic reasoning with discrete symbolic categories (e. g., *left*, *right*, *above*), but exhibit significant weaknesses regarding metric reasoning [9], part-level relations [12], and quantitative similarity measures between spatial configurations [6]. Their generative variants [1], [11]–[14] remain constrained to 2D image synthesis, lacking the capability to generate feasible 3D object configurations essential for robotic manipulation. Moreover, they fail to capture continuous subsymbolic transitions (e. g., "right" to "bottom-right"), limiting their utility in relation similarity estimation. Our approach addresses these gaps by learning spatial relations from a few demonstrations at both symbolic and subsymbolic levels, providing 3D generative reasoning capabilities and smoothly evolving probabilistic densities for interpretable decision-making.

III. Approach

We present a probabilistic framework for learning spatial relations from visual demonstrations. A spatial relation ρ of a *target* object $\mathcal{T}$ relative to a set of *reference* objects $\mathcal{R}$ is defined as a triplet $(\mathcal{R}, \mathcal{T}, \rho)$. For example, (plate, cup, right) describes "the cup to the right of the plate", Non-binary relations are expressed with multiple reference objects, e. g., ($\{$plate, apple$\}$, cup, between) for "the cup between a plate and an apple".

Spatial relations are represented at two levels. At the *subsymbolic* level, we encode continuous spatial object configurations that capture the full geometric richness of spatial relationships. At the *symbolic* level, we categorize these relationships into two discrete semantic categories following [3]: 1) *static relations*, determined by fixed object configurations (e. g., left, above, touching, close), and 2) *dynamic relations*, defined

by configurations before and after executing an action (e. g., closer, farther, other side). To bridge these representational levels, we introduce a novel *spotlight model* (Section III-A) that generates weighted samples, forming the basis for learning probabilistic subsymbolic distributions (Section III-B). These can be trained from demonstrations (Section III-C) and later applied to discriminative semantic spatial relation tasks for recognizing and classifying spatial relationships (Section III-D) and generative tasks for synthesizing new spatial configurations (Section III-E).

A. Spotlight Model

The spotlight model conceptualizes the reference object $\mathcal{R}$ as the source of a rotating spotlight that systematically illuminates the surrounding space (see Fig. 1.a-*bottom*). Imagine pointing a flashlight from every point on a plate toward a cup – some directions will illuminate more of the cup than others. Directions that align well with the cup's position relative to the plate receive higher weights, while directions pointing away receive lower weights. Technically, we represent objects as sets of points sampled from their volume, preserving geometric properties rather than reducing them to oversimplified abstractions. For each reference point, we cast rays in all directions and analyze how many points of the target object $\mathcal{T}$ are illuminated in each direction, while incorporating geometric properties such as shape, scale, distance, and orientation. For objects with containment properties, such as cups or bowls, we extend this by sampling candidate points from both solid geometry and containment volume to capture topological relations like *inside*.

For each reference point $\mathbf{r}_j \in \mathcal{R}$ on the reference object, we cast rays uniformly across all directions $\mathbf{x}_i \in \mathcal{S}^2$ on the unit sphere. Target points $\mathbf{t}_k \in \mathcal{T}$ that lie along each ray direction $\mathbf{x}_i$ contribute to the spatial representation with weights determined by three factors: 1) Euclidean distance between reference and target points $\|\mathbf{t}_k - \mathbf{r}_j\|$, 2) angular deviation of target points from ray direction $\mathbf{x}_i$, and 3) degree of overlap between the objects. By integrating these three geometric factors, the spotlight model yields a unified representation that simultaneously encodes directional and distance-based relations, spatial overlap, and geometric properties of both objects. This encoding forms the foundation for learning spatial relations that generalize across diverse object configurations. Unlike tangent-based methods [16], [22], [24] that restrict analysis to boundary tangents between object silhouettes, our model considers all pairwise connections between reference and target points and thus captures the full geometric richness of spatial relationships, including internal structure and volumetric properties that boundary-only methods necessarily discard (see Fig. 1.a).

The process consists of three steps: computing directional vectors from each reference point to all target points, aggregating weighted contributions, and estimating minimum distances between the reference and target objects.

1) Computing Directional Vectors: The spotlight model is computed in the following manner, beginning with the computation of directional vectors. Given an object configuration $(\mathcal{R}, \mathcal{T})$ with reference points $\mathcal{R} = \{\mathbf{r}_i\}_1^n$ and target points $\mathcal{T} = \{\mathbf{t}_i\}_1^n$, we compute the spotlight model as follows: First, we quantify *topological relations* by identifying reference and target points contained within the volume $\mathbb{V}(\cdot)$ of the other object: $\mathcal{R}_o = \{\mathbf{r}_i \mid \mathbf{r}_i \in \mathbb{V}(\mathcal{T})\}$, $\mathcal{T}_o = \{\mathbf{t}_i \mid \mathbf{t}_i \in \mathbb{V}(\mathcal{R})\}$. The overlap ratios are $o_{\text{ref}} = |\mathcal{R}_\text{o}|/|\mathcal{R}|$, $o_{\text{tgt}} = |\mathcal{T}_\text{o}|/|\mathcal{T}|$. To avoid noisy directions caused by neighboring overlapping points, we exclude the overlapping point sets: $\mathcal{R} \leftarrow \mathcal{R} \backslash \mathcal{R}_\text{o}$ if $o_{\text{tgt}} > o_{\text{ref}}$, otherwise $\mathcal{T} \leftarrow \mathcal{T} \setminus \mathcal{T}_\text{o}$. We then compute displacement vectors $\mathbf{d}_{jk} = \mathbf{t}_k - \mathbf{r}_j$, representing directions from each reference to each target point and normalize them for each point pair $(\mathbf{r}_j, \mathbf{t}_k)$ by $\hat{\mathbf{d}}_{jk} = \mathbf{d}_{jk}/d_{jk}$ for $d_{jk} = \|\mathbf{d}_{jk}\| > 0$.

In the presence of noise and outliers, the computed directional vectors $\hat{\mathbf{d}}_{jk}$ can be unreliable. Since individual directional vectors can be noisy, we compute the Fréchet mean of directional vectors over K nearest neighbors to obtain more robust directional estimates. For reference points, $\bar{\mathbf{d}}_j^{\text{ref}} = \text{FrMean}(\{\hat{\mathbf{d}}_{jn} \mid n \in \mathcal{N}_j^{\text{tgt}}\})$, and for target points $\bar{\mathbf{d}}_k^{\text{tgt}} = \text{FrMean}(\{\hat{\mathbf{d}}_{nk} \mid n \in \mathcal{N}_k^{\text{ref}}\})$, where $\mathcal{N}_j^{\text{tgt}}$ and $\mathcal{N}_k^{\text{ref}}$ are the neighbor indices on the target and reference objects, respectively.

2) Aggregating Weights: A directional vector $\hat{\mathbf{d}}_{jk}$ is more reliable if it aligns well with its corresponding mean directional vectors $\bar{\mathbf{d}}_j^{\text{ref}}$ and $\bar{\mathbf{d}}_k^{\text{tgt}}$. We compute alignment scores as the dot product between each directional vector and its corresponding mean directions $a_{jk}^{\text{ref}} = \hat{\mathbf{d}}_{jk} \cdot \bar{\mathbf{d}}_j^{\text{ref}}, a_{jk}^{\text{tgt}} = \hat{\mathbf{d}}_{jk} \cdot \bar{\mathbf{d}}_k^{\text{tgt}}$. To robustly distinguish well-aligned from noisy directional vectors, we transform the raw alignment scores using a nonlinear mapping that strongly rewards alignment and gradually penalizes deviations. This mapping, $\xi(a; \alpha_{\text{init}}, s, e)$, depends on an initial scaling factor α_{init}, an exponent e to scale the input, and a slope parameter s controlling steepness:

$$\xi(a; \alpha_{\text{init}}, s, e) = \frac{-\tanh(s \cdot (a^e - 0.5)) + \tanh(0.5 \cdot s)}{2 \cdot \tanh(0.5 \cdot s)} \cdot \alpha_{\text{init}} \quad (1)$$

The final alignment score a_{jk} is the product of the transformed reference and target terms with initial parameters $\alpha_\text{a}, s_\text{a}, e_\text{a}$:

$$a_{jk} = \xi(a_{jk}^{\text{ref}}; \alpha_\text{a}, s_\text{a}, e_\text{a}) \cdot \xi(a_{jk}^{\text{tgt}}; \alpha_\text{a}, s_\text{a}, e_\text{a}) \quad (2)$$

Spatial perception is strongly influenced by proximity, i. e., closer object parts typically dominate in relational reasoning. Thus, we penalize distant point pairs with a distance-dependent weight. However, in cases of strong overlap, distance becomes less informative; accordingly, we attenuate this penalty with an overlap-dependent factor $\lambda = \xi\left(\max\left(o_{\text{ref}}, o_{\text{tgt}}\right); \alpha_\text{o}, s_\text{o}, e_\text{o}\right)$. The final weight $\hat{w}_{jk}$ for each directional vector incorporates both alignment and distance penalties:

$$\hat{w}_{jk} = \exp\left(-(a_{jk} + \lambda) \cdot \frac{d_{jk} - \min_{j,k}\left(d_{jk}\right)}{d_{\text{scale}}}\right) \quad (3)$$

where d_{scale} denotes the minimum object dimension, ensuring scale-invariant transfer across objects of different sizes.

To approximate directional distributions, we uniformly sample N directions $\{\mathbf{x}_i\}_{i=1}^N$ on the unit sphere $\mathcal{S}^2$ via the Fibonacci lattice [29]:

$$\mathbf{x}_i = \left(\sqrt{1 - z_i^2}\cos(\theta_i), \sqrt{1 - z_i^2}\sin(\theta_i), z_i\right) \quad (4)$$

where $z_i = 1 - 2i/N$, $\theta_i = 2\pi(i-1)/\phi$, and $\phi = (1+\sqrt{5})/2$ is the golden ratio. A sampled direction $\mathbf{x}_i$ is considered aligned with $\hat{\mathbf{d}}_{jk}$ if their dot product exceeds a threshold τ_a, with alignment expressed by $m_{ijk}^{\text{align}} = \mathbb{1}(\mathbf{x}_i \cdot \hat{\mathbf{d}}_{jk} > \tau_\text{a})$. The weight w_i for each sampled direction $\mathbf{x}_i$ aggregates aligned contributions, and is min-max normalized for consistency across demonstrations (see Fig. 1.c for a visualization example):

$$w_i = \frac{\tilde{w}_i - \min_i(\tilde{w}_i)}{\max_i(\tilde{w}_i) - \min_i(\tilde{w}_i) + \epsilon}, \quad \tilde{w}_i = \textstyle\sum_{j,k} \hat{w}_{jk} \cdot m_{ijk}^{\text{align}}. \tag{5}$$

3) Computing Distances: To capture distance information, we compute the closest distance from the reference object to the target object along each sampled direction. For each direction $\mathbf{x}_i$, we take the minimum distance $\bar{d}_i$ over aligned directions, assigning infinity when no alignment exists. To match directional information in the distance distribution, the distances $\hat{d}_i$ are resampled according to the weights w_i:

$$\hat{d}_i \sim \textstyle\sum_{k=1}^{N} \frac{w_k}{\sum_{\ell=1}^{N} w_\ell} \delta_{\bar{d}_k}, \quad \bar{d}_i = \min\left(\left\{ d_{jk} \,\middle|\, m_{ijk}^{\text{align}} = 1 \right\} \cup \{\infty\}\right),$$

where $\delta_{\bar{d}_k}$ denotes a Dirac delta centerred at $\bar{d}_k$.

In summary, the spotlight model produces a compact representation $\mathcal{S}(\mathcal{R}, \mathcal{T}) = \left(\left\{\mathbf{x}_i, w_i, \hat{d}_i\right\}_{i=1}^{N}, o_\text{tgt}\right)$, capturing directional preferences, distance profiles, and topological overlap. This forms the foundation for our probabilistic relationship model.

B. Unified Probabilistic Representation

We now model spatial relations ρ as probabilistic representations that support both discriminative tasks (deciding whether an object configuration satisfies ρ) and generative tasks (synthesizing object configurations consistent with ρ).

At the subsymbolic level, each relation is described by three complementary distributions: 1) a *directional distribution* p_dir on the unit sphere $\mathcal{S}^2$, encoding relative directions from $\mathcal{R}$ to $\mathcal{T}$; 2) a *distance distribution* p_dist on $\mathbb{R}_+$, capturing relative distances; and 3) an *overlap distribution* p_o on $[0,1]$, quantifying the fraction of $\mathcal{T}$ contained within $\mathcal{R}$.

The *directional distribution* p_dir is estimated with Kernel Density Estimation (KDE) using the von Mises-Fisher kernel [30], $p_\text{dir}(\mathbf{d}) = \sum_{i=1}^{N} w_i \cdot \text{vMF}(\mathbf{d}; \mathbf{x}_i, \kappa)$, where $\mathbf{x}_i$ are sample directions, w_i are weights, and κ is the concentration parameter. Unlike parametric spherical GMMs [31], [32], KDE accommodates arbitrarily shaped distributions, essential for relations such as *inside* or *touching*, where feasible directions can span nearly the entire domain.

The *distance distribution* p_dist employs a two-component GMM, capturing diverse distance information without premature averaging, $p_\text{dist}(d) = \sum_{k=1}^{2} \pi_k \cdot \mathcal{N}(d; \mu_k, \sigma_k^2)$.

Finally, the *overlap distribution* $p_\text{o}(o) = \mathcal{N}(o; \mu_\text{o}, \sigma_\text{o}^2)$ models topological containment as the fraction $o \in [0,1]$ of target points within the reference volume. This distribution helps disambiguate relations with similar directional and distance profiles (e. g., *inside* versus *among*), thereby providing crucial topological discrimination.

C. Learning Spatial Relations from Demonstrations

Given a set of D demonstrations $\mathbf{D} = \left\{(\mathcal{R}_j, \mathcal{T}_j)\right\}_{j=1}^{D}$ fulfilling spatial relation ρ, we first apply the spotlight model to generate weighted spatial data (Section III-C.1) and then aggregate this data to estimate probabilistic distributions that capture the underlying spatial relationship (Section III-C.2).

1) Data Generation: To ensure consistent spatial analysis across all demonstrations, we establish a reference frame with the x-axis pointing right, the z-axis pointing up, and the origin at the reference centroid. This enables meaningful comparison and aggregation of spatial relationship patterns across different demonstrations. Following [3], our data generation procedure distinguishes between *static* and *dynamic* relations. For *static relations* that depend solely on fixed object configurations, we apply the spotlight model to each demonstration. The model computes directional weights, distance distributions, and alignment scores based on the instantaneous spatial arrangement between the reference and target objects. See examples in Fig. 1.b-e.

$$\mathcal{S}(\mathcal{R}_j, \mathcal{T}_j) = \left(\left\{\mathbf{x}_i^j, w_i^j, \hat{d}_i^j\right\}_{j=1}^{N}, o_\text{tgt}^j\right). \tag{6}$$

For *dynamic relations*, the spotlight models are obtained in the same way as Eq. (6) using initial and final configurations, respectively. As we model the final configuration relative to the initial configuration, the dominant initial direction and distance are extracted according to the maximum weight:

$$i_j^* = \arg\max_i w_i^j, \quad \mathbf{x}_j^* = \mathbf{x}_{i_j^*}^j, \quad d_j^* = \hat{d}_{i_j^*}^j \tag{7}$$

Final configurations are rotated such that the x-axis of the final configuration aligns with $\mathbf{x}_j^*$ by $\mathcal{R}_j \leftarrow \mathbf{M}_j \mathcal{R}_j$ and $\mathcal{T}_j \leftarrow \mathbf{M}_j \mathcal{T}_j$. Distances are normalized relative to the initial distance, $\hat{d}_i^j \leftarrow \hat{d}_i^j / (d_j^* + \epsilon)$. Here, $\hat{d}_i^j = 1.0$ corresponds to the initial distance, values > 1.0 indicate farther, and < 1.0 closer.

2) Estimation of Probabilistic Distributions: Since directions $\mathbf{x}_i^j$ align across demonstrations, we aggregate their associated weights by summation. In contrast, distance and overlap measurements are therefore concatenated to preserve the full distributional information from each demonstration.

$$\hat{\mathbf{x}}_i = \mathbf{x}_i^1, \;\; \hat{w}_i = \textstyle\sum_j w_i^j, \;\; \hat{\mathbf{D}} = \bigcup_{i,j} \hat{d}_i^j, \;\; \hat{\mathbf{O}} = \bigcup_j o_\text{tgt}^j$$

The probabilistic distributions are then estimated by:

$$p_\text{dir}(\mathbf{d}) \leftarrow \text{KDE}\left(\hat{\mathbf{G}}\right), \quad p_\text{dist}(d) \leftarrow \text{EM}(\hat{\mathbf{D}}), \quad p_\text{o}(o) \leftarrow \text{MLE}(\hat{\mathbf{O}}),$$

where $\hat{\mathbf{G}} = \{(\hat{\mathbf{x}}_i, \hat{w}_i)\}_{i=1}^{N}$.

D. Discriminative Model Capabilities

For classification tasks, we need to determine whether a configuration $(\mathcal{R}, \mathcal{T})$ satisfies a relation ρ. For example, a robot may need to decide whether the keys are *on* the table or the chair is *to the right* of the desk.

Since probabilistic densities lack universal scaling, we use percentile-based thresholds. Following [33], we sample t points $\{s_i\}_{i=1}^{t}$ with log-likelihoods $\{l_i\}_{i=1}^{t}$ where $l_i = \log p(s_i)$ from each distribution and set the threshold τ at the q-th percentile $\tau = \text{quantile}(\{l_i\}_{i=1}^{t}, q)$. This yields thresholds

$\tau_{\text{dir}}, \tau_{\text{dist}}, \tau_{\text{o}}$ for direction, distance, and overlap distributions, respectively. Note that q directly influences the acceptance region of a given spatial relation and thus could depend on personal preference. Examples of the acceptance region defined by p_{dir} are shown in Fig. 3.

For a new configuration, we compute its spotlight model $\mathcal{S}(\mathcal{R}, \mathcal{T})$. The distance or overlap relation holds if $\log p_{\text{dist}}(\min_i \hat{d}_i) > \tau_{\text{dist}}$ or $\log p_{\text{o}}(o_{\text{tgt}}) > \tau_{\text{o}}$. For directional relations, we sum up the spotlight weights of directions above a threshold, $w_{\text{T}} = \sum \mathbf{W}_{\text{T}}$, $\mathbf{W}_{\text{T}} = \{w_j \mid \log p_{\text{dir}}(d_j) > \tau_{\text{dir}}\}$. A directional spatial relation is considered satisfied if the aggregated weights exceed a threshold $w_{\text{T}} > \tau_{\text{w}} \cdot (|\mathbf{W}_{\text{T}}|/N)$, where τ_{w} is a scaling parameter.

Different relation types emphasize different distributions, i. e., directional relations depend on direction and overlap, while distance relations depend only on distance. In scenarios with limited training data (one- or few-shot learning), this semantic mapping between relation types and their distributions can be provided by human experts or extracted from LLMs. However, as the number of demonstrations increases, the model learns to de-emphasize irrelevant acceptance regions. Our model supports a broad range of semantic spatial relations, from basic categories (e. g., *inside*, *right*) to their fine-grained variants and combinations (e. g., *inside right*, *inside left*).

E. Generative Model Capabilities

A key advantage of our probabilistic representation is its ability to generate object placements that satisfy spatial relations. Given an initial object configuration $(\mathcal{R}, \mathcal{T})$ with reference object $\mathcal{R}$, target object $\mathcal{T}$ and a spatial relation represented as $(p_{\text{dir}}, p_{\text{dist}}, p_{\text{o}})$, we want to find a placement position $\hat{\mathbf{p}}$ such that both objects satisfy the relation when target object is moved to the position $\mathbf{p}_i$. To achieve this, we generate a set of candidate placement locations $\mathbf{P}$ and evaluate their suitability using our spotlight model. We sample a uniform set of placement locations $\mathbf{P} = \{\mathbf{p}_i\}_{i=0}^{N_P}$. The sampling strategy depends on the spatial relation type: For *containing relations* (e. g., *in*), the points are only sampled within the volumetric bounds of the reference object $\mathcal{R}$. For *non-containing relations*, points are sampled outside the boundary of $\mathcal{R}$ at distances drawn from the learned distance distribution p_{dist}. For dynamic relations, the distances are scaled by the initial minimum distance between the reference and the target object. When an explicit distance is requested (e. g., "place it 10 cm to the right"), the probabilistic distance sampling is replaced with the specified value.

For each candidate placement location $\mathbf{p}_i \in \mathbf{P}$, we compute a spotlight model using $\mathcal{R}$ as the reference and $\mathbf{p}_i$ as the target. The computation omits overlap-related steps and the uniform direction alignment following equation (4), yielding $|\mathcal{R}|$ weighted direction vectors $\{(\hat{\mathbf{d}}_{ij}, \hat{w}_{ij})\}_{j=1}^{|\mathcal{R}|}$ for each candidate point $\mathbf{p}_i$. For dynamic relations, we apply the inverse transformation described in Section III-C.1 based on the current configuration $(\mathcal{R}, \mathcal{T})$ to the direction vectors $\hat{\mathbf{d}}_{ij}$ to align the direction vectors with the current configuration.

Each placement location $\mathbf{p}_i$ receives a suitability score s_i that quantifies how well it satisfies the target spatial relation's directional distribution p_{dir}. For most spatial relations, suitability can be computed as the summed likelihood of its direction vectors: $s_i = \sum_j p_{\text{dir}}(\mathbf{d}_{ij})$. However, for relations like *inside center* or *between*, the overall shape of the distribution matters more than individual modes. To account for this, we estimate a kernel density estimate p_i from the weighted directions $\{(\hat{\mathbf{d}}_{ij}, \hat{w}_{ij})\}_j^{|\mathcal{R}|}$ and compute the suitability score as the Hellinger distance [34] $s_i = H(p_{\text{dir}}, p_i)$. Fig. 1.e shows exemplary placement locations and their suitability scores.

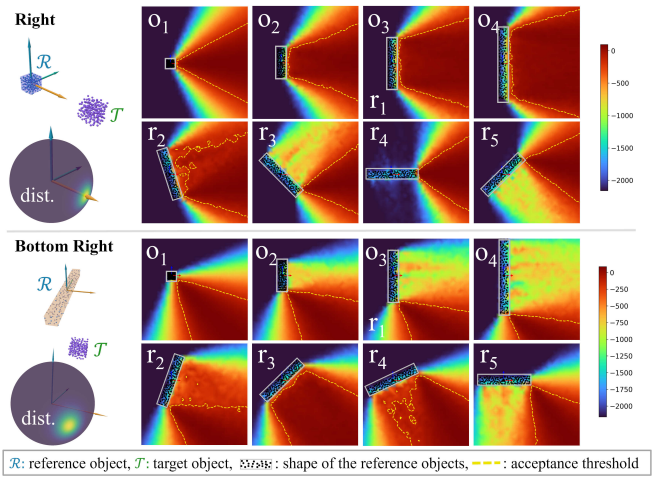

Fig. 3: Geometric properties of the *right* (top) and *bottom right* (bottom) relations. **Left**: the demonstration and resulting directional distribution. **Right**: The log-likelihood heatmap, generated by evaluating the learned directional distribution on direction vectors extracted via the spotlight model from points around the reference object, shown with the proposed acceptance threshold of Section III-D. For each relation, we varied the size of the reference object (o_1–o_4) and the rotation of o_3 (r_1–r_5).

We filter placement locations by retaining only the top 50 % with scores $s_i > 1/4\pi$. From these candidates, we draw c samples according to their squared suitability scores s_i^2, emphasizing high-quality locations. The point with the highest score is selected as the primary candidate. To mitigate the effects of noisy data and outliers, we compute the weighted average of this point and its n nearest neighbors to obtain the final placement candidate $\hat{\mathbf{p}}$.

To realize the placement, we identify the contact point $\mathbf{p}_{\mathcal{T}}$ on the target object $\mathcal{T}$ that aligns with $\hat{\mathbf{p}}$. Let $\hat{\mathbf{d}}_{\text{max}}$ be the maximum-weight direction vector from $\hat{\mathbf{p}}$. Since p_{dist} is the minimum inter-object distance, $\mathbf{p}_{\mathcal{T}}$ is the point on the target that is closest to the reference object along $\hat{\mathbf{d}}_{\text{max}}$. To account for irregular object geometries, directional uncertainties, and outliers, we compute $\mathbf{p}_{\mathcal{T}}$ as the centroid of the t closest points on the target object to this optimal location.

IV. Evaluation

We first illustrate in Section IV-A how the probability density fields of the learned directional relation smoothly evolve according to changes in object geometric properties. We then evaluate our approach in both generative (Section IV-B) and discriminative (Section IV-C) settings, demonstrating its effectiveness for object placement tasks on a real humanoid robot and for classification of spatial relations in annotated

TABLE I: Success Rates of Pick-and-Place Experiments

Relation	Cylindrical 1	Cylindrical 4	Cylindrical 10	RB –	RP –	Gemini –	Spotlight (Ours) 1	Spotlight (Ours) 4	Spotlight (Ours) 10
left	0.9	**1.0**	**1.0**	**1.0**	0.4	**1.0**	**1.0**	**1.0**	**1.0**
right	0.5	0.8	**1.0**	0.6	0.2	0.8	**1.0**	**1.0**	**1.0**
front	0.3	0.8	**1.0**	**1.0**	0.2	**1.0**	**1.0**	**1.0**	**1.0**
behind	**1.0**	**1.0**	**1.0**	0.8	0.2	0.6	**1.0**	**1.0**	**1.0**
on top	0.2	0.8	0.6	0.8	0.2	**1.0**	**1.0**	**1.0**	**1.0**
touching	0.8	0.6	0.4	0.2	0.4	0.4	**1.0**	**1.0**	0.8
inside	0.9	**1.0**	**1.0**	**1.0**	0.6	**1.0**	**1.0**	**1.0**	**1.0**
right inside	**1.0**	0.6	0.6	0.4	0.2	0.8	**1.0**	**1.0**	**1.0**
closer	**1.0**	**1.0**	**1.0**	0.8	0.8	0.8	**1.0**	**1.0**	**1.0**
other side	0.4	**1.0**	**1.0**	0.2	0.4	0.4	0.9	**1.0**	**1.0**

For the Cylindrical [3] and Spotlight approaches, the numbers 1, 4, and 10 indicate the number of demonstrations. Cylindrical and Spotlight were evaluated on 10 scenes in the single demonstration case; others were evaluated on 5 scenes. Legend see Section IV-B-*Baseline*.

datasets. While our approach supports a full range of relations, our evaluation focuses on a representative subset (listed in Tables I and II) that demonstrates its core capability.

A. Geometric Properties of Spatial Relations

We evaluate the geometric properties of the learned spatial relations. We first evaluate the model on a synthetic dataset, where we generate spatial relations between two objects of varying sizes and facing directions, as shown in Fig. 3. The results reveal a continuous evolution of the high-likelihood region as the reference object is scaled and rotated, while the demonstrated spatial relation remains valid. Moreover, in line with human spatial relation cognition, the acceptance region expands with increasing distance.

B. Object Placement Tasks

Our model enables one- and few-shot learning of spatial relations from visual demonstrations. We evaluate this capability on robot pick-and-place tasks using 10 semantic relations learned from 1, 4, and 10 demonstrations. The demonstrations and tasks involve a variety of household objects of different shapes and sizes, including *sponges*, *dustpans*, *brushes*, cutlery such as *plates, kettles, cups, trays*, containers like *boxes, baskets, bags*, and other common items. Object point clouds are extracted from RGB-D images using grounded SAM2 [35], [36], and 1000 points on each object are uniformly sampled inside the convex hull. We generate placement positions as described in Section III-E while keeping the same object orientation as before grasping, and execute the pick-and-place motions using Via-Point Movement Primitives [37] learned from demonstrations. Execution success is verified by human experts against the target spatial relation's requirements.

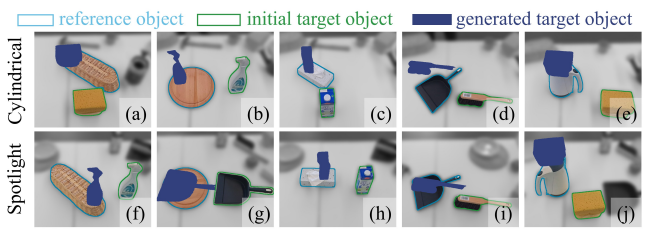

Fig. 4: Comparison of generated placement poses for the *on top* relation. Both models were trained on the same set of 10 demonstrations. The Cylindrical approach does not account for object geometry, and may generate placements near the reference object's edge ((a) and (e)) or fail to generate a valid placement (d). Our Spotlight approach always generates valid target poses.

Baselines: We compare against several baselines, i. e., the **cylindrical** distributions [3], estimated on the same set of demonstrations as our model, and the VLMs **RB**: Robobrain-2.0-3B [12], **RP**: RoboPoint-v1-vicuna-v1.5-13B [1], and **Gemini**: Gemini-3-Pro [14], which were trained on large-scale datasets. As the VLMs only predict placement points on 2D images, we project a randomly selected placement point to 3D space to obtain a placement position.

Results: Table I summarizes the success rates for the pick-and-place tasks. On average, our method outperforms [3] by $18.\overline{3}\%$, with the largest improvement (29%) in the single-demonstration case. This highlights the benefit of incorporating fine-grained geometric details in our spotlight model to learn generalizable spatial relations. While the performance gap narrows with more demonstrations, the baseline struggles with complex relations such as *on top*, *touching*, or *on the right inside* due to its oversimplified object representation, i. e., a single point. Examples are shown in Fig. 4.

Compared to RB, our method achieves more reliable placements for simple relations (*left*, *right*, *front*, *behind*) and substantially outperforms it on more complex relations (*touching*, *on the right inside*, *other side*), particularly those requiring fine-grained 3D information. Among the other baselines, RP does not generalize well to our cluttered test setup, missing even simple relations like *left*, *right*. While Gemini shows promising performance on most spatial relations, it falls short on the more complex ones (*touching*, *other side*, *behind*), due to limitations such as limited access to 3D information and difficulty understanding relative opposition, as *other side* requires.

Furthermore, qualitative examples show that our representation is less biased. For instance, the *basket* example depicted in Fig. 5 demonstrates that our probabilistic representation can generalize the relation *other side* in arbitrary directions, avoiding directional biases of simpler or learned representations in the baseline approaches.

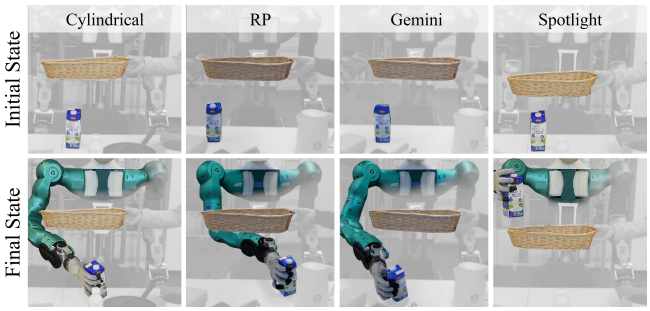

Fig. 5: Comparison of the *other side* task. While baseline methods (Cylindrical, RP, and Gemini) fail to relocate the milk box to the other side of the basket correctly, our Spotlight approach succeeds by accounting for the 3D geometry of the reference basket. Failures in the baselines stem from limited horizontal projections: Cylindrical defaults to the opposite side of the reference center, while RP and Gemini focus on the left/right or front/behind displacement.

TABLE II: Classification Accuracy per Spatial Relation Category on Rel3D Dataset.

Model	to the left •□△	to the right •□△	in front •□△	behind •□△	to the side •□△	around ○	near ○	far from ○	on □○△	on top □○△	over □△	below □△	in △	inside △	outside △	touching ○	**average**
2d	0.948	0.935	0.931	0.945	0.889	0.817	0.916	**0.872**	0.845	0.831	0.798	0.845	0.762	0.717	0.667	0.674	0.837
VipCNN	0.881	0.917	0.889	0.914	0.889	0.774	0.882	**0.872**	0.797	0.77	0.743	0.75	0.703	0.725	0.583	0.727	0.801
DRNet	0.952	0.957	**0.958**	0.961	0.875	0.848	0.849	0.723	0.832	0.858	0.775	0.777	0.752	0.75	0.708	0.667	0.828
PPFRCN	0.913	0.939	0.889	0.844	0.903	0.811	0.857	0.83	0.828	0.791	0.741	0.777	0.752	0.7	0.542	0.667	0.799
MLP-1	0.956	0.965	0.944	**0.977**	0.889	0.945	0.929	0.787	**0.909**	0.824	0.869	**0.946**	0.841	0.817	0.792	0.674	0.879
MLP-2	0.94	**0.987**	0.903	**0.977**	**0.917**	0.915	0.933	0.862	0.888	0.845	**0.874**	**0.946**	0.831	0.85	0.833	0.659	**0.885**
GPT-5.2	0.889	0.901	0.698	0.817	0.577	0.837	0.898	0.655	0.877	**0.929**	0.815	0.55	0.78	0.789	0.750	0.718	0.780
Gemini-3-Pro	0.951	0.946	0.825	0.844	0.648	0.894	**0.938**	0.655	0.904	0.922	0.846	0.807	0.871	0.867	**0.938**	0.798	0.853
Spotlight (#5)	0.951	0.977	0.952	0.936	0.789	0.844	0.769	0.774	0.781	0.809	0.841	0.857	0.463	0.889	**0.938**	**0.815**	0.836
Spotlight (#10)	0.959	0.982	0.937	0.917	0.732	**0.957**	0.818	0.762	0.826	0.809	0.844	0.836	**0.937**	**0.911**	**0.938**	0.806	0.873
Spotlight (#20)	**0.963**	0.986	0.952	0.899	0.817	0.901	0.876	0.702	0.904	0.844	0.854	0.836	0.882	0.900	**0.938**	0.806	0.879
Human	0.988	0.983	0.958	0.961	0.986	0.982	0.966	0.777	0.983	0.973	0.982	0.946	0.955	0.933	0.75	0.932	0.941

We evaluate our Spotlight models trained on different numbers of demonstrations (#5,10,20). Symbol • represents the first-person viewpoint; others with an object-centric viewpoint. Classification of direction (□), distance (○), and overlap (△) is marked by symbols. MLP-1 denotes MLP with aligned absolute features, and MLP-2 denotes raw absolute features.

C. Classification of Spatial Relations

We further evaluate discriminative performance (see Section III-D) on the Rel3D dataset [38], which provides annotated examples exclusively of static spatial relations, excluding dynamic ones. Each example includes an RGB-D image and object masks, from which we reconstruct partial point clouds for each object, sampling 1000 points inside their convex hulls. We exclude examples beyond the scope of our geometric model, specifically those involving complete occlusions or object-frame relations that necessitate pose estimation and sample a single view if multiple camera views are included. Our model is trained on 5, 10, and 20 positive examples randomly sampled from the training split, and tested on the corresponding test split. The task is to determine if a relation holds for a given object configuration, using a percentile threshold of $q = 50$ for the directional distribution ($\tau_{\text{w}} = 0.75$), $q = 20$ for the distance, and $q = 10$ for the overlap distribution.

Baselines: We compare against several baselines, including a 2D classifier based on bounding boxes, three CNN-based visual relationship detection models [39]–[41], and VLMs GPT-5.2 [13] and Gemini-3-Pro [14]. 3D baselines include two 5-layer MLPs [38] using 3D features such as mesh positions, orientations and scales which are not extracted from vision. One MLP operates on raw features (requiring explicit front/up encoding due to arbitrary axes), and one on axis-aligned features (pre-rotated to match each object's semantic orientation). Results for [39]–[41] and the MLPs [38] are reported in [38].

Results: As shown in Table II, our model achieves state-of-the-art performance on multiple representative spatial relations, including *left*, *right*, *front*, *around*, *in*, *inside*, *outside*, and *touching*, outperforming several 2D deep learning approaches, particularly for relations requiring fine-grained 3D information like *inside, in, touching, outside*.

For distance relations (*near*, *far*), inconsistent scaling in Rel3D limits the performance of both humans and our model. Dataset imbalance further impacts some categories. For instance, *to the side* is biased toward one side in training, preventing our model from learning a symmetric representation even with 20 demonstrations. This asymmetry negatively affects classification on the less demonstrated side.

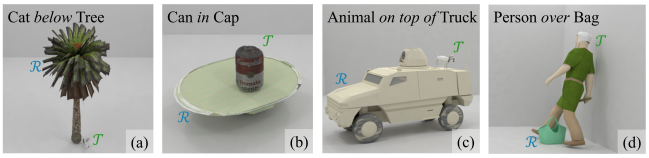

Fig. 6: Failure cases of spotlight classification on Rel3D. (a) Part-level reference (*below* the tree crown). (b) *In* without overlap. (c), (d) Convex hull approximation introducing false overlap. Reference ($\mathcal{R}$), Target ($\mathcal{T}$).

Performance decreases for relations like *on top*, *over*, and *below*. Our analysis reveals three key limitations of our purely geometric model: (1) some relations require semantic reasoning not captured geometrically; (2) others emphasize specific object parts rather than the object as a whole; and (3) convex hull approximations can introduce noise. Examples are shown in Fig. 6. These limitations suggest extensions, such as incorporating part-level representations or semantic priors, which we leave for future work. Nevertheless, our model achieves strong performance across most relations, demonstrating that geometry alone is a powerful and interpretable signal for spatial reasoning. Although the dataset contains no dynamic relations, similar results are expected, as the only difference lies in the coordinate transformation described in Section III-C.

V. Conclusion

We presented a unified probabilistic framework for modeling directional, distance-based, and topological spatial relations from a single or a few demonstrations, while incorporating object shape, size, orientation, and overlap. The model supports both generative tasks (robust placement synthesis) and discriminative tasks (relation classification). In robot experiments, our approach outperforms both a proba-

bilistic baseline [3] and VLMs [1], [12]–[14], demonstrating robust generalization to variations in object size, shape, and scale. For classification, it outperforms several deep learning approaches [13], [14], [39]–[41] in various relation categories on a large-scale dataset.

Current limitations include sensitivity to viewpoint changes, partial object observations, and the lack of semantic reasoning needed to disambiguate spatial relationships. Future work will integrate improved object perception, semantic reasoning, and interactive refinement, providing a unified foundation for spatial reasoning in robotics.

REFERENCES

[1] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali *et al.*, "Robopoint: A vision-language model for spatial affordance prediction in robotics," in *Conference on Robot Learning (CoRL)*, 2024, pp. 4005–4020.

[2] A. Singhal, J. Luo, and W. Zhu, "Probabilistic spatial context models for scene content understanding," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2003, pp. 235–241.

[3] R. Kartmann, D. Liu, and T. Asfour, "Semantic Scene Manipulation Based on 3D Spatial Object Relations and Language Instructions," in *Humanoids*, 2021, pp. 306–313.

[4] B. Kuipers, "Modeling spatial knowledge," *Cognitive Science*, vol. 2, no. 2, pp. 129–153, 1978.

[5] C. Hudelot, J. Atif, and I. Bloch, "Fuzzy spatial relation ontology for image interpretation," *Fuzzy Sets and Systems*, vol. 159, no. 15, pp. 1929–1951, 2008.

[6] Y. Wang, H. Peng, Y. Xiong, and H. Song, "Spatial relationship recognition via heterogeneous representation: A review," *Neurocomputing*, vol. 533, pp. 116–140, 2023.

[7] R. Kartmann, Y. Zhou, D. Liu, F. Paus, and T. Asfour, "Representing Spatial Object Relations as Parametric Polar Distribution for Scene Manipulation Based on Verbal Commands," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020, pp. 8373–8380.

[8] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas *et al.*, "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14 455–14 465.

[9] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz *et al.*, "SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models," in *Neural Information Processing Systems (NeurIPS)*, 2024.

[10] C. Ma, K. Lu, T.-Y. Cheng, N. Trigoni, and A. Markham, "SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors," in *Neural Information Processing Systems (NeurIPS)*, 2024.

[11] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, "RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 15 768–15 780.

[12] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang *et al.*, "RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 1724–1734.

[13] OpenAI, "Introducing GPT-5.2," https://openai.com/index/introducing-gpt-5-2/, 2025.

[14] Google, "A New Era of Intelligence with Gemini 3," https://blog.google/products-and-platforms/products/gemini/gemini-3/, 2025.

[15] R. Haar, "Computational models of spatial relations," Computer Science, University of Maryland, College Park, MD, Technical Report TR-478, 1976.

[16] E. Clementini, "Directional relations and frames of reference," *GeoInformatica*, vol. 17, no. 2, pp. 235–255, 2013.

[17] D. Papadias and Y. Theodoridis, "Spatial relations, minimum bounding rectangles, and spatial data structures," *International Journal of Geographical Information Science*, vol. 11, no. 2, pp. 111–138, 1997.

[18] D. J. Peuquet and Z. Ci-Xiang, "An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane," *Pattern Recognition*, vol. 20, no. 1, pp. 65–74, 1987.

[19] R. K. Goyal, "Similarity assessment for cardinal directions between extended spatial objects," Ph.D. dissertation, The University of Maine, 2000.

[20] A. Mukerjee and G. Joe, *A qualitative model for space*. Texas A and M University. Computer Science Department, 1990.

[21] B. R. Fajen and F. Phillips, "Spatial Perception and Action," in *Handbook of Spatial Relations*. Academic Press, 2013, pp. 67–80.

[22] K. Wang, "A computational model for direction relations between spatial objects in GIS," *Optik*, vol. 125, no. 23, pp. 6981–6986, 2014.

[23] K. Miyajima and A. Ralescu, "Organization in 2D segmented images: Representation and recognition of primitive spatial relations," *Fuzzy Sets and Systems*, vol. 65, no. 2-3, pp. 225–236, 1994.

[24] P. Matsakis and L. Wendling, "A new way to represent the relative position between areal objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 634–643, 1999.

[25] P. Matsakis, L. Wendling, and J. Ni, "A general approach to the fuzzy modeling of spatial relationships," in *Methods for Handling Imperfect Spatial Information*, J. Kacprzyk, R. Jeansoulin, O. Papini, H. Prade, and S. Schockaert, Eds. Springer Berlin Heidelberg, 2010, vol. 256, pp. 49–74.

[26] M. Deng and Z. Li, "A Statistical Model for Directional Relations Between Spatial Objects," *GeoInformatica*, vol. 12, no. 2, pp. 193–217, 2008.

[27] M. Casasola, L. B. Cohen, and E. Chiarello, "Six-month-old infants' categorization of containment spatial relations," *Child Development*, vol. 74, no. 3, pp. 679–693, 2003.

[28] E. Yiu, E. Kosoy, and A. Gopnik, "Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)," *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 874–883, 2024.

[29] Á. González, "Measurement of areas on a sphere using fibonacci and latitude–longitude lattices," *Mathematical geosciences*, vol. 42, no. 1, pp. 49–64, 2010.

[30] E. García–Portugués, "Exact risk improvement of bandwidth selectors for kernel density estimation with directional data," *Electronic Journal of Statistics*, vol. 7, no. none, 2013.

[31] A. SenGupta and B. C. Arnold, Eds., *Directional Statistics for Innovative Applications: A Bicentennial Tribute to Florence Nightingale*, 1st ed., ser. Forum for Interdisciplinary Mathematics. Springer Singapore, 2022.

[32] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, "Geodesic finite mixture models," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[33] P. Paalanen, J.-K. Kamarainen, J. Ilonen, and H. Kälviäinen, "Feature representation and discrimination based on gaussian mixture model probability densities—practices and algorithms," *Pattern Recognition*, vol. 39, no. 7, pp. 1346–1358, 2006.

[34] G. L. Yang and L. M. Le Cam, *Asymptotics in Statistics: Some Basic Concepts*. Berlin: Springer, 2000.

[35] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang *et al.*, "Grounding DINO: marrying DINO with grounded pre-training for open-set object detection," in *Euro. Conf. on Computer Vision (ECCV)*, vol. 15105, 2024, pp. 38–55.

[36] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma *et al.*, "SAM 2: Segment anything in images and videos," in *Intl. Conf. on Learning Representations (ICLR)*, 2025.

[37] Y. Zhou, J. Gao, and T. Asfour, "Learning Via-Point Movement Primitives with Inter- and Extrapolation Capabilities," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 4301–4308.

[38] A. Goyal, K. Yang, D. Yang, and J. Deng, "Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d," in *Neural Information Processing Systems (NeurIPS)*, 2020.

[39] Y. Li, W. Ouyang, X. Wang, and X. Tang, "ViP-CNN: Visual Phrase Guided Convolutional Neural Network," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7244–7253.

[40] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3298–3308.

[41] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang, "Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn," in *Intl. Conf. on Computer Vision (ICCV)*, 2017, pp. 4243–4251.