TOPOBOX: TOWARDS PHYSICALLY FEASIBLE 3D OBJECT RE-ARRANGEMENT VIA TOPOLOGY-AWARE BOX REASONING

Anonymous authors
Paper under double-blind review

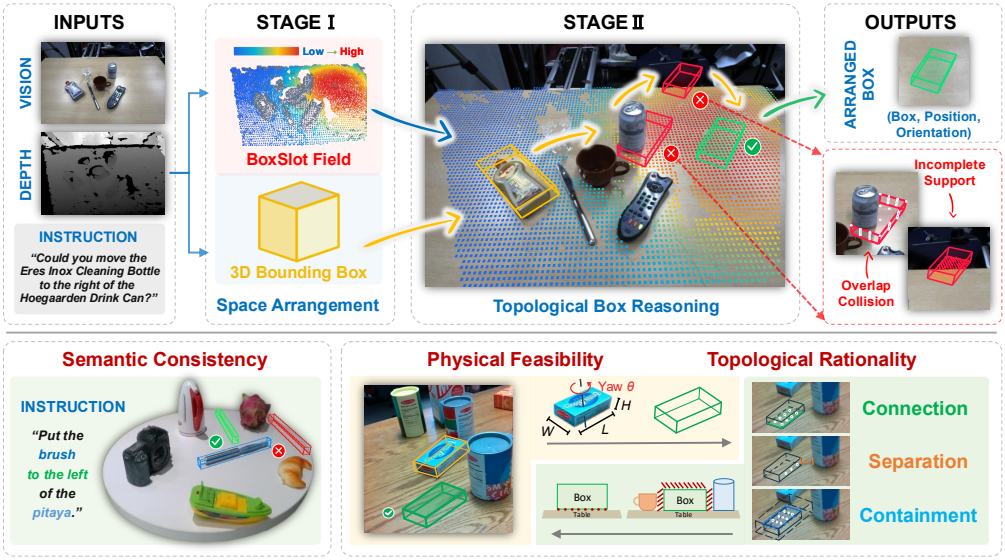

Figure 1: **TopoBox: a two-stage framework for physically feasible 3D Object Re-Arrangement.** Given the vision and depth inputs (i.e., RGB-D) and the textual instruction, **Stage I** first localizes the source object with a 3D bounding box and predicts the interaction status through a BoxSlot field. **Stage II** then refines the position and orientation with topological box reasoning, yielding results with **semantic consistency, physical feasibility, and topological rationality.**

ABSTRACT

3D object re-arrangement aims to discover vacant and well-suited spaces for object movement and placement, which is an essential task in embodied interaction scenarios. While existing methods have effectively maintained **semantic consistency** with instructive conditions, they still show limitations on arranging spaces with **physical feasibility** (e.g., 3D dimensions and orientation). Meanwhile, previous approaches ignore the **topological rationality** of the interaction between source objects and surrounding environments (e.g., structural support and collision-free connectivity), which also constrains the physical representation. To address these limitations, we propose **TopoBox**, a two-stage framework that leverages **Topo**logy-aware **Box** reasoning for physically feasible 3D object re-arrangement. At the first stage, we design a **Space Arrangement Predictor (Predictor)**, which explicitly localizes the 3D bounding box and concurrently forecasts an **interactive BoxSlot field**. At the second stage, we design a **Topological Box Reasoner (Reasoner)**, which utilizes the BoxSlot field as the core guidance, and realizes iterative topology refinement for final re-arrangement via multi-granularity and multi-surface feature sampling (i.e., bottom, side, and upper box surfaces). Furthermore, to address the lack of data for jointly modeling instruction-guided object re-arrangement and its physical feasibility, we introduce **ReArrange3D**, a dataset comprising 78K+ instruction-guided re-arrangement instances, with comprehensive annotations to support modeling and evaluation. Extensive experiments conducted on **ReArrange3D** demonstrate that TopoBox surpasses state-of-the-art (SOTA) counterparts on placement accuracy and physical feasibility.

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

1 INTRODUCTION

3D object re-arrangement focuses on identifying vacant and well-suited spaces for object movement and placement, which requires the target location to be semantically consistent with the instructive guidance and physically feasible within the complex 3D environment. With the rapid advancement of embodied intelligence, 3D object re-arrangement plays a crucial role in bridging the scene perception and embodied action, with potential applications spanning 3D scene construction Lin & Mu (2024); Yang et al. (2024); Boudjoghra et al. (2025); Tam et al. (2025), household robotics Wu et al. (2023); Shukla et al. (2025), and immersive AR/VR experiences Sharma et al. (2024); Bazargani et al. (2025).

Existing works on 3D object re-arrangement mainly focus on exploring **semantically consistent** placement from language instructions, typically represented as position-centric 2D/2.5D coordinates Tan et al. (2026); Zhu et al. (2026); Zhou et al. (2026), point-wise affordances Yuan et al. (2024), and probabilistic heatmaps Shao et al. (2025a). However, these representations mainly indicate where to place an object, ignoring to encode the 3D object explicitly (i.e. dimensions and orientation). As a result, they cannot explicitly capture the spatial occupancy of the placed object, limiting the ability for **physically feasible** placement. Meanwhile, other researchers Zhao et al. (2025); Qi et al. (2025); He et al. (2026) attempt to predict complete placement poses with implicit relation modeling between the arranged object and environment. However, these implicit methods ignore the **topological rationality**, which concerns the **interaction status** between the source object and the surrounding scene, i.e., whether the object remains **connected** to the supporting surface, stays **separated** from obstacles, and is **contained** within the available free space. Consequently, the predicted placement may still fail to ensure that the object is structurally supported and collision-free in the specific scene.

Therefore, we attempt to explore physically feasible 3D object re-arrangement via topology-aware box reasoning, which explicitly models the interaction status between objects and scenes. **Achieving this goal presents two main challenges: ❶** Existing position-centric representations show limitations in characterizing 3D object geometry and instruction-guided fields, which result in physical infeasibility. **❷** Reasoning the interaction status between object and scene is challenging, particularly in modeling structural support and collision-free connectivity during object placement.

To address these challenges, we propose **TopoBox**, a two-stage framework that leverages **Topo**logy-aware **Box** reasoning for physically feasible 3D object re-arrangement. Our key insight is to explicitly reason the topological relation between the re-arranged object and its surrounding geometry through box-centric spatial reasoning. Hence, our framework first characterizes the physical geometry of the source object and the instruction-conditioned spaces, and then progressively reasons about the interaction status toward topologically rational placements. To provide physical object representation and find instruction-relevant interaction spaces, we propose a **Space Arrangement Predictor (Predictor)**, which localizes the 3D bounding box of the source object explicitly, while concurrently forecasting an **interactive BoxSlot field**. The bounding box encodes the object dimensions and orientation, providing the geometric constraints to be accommodated during placement. Meanwhile, the BoxSlot is formulated as an instruction-guided voxel-wise affinity field that highlights candidate box-compatible interaction slots in the scene. Overall, the predictor characterizes both the object geometry and potential interaction regions, specifying what must be accommodated and where subsequent topology reasoning should be performed. Building on these predictions, we introduce a **Topological Box Reasoner (Reasoner)** to explore object–environment interactions through **iterative topology refinement**, with the BoxSlot field serving as the core guidance. Since position adjustments can alter support and collision relations, the Reasoner reassesses interaction evidence as the box position evolves. Specifically, the Reasoner performs multi-granularity and multi-surface feature sampling on the box bottom, side, and upper surfaces from the voxel grids at different spatial resolutions. This represents each interaction location at multiple levels of geometric detail, allowing structural support and collision-free connectivity to be reasoned from fine-grained, middle-grained, and coarse-grained spatial representations. The resulting surface evidence enables subsequent pose corrections to remain responsive to interaction changes that earlier adjustments induced.

Beyond modeling the spatial topological relations, advancing physically feasible 3D object re-arrangement also requires corresponding datasets that capture both instruction semantics and object–environment physical relations. Existing related datasets Yuan et al. (2024); Song et al. (2026); Zhou et al. (2026) focus primarily on whether object placements satisfy the semantic requirements

specified by instructions, while providing limited supervision for physical properties such as structural support and collisions with the surrounding environment. To address this gap, we introduce **ReArrange3D**, a dataset comprising 78,506 instruction-guided 3D object re-arrangement instances. Each instance provides annotations describing the source object, target placement, and the semantic and physical relations with the surrounding scenes. Therefore, ReArrange3D provides systematic data support for 3D object re-arrangement under both semantic consistency and physical feasibility.

Overall, our contributions are summarized as follows:

- We propose **TopoBox**, a two-stage framework leveraging topology-aware box reasoning for physically feasible 3D object re-arrangement, which explores the topological rationality of object-environment interactions.
- We design a **Space Arrangement Predictor** that explicitly localizes the 3D bounding box of the source object, and concurrently forecasts an interactive BoxSlot field, providing object geometry and candidate interaction spaces for topology reasoning.
- We design a **Topological Box Reasoner** that employs the BoxSlot field to guide iterative topology refinement via multi-granularity and multi-surface feature sampling, reasoning about structural support and collision-free connectivity for physically feasible re-arrangement.
- Extensive experiments on the ReArrange3D show that TopoBox achieves superior performance against competitors, displaying semantically consistent and physically feasible object re-arrangement.

2 Related Work

2.1 Object Re-Arrangement

Object re-arrangement aims to identify suitable target locations for object placement according to instructive conditions and scene observations. Early studies Tellex et al. (2011); Kollar et al. (2010) ground language instructions in spatial relations between objects and environments through **probabilistic graphical models**. Subsequent methods Mees et al. (2020); Zeng et al. (2021) formulate rearrangement as **dense spatial prediction** and incorporate pretrained vision-language representations for semantic grounding and language-conditioned action prediction Shridhar et al. (2022); Stengel-Eskin et al. (2022); Brohan et al. (2023); Driess et al. (2023). More recently, **vision-language foundation models** Yuan et al. (2024); Shao et al. (2025a); Zhou et al. (2026) have advanced embodied spatial reasoning and localization, improving generalization to unseen objects and instructions. However, point/region-based placement predictions do not explicitly represent the full 3D occupancy of the placed object, limiting reasoning about the physical feasibility of its target placement. To incorporate **3D geometry**, existing methods Liu et al. (2022b); Shridhar et al. (2023); Chen et al. (2023); Gervet et al. (2023); Zhang et al. (2023) perform rearrangement and manipulation using object-centric, voxel, or point cloud representations. Generative approaches Liu et al. (2022a); Xu et al. (2024) further capture the ambiguity of object arrangements through learned distributions, allowing **multiple arrangements** under the same instruction. Recent placement methods Abdelreheem et al. (2025); Zhao et al. (2025); Zhong et al. (2025) jointly consider scene geometry, object geometry, and language instructions, although object–environment interactions often remain **implicitly encoded** within learned features. Building on these advances, we adopt a box-centric formulation that explicitly represents object dimensions and orientation, providing geometric constraints for subsequent reasoning about physically feasible placement.

2.2 Spatial Interaction Representation

Spatial interaction representation aims to characterize how objects interact with the surrounding environment. Modeling these interactions requires geometric descriptions of objects and scenes, which continuous geometric fields Park et al. (2019); Mescheder et al. (2019); Chen & Zhang (2019) provide through signed distance functions and occupancy representations. Affordance-based methods Geng et al. (2023); Shao et al. (2025b); Gao et al. (2025); Wang et al. (2025) extend geometric descriptions toward task-relevant interaction regions, incorporating part-level geometry, language guidance, and learned affordance fields to identify where interactions may occur. However, such **local interaction regions** do not explicitly characterize the full spatial constraints imposed by

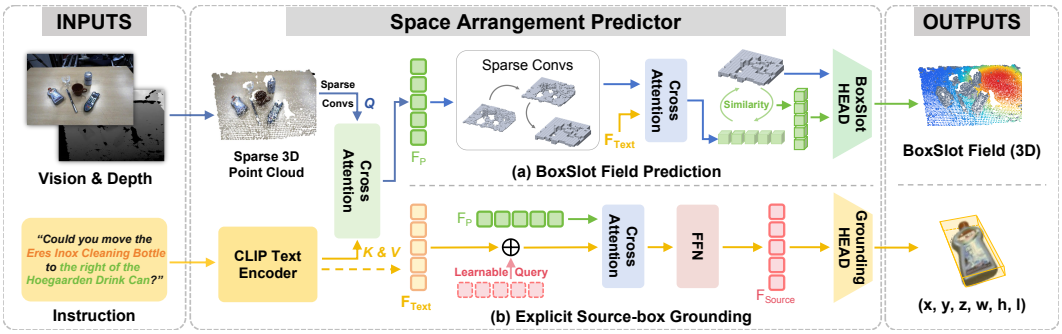

Figure 2: **Stage I: Space Arrangement Predictor**. Given an RGB-D observation and a textual instruction, the Predictor constructs language-conditioned 3D features, estimates the interactive BoxSlot field and predicts an explicit bounding box.

object dimensions and orientation. Further studies Tian et al. (2025); Cai et al. (2025); Liu et al. (2026) extend **interaction reasoning** to object-object relations and incorporate physical constraints, such as contact and collision awareness, highlighting the importance of both semantic compatibility and physical relations with the surrounding environment. Inspired by these advances, we introduce a topology-aware, box-centric representation of object–scene interactions.

3 Proposed Method

3.1 Problem Formulation

Given an RGB-D observation and a textual instruction, our **TopoBox** engages in accurate source object localization and physically feasible box re-arrangement, while satisfying the instructed relation, structural support, and collision constraints. We formulate it as:

$$(S, \mathcal{B}) = \text{TopoBox}(\mathcal{O}, \mathcal{I}), \tag{1}$$

where $\mathcal{O}$ and $\mathcal{I}$ denote observation and instruction. The source box $S = (\mathbf{c}, \mathbf{d})$ contains center $\mathbf{c}$ and dimensions $\mathbf{d} = (w, h, l)$, which indicate the width, height, and length, respectively. Here, $\mathcal{B} = \{B_k\}_{k=1}^{K}$ represents the set of K placement candidates, where each candidate $B = (\mathbf{p}, \mathbf{d}, \theta)$ is composed of the bottom center, source dimensions, and the yaw angle.

3.2 Stage I: Space Arrangement Predictor

To ensure the rearranged 3D bounding boxes are physically plausible and spatially consistent with the target interaction spaces, in **Stage I**, we explicitly localize the source object box and jointly generate candidate BoxSlot fields. Firstly, we back-project the input RGB-D observation into a point cloud and voxelize it into a sparse 3D representation, which provides spatial structural guidance for object placement. Then, we apply sparse convolutions on the voxelized point cloud to obtain aggregated geometric and visual features. Meanwhile, we leverage the encoded textual instruction as semantic cues to yield instruction-projected voxel features, which form a shared representation for BoxSlot field prediction and source-box grounding.

Interactive BoxSlot field. Considering that language instructions typically describe multiple plausible placement regions, we estimate a voxel-wise affinity field - termed **BoxSlot** - to model the interaction between 3D scene and candidate placements. In particular, each voxel is assigned an affinity score that encodes its dependency on the instructed placement, highlighting candidate interactive regions for subsequent box reasoning. Notably, the projected features are processed by sparse convolutions to reduce computational cost while capturing broader geometric context. Then, we combine the aggregated affinity with the paired voxel feature to predict the placement probability for each voxel at the sparsest granularity:

$$h_i = \mathcal{H}_{\text{BoxSlot}}\left(\mathbf{f}_i, \log\left[|\mathcal{T}|^{-1}\sum_{j\in\mathcal{T}}\exp(a_{ij})\right]\right), \tag{2}$$

where $\mathbf{f}_i$ denotes the i-th voxel feature at the coarsest granularity. a_{ij} is the similarity between the i-th voxel feature and the j-th updated text embedding. $|\mathcal{T}|$ is the number of updated text tokens. The prediction head $\mathcal{H}_{\text{BoxSlot}}$ includes a sigmoid output, yielding the placement probability h_i.

Explicit Source-box Grounding. To acquire the semantically consistent source object geometry, we exploit a learnable query concatenated with the text features to extract the source feature from the projected features. Subsequently, the Grounding Head maps the extracted features to the position and dimensions of the source box. The predicted dimensions determine the size of subsequent placement boxes and the sampling locations on their surfaces, while the source feature carries information about the object to be placed into subsequent topology reasoning.

3.3 Stage II: Topological Box Reasoner

Once we obtain the BoxSlot fields, we combine them with multi-granularity scene features to produce accurate topologically reasonable results in Stage II. In contrast to directly taking the one-shot object placement predictions as the final results, we formulate the re-arrangement as an **iterative refinement** process. By gradually refining the placement position and yaw of sampled scene features around the predicted box, we establish a topological relationship which is collision-free and support-stable.

Input Preliminary in Stage II. In stage II, we first apply farthest point sampling on the top-k (k=8) regions of the BoxSlot fields to gain 48 candidate points. Finest-grained features at these candidate points are gathered as the corresponding BoxSlot feature. Then, we combine the acquired BoxSlot, box size, and source object features to produce the placement embeddings, expressed as:

$$\mathbf{F}_{\text{place}} = MLP(\mathbf{F}_{\text{slot}} \oplus \mathbf{F}_{\text{size}} \oplus \mathbf{F}_{\text{source}}). \tag{3}$$

Where $\oplus$ denotes concatenation, $\mathbf{F}_{\text{slot}}$, $\mathbf{F}_{\text{size}}$, and $\mathbf{F}_{\text{source}}$ are features of the box slot, size, and source object, respectively. $\mathbf{F}_{\text{place}}$ represents the obtained initial placement embeddings.

Iterative Topology Refinement. To ensure the stability of the optimization and avoid the impact of position & yaw offset disturbances, we resample features from the surrounding scene and perform two coupled updates at each iteration. The **Implicit Feature Update** integrates semantic and topological information into placement features, while the **Explicit Pose Update** refines box position and yaw. Specifically, **for implicit feature update**, an attention block aggregates information across initial position candidates via placement features, and then outputs Source Placement Queries. Using these queries, the TopoBox Reasoning Module aggregates scene features sampled around the current boxes. Meanwhile, multi-head attention is adopted to reintroduce textual features, maintaining semantic consistency between placement adjustments and the instruction. After acquiring the optimized topology-aware and text instruction features, we introduce the Topology Semantic Fusion operation to adaptively fuse these two feature representations with a learned gate:

$$\boldsymbol{\alpha}(t) = \sigma\left(\text{MLP}\left([\mathbf{g}(t); \mathbf{s}(t)]\right)\right), \quad \mathbf{F}_{place}(t+1) = \text{LN}\left(\boldsymbol{\alpha}(t) \odot \mathbf{g}(t) + (1 - \boldsymbol{\alpha}(t)) \odot \mathbf{s}(t)\right). \tag{4}$$

Specifically, at iteration t, we use the learned gate $\boldsymbol{\alpha}(t)$ to adaptively combine the topology-aware feature $\mathbf{g}(t)$ and the instruction feature $\mathbf{s}(t)$ into the $\mathbf{F}_{place}$ of the next iteration, which encodes local object-environment relations and instruction semantics. Here, σ , $\odot$, and LN denote the sigmoid function, element-wise multiplication, and layer normalization, respectively.

For explicit pose update, the updated features predict the placement position, yaw, and probability of each candidate. The new position and yaw determine where scene features are sampled in the next iteration, allowing the model to reassess support and collision relations around the adjusted boxes. Together, the two updates progressively refine both the placement features and the box poses.

TopoBox Reasoning Module. To provide a physically reasonable re-arrangement 3D position within the topological relations, we design the multi-granularity and multi-surface sampling strategy. Sampling points are distributed across the bottom, side, and upper surfaces according to the predicted source dimensions. Each point is transformed into scene coordinates using the current placement position and yaw. Here, bottom samples characterize **contact** with the supporting surface. Side and upper samples provide complementary evidence for assessing potential obstacle **collision** and the box **containment** within the available free space. To capture local geometric detail and broader spatial

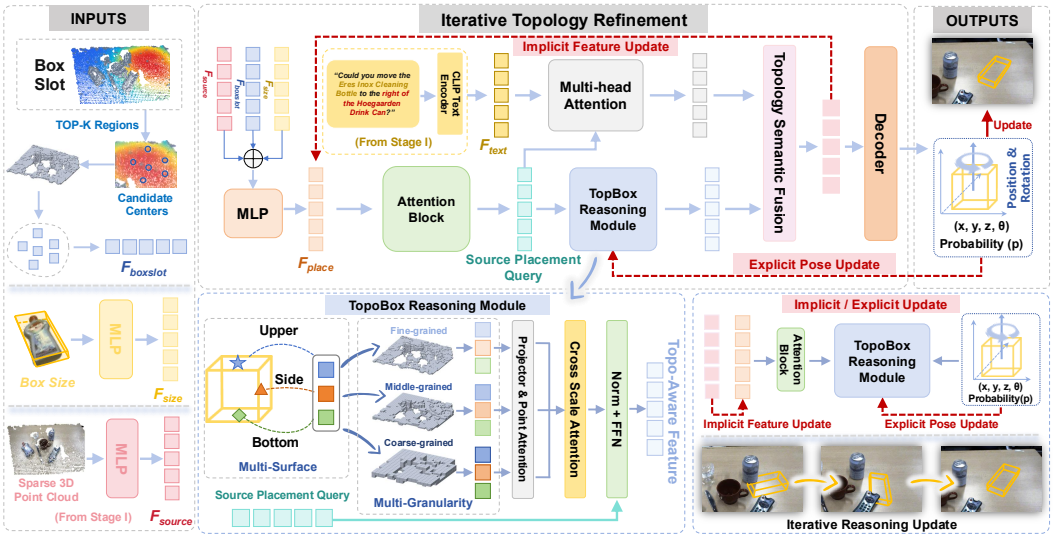

Figure 3: **Stage II: Topological Box Reasoner.** Guided by the BoxSlot field and source object geometry, the Reasoner iteratively refines candidate placements to update placement features, positions, and orientations, considering structural support and collision constraints.

context, we transfer the surface sampling points to scene coordinates and query Stage I features at fine-grained, middle-grained, and coarse-grained resolutions:

$$\begin{aligned} \mathcal{P}_u &= \mathbf{p} + \mathbf{R}(\theta)\,\Omega_u(\mathbf{d}), \qquad u \in \{\text{bottom}, \text{side}, \text{upper}\}, \\ \mathbf{F}_u^{(m)} &= \text{Sample}\left(\mathbf{F}_{\text{scene}}^{(m)}, \mathcal{P}_u\right), \quad m \in \{\text{fine}, \text{middle}, \text{coarse}\}. \end{aligned} \tag{5}$$

Specifically, we construct surface sampling points $\Omega_u(\mathbf{d})$ on the box surfaces with the predicted source dimensions $\mathbf{d}$, which defines relative positions of sampling points with respect to the box. Then these points are rotated by the current yaw θ and translated to the scene coordinates $\mathcal{P}_u$ according to the bottom-center position $\mathbf{p}$. At these coordinates, Sample retrieves voxel features from the Stage I scene features $\mathbf{F}_{\text{scene}}^{(m)}$, yielding topological evidence feature $\mathbf{F}_u^{(m)}$. Overall, we use the same sampling locations across all three resolutions. Fine-grained features capture local surface details, while coarser features encode broader spatial contexts. At each resolution, a projector integrates the sampled features with their box coordinates to distinguish evidence from different surface locations. We employ point attention, guided by the placement query, to aggregate relevant surface evidence. Then, we use cross-scale attention to adaptively fuse features across multiple granularities for each candidate, resulting in topology-aware features that guide subsequent feature and pose updates.

3.4 Training Objective

During training, re-arrangement refinement relies on source geometry. We first pretrain source-object localization and dimensions estimation. Then, we jointly optimize BoxSlot prediction and placement refinement while retaining source-box supervision. The pretraining objective is:

$$\mathcal{L}_{\text{src}} = \lambda_{\text{sc}}\mathcal{L}_{\text{sc}} + \lambda_{\text{sd}}\mathcal{L}_{\text{sd}} + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}}, \tag{6}$$

where $\mathcal{L}_{\text{sc}}$ and $\mathcal{L}_{\text{sd}}$ are smooth L1 constraints on source centers and dimensions, and $\mathcal{L}_{\text{IoU}}$ penalizes incomplete 3D box overlap.

For subsequent placement training, similar to Park et al. (2026), we utilize the Hungarian Kuhn (1955) matching to assign predicted placement candidates to valid centers or background using distances normalized by ground-truth source dimensions. The objective becomes:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \lambda_{\text{src}}\mathcal{L}_{\text{src}} + \lambda_{\text{slot}}\mathcal{L}_{\text{slot}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} \\ + \lambda_{\text{yaw}}\mathcal{L}_{\text{yaw}} + \lambda_{\text{corner}}\mathcal{L}_{\text{corner}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{slot}}$ applies focal loss to Gaussian targets restricted by support and language-direction masks. Confidence BCE $\mathcal{L}_{\text{cls}}$ uses targets based on valid-orientation proportions for matched candidates

and zero for background. Matched candidates receive size-normalized Smooth L1 loss $\mathcal{L}_{\text{center}}$, multi-hot yaw BCE $\mathcal{L}_{\text{yaw}}$, and symmetric Chamfer-L1 loss $\mathcal{L}_{\text{corner}}$, minimized over valid orientations and normalized by the ground-truth box diagonal. Auxiliary loss $\mathcal{L}_{\text{aux}}$ supervises intermediate refinements using the final-step matching; the coefficients λ balance the losses.

4 REARRANGE3D DATASET

Data Source and Canonicalization. We introduce **ReArrange3D** Dataset, which uses explicit 3D bounding boxes to represent both source-object geometry and the spatial extent of object on target placements, as illustrated in Fig.4. Following prior efforts Song et al. (2026); Zhou et al. (2026); Zhang et al. (2026), we construct our dataset from real RGB-D observations and obtain object annotations from five 6D-pose estimation datasets Xiang et al. (2017); Lin et al. (2021); Gouda et al. (2022); Jung et al. (2024); Zhang et al. (2024). Specifically, we first extract the object box-aware poses and the camera poses. Then, each scene is transformed into a canonical 3D space by standardizing units to centimeters, unifying coordinate conventions and aligning support-surface orientations.

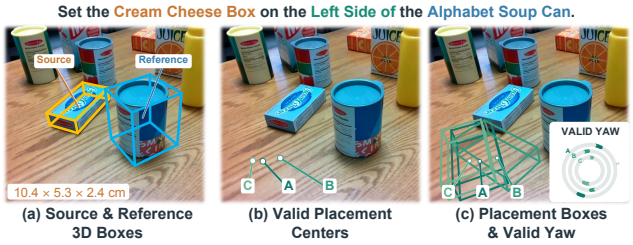

Figure 4: Sample illustrations of the ReArrange3D dataset and the multiple types of annotations.

Annotations: Multi-Valid Box Placement & Spatial Relation Instruction. To capture multiple physically feasible placements, we adopt an automatic annotation pipeline. We first use RANSAC Fischler & Bolles (1981) to detect support surfaces, and employ FFT-based Cooley & Tukey (1965) correlation to identify collision-free placements across positions and yaw angles, while preserving the box dimensions of the object. Then, a filtering strategy is used to remove candidates with insufficient support, excessive occlusion or outside the camera view. Finally, we use DBSCAN Ester et al. (1996) to group the remaining placements into spatial clusters. For each cluster, we record a representative box, dense valid placement centers, and their corresponding valid yaw angles. Given the placement positions and orientations in each cluster, we generate corresponding spatial relation instructions. To reduce subjectivity in human labeling, we use a fixed pipeline to derive spatial relations from the 3D scene geometry. Specifically, we use templates to describe the source object and target placement. Then we use Qwen3-VL Bai et al. (2025) to rewrite these instructions into more natural and diverse expressions thereby preserving the annotation correctness while enhancing the richness and diversity of semantic expressions. Furthermore, we also establish an objective and reproducible evaluation protocol for assessing the semantic accuracy of placement positions. These box placement annotations provide reliable references for object re-arrangement with multiple valid solutions.

5 EXPERIMENTS

5.1 DATASET AND EXPERIMENTAL SETTINGS

ReArrange3D Dataset To address the limitations in object geometry and placement feasibility annotations, we introduce ReArrange3D for evaluating language-guided, physically feasible placement. The dataset comprises 78,506 instances of language-annotated placement, constructed from 9,741 pairs of RGB-D images. Within each data source, we divided the RGB-D image pairs into training, validation, and test sets in a ratio of 8:1:1, with 63,384, 7,496, and 7,626 instances respectively. All instructions corresponding to the same pair of RGB-D images are grouped into the same set.

Evaluation Metrics. Metrics are listed for source localization, size estimation, instruction consistency, and physical feasibility.❶**3D Grounding.** Acc@0.25 and Acc@0.5 (Chen et al., 2020) measure source box accuracy at IoU thresholds of 0.25/0.5; mIoU measures average 3D IoU. ❷**Semantic Consistency.** Semantic Relation Correctness (SRC) measures the fraction of predicted placements satisfying instructed spatial relations. ❸**Physical Feasibility.** Support Rate (SR) measures the fraction of placements whose entire box footprint is covered by the connected support region containing

Table 1: Quantitative comparison of re-arrangement feasibility on ReArrange3D. SRC, SR, CFR, and PF denote Semantic Relation Correctness, Support Rate, Collision-Free Rate, and Placement Feasibility. SRC, SR and CFR evaluate the highest-ranked candidate. (Higher is better).

Methods	**Semantic Consistency**	**Physical Feasibility**		**Overall**	
	SRC↑	**SR↑**	**CFR↑**	**PF@1↑**	**PF@5↑**
RoboPoint Yuan et al. (2024) *CoRL*	56.49	77.00	82.54	39.47	-
RoboRefer Zhou et al. (2026) *NeurIPS*	60.90	73.63	72.40	40.13	-
RoboBrain2.5 Tan et al. (2026) *Tech.Rep.*	69.84	76.03	79.65	46.44	-
TopoBox (Ours)	**70.12**	**95.14**	**95.78**	**65.16**	**79.50**

the bottom center. Collision-Free Rate (CFR) measures the proportion with no volumetric overlap between the re-arranged box and any annotated scene object box. To evaluate the placement accuracy of semantic consistency and physical feasibility, we use Placement Feasibility (PF@K) to measure the fraction of instances where at least one top-K candidate meets all criteria: semantic consistency, stable support, and collision-free.

5.2 Quantitative Analysis

Comparison with Existing Methods. We evaluate TopoBox on source object grounding and placement feasibility. For source object grounding, we compare against MCLN Qian et al. (2024), EG-3DVG Park et al. (2026), and TSP3D Guo et al. (2025) on ReArrange3D. For placement feasibility, we fine-tune RoboPoint Yuan et al. (2024), RoboRefer Zhou et al. (2026) and RoboBrain2.5 Tan et al. (2026) on the ReArrange3D training set using LoRA. For point-based baselines, we retain their original point supervision and place the ground-truth source box at the predicted location during inference for geometric evaluation.

As shown in Table 1, TopoBox achieves state-of-the-art placement feasibility on ReArrange3D, exceeding the best-performing competing method in Placement Feasibility@1 by 18 percentage points while maintaining comparable Semantic Consistency. This result supports our approach of explicitly reasoning about structural support and collision-free connectivity. Table 2 reports the source object grounding results on ReArrange3D. TopoBox achieves performance comparable to specialized 3D grounding methods, demonstrating its ability to localize the source object while jointly addressing placement prediction. Accurate source object localization provides the geometric information needed for subsequent topology reasoning.

Table 2: Quantitative Comparison of Source Object Grounding Performance on ReArrange3D.

Methods	Acc@0.50 ↑	mIoU ↑
MCLN Qian et al. (2024) *ECCV*	65.29	56.09
TSP3D Guo et al. (2025) *CVPR*	72.91	58.74
EG-3DVG Park et al. (2026) *CVPR*	73.59	60.06
TopoBox (Ours)	**74.68**	**61.11**

Ablation Study. To further evaluate the effectiveness of different parts in TopoBox, we conduct ablation studies on the ReArrange3D dataset.

Effects of BoxSlot Field Guidance: To evaluate the contribution of learned region guidance, we replace the candidate locations using the predicted Top-k BoxSlot field with farthest point sampling. As shown in Table 3, this replacement reduces SRC by 5.06. This result indicates that the learned region scores provide more effective candidate initialization for subsequent placement refinement.

Effects of Topological Box Reasoning: To evaluate the contribution of the Topological Box Reasoner for physically feasible placement, we examine four aspects of its design. First, we disable the TopoBox Reasoning Module, testing the contribution of the topology-aware features to placement refinement. Second, we fix the center and yaw used for feature sampling, assessing the contribution of explicit pose update to placement refinement. Third, we replace multi-surface sampling with the same number of uniformly sampled points inside each box, assessing the role of explicit surface sampling. Finally, we restrict the module to the fine-grained level features, assessing

Table 3: Ablation Study of Different Components.

Methods	SRC ↑	SR ↑	CFR ↑	PF@1 ↑
w/o Explicit Pose Update	68.30	93.20	91.75	62.12
w/o Multi-Surface	69.77	93.38	95.04	63.60
w/o Multi-Granularity	69.16	92.48	94.44	63.88
w/o TopoBox Reasoning	70.01	89.50	91.23	60.49
w/o BoxSlot Guidance	65.06	93.89	94.38	58.63
TopoBox (Full)	**70.12**	**95.14**	**95.78**	**65.16**

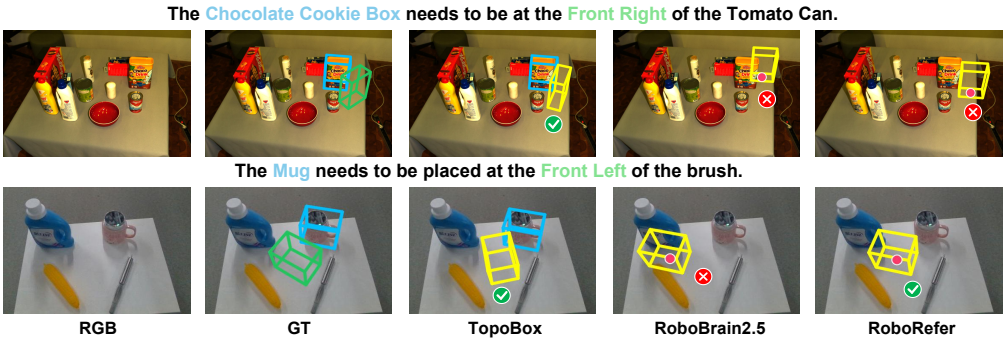

Figure 5: **Qualitative comparisons with counterparts.** TopoBox predicts physically feasible results that better satisfy the spatial relations with instructions. **(More results in Appendix.)**

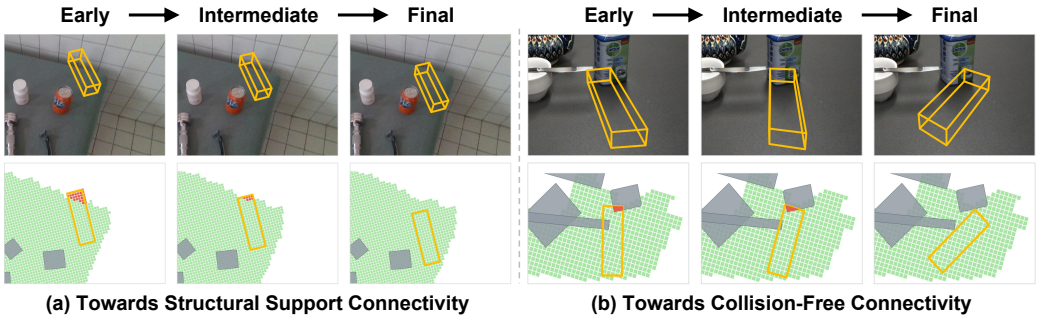

Figure 6: **Illustrations of iterative topology refinement.** TopoBox progressively updates the re-arrangements towards (a) structural support and (b) collision-free. **(More results in Appendix.)**

the contribution of multi-granularity features. As shown in Table 3, all four variants exhibit lower physical feasibility, whereas the semantic consistency remains largely unchanged. These results support the contribution of topological box reasoning to stable support and collision constraints.

5.3 Qualitative Analysis

Comparison with Existing Methods. In Figure 5, the counterparts can identify semantically consistent point locations of free space, but ignore the full object extent. Consequently, the resulting re-arrangement boxes may overlap neighboring objects or extend beyond the supporting surface. In contrast, TopoBox produces re-arrangements that remain within supported and collision-free space. Additional qualitative comparisons are provided in the supplementary materials.

Iterative Topology Refinement. Figure 6 illustrates how the interactions between object and environment update during refinement. Pose refinement improves support coverage of the bottom footprint in (a) and reduces overlap with nearby obstacles in (b), promoting structural support connectivity and collision-free connectivity, respectively. These examples show how iterative topology refinement improves the topological relations between the object and its surroundings: the box maintains contact with the supporting surface, remains separated from obstacles, and is contained within the available free space, yielding more physically feasible re-arrangements. More results of the refinement process are included in the supplementary materials.

6 Conclusion

In this paper, we propose TopoBox, a two-stage framework for physically feasible 3D object re-arrangement via topology-aware box reasoning. By explicitly modeling source-object geometry and object–environment interactions, TopoBox improves physical feasibility while maintaining semantic consistency. For object re-arrangement verification, we further introduce ReArrange3D, a language-guided dataset with source-object geometry and physically feasible re-arrangement annotations, on which TopoBox achieves SOTA placement feasibility. Future work will explore richer physical constraints in real-world robotic interaction.

AI USE STATEMENT

In this work, we used generative AI tools for dataset processing, including data cleaning, reformatting, and annotation organization. We did not use generative AI tools for other tasks requiring disclosure; relevant details are provided in the main paper and appendix. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ahmed Abdelreheem, Filippo Aleotti, Jamie Watson, Zawar Qureshi, Abdelrahman Eldesokey, Peter Wonka, Gabriel Brostow, Sara Vicente, and Guillermo Garcia-Hernando. Placeit3d: Language-guided object placement in real 3d scenes. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6645–6655. IEEE, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jalal Safari Bazargani, Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi. Integrating user input in automated object placement for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):8900–8914, 2025.

Mohamed Boudjoghra, Ivan Laptev, and Angela Dai. Scanedit: Hierarchically-guided functional 3d scan editing, 2025. URL https://arxiv.org/abs/2504.15049.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Xinhao Cai, Minghang Zheng, Xin Jin, and Yang Liu. Interactmove: Text-controlled human-object interaction generation in 3d scenes with movable objects. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9491–9499, 2025.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

Shizhe Chen, Ricardo Garcia, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. *arXiv preprint arXiv:2309.15596*, 2023.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5932–5941. IEEE, 2019.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, and Bin Zhao. Learning 2d invariant affordance knowledge for 3d affordance grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3095–3103, 2025.

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7081–7091. IEEE, 2023.

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. *arXiv preprint arXiv:2306.17817*, 2023.

Anas Gouda, Abraham Ghanem, and Christopher Reining. Dopose-6d dataset for object segmentation and 6d pose estimation. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 477–483. IEEE, 2022.

Wenxuan Guo, Xiuwei Xu, Ziwei Wang, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Text-guided sparse voxel pruning for efficient 3d visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3666–3675. IEEE, 2025.

Lyuxing He, Eric Cai, Shobhit Aggarwal, Jianjun Wang, and David Held. Disentangled point diffusion for precise object placement, 2026. URL https://arxiv.org/abs/2604.11793.

HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2024.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 259–266. IEEE, 2010.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior, 2024. URL https://arxiv.org/abs/2402.04717.

Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Multi-view fusion for multi-level robotic scene understanding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6817–6824, 2021. doi: 10.1109/IROS51168.2021.9635994.

Jie Liu, Yu Sun, Alpar Cseke, Yao Feng, Nicolas Heron, Michael J Black, and Yan Zhang. Open-vocabulary functional 3d human-scene interaction generation. *arXiv preprint arXiv:2601.20835*, 2026.

Weiyu Liu, Yilun Du, Tucker Hermans, Sonia Chernova, and Chris Paxton. Structdiffusion: Language-guided creation of physically-valid structures using unseen objects. *arXiv preprint arXiv:2211.04604*, 2022a.

Weiyu Liu, Chris Paxton, Tucker Hermans, and Dieter Fox. Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6322–6329. IEEE, 2022b.

Oier Mees, Alp Emek, Johan Vertens, and Wolfram Burgard. Learning object placements for relational instructions by hallucinating scene representations. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 94–100. IEEE, 2020.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

GwangWook Park, Hyo-Jun Lee, Jong-Hyeon Baek, Hanul Kim, and Yeong Jun Koh. Eg-3dvg: Expression and geometry aware grounding decoder for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2026.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhao Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation, 2025. URL https://arxiv.org/abs/2502.13143.

Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch collaborative learning network for 3d visual grounding. In *European conference on computer vision*, pp. 381–398. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Xinyu Shao, Yanzhe Tang, Pengwei Xie, Kaiwen Zhou, Yuzheng Zhuang, Xingyue Quan, Jianye Hao, Long Zeng, and Xiu Li. More than a point: Capturing uncertainty with adaptive affordance heatmaps for spatial grounding in robotic tasks, 2025a. URL https://arxiv.org/abs/2510.10912.

Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17326–17336. IEEE, 2025b.

Aditya Sharma, Luke Yoffe, and Tobias Höllerer. Octo+: A suite for automatic open-vocabulary object placement in mixed reality, 2024. URL https://arxiv.org/abs/2401.08973.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pp. 894–906. pmlr, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Arth Shukla, Stone Tao, and Hao Su. Maniskill-hab: A benchmark for low-level manipulation in home rearrangement tasks, 2025. URL https://arxiv.org/abs/2412.13211.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics, 2026. URL https://arxiv.org/abs/2411.16537.

Elias Stengel-Eskin, Andrew Hundt, Zhuohong He, Aditya Murali, Nakul Gopalan, Matthew Gombolay, and Gregory Hager. Guiding multi-step rearrangement tasks with natural language instructions. In *Conference on Robot Learning*, pp. 1486–1501. PMLR, 2022.

Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Angel X. Chang, and Manolis Savva. Scenemotifcoder: Example-driven visual program learning for generating 3d object arrangements, 2025. URL https://arxiv.org/abs/2408.02211.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, Yibo Li, Yi Han, Shaoxuan Xie, Guocai Yao, Songjing Wang, Leiduo Zhang, Xi Yang, Yance Jiao, Donghai Shi, Kunchang Xie, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. Robobrain 2.5: Depth in sight, time in mind, 2026. URL https://arxiv.org/abs/2601.14352.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pp. 1507–1514, 2011.

Tongxuan Tian, Xuhui Kang, and Yen-Ling Kuo. O^3 afford: One-shot 3d object-to-object affordance grounding for generalizable robotic manipulation. *arXiv preprint arXiv:2509.06233*, 2025.

Hanqing Wang, Zhenhao Zhang, Kaiyang Ji, Mingyu Liu, Wenti Yin, Yuchao Chen, Zhirui Liu, Xiangyu Zeng, Tianxiang Gui, and Hangxing Zhang. Dag: Unleash the potential of diffusion model for open-vocabulary 3d affordance grounding. *arXiv preprint arXiv:2508.01651*, 2025.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, November 2023. ISSN 1573-7527. doi: 10.1007/s10514-023-10139-z. URL http://dx.doi.org/10.1007/s10514-023-10139-z.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

Yiqing Xu, Jiayuan Mao, Linfeng Li, Yilun Du, Tomas Lozáno-Pérez, Leslie Pack Kaelbling, and David Hsu. "set it up": Functional object arrangement with compositional generative models. *The International Journal of Robotics Research*, pp. 02783649251378198, 2024.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai, 2024. URL https://arxiv.org/abs/2404.09465.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics, 2024. URL https://arxiv.org/abs/2406.10721.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pp. 726–747. PMLR, 2021.

Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pp. 199–216. Springer, 2024.

Jiyao Zhang, Mingxu Zhang, Yitong Peng, Haoxuan Liu, Chenshuo Wang, Yuxing Long, Haoyang Huang, Dongjiang Li, Nan Duan, Hui Shen, et al. Embodied3dbench: Benchmarking low-level embodied spatial intelligence of vision language models. *arXiv preprint arXiv:2605.29074*, 2026.

Tong Zhang, Yingdong Hu, Hanchen Cui, Hang Zhao, and Yang Gao. A universal semantic-geometric representation for robotic manipulation. *arXiv preprint arXiv:2306.10474*, 2023.

Yuchi Zhao, Miroslav Bogdanovic, Chengyuan Luo, Steven Tohme, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Anyplace: Learning generalized object placement for robot manipulation, 2025. URL https://arxiv.org/abs/2502.04531.

Yao Zhong, Hanzhi Chen, Simon Schaefer, Anran Zhang, and Stefan Leutenegger. Gopla: Generalizable object placement learning via synthetic augmentation of human arrangement. *arXiv preprint arXiv:2510.14627*, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics, 2026. URL https://arxiv.org/abs/2506.04308.

Qiming Zhu, Zhirui Fang, Tianming Zhang, Chuanxiu Liu, Xiaoke Jiang, and Lei Zhang. Spatialpoint: Spatial-aware point prediction for embodied localization, 2026. URL https://arxiv.org/abs/2603.26690.

A APPENDIX

We provide details on the construction of ReArrange3D and the experimental evaluation of TopoBox, together with additional qualitative results.

A.1 DATASET CONSTRUCTION

As shown in Figure 7, ReArrange3D combines observations from existing RGB-D datasets with automatically generated placement annotations and spatial relation instructions. The following sections describe its construction and experimental splits.

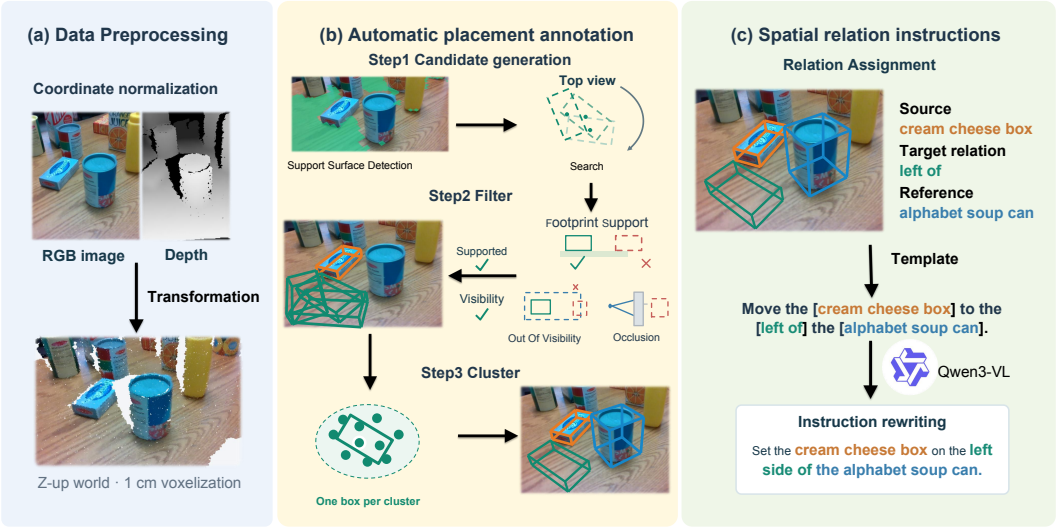

Figure 7: **Overview of the ReArrange3D construction pipeline.** RGB-D scenes are automatically processed to identify and cluster physically feasible placements and generate corresponding language instructions based on spatial relations.

A.1.1 DATA SOURCES AND PREPROCESSING

Data Sources and Frame Selection. To obtain diverse scenes for object re-arrangement, we use YCB-Video Xiang et al. (2017), HOPE Lin et al. (2021), DOPose Gouda et al. (2022), HouseCat6D Jung et al. (2024), and Omni6DPose Zhang et al. (2024). Each selected frame provides an RGB-D observation, camera parameters, and object poses. We obtain object boxes from the available geometry or annotated dimensions and locate them in the scene using the corresponding poses. Frames are sampled at dataset-specific intervals to reduce redundancy. For Omni6DPose, we additionally discard frames whose annotated poses fail an alignment check against visible surfaces recovered from depth and instance masks.

Coordinate normalization. Consistent geometric processing requires a shared spatial representation. We convert spatial units to centimeters and, for DOPose Gouda et al. (2022), HouseCat6D Jung et al. (2024), and Omni6DPose Zhang et al. (2024), align the upward support-plane normal estimated by RANSAC Fischler & Bolles (1981) with the positive world-Z axis. The same transformation is applied to scene points, camera poses, and object poses. HOPE Lin et al. (2021) and YCB-Video Xiang et al. (2017) use their provided coordinate transformations. Colored point clouds reconstructed from filtered depth observations are resampled to 50,000 points and voxelized at 1 cm resolution for placement annotation.

A.1.2 AUTOMATIC PLACEMENT ANNOTATION

Candidate generation. We search for feasible placements using an upright box with the source object dimensions. Support surfaces are detected using RANSAC Fischler & Bolles (1981), with regions occupied by annotated objects excluded from the candidate surface. Positions are searched at

1 cm resolution over 24 yaw angles at 15° intervals. FFT-based correlation identifies placements with no overlap with scene obstacles while permitting contact with the supporting surface.

Support and visibility filtering. Collision-free candidates must also provide sufficient support and remain observable. We require the entire box footprint to lie within the connected support region containing its bottom center. Morphological closing and hole filling reduce sensitivity to gaps in the observed surface. Candidates are retained only when their boxes lie fully within the camera view and pass the occlusion filter.

Spatial clustering. The remaining placements often contain several valid positions and orientations. We group nearby placements using DBSCAN and store a representative box, all valid centers, and the associated yaw angles for each cluster. The representative favors centers with more valid candidates; ties are resolved using support area, distance to the cluster centroid, and obstacle clearance. Keeping the full center and yaw annotations preserves multiple solutions for subsequent instruction generation.

A.1.3 Spatial Relation Instructions

Relation assignment. We derive spatial relations directly from scene geometry to keep instructions consistent with the placement annotations. Horizontal relations use box footprint centers and axes aligned with the camera view, with eight 45° sectors defining the directional categories. Reference selection considers visibility and spatial proximity. The template "Move [object] to [target relation] [reference B]" describes both the source location and intended placement. Within each cluster, only centers satisfying the target relation are retained.

Instruction rewriting. We use Qwen3-VL Bai et al. (2025) to rewrite these templates into varied, natural instructions. The prompt requires the source object, target reference object, and target relation to be preserved while varying the expression style.

A.1.4 Dataset Statistics and Splits

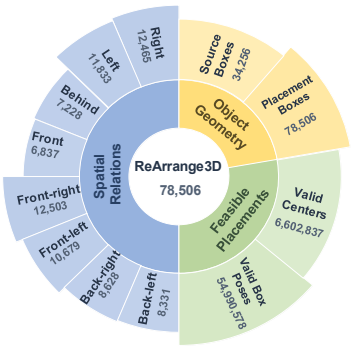

Figure 8: Statistics of ReArrange3D across spatial relations, object geometry and feasible placements

As shown in Figure 8, we summarize ReArrange3D across three aspects: spatial relations, object geometry, and feasible placements. The dataset contains 78,506 instruction-guided instances. Spatial relations cover eight horizontal directions: left, right, front, behind, front-left, front-right, back-left, and back-right. These include both cardinal and diagonal directions, allowing evaluation of instruction following across different relative placement requirements. Object geometry is represented by 34,256 source object boxes and 78,506 representative placement boxes. These annotations describe the spatial extent of the source objects and their target placements, supporting explicit reasoning about whether an object can fit within the available space. Feasible placements comprise 6,602,837 valid centers and 54,990,578 valid box poses, with each pose specified by a center and a valid yaw angle. By providing multiple feasible solutions rather than a single target pose, these annotations capture the freedom in placement position and orientation while accounting for structural support and collisions with the surrounding scene. Together, these three aspects support joint evaluation of instruction consistency and physical feasibility.

The current experimental split divide the dataset into 63,384 training, 7,496 validation, and 7,626 testing instances, drawn from 7,793, 974, and 974 frames, respectively. All instructions from the same frame and source dataset belong to one subset. Validation is used for model selection, and the test set is reserved for final evaluation.

A.2 Experimental Details and Additional Results

We describe the implementation and evaluation settings below, then expand on the method comparisons and ablations reported in the main paper.

A.2.1 Implementation Details

Model Configuration. We use frozen CLIP ViT-B/16 Radford et al. (2021) encoders for visual and language features and sparse 3D convolutions for scene encoding at resolutions of 1, 2, and 4 cm. The hidden dimension is 256, with eight attention heads and dropout of 0.1. Source grounding and BoxSlot field interaction each use four attention layers. The eight highest-scoring regions initialize up to 48 placement queries, refined through four Reasoner layers. Each layer samples 64 locations per candidate and retrieves features at all three resolutions. The first layer uses a cylindrical template, and subsequent layers use oriented box surfaces.

Training Procedure. We pretrain source grounding for 50 epochs and jointly train grounding and placement prediction for up to 65 epochs, with CLIP frozen throughout. AdamW uses an initial learning rate of 10^{-4} and weight decay of 0.01. During joint training, pretrained parameters receive one-tenth of the learning rate applied to new parameters. We use eight NVIDIA RTX 4090 GPUs, with per-GPU batch sizes of 15 and eight for the two stages. A plateau scheduler halves the learning rates when validation PF@1 stops improving. Checkpoints are selected by validation source box mIoU during pretraining and validation PF@1 during joint training.

Inference and Candidate Selection. Each placement box combines a predicted center and yaw with the predicted source dimensions. Yaw uses 12 bins over $[0, \pi)$ because box orientations separated by 180° are equivalent. Candidates are ranked by placement confidence multiplied by the highest yaw score. We retain scores of at least 0.5, falling back to the highest-scoring valid candidate if none pass. Duplicate poses are suppressed when the center distance, normalized along each axis by the box dimensions, is below 0.25 and the circular yaw difference is at most 15°. At most 16 candidates are retained for Top-K evaluation.

A.2.2 Evaluation Protocol

To assess instruction consistency and physical feasibility, we introduce Semantic Relation Correctness (SRC), Support Rate (SR), and Collision-Free Rate (CFR). Each evaluates the highest-ranked prediction. PF@K measures the percentage of instances with at least one top-K candidate satisfying all three criteria simultaneously.

- **Semantic Relation Correctness (SRC).** We compute the relation between the predicted placement box and the annotated reference object using the geometric rules and coordinate conventions used for instruction generation. A prediction is correct when this relation matches the instruction. Evaluating the relation allows different positions to satisfy the same instruction, consistent with the multiple valid solutions provided by ReArrange3D.
- **Support Rate (SR).** For SR and CFR, the evaluation box combines the predicted center and yaw with ground-truth dimensions, separating physical assessment from size prediction. SR requires the full footprint to lie within the eight-connected support region containing its bottom center. This region is obtained from scene voxels between 3 cm below and 1 cm above the box bottom, followed by horizontal projection, 3×3 closing, and hole filling. These operations reduce sensitivity to observation gaps while the full-footprint requirement rejects placements extending beyond the supporting region. The original source location remains in the scene point cloud.
- **Collision-Free Rate (CFR).** We test the evaluation box against all annotated scene object boxes, including the source box at its original location, using the separating axis theorem. A placement passes when no pair has positive-volume overlap; boundary contact alone is permitted. Annotated boxes account for object extent that may be partially missing from depth observations. SR and CFR report the percentages of instances passing their respective checks.

A.2.3 Comparison with Existing Methods

We fine-tune RoboPoint Yuan et al. (2024), RoboRefer Zhou et al. (2026), and RoboBrain2.5 Tan et al. (2026) on the same ReArrange3D training split using LoRA, with point supervision derived from annotated box bottom centers. For geometric evaluation, predicted image points are back-projected using observed depth and camera calibration, then combined with ground-truth source dimensions and the observed source yaw. This allows point predictions to be assessed under the same support and collision criteria. Source grounding is compared against MCLN, TSP3D, and EG-3DVG.

Table 4: Quantitative comparison of placement feasibility across the five data sources in ReArrange3D. Values denote PF@1 (%). Overall aggregates the evaluated instances across all five sources. Higher is better.

Methods	DOPose	HOPE	HouseCat6D	Omni6DPose	YCB-Video	Overall
RoboPoint-7B	32.80	38.82	43.47	39.32	31.72	39.47
RoboRefer-2B	35.59	35.53	46.26	41.28	41.52	40.13
RoboBrain2.5-4B	39.08	41.72	46.11	47.34	40.27	46.44
TopoBox (Ours)	**65.18**	**62.82**	**69.81**	**63.89**	**72.46**	**65.16**

Re-arrangement Feasibility. Table 1 shows that similar semantic accuracy can accompany substantially different physical feasibility. RoboBrain2.5 achieves 69.84% SRC, close to the 70.12% of TopoBox, but its PF@1 is 46.44%, compared with 65.16%. TopoBox also improves SR and CFR by 19.11 and 16.13 percentage points. Even with ground-truth dimensions supplied during evaluation, the baseline positions frequently fail support or collision checks. These results support the need to account for the complete object extent and its interaction with surrounding geometry when selecting a placement.

Source Object Grounding. Source grounding supplies the geometry used for placement prediction and topology reasoning: errors in box dimensions alter the spatial extent considered during refinement. Table 2 examines this prerequisite. TopoBox achieves 74.68% Acc@0.5 and 61.11% mIoU, exceeding EG-3DVG by 1.09 and 1.05 percentage points. The comparison demonstrates competitive source box estimation against dedicated grounding methods while jointly addressing placement prediction.

Results by Data Source. Table 4 compares placement feasibility across the five data sources in ReArrange3D. TopoBox exceeds the strongest baseline on each source by 16.55–30.94 percentage points in PF@1. The relative ranking of the baselines varies across sources, whereas TopoBox maintains its advantage throughout. The largest gain occurs on YCB-Video, where TopoBox reaches 72.46% compared with 41.52% for the strongest baseline. Even on Omni6DPose, where the gap is smallest, TopoBox improves PF@1 from 47.34% to 63.89%. Because PF@1 requires semantic correctness, stable support, and collision constraints to hold simultaneously, these gains show that TopoBox more reliably satisfies the combined re-arrangement requirements across all five sources.

A.2.4 Ablation Design and Analysis

We construct five variants to examine the contribution of each component. Replacing BoxSlot ranking with spatial farthest point sampling retains eight regions and up to 48 queries, testing the value of learned region selection. Disabling local geometry while retaining language features and prediction updates tests whether refinement needs object–environment interaction evidence. Fixing sampling centers and yaw while allowing predictions to evolve tests whether features must be sampled again around updated boxes. Using the same number of interior samples instead of surface samples tests the value of observing object boundaries. Finally, fine-grained retrieval examines the contribution of multiple spatial resolutions.

Table 3 shows the largest PF@1 decrease when BoxSlot guidance is replaced, from 65.16% to 58.63%. Removing local geometry reduces SR and CFR by 5.64 and 4.55 percentage points, while SRC changes by only 0.11 points, linking the geometry branch primarily to physical feasibility. Fixed sampling lowers PF@1 to 62.12%, supporting the use of updated poses to gather relevant geometric evidence. Interior sampling and fine-grained retrieval yield 63.60% and 63.88%, respectively. The latter also reduces feature samples per candidate per layer from 192 to 64, so this difference cannot be attributed solely to spatial resolution.

A.2.5 Additional Qualitative Results

We provide additional qualitative comparisons, examples of iterative topology refinement, and visualizations of BoxSlot predictions, as shown in figure 9, 10 and 11. The comparisons cover all eight horizontal spatial relations and show complete predicted boxes to expose failures in structural support and collision constraints. The refinement examples illustrate how successive pose updates change contact with the supporting surface, overlap with obstacles, and containment within the available free space. The BoxSlot visualizations show how high-scoring locations guide placement candidate initialization within instruction-relevant regions for subsequent topology reasoning.

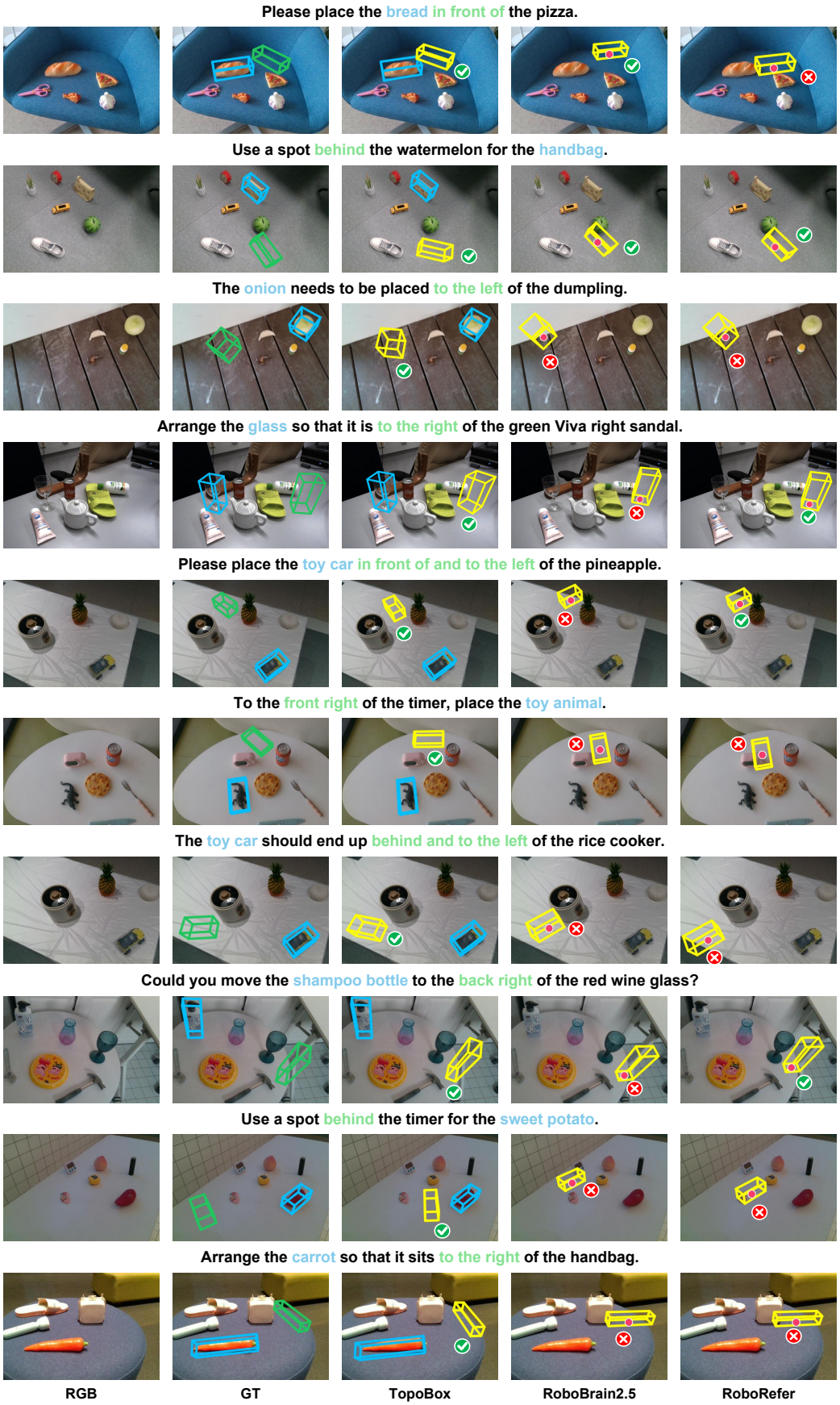

Figure 9: **Additional qualitative comparisons with counterparts across all eight spatial relations.**

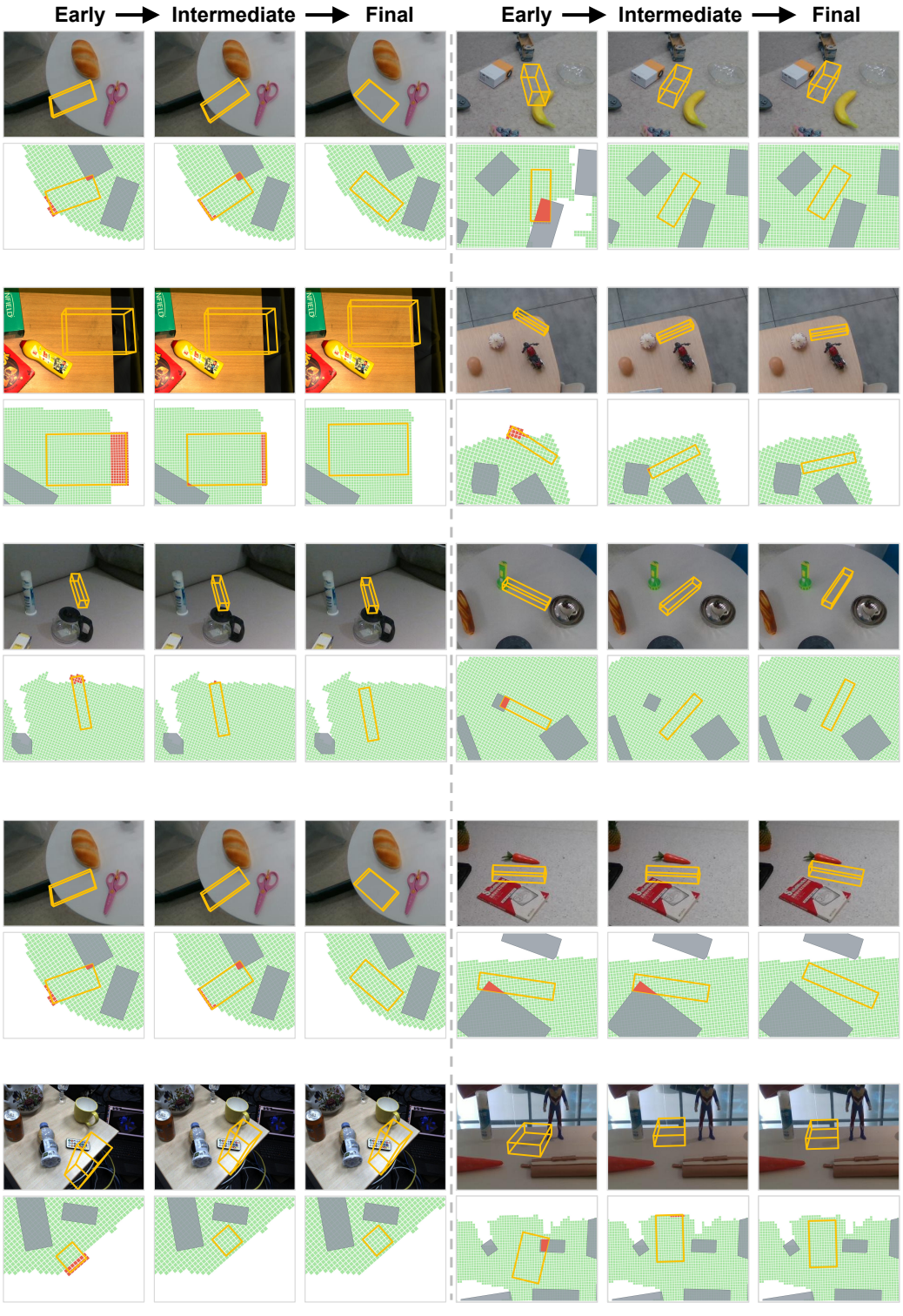

Figure 10: **Additional qualitative illustration of iterative topology refinement**. (a) Towards Structural Support Connectivity illustrates the evolution of contact with the supporting surface. (b) Towards Collision-Free Connectivity illustrates the evolution of obstacle collision and box containment within the available free space.



Figure 11: **Additional qualitative visualization of BoxSlot predictions and their guidance for placement candidate initialization.** Blue and green boxes show ground-truth source and placement boxes. Warmer colors indicate higher BoxSlot scores. Yellow boxes are centered in the eight highest-scoring regions and use the predicted source dimensions and orientation to visualize the corresponding slots. These slots guide placement candidate initialization within instruction-relevant regions for subsequent topology reasoning.