DEPTH-AWARE VISUAL MANIPULATION RELATIONSHIP REASONING WITH GATED DECODER-LAYER AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object retrieval in cluttered scenes depends on removal order, since an object that appears graspable may still support or constrain its neighbors. Inferring this order from a single observation requires accurate object localization and recovery of a sparse, directed relationship graph over many possible ordered pairs. We propose a depth-aware model for visual manipulation relationship reasoning that jointly detects objects and predicts this graph from aligned RGB-D input. Specifically, DFormerV2-S produces multi-scale RGB-D features, while Deformable DETR generates object queries that are progressively refined across six decoder layers. For each ordered query pair, our relation head constructs a directed representation from every layer and combines the resulting evidence with pair-dependent sigmoid gates. Under the unified end-to-end protocol on the constructed ManipRelaD dataset, the model achieves 31.68% triplet F1 and 13.91% full-scene accuracy, improving upon the strongest compared baselines by 16.34 and 3.21 percentage points, respectively.

1 INTRODUCTION

Object retrieval in cluttered scenes depends on more than object visibility or graspability. An apparently accessible object may support or mechanically constrain neighboring objects, making removal order essential for reliable manipulation. VMRN Zhang et al. (2018) represents such dependencies as a directed graph whose nodes denote objects and whose edges encode pairwise removal precedence. This representation complements conventional object detection because a robot must infer not only which objects are present and where they are, but also how their physical arrangement constrains the order of manipulation.

Inferring such a graph from a single RGB-D observation presents two coupled challenges. Occlusion complicates object localization, while relationship reasoning must recover a sparse, directed graph. Only a small fraction of ordered object pairs form valid edges, and reversing their order yields a distinct candidate edge. Aligned depth provides complementary geometric information about object configurations. DFormerV2 Yin et al. (2025), for example, uses depth as a geometric prior for visual representation learning. Exploiting this information for relationship prediction requires object representations that preserve fused RGB-D information. Existing approaches use region-pair features Zhang et al. (2018), explicit graph propagation Zuo et al. (2021); Ding et al. (2022), or Transformer decoder features Rabino & Tommasi (2025); Im et al. (2024) for relationship reasoning. In object-query detectors, decoder layers progressively refine representations for localization and contextual reasoning. The relevance of the resulting layer-wise relation evidence can vary across ordered object pairs, motivating pair-dependent gates that aggregate directed pair representations across decoder layers.

We therefore propose a depth-aware object-query model that jointly detects objects and predicts a directed manipulation relationship graph from a single RGB-D observation. DFormerV2-S Yin et al. (2025) encodes the aligned RGB-D input, and a Deformable DETR decoder Zhu et al. (2021) progressively refines a fixed set of object queries. At each decoder layer, layer-specific subject and object projections construct a directed representation from the hidden states of every ordered query pair. Pair-dependent gates weight these layer-wise representations, and a lightweight relation MLP predicts each edge from their gated sum. The shared decoder states thereby propagate fused RGB-D cues to the relation head.

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.

Our main contributions are summarized as follows.

(1) Depth-aware object-query framework. We develop a joint framework that encodes aligned RGB-D observations, detects objects with learned queries, and predicts directed manipulation relationships from the resulting decoder states. These depth-conditioned states provide shared representations for object localization and directed edge prediction.

(2) Gated decoder-layer aggregation. We introduce a relation reasoning mechanism that uses layer-specific subject and object projections to construct directed pair representations from the hidden query states at every decoder layer. Pair-dependent gates aggregate these representations across decoder layers before edge prediction.

2 Related Work

2.1 RGB-D Object Detection and Representation

Transformer-based object detectors formulate detection as set prediction over a fixed collection of object queries. DETR Carion et al. (2020) matches learned queries to ground-truth objects through bipartite matching and predicts the complete detection set in parallel. Deformable DETR Zhu et al. (2021) replaces dense attention over feature maps with sparse attention around reference points and operates on multi-scale features, improving convergence and small-object detection. Successive decoder layers refine the queries used for classification and localization, yielding a sequence of instance representations at different stages of the decoding process.

Depth complements RGB appearance with geometric information, and RGB-D methods differ primarily in how they encode and fuse the two modalities. Gupta et al. (2014) encode depth using a hand-crafted geocentric representation and combine it with RGB features for object detection and segmentation. DFormer Yin et al. (2024) instead learns joint RGB-D representations through RGB-D pretraining and dedicated multimodal blocks. DFormerV2 Yin et al. (2025) treats depth as a geometric prior and uses depth cues and spatial distances between image tokens to guide self-attention. Our model employs its small variant, DFormerV2-S, as the RGB-D backbone of a Deformable DETR detector and uses the hidden query states from all decoder layers to construct directed object-pair representations for manipulation relationship reasoning.

2.2 Visual Manipulation Relationship Reasoning

Visual manipulation relationship reasoning has progressed from pairwise feature extraction to graph-based scene modeling. VMRN Zhang et al. (2018) uses object-pairing pooling to learn directed relationships jointly with object detection, while fully connected CRF methods Yang et al. (2018); Ding et al. (2025b) impose scene-level consistency on pairwise predictions. GVMRN Zuo et al. (2021) applies graph convolution, GGNN-VMRN Ding et al. (2022) uses gated graph propagation, and GSAGED Luo et al. (2024) combines neighborhood sampling with feature aggregation to capture local and global context. Multi-view methods use complementary observations and, in some cases, explicit 3D geometry. MMRDN Wang et al. (2023) combines multi-view image and point-cloud representations, while Dual-GAT Ding et al. (2025a) models object- and relation-level interactions after multi-view feature fusion. Our model instead predicts directed manipulation relationships from a single aligned RGB-D observation and shares depth-aware object-query representations between object detection and relationship prediction.

Recent Transformer-based methods integrate object and relationship prediction within query-based detectors. Adj-Net in DUQIM-Net Tchuiev et al. (2022) augments a Transformer detector with an adjacency head for hierarchical relationship prediction. D3G Rabino & Tommasi (2025) constructs directed edge features from the final decoder queries and refines them with a graph transformer, whereas EF-VMR^{2}N Chen et al. (2026) predicts a fixed-size set of sparse relation triplets. O2O-I-Grasp Liu et al. (2026) uses interaction queries to jointly predict two objects and their relationship. For scene graph generation, EGTR Im et al. (2024) extracts relation features from the self-attention query and key representations of every decoder layer together with the final hidden states. D2TriPO-DETR Pu et al. (2026) gates query–key (Q–K) pair features and output-query pair features within each decoder layer. In contrast, our model constructs directed pair representations solely from the

hidden object-query states of all decoder layers and applies pair-dependent gates across layers to aggregate their relation evidence.

3 METHOD

Figure 1 summarizes the complete architecture. The model shares RGB-D-conditioned object-query representations between object detection and directed relationship prediction. After detector pretraining, the RGB-D backbone, object detector, gated decoder-layer aggregation module, and relation head are jointly optimized.

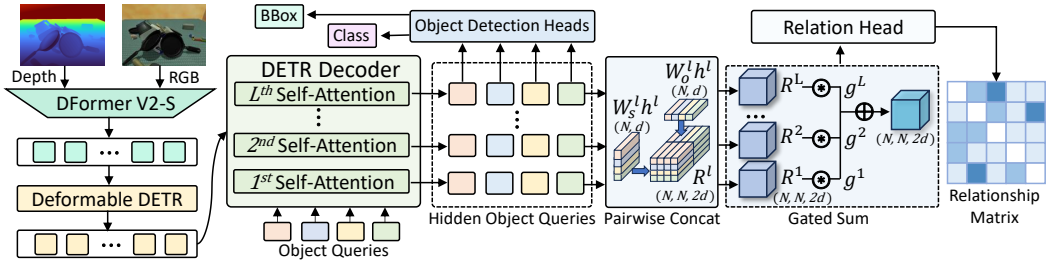


Figure 1: Overall framework. DFormerV2-S and Deformable DETR encode an aligned RGB-D observation and refine a fixed set of object queries. The final-layer queries feed the object detection heads, while hidden queries from every decoder layer form ordered pair representations R^ℓ. Pair-dependent gates g^ℓ combine these representations into $\bar{R}$, from which the relation head predicts the $N \times N$ directed manipulation relationship matrix.

3.1 PROBLEM FORMULATION

Let $I \in \mathbb{R}^{H\times W\times 3}$ and $D \in \mathbb{R}^{H\times W}$ denote an aligned RGB image and depth map, where H and W are the image height and width. A scene containing M object instances is represented by a directed manipulation relationship graph $G = (V, E)$, where $V = \{v_i\}_{i=1}^{M}$ is the set of object nodes and E is the set of directed edges. For two distinct object indices $i, j \in \{1, \dots, M\}$, an edge $(i, j) \in E$ indicates that object i should be removed before object j. The graph is encoded by a binary adjacency matrix $A \in \{0, 1\}^{M\times M}$, where $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ii} = 0$. Given (I, D), the task is to detect the object instances and predict their manipulation relationship graph.

3.2 RGB-D OBJECT QUERY EXTRACTION

RGB-D Feature Encoding. We use DFormerV2-S Yin et al. (2025) to encode the aligned RGB image and depth map. RGB is processed by the hierarchical backbone, while the aligned depth map is resized at each stage to construct spatial-depth geometry priors that guide attention. We retain the outputs of the last three stages, with strides 8, 16, and 32 and channel dimensions 128, 256, and 512, respectively. Each output is projected to a common dimension $d = 256$. An additional feature level with stride 64 is generated from the lowest-resolution backbone output, yielding four feature levels for object detection.

Object Query Decoding. The projected feature maps are processed by a Deformable DETR encoder–decoder Zhu et al. (2021). The detector contains six encoder layers, $L = 6$ decoder layers, and $N = 200$ object queries. Each decoder layer updates the query representations through query self-attention, multi-scale deformable cross-attention, and a feed-forward network. The final decoder outputs are mapped to category logits and normalized bounding boxes. We denote the query hidden states produced by decoder layer ℓ as

$$H^\ell = \left[h_1^\ell, \dots, h_N^\ell\right] \in \mathbb{R}^{N\times d}, \qquad \ell = 1, \dots, L, \tag{1}$$

where $h_q^\ell \in \mathbb{R}^d$ is the hidden state of query $q \in \{1, \dots, N\}$ at decoder layer ℓ.

3.3 GATED DECODER-LAYER AGGREGATION

Layer-Wise Directed Pair Construction. For every ordered query pair $(q, t) \in \{1, \ldots, N\}^2$, we construct one directed pair representation from each decoder layer. Layer-specific subject and object projections map the corresponding query states to a shared feature space:

$$r_{qt}^{\ell} = \left[W_s^{\ell} h_q^{\ell} \parallel W_o^{\ell} h_t^{\ell}\right] \in \mathbb{R}^{2d}, \tag{2}$$

where $W_s^{\ell}, W_o^{\ell} \in \mathbb{R}^{d \times d}$ are learned subject and object projections, respectively, and $\parallel$ denotes concatenation. Stacking the representations of all ordered pairs at decoder layer ℓ gives $R^{\ell} \in \mathbb{R}^{N \times N \times 2d}$, whose (q, t)-indexed feature vector is r_{qt}^{ℓ}. Repeating this construction across the decoder produces $\{R^{\ell}\}_{\ell=1}^{L}$. The distinct subject and object projections preserve edge direction, so reversing the order of a query pair generally produces a different representation.

Pair-Dependent Gated Sum. A shared linear gate maps each layer-wise pair tensor to a gate tensor:

$$g^{\ell} = \sigma\big(R^{\ell} W_g + b_g\big) \in (0, 1)^{N \times N \times 1}, \tag{3}$$

where $W_g \in \mathbb{R}^{2d \times 1}$ and $b_g \in \mathbb{R}$ are the learnable weight matrix and bias of a linear projection shared across all decoder layers and ordered query pairs. The sigmoid $\sigma(\cdot)$ is applied element-wise, and g_{qt}^{ℓ} denotes the gate value for ordered query pair (q, t). The layer-wise tensors are aggregated as

$$\bar{R} = \sum_{\ell=1}^{L} g^{\ell} \odot R^{\ell} \in \mathbb{R}^{N \times N \times 2d}, \tag{4}$$

where $\odot$ denotes element-wise multiplication with g^{ℓ} broadcast along the feature dimension. The (q, t)-indexed feature vector of $\bar{R}$ is $\bar{r}_{qt} = \sum_{\ell=1}^{L} g_{qt}^{\ell} r_{qt}^{\ell} \in \mathbb{R}^{2d}$. The decoder layers represent successive stages of query refinement, so a single fixed layer need not provide the most informative evidence for every ordered pair. The layer-specific projections align these stage-dependent states in a common directed pair space, while the shared gate evaluates the relevance of each source from the pair representation itself. Unlike a softmax across layers, the independent sigmoid coefficients allow several decoder stages to contribute simultaneously and suppress sources that are less informative for a particular pair. The gated sum also preserves a fixed $2d$-dimensional representation regardless of the number of decoder layers, enabling the relation head to use the full query-refinement trajectory without introducing additional relation queries.

Directed Relation Prediction. The aggregated representation is processed by a three-layer MLP f_{rel}, with ReLU activations in the first two layers and a scalar relation logit produced by the last layer. The directed relation score and probability are

$$s_{qt} = f_{\text{rel}}(\bar{r}_{qt}), \qquad p_{qt} = \sigma(s_{qt}), \tag{5}$$

where $s_{qt} \in \mathbb{R}$ and $p_{qt} \in (0, 1)$ are the logit and probability for the directed relation from query q to query t, respectively. Applying the relation head to all ordered query pairs produces a matrix of directed relationship probabilities. At inference, the manipulation relationship matrix over the retained object detections is obtained with a relation threshold η:

$$\hat{A}_{qt} = \begin{cases} \mathbb{I}(p_{qt} \geq \eta), & q \neq t, \\ 0, & q = t, \end{cases} \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\hat{A}_{qt} = 1$ denotes the predicted directed relation from the object associated with query q to the object associated with query t.

3.4 TRAINING OBJECTIVE

After detector pretraining with object-category and bounding-box annotations, relationship training jointly updates the RGB-D backbone, detector, and relation head. A Hungarian assignment matches ground-truth objects to object queries using classification, box ℓ_1, and generalized IoU (GIoU) costs Carion et al. (2020); Rezatofighi et al. (2019). Let $\pi : \{1, \ldots, M\} \to \{1, \ldots, N\}$ map each

ground-truth object index to its matched query index. For one scene, the relationship loss is binary cross-entropy over the matched, non-diagonal object pairs:

$$\mathcal{L}_{\text{rel}} = \frac{1}{\max\{1, M(M-1)\}} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \text{BCEWithLogits}\big(s_{\pi(i)\pi(j)}, A_{ij}\big)\,. \tag{7}$$

The denominator is set to one when $M = 1$; the empty sum then gives zero relationship loss because the scene has no valid off-diagonal pair. The complete objective combines the focal classification loss Lin et al. (2017), box ℓ_1 loss, GIoU loss, and relationship loss:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{focal}} + \lambda_{\text{box}}\mathcal{L}_{\ell_1} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}. \tag{8}$$

Here, $\mathcal{L}_{\text{focal}}$ supervises object categories, $\mathcal{L}_{\ell_1}$ and $\mathcal{L}_{\text{giou}}$ supervise normalized bounding boxes, and the four λ terms are their loss coefficients. Their values for the two training stages are provided in Appendix B.

4 Experiments and Evaluation

4.1 Experimental Setup and Dataset Statistics

The ManipRelaD dataset contains 3,036 simulated cluttered tabletop scenes with 1 to 30 objects. The scenes are constructed from 142 OmniObject3D assets spanning 140 semantic categories. Each scene is observed from ten surrounding viewpoints, yielding 30,360 aligned RGB-D image pairs with instance segmentation, 6D object poses, and a directed manipulation relationship graph (MRG). The MRGs are derived automatically from inter-object contact forces recorded after each scene has settled; contact-force measurements are used only to construct the graph annotations and are not required at test time. We partition the dataset by scene into approximately 80%, 10%, and 10% training, validation, and test splits, respectively, and assign all ten views of each scene to the same split. Further details on dataset construction and annotation are provided in Appendix A.

4.2 Benchmark Setup and Baseline Evaluation

4.2.1 Implementation Details

We implement the model in PyTorch and train it in two stages. The DFormerV2-S and Deformable DETR detector is first optimized for at most 50 epochs with an effective batch size of 2. The complete model is then initialized from the pretrained detector and jointly optimized for up to 30 epochs, updating the RGB-D backbone, detector, and relation head together. Both stages use AdamW. Detailed optimization settings, initialization, preprocessing, and checkpoint selection are provided in Appendix B.

4.2.2 Evaluation Metrics

We use a unified end-to-end protocol that evaluates both object detection and directed relationship prediction. Detection is measured by class-averaged AP at an IoU threshold of 0.5 using continuous all-point interpolation (Det. mAP@0.5). After one-to-one object matching, object-pair precision (OP), object-pair recall (OR), and F1 are computed over positive directed subject–relation–object triplets. A predicted triplet is correct only when both object detections are matched to the corresponding ground-truth instances and the directed relationship is correct. Image accuracy (IA) measures the fraction of images whose predicted positive triplet set exactly matches the ground truth. Full-scene accuracy (SceneAcc) additionally requires all objects in an image to be matched one-to-one and every ordered-pair relationship label to be correct. Appendix C.1 reports complementary results under a geometry-aligned pairwise edge protocol.

4.2.3 Comparison with Existing Methods

Evaluation Protocol All evaluated methods are trained on the same training partition and evaluated on the same balanced test manifest of 3,300 RGB-D samples, with 110 samples for each object count from 1 to 30. For Det. mAP@0.5, detections are filtered at a confidence threshold of 0.001, followed

by class-agnostic non-maximum suppression (NMS) at an IoU threshold of 0.7, and at most 200 detections are retained per image. For relationship evaluation, the confidence threshold is 0.1 and at most 30 detections are retained after the same NMS. These predictions are sorted by detection confidence and greedily matched one-to-one to unused ground-truth instances of the same class, with box IoU strictly greater than 0.5. Relationships are evaluated at a threshold of 0.5 over the matched off-diagonal directed object pairs.

Adaptation of Comparison Methods We reproduce three representative manipulation relationship detection methods: VMRN Zhang et al. (2018), D3G Rabino & Tommasi (2025), and EGTR Im et al. (2024). Each method retains its native architecture and input modality, while its outputs are converted to the same binary directed-relation representation. All methods use the common test-time filtering, matching, and relation threshold described above. Appendix B provides training and baseline configuration details.

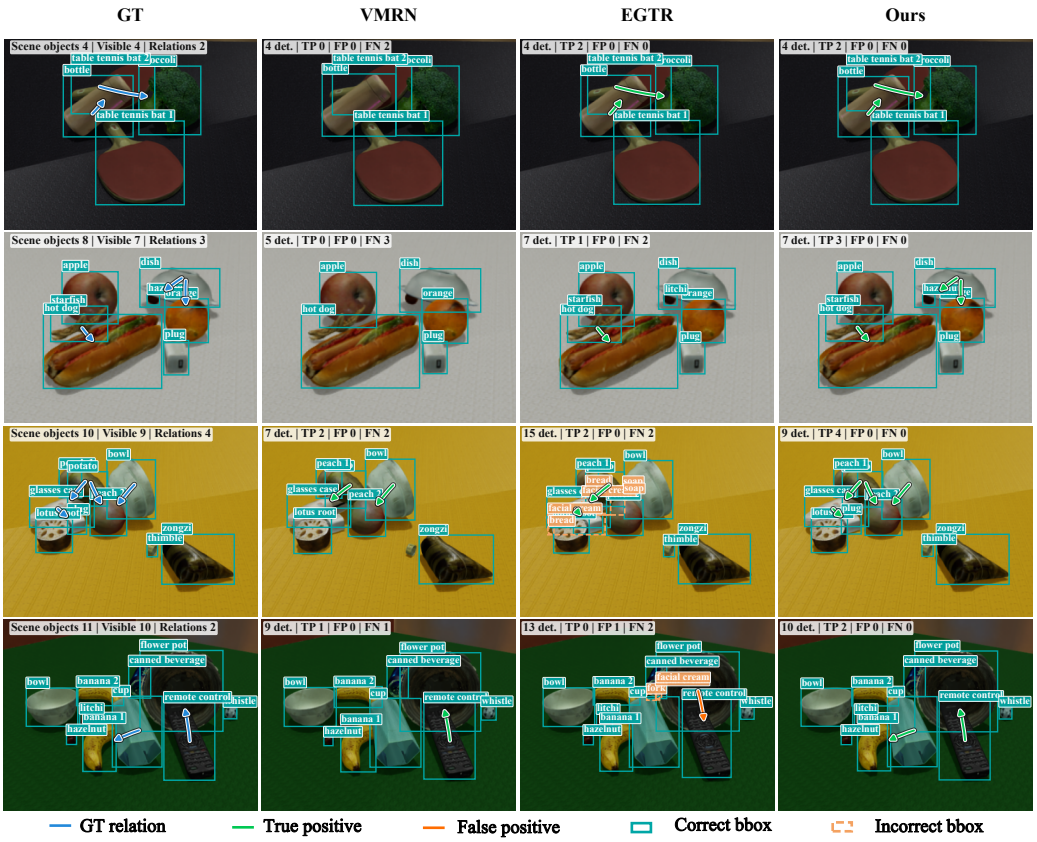


Figure 2: Qualitative comparison on four selected ManipRelaD test scenes containing 4, 8, 10, and 11 objects. Columns show the ground truth, VMRN, EGTR, and our method. Ground-truth headers report the numbers of scene objects, visible objects, and annotated relationships; prediction headers report detections and relation true positives (TP), false positives (FP), and false negatives (FN).

Overall Performance Table 1 shows that our method achieves the strongest overall end-to-end result, with 90.60% Det. mAP@0.5, 31.68% triplet F1, and 13.91% SceneAcc. Relative to the strongest baseline for each metric, our method improves triplet F1 by 16.34 percentage points over EGTR and SceneAcc by 3.21 points over VMRN. The F1 gain is driven by a triplet recall of 23.58%, compared with 8.95% for EGTR, while maintaining 48.23% precision. VMRN attains 59.05% precision with 8.76% recall. These results show that our method recovers substantially more valid directed relationships and provides a stronger precision–recall balance under strict end-to-end matching. Qualitative examples of the corresponding object and relationship predictions are shown in Figure 2. The D3G result also illustrates why IA should be interpreted together with SceneAcc. At the fixed relation threshold, D3G produces no positive triplets but obtains 14.79% IA because IA is

satisfied for images in which both the predicted and ground-truth positive triplet sets are empty. Its 0% SceneAcc shows that these cases do not constitute complete end-to-end scene predictions.

Table 1: End-to-end comparison on the ManipRelaD dataset.

Method	mAP@0.5 (%)	OP (%)	OR (%)	F1 (%)	IA (%)	SceneAcc (%)
VMRN	61.06	**59.05**	8.76	15.25	**16.24**	10.70
D3G	0.00	0.00	0.00	0.00	14.79	0.00
EGTR	78.27	53.43	8.95	15.34	15.52	9.67
Ours	**90.60**	48.23	**23.58**	**31.68**	**16.24**	**13.91**

Qualitative Comparison Figure 2 mirrors the aggregate recall differences reported in Table 1. In the four-object scene, EGTR and our method recover both annotated relationships, whereas VMRN misses both. As the scenes become denser, our method recovers all three relationships in the eight-object scene and all four in the ten-object scene; EGTR recovers one and two, respectively, while VMRN recovers none and two. In the ten-object example, our method also matches all nine visible objects, whereas six of EGTR's detections remain unmatched. In the eleven-object scene, our method recovers both annotated relationships without a false positive, while EGTR misses both and predicts one spurious edge. These examples show that the recall gain corresponds to recovering more complete sparse relation sets, while also including a simpler case in which EGTR is already correct.

Object-Count Sensitivity Figure 3 compares exact-scene performance across the 30 equally sized object-count subsets. For scenes with 2–10 objects, our method achieves an average SceneAcc of 33.94%, compared with 24.65% for VMRN and 21.41% for EGTR. Over the 11–20 object range, our method retains the highest mean SceneAcc at 1.18%; all methods reach zero SceneAcc from 16 objects onward under this exact criterion. IA compares positive triplet sets rather than full object-and-relation correctness, which explains why it can remain higher than SceneAcc. The decline in SceneAcc follows from its exactness: as the number of objects grows, more detections and ordered-pair labels must all be correct simultaneously, and any object or relationship error invalidates the scene. The trend therefore reflects the increasing difficulty of the complete prediction task rather than an isolated change in relation classification.

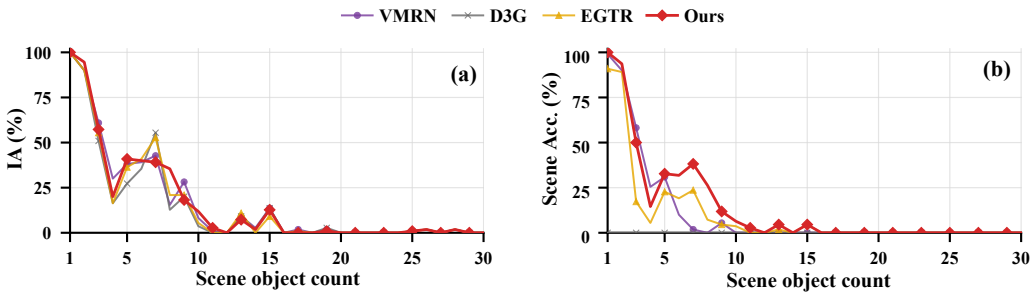


Figure 3: IA and SceneAcc by object count under the unified end-to-end protocol. Each point aggregates 110 ManipRelaD test samples.

4.2.4 Ablation Studies

We evaluate how relation-source construction, decoder-layer aggregation, relation-head capacity, and RGB-D encoding affect the final performance. Table 2 summarizes the results, with the full model included as the reference.

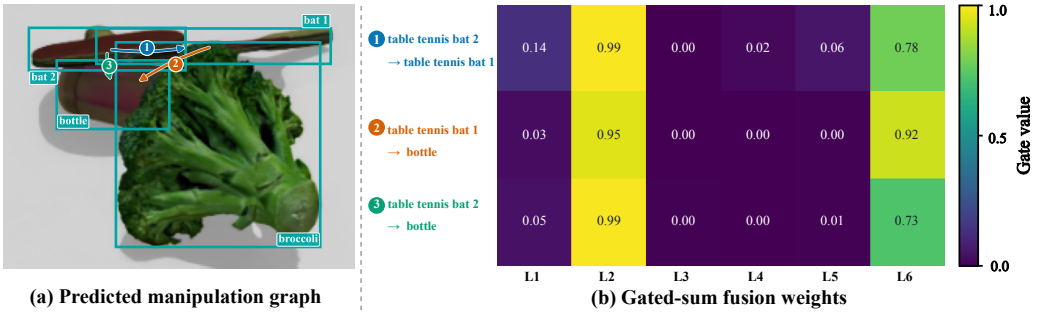
RGB-D Encoding and Relation Head The DFormerV2-S RGB-D configuration outperforms the ResNet-50 RGB alternative by 13.04 points in Det. mAP@0.5 and 14.40 points in triplet F1. Because this comparison changes both the input modality and the backbone, it supports the complete depth-aware RGB-D encoder configuration rather than isolating the effect of depth alone. For relation

Table 2: Component ablations on the ManipRelaD dataset.

Ablation	mAP@0.5 (%)	OP (%)	OR (%)	F1 (%)	IA (%)
ResNet-50 RGB encoder	77.56	27.94	12.50	17.28	14.18
Q–K pairs + final query pair	90.32	45.61	20.54	28.32	15.88
Final decoder layer only	90.65	49.36	22.22	30.65	15.97
Linear relation head	90.63	42.63	20.88	28.03	15.39
Full model	90.60	48.23	23.58	31.68	16.24

prediction, the three-layer relation MLP improves both precision and recall over a linear head, increasing OP by 5.61 points, OR by 2.70 points, and F1 by 3.64 points. Together, these ablations show that the depth-aware DFormerV2-S backbone provides a stronger basis for joint detection and relationship prediction, while the nonlinear relation head further improves the precision–recall balance.

Relation Sources and Decoder-Layer Aggregation Combining layer-wise Q–K pair features with an additional final-query pair yields a triplet F1 of 28.32%. Our layer-wise hidden-state construction improves OP, OR, and F1 by 2.62, 3.04, and 3.36 percentage points, respectively, while Det. mAP@0.5 changes by only 0.28 points. Compared with using only the final decoder layer, aggregating hidden-state pair representations from all decoder layers further increases F1 from 30.65% to 31.68% and OR from 22.22% to 23.58%, with a 0.06-point difference in detection mAP. These comparisons support the layer-wise hidden-state relation sources used in the full model and show that the relation gains are not explained by improved object detection.



(a) Predicted manipulation graph (b) Gated-sum fusion weights

Figure 4: Visualization of gated decoder-layer aggregation in one ManipRelaD test scene. (a) Predicted manipulation graph. (b) Raw sigmoid gate coefficients for three correctly predicted relationships across six decoder layers.

Figure 4 visualizes the independent sigmoid gate coefficients for three correctly predicted relationships in one test scene. The gates emphasize layers 2 and 6 while largely suppressing layers 3–5, illustrating non-uniform decoder-layer aggregation in this example.

We aggregate the six independent sigmoid gates over all 1,006,248 retained directed object pairs across the 3,300 test samples. The mean gate values from decoder layers 1 through 6 are 0.317, 0.997, 0.029, 0.059, 0.219, and 0.587, respectively. Layer 2 has the largest gate value for 99.20% of the pairs, showing a strong global preference for this decoder layer. Nevertheless, the broad percentile ranges for layers 1, 5, and 6 in Figure 5 indicate that the independently gated sources retain pair-dependent variation. The largest-gate frequency should not be interpreted as a decomposition of prediction importance because the effect of each layer also depends on its projected pair representation and the downstream relation MLP.

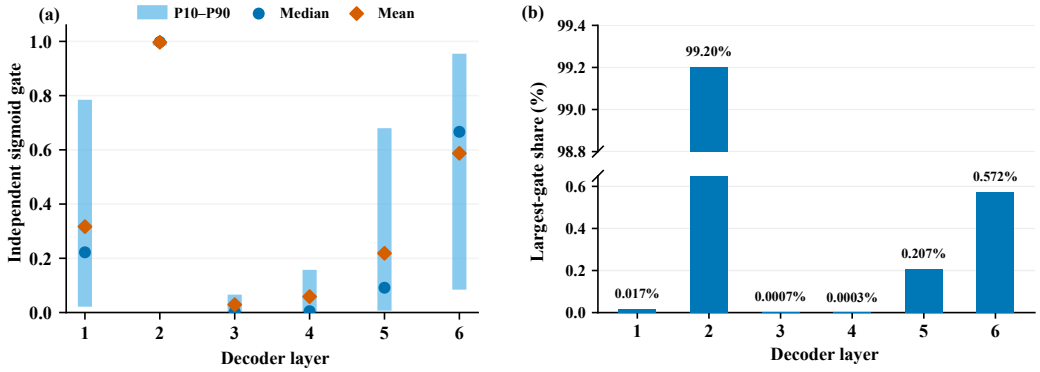


Figure 5: Dataset-level statistics of the six independent sigmoid gates across the 3,300 ManipRelaD test samples. (a) Mean, median, and 10th–90th percentile range for each decoder layer. (b) Percentage of directed object pairs for which each layer has the largest gate value, shown with a broken y-axis. Labels report the exact percentages.

5 Conclusion

This paper presents a depth-aware object-query framework for visual manipulation relationship reasoning from a single RGB-D observation. The framework uses shared RGB-D-conditioned object-query states for detection and relationship prediction, constructs directed pair representations from every decoder layer, and combines them with pair-dependent gates before edge prediction. Under the unified end-to-end protocol on the ManipRelaD dataset, the model achieves 31.68% triplet F1 and 13.91% SceneAcc, improving upon the strongest compared baseline for each metric by 16.34 and 3.21 percentage points, respectively. The F1 improvement is driven by higher relationship recall, while precision remains below that of VMRN and EGTR. The ablation results further show that aggregating hidden-state pairs from all decoder layers increases F1 from 30.65% to 31.68% over final-layer-only reasoning, supporting the use of layer-wise relation evidence.

The current study leaves two practical challenges open. First, exact full-scene prediction becomes increasingly difficult as the number of objects grows, with all evaluated methods reaching zero SceneAcc from 16 objects onward. Second, the experiments use simulated, static single-view RGB-D observations and therefore do not evaluate robustness to real-sensor noise and domain shift or the effect of relationship prediction on downstream grasp execution. Future work will evaluate the method with real RGB-D sensors and robot trials, update the predicted relationship graph after each manipulation action, and investigate temporal or actively acquired multi-view observations for highly occluded scenes.

AI use statement

Generative AI tools were used to polish the manuscript for clarity and readability. They were also used to identify potentially relevant literature for the related work. The authors manually screened all candidate papers, read the original publications, and verified the descriptions of the selected studies. In addition, generative AI tools assisted in refining the color palettes used to visualize experimental results. All plotted values were obtained from verified experimental outputs and were not altered during this process. The authors reviewed all AI-assisted text and visual artifacts and take full responsibility for the final content of the paper.

References

Seunghyeok Back, Joosoon Lee, Kangmin Kim, Heeseon Rho, Geonhyup Lee, Raeyoung Kang, Sangbeom Lee, Sangjun Noh, Youngjin Lee, Taeyeop Lee, and Kyoobin Lee. Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes. *IEEE Robotics and Automation Letters*, 10:10498–10505, 2025. doi: 10.1109/LRA.2025.3601045.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.

Lu Chen, Zhuomao Li, Zhou Zhao, Zhenyu Lu, Huaiyao Wang, and Chenguang Yang. Efficient visual manipulation relationship reasoning with relationship attention and sparse graph in robotic grasping. *IEEE Transactions on Automation Science and Engineering*, 23:5714–5731, 2026. doi: 10.1109/TASE.2026.3667740.

Lennart Demes. AmbientCG. `https://ambientcg.com`, 2022. Online; accessed 2025.

Mengyuan Ding, Yaxin Liu, Chenjie Yang, and Xuguang Lan. Visual manipulation relationship detection based on gated graph neural network for robotic grasping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1404–1410, 2022. doi: 10.1109/IROS47612.2022.9981077.

Mengyuan Ding, Yaxin Liu, Yaorui Shi, Xuguang Lan, and Nanning Zheng. Dual graph attention networks for multi-view visual manipulation relationship detection and robotic grasping. *IEEE Transactions on Automation Science and Engineering*, 22:13694–13705, 2025a. doi: 10.1109/TASE.2025.3555206.

Mengyuan Ding, Chenjie Yang, Jianyi Liu, Xuguang Lan, and Nanning Zheng. Relationship detection for manipulation in object stacking scene with fully connected crf. *Neurocomputing*, 611:128661, 2025b. doi: 10.1016/j.neucom.2024.128661.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 11444–11453, 2020.

Maximilian Gilles, Yuhao Chen, Emily Zhixuan Zeng, Yifan Wu, Kai Furmans, Alexander Wong, and Rania Rayyes. Metagraspnetv2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping. *IEEE Transactions on Automation Science and Engineering*, 21(3):2302–2320, 2024. doi: 10.1109/TASE.2023.3328964.

Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

Tomas Hodan, Pavel Haluza, Stepan Obdrzalek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects . In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–888. IEEE, March 2017. doi: 10.1109/WACV.2017.103.

Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. EGTR: Extracting graph from transformer for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24229–24238, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Zhonglei Liu, Kechen Song, Xiaozheng Liu, Fuqi Sun, Jinhai Wang, Lufeng Luo, and Yunhui Yan. O2O-I-Grasp: An efficient multi-task network for sequential robotic grasping in stacked-object scenes. *Journal of Visual Communication and Image Representation*, 121:104967, 2026. doi: 10.1016/j.jvcir.2026.104967.

Jiayuan Luo, Yaxin Liu, Han Wang, Mengyuan Ding, and Xuguang Lan. Grasp manipulation relationship detection based on graph sample and aggregation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4098–4104, 2024. doi: 10.1109/ICRA57147.2024.10611627.

Mayank Mittal et al. Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. doi: 10.48550/arXiv.2511.04831.

NVIDIA. NVIDIA Isaac Sim, version 5.1.0. Software, 2025. [Online]. Available: https://github.com/isaac-sim/IsaacSim/releases/tag/v5.1.0.

Menghao Pu, Chaoqun Han, Zhiping Chai, Yifan Zhou, Pu Wen, Jihong Zhu, Chao Wang, Han Ding, Xuguang Lan, and Zhigang Wu. D2TriPO-DETR: Dual-decoder triple-parallel-output detection transformer. *IEEE Transactions on Industrial Informatics*, 22(6):4837–4848, 2026. doi: 10.1109/tii.2026.3668360.

Paolo Rabino and Tatiana Tommasi. A modern take on visual relationship reasoning for grasp planning. *IEEE Robotics and Automation Letters*, 10(2):1712–1719, 2025.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019. doi: 10.1109/CVPR.2019.00075.

Vladimir Tchuiev, Yakov Miron, and Dotan Di Castro. DUQIM-Net: Probabilistic object hierarchy representation for multi-view manipulation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10470–10477, 2022. doi: 10.1109/iros47612.2022.9981406.

Paolo Tripicchio, Salvatore D'Avella, and Carlo Alberto Avizzano. Cepb dataset: A photorealistic dataset to foster the research on bin picking in cluttered environments. *Frontiers in Robotics and AI*, 11:1222465, 2024. doi: 10.3389/frobt.2024.1222465.

Han Wang, Jiayuan Zhang, Lipeng Wan, Xingyu Chen, Xuguang Lan, and Nanning Zheng. MMRDN: Consistent representation for multi-view manipulation relationship detection in object-stacked scenes. *arXiv preprint arXiv:2304.12592*, 2023. doi: 10.48550/arXiv.2304.12592.

Ke-Jyun Wang, Yun-Hsuan Liu, Hung-Ting Su, Jen-Wei Wang, Yu-Siang Wang, Winston Hsu, and Wen-Chin Chen. OCID-ref: A 3D robotic dataset with embodied language for clutter scene grounding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5333–5338. Association for Computational Linguistics, June 2021. doi: 10.18653/v1/2021.naacl-main.419.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. doi: 10.15607/RSS.2018.XIV.019.

Yuxiang Yan, Zhiyuan Zhou, Xin Gao, Guanghao Li, Shenglin Li, Jiaqi Chen, Qunyan Pu, and Jian Pu. Learning spatial-aware manipulation ordering. *Advances in Neural Information Processing Systems*, 38:164727–164746, 2026.

Chenjie Yang, Xuguang Lan, Hanbo Zhang, Xinwen Zhou, and Nanning Zheng. Visual manipulation relationship detection with fully connected CRFs for autonomous robotic grasp. In *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 393–400, 2018. doi: 10.1109/ROBIO.2018.8665225.

Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. DFormerv2: Geometry self-attention for RGBD semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19345–19355, 2025.

Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. DFormer: Rethinking RGBD representation learning for semantic segmentation. In *International Conference on Learning Representations*, 2024.

Yang You, Kai Xiong, Zhening Yang, Zhengxiang Huang, Junwei Zhou, Ruoxi Shi, Zhou Fang, Adam W. Harley, Leonidas Guibas, and Cewu Lu. Pace: A large-scale dataset with pose annotations in cluttered environments. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 473–489, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72983-6.

Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship network for autonomous robotics. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pp. 118–125, 2018. doi: 10.1109/HUMANOIDS.2018.8625071.

Hanbo Zhang, Deyu Yang, Han Wang, Binglei Zhao, Xuguang Lan, Jishiyu Ding, and Nanning Zheng. Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter. *IEEE Robotics and Automation Letters*, 7(2):2929–2936, 2022. doi: 10.1109/LRA.2022.3142401.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Guoyu Zuo, Jiayuan Tong, Hongxing Liu, Wenbai Chen, and Jianfeng Li. Graph-based visual manipulation relationship reasoning network for robotic grasping. *Frontiers in Neurorobotics*, 15: 719731, 2021. doi: 10.3389/fnbot.2021.719731.

A MANIPRELAD DATASET CONSTRUCTION AND ANNOTATIONS

Dataset Contents. The ManipRelaD dataset contains 3,036 cluttered tabletop scenes with 1 to 30 objects and ten views per scene, yielding 30,360 aligned RGB-D image pairs. Each observation is paired with instance segmentation, 6D object poses, and a directed MRG. Figure 6 illustrates the simulation environment together with the observation modalities and annotations.

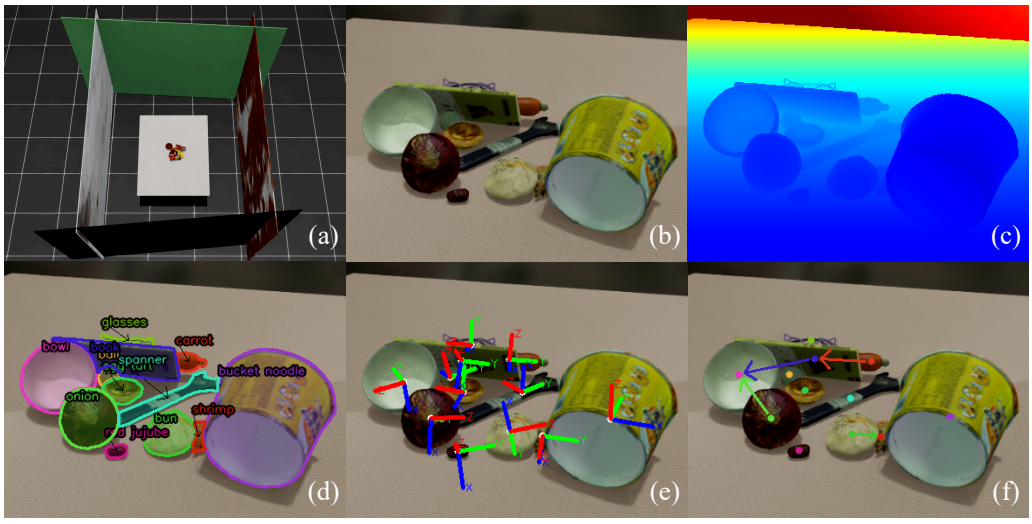


Figure 6: Overview of the ManipRelaD dataset. Panels (a)–(f) show the simulation environment, RGB image, depth map, instance segmentation, 6D object poses, and contact-force-derived MRG, respectively.

Object Assets and Scene Generation. The ManipRelaD dataset is generated in NVIDIA Isaac Sim NVIDIA (2025) using 142 textured object assets from OmniObject3D Wu et al. (2023), spanning 140 semantic categories. The assets cover six semantic domains: food (48.6%), household items (16.9%), personal care objects (11.3%), toys (8.5%), office items (7.7%), and tools (7.0%). For each scene, objects are sampled and initialized in vertical layers above a bounded tabletop region, with

a minimum separation between their initial positions. Their poses and linear velocities are slightly perturbed before the objects are released under gravity and allowed to settle into a stable cluttered arrangement. The tabletop and background appearances are randomized using scanned physically based rendering materials from ambientCG Demes (2022). Ten cameras are sampled from a spherical band around the workspace, subject to a minimum angular separation and small roll perturbations, yielding ten aligned RGB-D observations of each settled scene. Figure 7 shows representative RGB observations across the complete range of scene object counts from 1 to 30.



Figure 7: Representative RGB observations of the ManipRelaD dataset for scenes containing 1 to 30 objects.

Pairwise Contact-Force Acquisition. After a scene settles, we construct its MRG from pairwise inter-object contact forces. We extend the Isaac Lab contact sensor Mittal et al. (2025) with the low-level PhysX contact-reporting interface in Isaac Sim NVIDIA (2025) to recover pair-specific normal-force measurements. Separate contact-filter entries isolate object–object interactions from object–table support, and the normal-force vectors over all contact points are summed for each object pair at every simulation step.

MRG Annotation from Contact Forces. Let $\mathbf{F}_{ij}^{(t)}$ denote the aggregated normal contact force exerted by candidate supporter i on object j at simulation step t. The upward unit vector is $\mathbf{n}_{\text{up}} = -\mathbf{g}/\|\mathbf{g}\|_2$, where $\mathbf{g}$ is the gravitational-acceleration vector. We retain forces with positive upward components and average their contributions over $T = 6$ consecutive simulation steps:

$$\chi_{ij}^{(t)} = \mathbb{I}\left(\mathbf{F}_{ij}^{(t)} \cdot \mathbf{n}_{\text{up}} > 0\right), \tag{9}$$

$$u_{ij} = \frac{1}{T}\sum_{t=1}^{T} \chi_{ij}^{(t)} \left(\mathbf{F}_{ij}^{(t)} \cdot \mathbf{n}_{\text{up}}\right), \tag{10}$$

$$m_{ij} = \frac{1}{T}\sum_{t=1}^{T} \chi_{ij}^{(t)} \left\|\mathbf{F}_{ij}^{(t)}\right\|_2. \tag{11}$$

Here, u_{ij} is the time-averaged upward contribution, while m_{ij} averages the per-step force magnitudes after the upward-component test. Let $\mathcal{S}_j$ be the set of active candidate supporters of object j, excluding

j itself, and define $U_j = \sum_{k \in \mathcal{S}_j} u_{kj}$ and $M_j = \sum_{k \in \mathcal{S}_j} m_{kj}$. Their normalized contributions are

$$r_{ij}^{\mathrm{up}} = \begin{cases} u_{ij}/U_j, & U_j > \epsilon_{\mathrm{norm}}, \\ 0, & \text{otherwise}, \end{cases}$$
$$r_{ij}^{\mathrm{mag}} = \begin{cases} m_{ij}/M_j, & M_j > \epsilon_{\mathrm{norm}}, \\ 0, & \text{otherwise}, \end{cases} \qquad \epsilon_{\mathrm{norm}} = 10^{-8}. \tag{12}$$

The manipulation relationship is then assigned by

$$A_{ji} = \mathbb{I}\big(r_{ij}^{\mathrm{up}} \geq \tau \ \vee \ r_{ij}^{\mathrm{mag}} \geq \tau\big), \qquad \tau = 0.1. \tag{13}$$

This criterion maps the physical interaction of supporter i on object j to the manipulation edge $j \to i$, indicating that j should be removed before i. The upward term measures vertical support, while the magnitude term retains the lateral components of qualifying normal-force vectors.

Dataset Comparison. Table 3 compares ManipRelaD with related object-centric datasets in terms of annotations, scale, scene complexity, and data source. ManipRelaD jointly provides RGB-D observations, instance segmentation, 6D object poses, and automatically generated pairwise MRG annotations derived from inter-object contact forces.

Table 3: Comparison with Related Object-Centric Datasets

Dataset	Annotations				Scale			Scene Complexity		Dataset Type
	Modality	Segm.	Obj. Pose	MRG (Labeling)	Object Set	Scenes	Images	Objects per Scene	Environment Type(s)	
CEPB Tripicchio et al. (2024)	RGB-D	✓	✓	✗	40	50K	1.5M	≤40	bin	sim
OCID-Ref Wang et al. (2021)	RGB + point cloud	✓	✗	✗	58 cat.	2.3K	2.3K	∼7.7	table, floor	real
YCB-Video Xiang et al. (2018)	RGB-D	✓	✓	✗	21	92	130K	∼4.5	table	real
T-LESS Hodan et al. (2017)	RGB-D	✓	✓	✗	30	30 train seq. 20 test scenes	∼39K train ∼10K test per sensor	1 (train) ∼7 (test)	isolated (train) table (test)	real
PACE You et al. (2025)	RGB-D	✓	✓	✗	238 (real) 931 (sim.)	300 videos (real) – (sim.)	55K (real) 100K (sim.)	∼4.7 (real) ∼24 (sim.)	indoor (real) synthetic clutter (sim.)	real, sim
GraspNet-1B Fang et al. (2020)	RGB-D	✓	✓	✗	88	190	97K	∼8.9	table	real
GraspClutter6D Back et al. (2025)	RGB-D	✓	✓	✗	200	1000	52K	∼14.1	bin, shelf, table	real
VMRD Zhang et al. (2018)	RGB	✗	✗	✓ (Manual)	∼200	–	5.2K	∼3.4	table	real
REGRAD Zhang et al. (2022)	RGB-D	✓	✓	✓ (Automatic,removal)	50K	111.8K	900K	1–20	table	sim
MetaGraspNetV2 Gilles et al. (2024)	RGB-D	✓	✓	△[a] (Rule-based)	82	8K (sim.) 800 (real)	296K (sim.) 3.2K (real)	10 (sim.) 15 (real)	bin	sim, real
Yan et al. Yan et al. (2026)	RGB-D	✗	✗	△[b] (VLM-generated)	5 cat. (sim.) 2 cat. (real)	–	163K (sim.) 33K (real)	24, 36, 60 (sim.) – (real)	table	sim, real
ManipRelaD dataset	RGB-D	✓	✓	✓ (Automatic, **contact-force**)	142	3.0K	30.4K	1–30	table	sim

✓, △, and ✗ indicate explicit pairwise MRG annotations, related but non-equivalent supervision, and no MRG-related supervision, respectively. [a] MetaGraspNetV2 uses rule-based, occlusion-derived object-layout graphs. [b] Yan et al. use a VLM with spatial priors for global removal-order supervision.

B Implementation and Reproducibility

Model Configuration. The Deformable DETR detector uses eight attention heads, four sampling points in each deformable attention module, and feed-forward layers with dimension 1024. The relation predictor is the three-layer MLP described in Section 3.3. During the second training stage, the RGB-D backbone, detector, gated decoder-layer aggregation, and relation predictor are optimized jointly.

Input Processing and Augmentation. The model does not assume a fixed input height and width. At test time, each RGB-D observation is resized while preserving its aspect ratio so that the shorter side is 800 pixels and the longer side does not exceed 1333 pixels; samples are dynamically padded within each batch. Stage-2 training applies synchronized random horizontal flipping and multi-scale resizing to RGB and depth. The shorter side is sampled between 480 and 800 pixels, either directly or after an intermediate resize to 400, 500, or 600 pixels. RGB channels are normalized with ImageNet mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. Invalid, non-finite, or nonpositive depth values are set to zero. Depth is then clipped to $[0, 2.0]$ m, divided by 2, and standardized with mean 0.48 and standard deviation 0.28 before being supplied as the fourth input channel.

Optimization and Checkpoint Selection. Both stages use AdamW. The first stage uses an effective batch size of 2 for at most 50 epochs. The backbone, reference-point, and sampling-offset parameter groups use a learning rate of 1×10^{-5}, while the remaining detector parameters use 1×10^{-4}. The second stage uses a batch size of 2, 32-bit precision, eight data-loading workers, and no gradient accumulation for at most 30 epochs. Its low-learning-rate parameter group, which includes the backbone, uses 2×10^{-7}; the remaining pretrained detector parameters use 2×10^{-6}; and the newly initialized relation modules use 2×10^{-4}. Both stages use a weight decay of 1×10^{-4}, gradient clipping at 0.1, and no learning-rate scheduler. The focal classification, box ℓ_1, and GIoU loss coefficients are $\lambda_{\text{cls}} = 2$, $\lambda_{\text{box}} = 5$, and $\lambda_{\text{giou}} = 2$ in both stages. The second stage additionally uses the connectivity binary cross-entropy with $\lambda_{\text{rel}} = 30$.

Stage 1, Stage 2, and evaluation use random seed 42 with fixed worker seeds. During Stage 2, validation loss is monitored after every epoch, and early stopping uses a patience of ten epochs. Training terminates at epoch 18, and the epoch-8 checkpoint, which attains the minimum validation loss of 3.5522, is used for evaluation.

Baseline Configuration. The comparison models are trained on the same scene-level training split and evaluated by the same end-to-end evaluator. VMRN uses the F-VMRN configuration with a ResNet-101 detector. D3G retains its object-query graph formulation, and EGTR retains its RGB object-query detector and decoder-based relation extraction branch. We preserve the native input modality and architecture of each method rather than replacing its encoder with ours. For evaluation, all model outputs are converted to binary directed relation scores and processed with the metric-specific detection thresholds and maximum detection counts, class-agnostic NMS, object-matching rule, and relation threshold described in the main paper.

C Additional Evaluation and Model Analysis

C.1 Geometry-Aligned Pairwise Edge Evaluation

Protocol and Metrics. To examine directed relationship scores independently of class-aware hard-IoU matching, we complement the main end-to-end evaluation with a geometry-aligned pairwise edge protocol. The protocol uses the same balanced manifest of 3,300 test samples and the same detection and relation post-processing settings. Predicted and ground-truth objects are assigned one-to-one by a DETR-style Hungarian cost that combines normalized-box ℓ_1 distance and GIoU with weights of 5 and 2, respectively, without a classification cost or a hard IoU cutoff. Edge AP pools the continuous scores and binary labels of all valid off-diagonal directed pairs. Edge OP, Edge OR, and Edge F1 are computed from predictions thresholded at 0.5, while Edge IA measures the percentage of images whose complete valid directed-edge graph is correct. GT/Pred denotes the ground-truth and predicted positive-edge ratios. These metrics characterize relationship prediction after spatial alignment and complement the unified end-to-end results in the main paper.

Overall Comparison. Our method leads every reported edge metric in Table 4, reaching 43.32% Edge AP and 35.80% Edge F1. Relative to EGTR, these results improve Edge AP by 14.00 points and Edge F1 by 19.20 points. The method also raises Edge OR from 9.65% to 25.22% while attaining the highest Edge OP of 61.64% and the highest Edge IA of 17.58%. Its predicted positive-edge ratio of 1.22% is the closest to the 2.98% ground-truth ratio among the evaluated methods, combining sparse graph prediction with substantially stronger edge recovery.

Table 4: Pairwise edge evaluation on ManipRelaD.

Method	Edge AP (%)	F1 (%)	IA (%)	OP (%)	OR (%)	GT/Pred (%)
VMRN	20.28	15.90	16.30	61.58	9.13	2.98/0.44
D3G	8.06	0.00	14.79	0.00	0.00	2.98/0.00
EGTR	29.32	16.60	15.73	59.19	9.65	2.98/0.49
Ours	**43.32**	**35.80**	**17.58**	**61.64**	**25.22**	2.98/1.22

Object-Count Sensitivity. Figure 8 shows that our method obtains the highest Edge AP, Edge F1, and Edge OR in all 29 nontrivial object-count subsets from 2 to 30 objects, and the highest Edge IA in 24 of them. VMRN and EGTR often retain high Edge OP but recover fewer positive relationships as scene complexity increases. The consistent lead in ranking quality and edge recovery shows that the advantage extends across the full range of scene sizes. The one-object subset contains no valid off-diagonal pair, so its edge-level metrics are zero and its Edge IA is 100%.

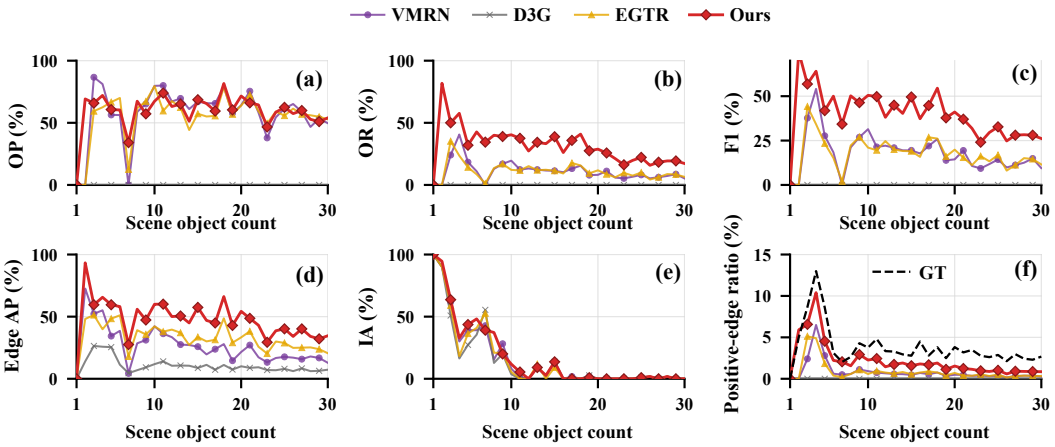


Figure 8: Pairwise edge metrics by scene object count on ManipRelaD. Panels (a)–(f) report OP, OR, F1, Edge AP, IA, and the predicted positive-edge ratio, respectively; the dashed curve in panel (f) denotes the ground-truth ratio. Each point aggregates 110 test samples.

C.2 Relation-Threshold Sensitivity

Protocol. We measure threshold sensitivity under the unified end-to-end protocol used for the main comparison. Detection outputs, post-processing, and class-aware one-to-one object matches are held fixed while the relation threshold is swept from 0 to 1 in increments of 0.005.

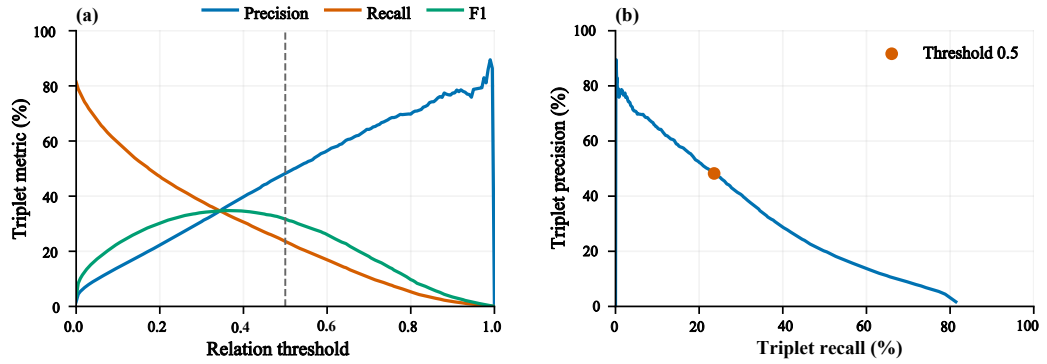


Figure 9: Relation-threshold sensitivity on ManipRelaD with fixed detections and object matches. (a) Triplet precision, recall, and F1 as the relation threshold varies; the dashed line marks the reported threshold of 0.5. (b) Triplet precision–recall curve, with the same operating point highlighted.

Precision–Recall Trade-off. At the fixed threshold of 0.5, the sweep exactly reproduces the reported OP, OR, and F1 of 48.23%, 23.58%, and 31.68%, respectively. Figure 9 shows that lowering the threshold increases relation recall at the cost of precision. The test-set curve reaches its maximum F1 of 34.77% at a threshold of 0.355, with an OP of 35.80% and an OR of 33.80%. Because this point is identified by scanning the test set, it is used only to characterize score sensitivity; all reported model comparisons retain the fixed threshold of 0.5.

C.3 Systematic Qualitative Error Analysis

We audit the complete test cache and select one distinct physical scene from each of three predefined error categories, restricting candidates to scenes with 10–14 ground-truth objects and 2–6 positive ground-truth relations. This selection separates relation-only errors under complete object matching from errors coupled with imperfect detection.

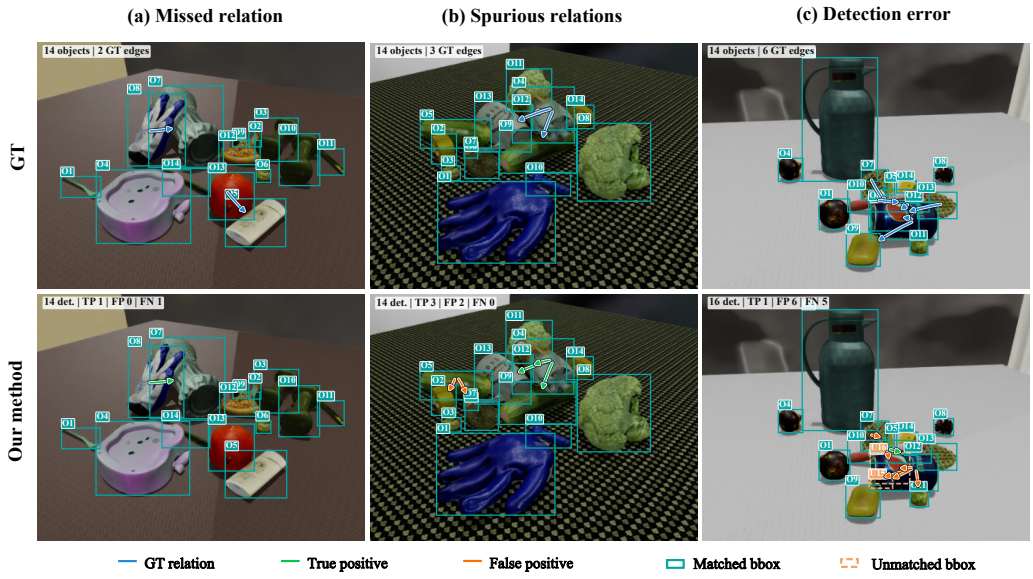


Figure 10: Systematically selected error cases. Columns (a)–(c) show a missed relation with all objects matched, spurious relations with all objects matched, and relation errors coupled with unmatched detections, respectively. Ground truth is shown above the corresponding prediction.

In Figure 10(a), all 14 objects are matched, and the model recovers one of the two ground-truth relations without introducing a false positive. In Figure 10(b), all 14 objects are again matched, and all three ground-truth relations are recovered, together with two false-positive edges. Figure 10(c) contains 14 ground-truth objects, of which 13 are matched, as well as three unmatched predictions; the resulting graph contains one true-positive, five false-negative, and six false-positive relations. The first two cases isolate missed and spurious relation predictions, whereas the third illustrates how object-detection errors propagate to the predicted relationship graph.