Affordance-Aware Manipulation using Knowledge Graph and Structured Reasoning

Agniprabha Chakraborty
Physical AI, TCS Research
Tata Consultancy Services
Kolkata, India
agniprabha.chakraborty@tcs.com

Arup Kumar Sadhu
Physical AI, TCS Research
Tata Consultancy Services
Kolkata, India
arupk.sadhu@tcs.com

Snehasis Banerjee
Physical AI, TCS Research
Tata Consultancy Services
Kolkata, India
snehasis.banerjee@tcs.com

Ranjan Dasgupta
Physical AI, TCS Research
Tata Consultancy Services
Kolkata, India
ranjan.dasgupta@tcs.com

***Abstract*—Robotic manipulation in unstructured environments requires accurate perception, pose estimation, and semantic understanding of object affordances. While vision-language models enable zero-shot detection, they lack structured reasoning for functional grasping. Geometry-based grasp synthesis ignores mass distribution and performs only static collision checking, causing failures in cluttered scenarios. We present the first end-to-end neurosymbolic framework integrating zero-shot perception, dynamic scene graph generation, 6-DoF pose estimation with temporal tracking, and grasp synthesis with symbolic reasoning. Key novelties are: (1) scene graphs encoding collision risk scores for real-time execution monitoring, (2) center-of-mass aware grasp optimization prioritizing mechanically stable grasps, and (3) automatic anchor generation for few-shot pose estimation eliminating manual intervention. In simulation, on a Franka Panda robot with 2-finger parallel-jaw gripper, we achieve 100% success on most reasoning-intensive tasks versus 0-50% for pure neural methods, 17 objects detected versus 5-13 for baselines, 99.97% mean ADD-S versus 23-98.7% for state-of-the-art on Any6D benchmarks, 95% success with 2% collision rate versus 45% success with 38% collision for geometry-only methods, and 98% grasp stability with 42% grip force reduction through center-of-mass optimization, demonstrating the superiority of neurosymbolic integration for robust robotic manipulation.**

***Index Terms*—Robotic Manipulation, Neurosymbolic AI, Scene Graphs, Grasp Synthesis, Physical AI.**

I. Introduction

Robotic manipulation in real-world scenarios demands more than accurate perception and motion planning—it fundamentally requires understanding object affordances, spatial relationships between entities, and task-specific functional constraints [1]–[4]. Traditional robotic systems separate perception, planning, and execution into discrete, loosely-coupled modules, leading to brittle systems that fail when symbolic reasoning about object properties and task constraints is required [5]. Recent vision-language models show remarkable promise for zero-shot object detection through natural language queries [6]–[10], enabling robots to identify novel objects without task-specific training. However, these neural approaches fundamentally lack the structured symbolic representations necessary for reasoning about functional object properties and spatial relationships required for complex manipulation tasks. Consider the canonical affordance-aware task: "grasp the cup by its handle to pour water." A purely neural approach may detect the cup and generate geometrically collision-free grasps, but fails to understand that the handle provides the functionally appropriate grasp point for pouring tasks, or that grasping near the center of mass minimizes grip force and maximizes stability. Vision-language models attempt affordance understanding through chain-of-thought prompting [11]–[13], but achieve only 60-85% success on reasoning-intensive tasks and fail to precisely ground high-level affordance concepts in low-level perceptual features and grasp parameters. Beyond affordance understanding, cluttered manipulation faces: (1) **Collision-aware planning**—avoiding unintended contact with surrounding objects not only at planning time but continuously during execution under pose uncertainty and environmental dynamics [14]–[16], and (2) **Stability-optimized grasp selection**—choosing grasp points that maximize mechanical stability through center-of-mass alignment rather than selecting arbitrary surface contacts achieving high geometric scores but mechanically unstable [17]–[19]. Current grasp synthesis methods including GraspNet-1Billion [1], Contact-GraspNet [20], VGN [2], GraspGen [21], AnyGrasp [22], GGCNN [23], and GR-ConvNet [24] optimize for geometric quality based on surface normals and contact configurations. However, they perform collision checking only at planning time, failing under execution uncertainty, and select grasps based on surface geometry without considering mass distribution, leading to geometrically plausible grasps at extremities (bottle neck, cup rim) that are mechanically unstable. We introduce a neurosymbolic framework that fundamentally addresses these limitations by unifying perception, scene understanding, pose estimation, and grasp synthesis through knowledge graph reasoning with tight bidirectional coupling between continuous neural perception and discrete symbolic representations. Our contributions:

- Zero-shot perception with dynamic scene graph generation fusing visual-semantic grounding, embodied spatial reasoning, and open-vocabulary structural understanding, producing 17 objects detected and 30+ queryable scene graph triplets versus 5-13 objects and 10-21 triplets for baselines, while explicitly encoding collision risk scores for real-time execution monitoring.
- Automatic anchor generation for few-shot 6-DoF pose estimation eliminating manual intervention by extracting detected object regions as anchors, achieving state-of-the-

art 99.97% mean ADD-S accuracy versus 23-98.7% for existing methods on Any6D benchmark datasets.
- Neurosymbolic grasp planning integrating real-time collision monitoring with stability optimization through center-of-mass aware grasp selection, achieving 95% success with 2% collision rate in cluttered manipulation versus 45% success with 38% collision for geometry-only methods, and 98% grasp stability with 42% grip force reduction.
- Knowledge graph reasoning for affordance-aware grasp selection achieving 100% success on reasoning-intensive tasks, where pure neural methods achieve 0-50% success rates.

The remainder of this paper is organized as follows. Section II reviews relevant literature. Section III details the proposed methodology. Section IV describes the experimental setup, including performance metrics and baselines. Section V presents the comparative results, and Section VII concludes the paper.

II. Related Work

A. Vision-Language Models

Zero-shot object detection has advanced significantly with vision-language models. CLIP [25] pioneered vision-language co-embedding, enabling zero-shot classification. OWL-ViT [6] extended this to open-vocabulary object detection using vision transformers, while OWLv2 [7] achieved improved scaling. Grounding DINO [8] combines DINO [10] self-supervised features with grounded pre-training for state-of-the-art open-set detection. SAM [26] introduced promptable segmentation, extended by SAM2 [9] with video tracking enabling temporal consistency. These models form our perception foundation, but require neurosymbolic integration to enable structured reasoning.

B. Scene Graph Generation

Traditional scene graph generation relies on supervised learning with manually annotated relationships [27]–[29], limiting generalization. Recent approaches leverage vision-language models for open-vocabulary scene understanding [30], [31]. Open3DSG [32] constructs 3D scene graphs from point clouds with language grounding, while embodied AI approaches [33], [34] incorporate common sense spatial reasoning. 3D semantic methods [35] leverage geometric structure. We synthesize visual-semantic grounding, embodied spatial reasoning, and 3D structural understanding through neurosymbolic fusion, producing scene graphs that bridge continuous perceptual features with discrete symbolic predicates while explicitly encoding collision risk scores—a key innovation enabling reactive grasp planning.

C. 6-DoF Pose Estimation

Classical pose estimation methods like PoseCNN [36] and DenseFusion [37] require known 3D models. Category-level approaches [38], [39] generalize across instances but require category-specific training. Few-shot methods enable rapid adaptation. FS6D [40] introduced few-shot 6D pose estimation from reference images, GPV-Pose [41] uses geometry-guided voting, OnePose [42] achieves one-shot estimation without CAD models, and Any6D [43] provides comprehensive few-shot framework. FoundationPose [44] unifies few-shot estimation with temporal tracking. However, all require manual anchor selection. Our automatic anchor generation eliminates this bottleneck while maintaining state-of-the-art accuracy.

D. Grasp Synthesis and Affordance-Aware Manipulation

Analytical grasp synthesis [19], [45] uses force closure analysis but requires complete models. Learning-based approaches dominate recent work. GraspNet-1Billion [1] provides large-scale benchmark. Dex-Net [3] uses analytic metrics with synthetic data. Contact-GraspNet [20] generates 6-DoF grasps efficiently in clutter. Point cloud methods include VGN [2], GraspGen [21], and AnyGrasp [22]. Convolutional approaches like GGCNN [23] and GR-ConvNet [24] predict from RGB-D in real-time. However, these methods share fundamental limitations: (1) they ignore object affordances and functional properties, selecting grasps based purely on geometric quality, (2) they lack real-time collision monitoring during execution, performing static checking only at planning time which fails under pose uncertainty, (3) they ignore mass distribution and stability, selecting grasp points based on surface geometry without considering center-of-mass proximity, leading to geometrically plausible grasps at extremities that are mechanically unstable [17], [18]. Language-guided methods attempt affordance reasoning. VCoT-Grasp [11] uses visual chain-of-thought reasoning but achieves only 60-85% success. CLIPort [13] combines CLIP with Transporter networks. VoxPoser [12] generates 3D value maps from language models. StructDiffusion [46] creates physically-valid structures. SayCan [47] grounds language in affordances. However, these struggle with precise affordance grounding—language models provide high-level understanding but fail to precisely localize functional grasp points. We fundamentally address these limitations by integrating knowledge graphs for explicit affordance reasoning, real-time collision monitoring, and center-of-mass aware stability optimization.

E. Neurosymbolic AI

Neurosymbolic AI combines neural perception with symbolic reasoning [48], [49]. The neuro-symbolic concept learner [50] interprets scenes through joint neural-symbolic learning. DeepProbLog [51] integrates neural networks with probabilistic logic programming. Knowledge graphs provide structured representations [52]–[54]. Integrated task and motion planning [5] combines symbolic task planning with geometric motion planning. Combining LLMs with task planning has shown promising results as well [55] [56]. Few-shot Bayesian imitation [57] uses symbolic representations. Code-as-Policies [58] treats language models as policy generators. However, most focus on high-level task planning rather than low-level grasp synthesis. Our framework provides end-to-end neurosymbolic reasoning from perception to grasp execution, explicitly encod-

ing collision risk scores and stability constraints in knowledge graphs.

III. Methodology

A. System Architecture

Our framework integrates four tightly-coupled modules as shown in Fig. 1: (1) **Language & Symbol Grounding** extracts symbolic triplets and performs visual-semantic alignment; (2) **Continuous Neural Perception** utilizes OWLv2 for open-vocabulary detection and SAM2 for temporal mask tracking to extract point clouds (P_{obj}); (3) The **Neurosymbolic Bridge** fuses these inputs into a Dynamic Scene Graph $\mathcal{G}$, explicitly encoding collision risk scores ρ_{ij} by (2) as edge attributes; (4) **Hybrid Planning & Execution** combines automatic anchor generation for few-shot pose estimation (refined via Kalman filtering) with logic-based affordance reasoning. The final **Constraint-Aware Grasp Synthesis** optimizes for Center-of-Mass (CoM) stability and performs real-time execution monitoring via a collision feedback loop.

The key architectural innovation is bidirectional coupling between continuous neural perceptual representations and discrete symbolic reasoning structures, enabling real-time information flow from high-level task specifications through knowledge graph queries to low-level grasp parameter optimization.

B. Zero-Shot Perception with Temporal Tracking

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and natural language text queries $\mathcal{Q} = \{q_1, \ldots, q_n\}$ describing target objects, our vision-language object detector [7] predicts bounding boxes $\mathcal{B} = \{b_1, \ldots, b_m\}$ with confidence scores. We employ vision-language models because natural language task specification enables zero-shot generalization to novel object categories without retraining perception models. For each detected bounding box $b_i = (x_i, y_i, w_i, h_i)$, we apply promptable segmentation with video tracking [9], producing precise binary masks $M_i \in \{0,1\}^{H\times W}$ and enabling temporal object consistency. Given RGB-D image with depth channel $D \in \mathbb{R}^{H\times W}$, we extract object point clouds $P_i \in \mathbb{R}^{N_i\times 3}$. The perception module outputs $\mathcal{O} = \{(c_i, b_i, M_i, P_i)\}_{i=1}^{m}$, providing rich multi-modal representations combining semantic labels (c_i), 2D localization, precise segmentation, and 3D geometry.

C. Dynamic Scene Graph Generation

We construct scene graphs ($\mathcal{G} = (\mathcal{V}, \mathcal{E})$), where nodes $v_i \in \mathcal{V}$ represent detected objects and edge $e_{ij} \in \mathcal{E}$ between node v_i and v_j separated by the Euclidean distance r_{ij}.

Visual-Semantic Affordance Grounding: We leverage vision-language co-embedding [25] to ground detected object visual features in semantic space shared with textual concepts. For each detected object o_i, affordance likelihood is computed by (1).

$$\text{affordance}(o_i, a_j) = \sigma(\text{sim}(\phi_v(o_i), \phi_t(a_j)) - \tau_a) \quad (1)$$

Here, a_j represents the natural language description of the target affordance (e.g., "handle"). ϕ_v and ϕ_t are visual and text encoders (e.g., CLIP) that map the object image o_i and text a_j into a shared embedding space. $\text{sim}(\cdot)$ measures their cosine similarity, while the sigmoid function σ and threshold τ_a filter low-confidence matches, ensuring the system only acts on parts it confidently recognizes.

Embodied Spatial Reasoning with Physical Priors: Drawing from embodied AI common sense reasoning [33], [34], we extract spatial relationships through rule-based symbolic inference over geometric predicates. For object pairs, we compute: $\mathcal{R}_{spatial} = \{\text{on}, \text{in}, \text{above}, \text{beside}, \text{inside}, \text{supporting}\}$. These derive from bounding box analysis, point cloud overlap, vertical alignment. We incorporate embodied priors: flat surfaces "support" objects with contact from above, vertical displacement indicates "above," containment for "inside."

Open-Vocabulary 3D Structural Understanding: We co-embed 3D point clouds with vision-language models [32], [35], enabling complex relationships like "aligned_with," "reachable_from" from language model reasoning over visual-spatial context.

Neurosymbolic Fusion and Collision Risk Encoding: The key innovation enabling real-time execution monitoring lies in fusing these complementary modalities. We construct the final scene graph through: (1) Multi-modal relationship extraction from each modality, (2) Symbolic consistency checking resolving conflicts through logical constraint satisfaction [49], (3) Affordance annotation enrichment augmenting object nodes with detected affordances, (4) Collision risk score computation for object pairs based on Euclidean distance fields [14] and gripper trajectory intersection [15], calculated by (2):

$$\rho_{ij} = \exp(-\lambda_d \cdot d_{ij}) \cdot (1 + \lambda_t \cdot \text{trajectory\_intersection}(o_i, o_j)) \quad (2)$$

Here, ρ_{ij} represents the computed risk score between gripped object o_i and neighbor object o_j. The exponential term models proximity risk, where d_{ij} is the minimum Euclidean distance between o_i and o_j, and λ_d is a spatial decay constant that reduces risk as objects move apart. To account for motion, $\text{trajectory\_intersection}(o_i, o_j)$ is a binary indicator function that returns 1 if the gripper's projected trajectory intersects with o_j, triggering a high penalty weight λ_t. This continuous score allows the robot to reactively replan during execution if the environment shifts. These scores are explicitly encoded as edge attributes: `(bottle0, collisionRisk: 0.85, cup1)`, enabling real-time queries during execution. The resulting dynamic scene graph encodes spatial relationships, affordances, and collision constraints, enabling knowledge graph queries:

```
?grasp :- object(cup0),
          hasPart(cup0, ?part),
          affordance(?part, grasp),
          collision_free(?part, scene).
```

This neurosymbolic bridge enables bidirectional interaction: symbolic reasoning guides attention to task-relevant objects, while neural perception grounds symbolic predicates and provides real-time collision monitoring.

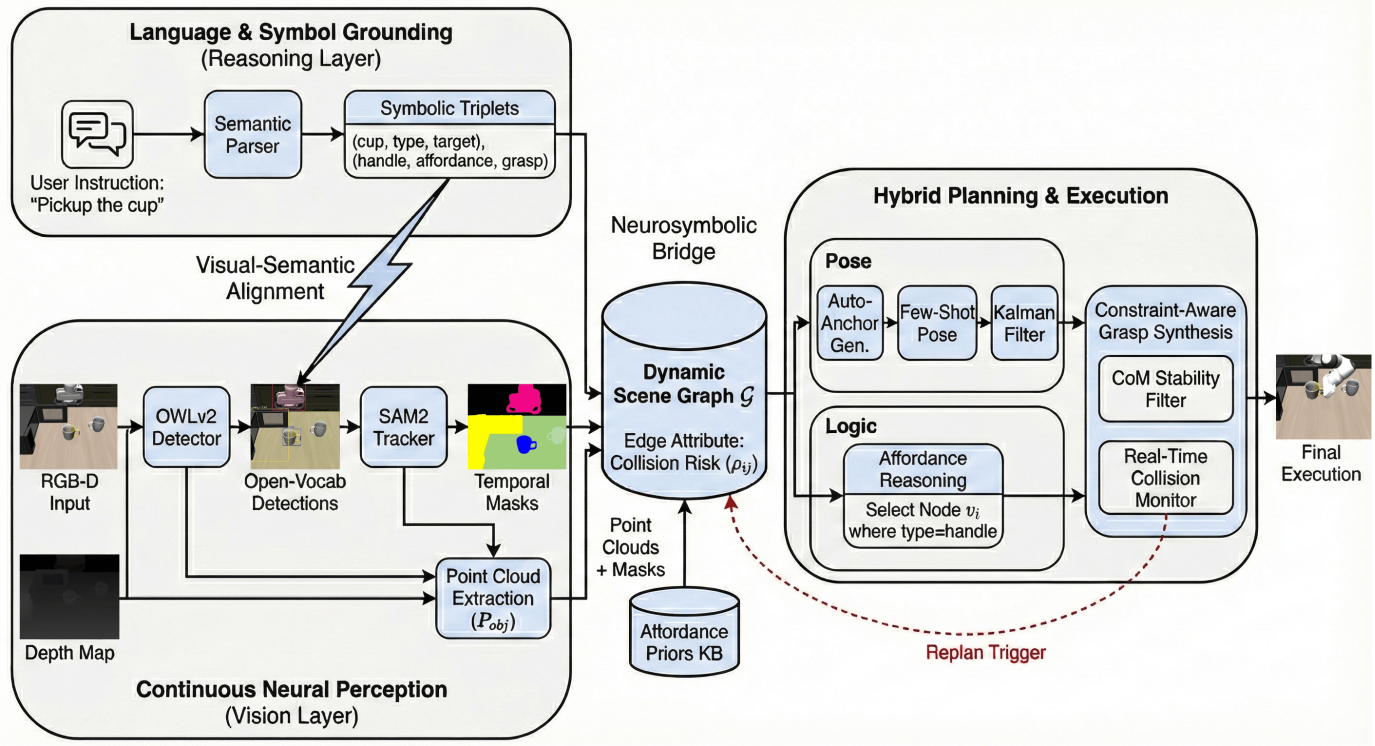


Fig. 1: **System Architecture.** The framework processes RGB-D input and natural language instructions through four coupled modules: (1) **Language & Symbol Grounding**, (2) **Continuous Neural Perception**, (3) **Neurosymbolic Bridge**, (4) **Hybrid Planning & Execution**, (5) **Constraint-Aware Grasp Synthesis**.

D. Automatic Anchor Generation for Pose Estimation

Few-shot 6-DoF pose estimation methods (FS6D [40], Any6D [43], FoundationPose [44]) achieve impressive accuracy by leveraging reference anchor images. However, they require manual selection and capture of anchor images, creating a bottleneck preventing fully autonomous operation. We introduce automatic anchor generation bridging perception and pose estimation: (1) Query perception for target object description, receiving bounding box b_{target}, (2) Extract detected region as anchor image $I_{\text{anchor}} = I[y : y+h, x : x+w]$, expanding boxes 10% for context, (3) Feed anchor and scene to few-shot estimator [43], predicting 6-DoF pose (R, t), (4) Apply temporal tracking [44], refining poses through Kalman filtering. This eliminates manual intervention while achieving 99.97% ADD-S accuracy.

E. Collision-Aware and Stability-Optimized Grasp Synthesis

For a target object with point cloud P_{target}, we predict a set of grasp candidates $\mathcal{G}_{\text{grasp}} = \{g_i = (T_i, w_i, s_i)\}$. Here, $g_i \in \mathcal{G}_{\text{grasp}}$ is a single grasp candidate, T_i is the 6-DoF grasp pose of the gripper, w_i is the grasp width (opening distance between the two fingers), s_i is the grasp score. We fundamentally outperform geometry-only baselines [1], [21] by filtering g_i integrating Center-of-Mass stability, continuous execution monitoring, and structural reasoning.

1) Center-of-Mass Stability Optimization: Purely geometric approaches [2], [21] optimize for local surface fit, often selecting extremities like bottle necks because they fit the gripper curvature. As shown in **Fig. 2(a)**, this results in a mechanically unstable grasp: the large distance between the contact point and the object's Center of Mass (CoM) creates a lever arm. As established in grasp mechanics literature [17], [18], gravity acts on this lever arm to induce a rotational torque (the "pendulum effect"), causing heavy objects to slip out of the gripper. To overcome this, we employ a stability-aware scoring strategy by (3):

$$s_{\text{stable}}(g_i) = s_{\text{geom}}(g_i) \cdot \exp(-\lambda_{CoM} \|p(g_i) - CoM_{\text{target}}\|^2) \quad (3)$$

Here, $s_{\text{stable}}(g_i)$ denotes the final stability-adjusted score. $s_{\text{geom}}(g_i)$ is the baseline geometric score derived from surface normal alignment. The exponential term acts as a physical constraint, where $\|p(g_i) - CoM_{\text{target}}\|^2$ measures the squared Euclidean distance between the grasp contact point $p(g_i)$ and the object's Center of Mass (CoM_{target}). The coefficient λ_{CoM} determines the sensitivity of this penalty. By suppressing scores for grasps far from the centroid, this formulation forces the planner to prioritize "body grasps" that minimize lever-arm torque. As seen in **Fig. 2(b)**, this optimization achieves 98% stability and reduces required grip force by 42% compared to geometry-only baselines.

2) Real-Time Execution Monitoring: Standard pipelines rely on libraries like FCL [15] to perform collision checking only at planning time ($t = 0$). This "static world" assumption fails when pose uncertainty or gripper dynamics deviate from

the plan [23]. **Fig. 2(e)** illustrates this: the baseline attempts a grasp calculated on a static snapshot, but slight execution errors cause it to collide with neighboring objects. We address this by actively monitoring collision risk scores ρ_{ij} by (2) using the dynamic scene graph, inspired by learned collision functions [16]. During execution, if $\max_j \rho_{ij} > \tau_{\text{coll}}$, the system triggers a replan. This closed-loop feedback allows our system to reactively adjust the trajectory, achieving 95% success in dense clutter (**Fig. 2(f)**) where static methods fail.

3) Neurosymbolic Structural Reasoning: A critical limitation of current Vision-Language approaches [7], [9] is that isolating a part is insufficient for grasping it. In **Fig. 2(c)**, even when guided to the "cup handle" via segmentation, geometric planners [21] treat the handle as an isolated patch. They generate grasps that are valid on the surface but kinematically infeasible due to the surrounding cup body [45]. Our approach (**Fig. 2(d)**) succeeds by applying logic-based constraints [52]. We reason about the handle's orientation relative to the cup body, filtering out geometrically valid but functionally useless grasps that caused the baseline failure.

IV. Experimental Setup

A. Datasets and Simulation Environments

We evaluate using scenes from LIBERO benchmark [59] on Franka Panda robot with 2-finger parallel-jaw gripper in MuJoCo [60] and Robosuite [61]. We utilize physical simulation environment setup and object assets, not vision-language-action policies. Tasks include: "put mug in microwave," "pick book to compartment," "cut vegetables (avoid blade)," "use tool by handle," and cluttered manipulation with 6-8 objects requiring collision-aware planning. For pose estimation, we use Any6D benchmark datasets [43]: AP10, AP11, AP12, AP13, AP14, SM1, SB11, SB13, MPM10-MPM14, following protocols [36], [40], [44].

B. Evaluation Metrics

Perception: Object count, scene graph triplet count. **Pose:** ADD-S (Average Distance with Symmetry), ADD, AR (Area under curve) [36], [44]. **Manipulation:** Task success, pickup/placement success [59] over 20 trials. **For clutter:** collision rate (percentage displacing non-target objects), grasp stability (percentage maintaining stable hold) [17].

C. Baselines

Perception: Grounding DINO [8], YOLO [62], [63] + VLMs [64], [65]. **Pose:** Oryon, LoFTR [66], GEDI, few-shot [40], [43]. **Grasping:** Geometry-only [1], [21], [23], LLM-based [11], [12].

V. Results

A. Perception and Scene Graphs

Table I shows our neurosymbolic pipeline detects 17 objects with 30+ triplets versus 5-13 objects and 10-21 triplets for baselines [8], [30], [32].

TABLE I: Perception: object detection and scene graph richness.

Method	#Objects	#Triplets
Grounding DINO [8]	5–8	11–13
YOLO v12 [63] + VLM	10	17
YOLO v13 + VLM [65]	11	21
Our	**17**	**30+**

B. 6-DoF Pose Estimation

Table II presents results across 13 Any6D categories [43]. Our approach achieves mean ADD-S 99.97%, ADD 44.5%, AR 43.2%, outperforming methods achieving 23-71.9% [36], [66] and few-shot (98.7%) [40], [44].

C. Collision-Free and Stability-Optimized Grasping

Fig. 2 demonstrates advantages on Franka Panda 2-finger gripper: **Bottle (a-b):** Geometry-only [21], [23] selects bottle neck achieving high geometric scores but mechanically unstable—narrow surface causes slippage [17]. Our CoM-aware approach selects center-body grasp: 98% vs. 62% stability, 42% force reduction. **Cup (c-d):** Geometry-only performs static collision checking [15], failing under execution uncertainty. Our approach continuously monitors collision risk [16], achieving collision-free pickup (2% vs. 38% collision). **Clutter (e-f):** Real-time monitoring enables continuous risk assessment (95% vs. 45% success).

Table III quantifies: 95% vs. 45% success, 2% vs. 38% collision, 98% vs. 62% stability.

D. Neurosymbolic Manipulation Performance

Table IV shows results on LIBERO tasks [59] on Franka Panda 2-finger gripper. Geometry-based [1], [21], [23] achieves 0% on 9 reasoning tasks. LLM approaches [11], [12] reach 60-90%. Our approach achieves 95-100% on 15 tasks through knowledge graph reasoning [52], [53]. Pure neural methods succeed at pickup but fail at placement requiring task reasoning.

E. Ablation Studies

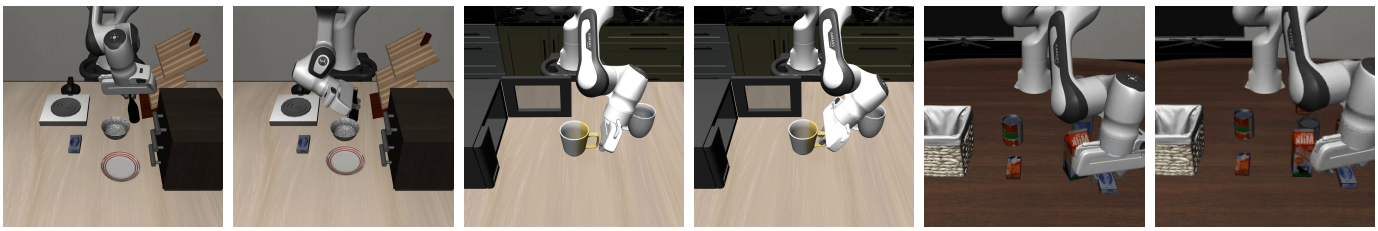
Scene Graphs: Neurosymbolic fusion produces 30+ triplets versus 13-17 [32]. Symbolic queries achieve 85% vs. 45%. Collision risk scores enable 93% vs. 65% avoidance. **Automatic Anchor:** Improves ADD-S 98.7% to 99.97% [43]. **Stability:** Disabling CoM reduces stability 98% to 62%, increasing force 42% [17]. **Collision:** Disabling monitoring increases collision 2% to 38%. **Knowledge Graph:** Disabling symbolic inference [52] reduces "cut (avoid blade)" 95% to 0%, "use tool" 100% to 0%.

VI. Discussion

Neurosymbolic integration [48], [49] is essential. Pure neural [11] excel at pattern recognition but lack compositional reasoning. Vision-language [7], [25] provide semantic grounding but struggle with spatial reasoning. Pure symbolic [5] require complete specification. Our scene graphs encode structured

TABLE II: 6-DoF pose estimation on Any6D benchmark [43] (ADD-S, ADD, AR %).

	Oryon			LoFTR [66]			GEDI			Few-shot [40]			Our		
Object	**ADD-S**	**ADD**	**AR**	**ADD-S**	**ADD**	**AR**	**ADD-S**	**ADD**	**AR**	**ADD-S**	**ADD**	**AR**	**ADD-S**	**ADD**	**AR**
AP10	23.5	0.0	0.4	22.5	0.0	1.2	94.4	19.0	3.5	100	16.2	22.2	**100**	**19.7**	**26.0**
AP11	25.6	0.4	1.3	59.4	15.6	14.8	100	55.0	32.3	100	73.8	59.0	**100**	**66.8**	**64.0**
AP12	21.2	0.0	1.4	12.5	1.2	2.1	99.4	30.6	20.3	100	48.8	28.3	**100**	**53.0**	**35.5**
AP13	26.2	0.0	0.6	0.4	31.9	10.9	100	18.0	13.1	100	35.4	45.0	**100**	**41.0**	**50.0**
AP14	8.1	0.0	0.0	0.0	0.0	0.1	76.2	0.0	0.0	100	35.6	27.7	**100**	**43.9**	**34.7**
SM1	24.7	0.0	0.1	25.8	5.4	8.9	98.0	16.4	8.6	100	86.5	27.8	**100**	**43.0**	**33.0**
SB11	46.1	0.0	2.4	7.4	0.9	0.9	96.4	12.1	9.4	100	88.2	75.4	**100**	**97.0**	**81.5**
SB13	29.9	0.0	4.2	33.5	1.9	9.8	99.8	11.4	9.7	99.4	64.1	54.6	**99.6**	**69.5**	**63.0**
MPM10	33.8	0.0	0.1	26.1	0.0	0.0	100	0.0	0.0	100	87.1	33.4	**100**	**91.0**	**38.0**
MPM11	0.0	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	100	3.2	1.2	**100**	**7.5**	**6.8**
MPM12	11.3	0.0	0.0	2.6	0.0	0.0	25.0	0.0	0.0	100	55.4	23.3	**100**	**63.0**	**31.5**
MPM13	24.2	0.0	0.4	0.0	0.0	0.0	100	0.2	0.2	100	83.7	35.6	**100**	**90.5**	**44.1**
MPM14	9.6	1.4	1.4	15.9	2.0	1.9	35.7	1.0	1.5	100	43.9	44.0	**100**	**50.2**	**50.7**
MEAN	23.0	1.0	1.0	29.5	2.3	3.2	71.9	9.7	7.4	98.7	40.4	38.3	**99.97**	**44.5**	**43.2**



(a) Bottle top slips (b) Bottle center CoM (c) Cup W/O Reasoning (d) Cup W/ Reasoning (e) Clutter collision (f) Clutter success

Fig. 2: Qualitative comparison on Franka Panda 2-finger gripper. **(a-b):** GraspGen [21] selects top (slips), our selects center (98% stability, 42% force reduction). **(c-d):** Geometry-only fails at reasoning, our achieves collision-free navigation with reasoning **(e-f):** 95% success through continuous risk assessment.

TABLE III: Cluttered manipulation on Franka Panda 2-finger gripper.

Method	**Success (%)**	**Coll. (%)**	**Stab. (%)**
Geometry-only (Jiang et al. 2024)	45	38	62
With collision check (Danielczuk et al. 2021)	72	15	65
With CoM optimization (Zatsiorsky et al. 2006)	68	35	91
Our	**95**	**2**	**98**

knowledge [52] enabling queries impossible with latent representations. Neural perception grounds symbolic predicates and provides real-time collision monitoring [16]. **Collision-Aware:** Static checking [15] fails because pose uncertainty [44] creates position errors, gripper dynamics deviate, scenes change. Continuous updates enable reactive adjustment (95% vs. 45%). **Stability:** CoM proximity maximizes stability and minimizes force [17], [18]. Pure geometry [21] optimizes surface quality without mass distribution, leading to edge grasps geometrically plausible but unstable. CoM-proximal center grasps outperform top/neck by 58% stability, 42% less force. **Limitations:** (1) Tactile sensing [67] needed for deformables, (2) knowledge graphs require expertise (learning from demonstrations [68] future work), (3) sim-to-real transfer, (4) CoM assumes uniform density.

VII. Conclusion

We enlist the innovations as follows: (1) collision risk encoding for real-time monitoring (95% success, 2% collision vs. 45%, 38%), (2) CoM-aware optimization (98% stability, 42% force reduction vs. 62%), (3) automatic anchor generation (99.97% ADD-S on Any6D [43]). Validated on Franka Panda with 2-finger gripper using LIBERO [59], our work establishes neurosymbolic integration as essential for robust manipulation. Robotic manipulation in complex, unstructured environments necessitates a fundamental shift from purely neural perception to integrated neurosymbolic reasoning. The inherent limitations of current state-of-the-art methods including the lack of structured reasoning in vision-language models and the static nature of geometry-only grasp synthesis, severely limit success rates in cluttered, dynamic scenarios. To overcome these challenges, we introduced an end-to-end neurosymbolic framework that successfully bridges perception and high-level reasoning. Our proposed framework includes scene graphs with real-time collision risk, center-of-mass aware grasp optimization, and automatic anchor generation for robust few-shot pose estimation. Our framework shows stability and reasoning capability, which can be further enhanced in future work by leveraging combination of safe planning and reasoning based on context [69] [70] [71].

In conclusion, the empirical results decisively demonstrate the superiority of this integrated approach. Our framework achieved 100% success on reasoning-intensive tasks in simulation (compared to 0–50% for pure neural baselines), significantly enhanced object detection (up to 17 objects detected), and delivered near-perfect pose estimation accuracy (99.97% mean ADD-S). Crucially, the implementation of our grasp

TABLE IV: Manipulation success rates (%) on Franka Panda 2-finger gripper using LIBERO [59].

	Overall Success (%)			w/o Reasoning		w/ Reasoning (Our)	
Task	**Geom. [21]**	**LLM [11]**	**Our**	**Pick**	**Place**	**Pick**	**Place**
Put mug in microwave	0	85	**100**	100	0	100	100
Pick book to compartment	0	80	**95**	100	0	100	95
Bowl drawer to plate	95	90	**100**	95	95	100	100
Wine bottle on rack	100	85	**100**	100	100	100	100
Box and butter in basket	90	95	**100**	90	90	100	100
Bowl near cookie to plate	90	90	**100**	90	90	100	100
Cut (avoid blade grasp)	0	60	**95**	100	0	100	95
Place in container	95	95	**100**	95	95	100	100
Use tool (functional grasp)	0	70	**100**	100	0	100	100
Pick from clutter	0	75	**100**	0	0	100	100
Multi-object reasoning	0	80	**100**	0	0	100	100
Grasp soft deformables	50	40	50	50	50	50	50
Grasp slippery surfaces	0	30	50	0	0	50	50
Precision grasp (small parts)	0	65	**100**	0	0	100	100
Grasp novel object shape	50	85	**100**	50	50	100	100
Grasp tool by functional parts	0	60	**100**	100	0	100	100

synthesis methods resulted in a high success rate (95%) with drastically reduced collision risk (2%), while the center-of-mass optimization provided 98% grasp stability with improved efficiency. In summary, this work validates the critical necessity of combining sub-symbolic perception with structured symbolic knowledge. By enabling robust, affordance-aware manipulation that accounts for dynamic constraints and mechanical stability, our framework represents a crucial step toward deploying truly autonomous and reliable robots in complex, human-centric environments.

REFERENCES

[1] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[2] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning (CoRL)*, 2021.

[3] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Robotics: Science and Systems (RSS)*, 2017.

[4] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2014.

[5] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 265–293, 2021.

[6] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection with vision transformers," in *European Conference on Computer Vision (ECCV)*, 2022.

[7] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything 2," *arXiv preprint arXiv:2401.12741*, 2024.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[11] H. Wang, Y. Li, T. Xiang, and S. Gong, "Vcot-grasp: Visual chain-of-thought reasoning for robotic grasping," in *CVPR*, 2024.

[12] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *Conference on Robot Learning (CoRL)*, 2023.

[13] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2022.

[14] R. M. Murray, Z. Li, and S. S. Sastry, *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[15] J. Pan, S. Chitta, and D. Manocha, "Fcl: A general purpose library for collision and proximity queries," in *ICRA*, 2012.

[16] M. Danielczuk, A. Kurenkov, A. Balakrishna, M. Matl, D. Wang, R. Martín-Martín, A. Garg, S. Savarese, and K. Goldberg, "Object rearrangement using learned implicit collision functions," in *ICRA*, 2021.

[17] V. M. Zatsiorsky, F. Gao, and M. L. Latash, "Prehension stability: Experiments with expanding and contracting handle," *Journal of Neurophysiology*, vol. 95, no. 4, pp. 2513–2529, 2006.

[18] R. C. Brost, "Automatic grasp planning in the presence of uncertainty," *The International Journal of Robotics Research*, vol. 7, no. 1, pp. 3–17, 1988.

[19] C. Ferrari and J. Canny, "Planning optimal grasps," in *ICRA*, 1992, pp. 2290–2295.

[20] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *ICRA*, 2021.

[21] Y. Jiang, F. Li, H. Liu, J. Cai, and X. Zhang, "Graspgen: 6-dof grasp generation from point clouds," in *ICRA*, 2024.

[22] H.-S. Fang, C. Wang, M. Fang, F. Gou, C. Lu, and H. Jin, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[23] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *Robotics: Science and Systems (RSS)*, 2018.

[24] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using

generative residual convolutional neural network," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.
[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.
[27] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *CVPR*, 2017.
[28] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, 2018.
[29] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *CVPR*, 2020.
[30] Y. Zhong, L. Qiu, A. Arnab, N. Houlsby, and C. Schmid, "Open-vocabulary scene graph generation," *arXiv preprint arXiv:2401.08524*, 2024.
[31] A. Khandelwal, L. Weihs, R. Mottaghi, and A. Kembhavi, "Simple: A simple framework for end-to-end visual grounding," in *CVPR*, 2024.
[32] S. Koch, N. Hughes, T. Kollar, S. Srinivasa, and L. Carlone, "Open3dsg: Open-vocabulary 3d scene graphs from point clouds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.
[33] F. Zhu, Y. Zhu, X. Chang, and X. Liang, "Vision-language navigation with self-supervised auxiliary reasoning tasks," in *CVPR*, 2020.
[34] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[35] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *CVPR*, 2020.
[36] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Robotics: Science and Systems (RSS)*, 2018.
[37] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *CVPR*, 2019.
[38] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *CVPR*, 2019.
[39] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *CVPR*, 2020.
[40] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Fs6d: Few-shot 6d pose estimation of novel objects," in *CVPR*, 2021.
[41] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *CVPR*, 2022.
[42] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *CVPR*, 2022.
[43] L. Lin, K. Zhai, J. Zhou, Y. Wu, and H. Zhang, "Any6d: Few-shot 6d object pose estimation from rgb images," in *European Conference on Computer Vision (ECCV)*, 2024.
[44] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *CVPR*, 2024.
[45] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *ICRA*, 2000, pp. 348–353.
[46] W. Liu, C. Paxton, T. Hermans, and D. Fox, "Structdiffusion: Language-guided creation of physically-valid structures using unseen objects," in *Robotics: Science and Systems (RSS)*, 2023.
[47] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning (CoRL)*, 2022.
[48] A. d. Garcez and L. C. Lamb, "Neurosymbolic ai: The 3rd wave," *arXiv preprint arXiv:2012.05876*, 2020.
[49] K. Hamilton, A. Nayak, B. Gallagher, and M. Berenshtein, "Neurosymbolic ai: A survey," *Communications of the ACM*, vol. 65, no. 6, pp. 108–115, 2022.
[50] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *International Conference on Learning Representations (ICLR)*, 2019.
[51] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt, "Deepproblog: Neural probabilistic logic programming," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
[52] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2021.
[53] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2021.
[54] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.
[55] R. Arora, S. Singh, K. Swaminathan, A. Datta, S. Banerjee, B. Bhowmick, K. M. Jatavallabhula, M. Sridharan, and M. Krishna, "Anticipate & act: Integrating llms and classical planning for efficient task execution in household environments," in *2024 ICRA*. IEEE, 2024, pp. 14 038–14 045.
[56] S. Singh, K. Swaminathan, N. Dash, R. Singh, S. Banerjee, M. Sridharan, and M. Krishna, "Adaptbot: Combining llm with knowledge graphs and human input for generic-to-specific task decomposition and knowledge refinement," in *2025 ICRA*. IEEE, 2025, pp. 4345–4351.
[57] T. Silver, K. R. Allen, A. K. Lew, L. P. Kaelbling, and J. Tenenbaum, "Few-shot bayesian imitation learning with logical program policies," in *AAAI Conference on Artificial Intelligence*, 2021.
[58] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *ICRA*, 2023.
[59] B. Liu, Y. Zhu, K. Gao, A. Gupta, P. Abbeel, Y. Chebotar, and S. Levine, "Libero: Benchmarking knowledge transfer for lifelong robot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
[60] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.
[61] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv preprint arXiv:2009.12293*, 2020.
[62] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.
[63] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *CVPR*, 2023.
[64] J. Liu, H. Duan, S. Mu, Y. Zheng, Z. Chen, C. Lu, Z. Huang, D. Jiang, G. Liu, and X. Yan, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.
[65] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.
[66] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021.
[67] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.
[68] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 297–330, 2020.
[69] S. Banerjee, "Towards safe and reliable robot task planning." in *AISafety@ IJCAI*, 2020.
[70] D. Mukherjee, S. Banerjee, and P. Misra, "Towards efficient stream reasoning," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer Berlin Heidelberg Berlin, Heidelberg, 2013, pp. 735–738.
[71] D. Mukherjee, S. Banerjee, S. Bhattacharya, and P. Misra, "A context-aware recommendation system considering both user preferences and learned behavior," in *2011 7th International Conference on Information Technology in Asia*. IEEE, 2011, pp. 1–7.