# Multi-Modal Large Model for Robotic Grasping with Multi-Manifold Priors

Haibo Li[1,2,†], Qin Lai[1,2,†], Yaming Yang[1,2], Yaoxin Chen[3,2], and Qicong Wang[1,2,*]
[1]*Department of Computer Science and Technology, School of Informatics*, Xiamen University, Xiamen 361005, China
[2]*Shenzhen Research Institute*, Xiamen University, Shenzhen 518000, China
[3]*Institute of Artificial Intelligence*, Xiamen University, Xiamen 361000, China
Email: {lihaibo, 31520221154207, jaminyang13, chenyaoxin}@stu.xmu.edu.cn, qcwang@xmu.edu.cn

***Abstract*—Executing instruction-conditioned grasping in cluttered scenes remains challenging because semantic grounding and physical feasibility must be handled jointly. We propose a multi-modal 3D perception framework guided by multi-manifold priors that impose geometric constraints for robust grasp prediction. To mitigate error accumulation in cascaded pipelines, we adopt a parallel dual-branch architecture: a semantic branch grounds language in vision, and a geometric branch generates physically plausible grasp candidates from point clouds. Our key idea is to encode grasp representations in non-Euclidean spaces: hyperbolic space captures hierarchical organization of grasp regions, while spherical space preserves directional consistency for grasp orientation. Building on this principle, we introduce: (i) a shared-attention point cloud encoder to strengthen cross-layer feature consistency, (ii) a scene-aware and geometry-enhanced grasp pose generation network, and (iii) multi-manifold consistency learning with geodesic constraints. Experiments demonstrate improved accuracy and robustness in complex environments.**

***Index Terms*—Embodied AI, Multi-modal Large Models, Riemannian Manifold**

## I. Introduction

Accurately executing embodied actions in cluttered environments from human intent remains a central challenge in robotics. Embodied agents must jointly perceive 3D geometry, interpret language goals, and output actions that satisfy physical constraints. Despite rapid progress in vision-language models and robotic manipulation, many systems still decouple semantic grounding from geometric reasoning, which often leads to brittle behavior in cluttered, contact-rich scenes.

Traditional geometry-based grasp planning methods [1]–[6] rely on precise geometric models, which struggle to generalize to novel objects or open-world semantic instructions. To overcome this, deep learning approaches [7]–[10] improve generalization by learning grasp policies from data. However, most deep methods learn in Euclidean spaces—suboptimal for representing hierarchical and directional affordance structures—and often predict grasps without explicit language grounding, limiting their use in instruction-driven settings.

Recent open-vocabulary methods [11], [12] use large language models (LLMs) and vision-language models to enable instruction-driven manipulation. In many cases, however, they rely on cascaded pipelines in which errors propagate from perception to grounding and then to action selection. Information bottlenecks between modules and limited geometric reasoning further reduce reliability.

To address these challenges, we propose a multimodal framework guided by multi-manifold priors for embodied grasping. Our key insight is that grasp representations exhibit intrinsic geometric structure that is better modeled in non-Euclidean spaces. We use hyperbolic space to represent hierarchical relationships among grasp regions and spherical space to preserve directional consistency for grasp orientations. By mapping features to these Riemannian manifolds, we introduce geometric priors that regularize feature learning and improve grasp prediction.

Our framework adopts a parallel dual-branch architecture to reduce error accumulation. The semantic branch grounds language instructions in visual scenes, while the geometric branch generates physically plausible grasp candidates from point cloud features. A cross-attention module aligns the two representations to enable instruction-aware grasp selection. Experiments on GraspNet-1Billion show that our approach achieves 39.64 AP on unknown objects, surpassing prior best results by 48.5%. In instruction-driven simulation, we achieve an 81.3% success rate, validating robust semantic grounding and geometric reasoning.

The main contributions of this work are:

- A parallel dual-branch framework that integrates semantic understanding with geometric reasoning while avoiding cascaded error accumulation.
- A shared-attention point cloud encoder that leverages cross-layer correlations as structural priors for improved feature quality.
- A scene-aware grasp pose generation network combining utility-aware sampling with explicit geometric primitives.
- Multi-manifold feature learning that maps representations to hyperbolic and spherical spaces, providing strong geometric priors through geodesic constraints.

## II. Related Work

### A. Grasp Detection Methods

Early geometry-based methods evaluated grasp stability via force closure [1] or grasp wrench space [2]. While offering interpretable guarantees, they rely on accurate object models

* Corresponding author
† Equal contribution

and struggle in cluttered scenes, even when extended to task-oriented planning [3] and shape reconstruction [4]–[6]. Data-driven methods improve generalization by learning from data. Simulation-based approaches [7] suffer from domain gaps, and heuristic sampling [8], [9] heavily depends on hand-designed candidates. End-to-end methods directly process point clouds using graspability guidance [10], generative models [13], or multi-scale fusion [14]. However, these typically operate in Euclidean spaces and rarely incorporate non-Euclidean geometric priors or instruction grounding, limiting robustness to novel objects and language tasks.

### B. Embodied Intelligence and Point Cloud Processing

Semantic-guided interaction integrates perception and language for embodied agents. Early multi-task [11] and multi-modal networks [12] couple vision with language but struggle with novel instructions and complex scenes. While Large Language Models [15], [16] offer strong semantic priors, they lack visual grounding. Multi-modal Large Models [17]–[19] address this but often suffer from error accumulation across cascaded pipelines [20]. For point cloud processing, early voxelization [21] and PointNet-style models [22], [23] face discretization artifacts or independent feature learning. PointCNN [24] and graph networks [25], [26] improve local modeling but encounter scalability issues. Recent Transformers [27], [28] capture global dependencies but operate in Euclidean space, failing to explicitly encode the hierarchical and directional structure of grasp affordances or tightly integrate non-Euclidean priors with instruction-conditioned grasping.

## III. Methodology

### A. Overall Network Framework

We propose a multimodal framework that integrates semantic grounding and geometric reasoning. As shown in Fig. 1, given a point cloud, an RGB image, and an instruction, the model outputs a 6-DoF grasp pose through (a) semantic understanding, (b) geometric candidate generation, (c) cross-attention alignment, and (d) policy optimization.

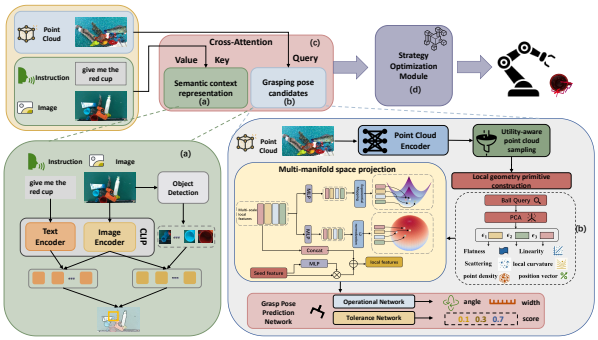


Fig. 1. Overview of the proposed framework. A semantic branch grounds the instruction in the visual scene, while a geometric branch generates physically plausible grasp candidates from point clouds. Cross-attention aligns the two representations, and a policy module selects the final command-consistent 6-DoF grasp.

The semantic branch grounds the instruction in the visual scene. Given an RGB image $I$ and a text instruction $T$, we first detect object bounding boxes $\mathbf{B} = \{b_i\}_{i=1}^{N}$ and embed them as:

$$\mathbf{F}_b = \text{MLP}(\mathbf{B}) \tag{1}$$

We also extract visual features $\mathbf{F}_v = \text{Enc}_v(I)$ and text features $\mathbf{F}_t = \text{Enc}_t(T)$ using pre-trained encoders.

We apply cross-modal attention to align the language and visual representations. First, text attends to image features:

$$\mathbf{F}_{tv} = \text{CrossAttn}(\mathbf{F}_t, \mathbf{F}_v, \mathbf{F}_v) \tag{2}$$

Second, image features attend to bounding-box embeddings to refine spatial localization:

$$\mathbf{F}_{vb} = \text{CrossAttn}(\mathbf{F}_v, \mathbf{F}_b, \mathbf{F}_b) \tag{3}$$

Finally, we fuse the two streams to obtain the visual-language representation:

$$\mathbf{F}_{vl} = \text{CrossAttn}(\mathbf{F}_{tv}, \mathbf{F}_{vb}, \mathbf{F}_{vb}) \tag{4}$$

This fusion preserves fine-grained spatial cues while maintaining semantic alignment across modalities.

The geometric branch encodes the point cloud $\mathbf{P}$ with shared-attention to obtain features $\mathbf{F}_p = \text{Enc}_p(\mathbf{P})$, then generates $K$ geometry-aware grasp candidates $\mathcal{A} = \{a_k\}_{k=1}^{K}$. Each candidate $a_k = (\mathbf{t}_k, \mathbf{R}_k, w_k)$ specifies translation, rotation, and gripper width.

We then align semantic and geometric features via cross-attention for instruction-aware grasp selection:

$$\mathbf{H} = \text{CrossAttn}(\mathbf{F}_p, \mathbf{F}_{vl}, \mathbf{F}_{vl}) \tag{5}$$

where point-cloud features $\mathbf{F}_p$ are queries and visual-language features $\mathbf{F}_{vl}$ are keys/values, enabling geometry features to attend to task-relevant regions.

**Grasp Prediction Network.** The fused representation $\mathbf{H}$ is fed into a grasp prediction network with two heads: operation and tolerance. The operation head predicts in-plane rotation $\theta \in [0, 180°]$, approach distance $d$, gripper width $\omega$, and grasp confidence $s$. To encourage pose consistency, we group candidate grasps from the same object into cylindrical perception regions. The tolerance head estimates robustness by searching a local neighborhood on the sphere and computing the maximum perturbation $T$ that keeps the grasp score above a threshold (e.g., 0.5). The final grasp is represented as:

$$G = \{\theta, d, \omega, s, T\} \tag{6}$$

where $T$ serves as a tolerance metric quantifying grasp robustness.

**Baseline Loss Functions.** Following [29], the baseline grasp detection loss $\mathcal{L}_{\text{grasp}}$ combines classification and regression objectives for grasp operation and robustness tolerance:

$$\mathcal{L}_{\text{grasp}} = \mathcal{L}_{cls}^{rot} + \mathcal{L}_{reg}^{d} + \mathcal{L}_{reg}^{\omega} + \mathcal{L}_{reg}^{s} + \lambda_T \mathcal{L}_{reg}^{tol} \tag{7}$$

where $\mathcal{L}_{cls}^{rot}$ is the rotation cross-entropy loss, $\mathcal{L}_{reg}^{d,\omega,s,tol}$ are smooth L1 losses for approach distance, gripper width, confidence, and tolerance, respectively. We set $\lambda_T = 0.1$.

However, this baseline provides limited geometric constraints for representation learning. To address this limitation, we augment $\mathcal{L}_{\text{grasp}}$ with multi-manifold regularization (Sec. III-D) to inject stronger geometric priors for robust grasp prediction.

**Two-Stage Training Strategy.** We adopt a two-stage training procedure. In Stage 1, the operation and tolerance heads are trained with $\mathcal{L}_{\text{SGN}}$ to generate grasp candidates under multi-manifold priors. In Stage 2, the policy module is trained on simulated expert demonstrations to select an instruction-consistent grasp. We minimize cross-entropy between predicted action probabilities and expert labels:

$$\mathcal{L}_{ce} = -\sum_{k=1}^{K} y_k \log(\omega_k) \tag{8}$$

where $y_k$ is the one-hot encoding of expert actions, and $\omega_k$ is the predicted probability for the $k$-th candidate. This design combines vision-language semantic guidance with physics-aware grasp constraints, enabling stable, target-specific grasping in cluttered scenes.

### B. Shared-Attention-based Point Cloud Encoder

Point cloud feature extraction is critical for grasp prediction. In cluttered scenes with unknown objects, robust grasping often depends on subtle correlations between fine-scale contact cues and coarse global context. However, many encoders parameterize layers independently and underutilize cross-layer correlations, so shallow and deep features may focus on inconsistent regions, which makes it difficult to form stable 6-DoF grasps. We introduce a shared-attention mechanism that injects learnable bias vectors as structural priors to improve cross-layer consistency.

Motivated by the observation that attention patterns across layers are often correlated, we introduce a shared learnable vector $\phi_{\text{share}}$ for each skip-connected encoder-decoder pair. The attention computation is:

$$\begin{aligned} f_i = \sum_{x_j \in \mathcal{X}(i)} & \rho(\psi_4(\psi_1(x_i) - \psi_2(x_j) + \Delta) \\ & + \varphi(\phi_{\text{share}})) \odot (\psi_3(x_j) + \Delta) \end{aligned} \tag{9}$$

where $f_i$ is the output feature at center point $x_i$, $\mathcal{X}(i)$ denotes its local neighborhood, $\psi_{1,2,3,4}$ are feature transformations, $\Delta$ is positional encoding, and $\rho$ is Softmax normalization. The shared vector $\phi_{\text{share}}$ is transformed by $\varphi$ to form a bias term, improving cross-layer consistency.

### C. Scene-Aware Grasp Pose Generation

Efficient grasp pose generation in cluttered scenes requires informative seed points and reliable direction estimation. Uniform sampling produces many low-quality candidates, while purely implicit features offer limited geometric interpretability. In practice, we observe that such designs tend to sample grasps on background or severely occluded points and often predict plausible-looking but physically unstable orientations. We propose a scene-aware generation network that combines utility-aware sampling with explicit geometric primitives to directly address these issues.

As shown in Fig. 2, we propose: (a) utility-aware sampling, (b) explicit geometric primitives, and (c) adaptive aggregation.

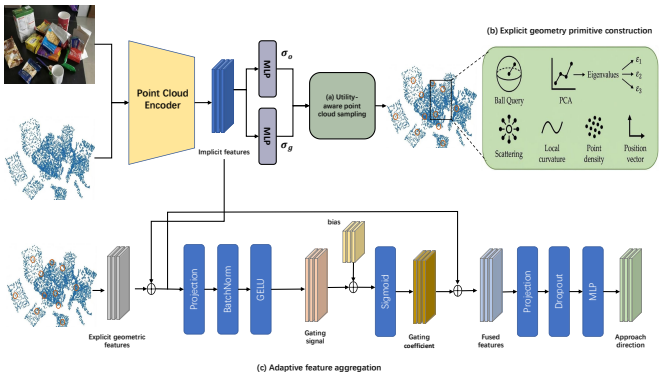


Fig. 2. This module integrates: (a) utility-aware point cloud sampling for seed points; (b) explicit geometric primitive extraction for features; (c) adaptive feature aggregation for direction prediction.

Our utility-aware sampling selects points with highest grasp potential using objectness $\sigma_o$ and grasp likelihood $\sigma_g$:

$$\eta_{wg}(\varsigma, \Gamma) = \eta(\varsigma, \Gamma) \cdot \exp(\sigma_o + \sigma_g) \tag{10}$$

where $\eta(\varsigma, \Gamma)$ is the Euclidean distance from point $\varsigma$ to the selected set $\Gamma$, $\sigma_o$ is the objectness score, and $\sigma_g$ is the grasp likelihood score.

Higher $\sigma_o$ and $\sigma_g$ prioritize graspable points. We apply PCA to the neighborhood $N(\varsigma^*)$ to obtain eigenvalues $(\epsilon_1 \geq \epsilon_2 \geq \epsilon_3)$:

$$\chi_p = \frac{\epsilon_3}{\epsilon_1}, \quad \chi_l = \frac{\epsilon_1 - \epsilon_2}{\epsilon_1}, \quad \chi_s = \frac{\epsilon_3}{\epsilon_1 + \epsilon_2 + \epsilon_3} \tag{11}$$

representing planarity, linearity, and scattering, respectively. These descriptors capture local geometry: $\chi_p$ near 0 indicates planar regions (often suitable for stable grasps), $\chi_l$ close to 1 suggests linear structures (e.g., edges), and $\chi_s$ reflects neighborhood isotropy. We compute the surface normal $\mathbf{n}$ as the eigenvector associated with $\epsilon_3$, which provides orientation cues for grasp alignment.

Beyond these descriptors, we compute additional geometric features. Curvature $\kappa$, point density $\zeta$, and relative offset $\mathbf{f}_{\text{rel}}$ are defined as:

$$\kappa = \frac{\epsilon_3}{\epsilon_1 + \epsilon_2 + \epsilon_3}, \quad \zeta = \frac{|N_r(\varsigma^*)|}{V_r}, \quad \mathbf{f}_{\text{rel}} = \varsigma^* - \bar{\varsigma} \tag{12}$$

where $\kappa$ measures surface curvature, $\zeta$ quantifies point density within radius $r$ and search volume $V_r = \frac{4}{3}\pi r^3$, and $\mathbf{f}_{\text{rel}}$ encodes the offset to the local centroid $\bar{\varsigma}$ of $N(\varsigma^*)$.

All features are concatenated into a comprehensive explicit feature vector:

$$\mathbf{f}_{\text{exp}} = [\chi_p, \chi_l, \chi_s, \mathbf{n}, \kappa, \zeta, \mathbf{f}_{\text{rel}}] \tag{13}$$

This explicit representation provides interpretable geometric cues that complement implicit features learned by neural

networks. Adaptive fusion then combines the explicit features with deep features from the point cloud encoder, enabling direction prediction that is both geometrically plausible and semantically informed.

### D. Multi-Manifold Feature Learning

Euclidean features often fail to capture the hierarchical and directional structure of grasp representations and offer limited inductive bias for multi-scale consistency. In particular, graspable regions naturally form a hierarchy from objects to parts and local patches, while approach directions lie on a sphere. Forcing all these patterns into a single Euclidean space makes the network learn such structure implicitly, which is data-inefficient and harms generalization to novel objects. We address this by projecting features into hyperbolic and spherical spaces and enforcing consistency using geodesic distances.

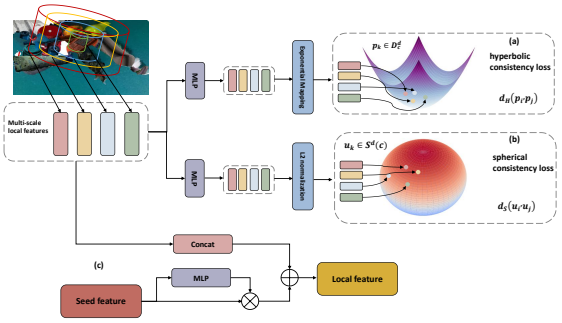


Fig. 3. Multi-manifold feature learning maps features to hyperbolic (a) and spherical (b) spaces for hierarchical and directional modeling, enforced by geodesic consistency. Features are then fused with seed points (c) for grasp prediction.

As shown in Fig. 3, we map multi-scale features to hyperbolic space to model hierarchy and to spherical space to model directional cues. For the hyperbolic branch, we map features to the Poincaré ball $\mathcal{D}_c^d = \{x \in \mathbb{R}^d \mid \|x\| < \frac{1}{\sqrt{c}}\}$:

$$p = Exp_{\mathbf{0}}^c(\mathbf{z}) = \tanh(\sqrt{c}\|\mathbf{z}\|)\frac{\mathbf{z}}{\sqrt{c}\|\mathbf{z}\|} \tag{14}$$

For the spherical branch, we map features to the sphere $\mathcal{S}^d(c) = \{x \in \mathbb{R}^{d+1} \mid \|x\| = \frac{1}{\sqrt{c}}\}$:

$$u = Exp_{\mathbf{n}}^c(\mathbf{z}) = \cos(\sqrt{c}\|\mathbf{z}\|)\mathbf{n} + \sin(\sqrt{c}\|\mathbf{z}\|)\frac{\mathbf{z}}{\sqrt{c}\|\mathbf{z}\|} \tag{15}$$

where $\mathbf{n}$ is the north pole of the sphere. We extract features at $N$ perception radii and project them to obtain multi-scale embeddings $\{p_1, \ldots, p_N\}$ (hyperbolic) and $\{u_1, \ldots, u_N\}$ (spherical). By constraining geodesic distances across scales, the model learns a scale-consistent representation. The overall objective $\mathcal{L}_{\text{SGN}}$ for grasp generation is:

$$\mathcal{L}_{\text{SGN}} = \mathcal{L}_{\text{grasp}} + \mathcal{L}_{\text{reg}} \tag{16}$$

where the regularization term $\mathcal{L}_{\text{reg}}$ is defined as:

$$\mathcal{L}_{\text{reg}} = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} [\lambda_{\text{hyp}} d_{\mathcal{H}}(p_i, p_j) + \lambda_{\text{sph}} d_{\mathcal{S}}(u_i, u_j)] \tag{17}$$

where $\mathcal{L}_{\text{grasp}}$ is the grasp prediction loss, $\lambda_{\text{hyp}}$ and $\lambda_{\text{sph}}$ are balancing weights. $d_{\mathcal{H}}(\cdot,\cdot)$ and $d_{\mathcal{S}}(\cdot,\cdot)$ are geodesic distances in hyperbolic and spherical spaces, measuring feature consistency across scales, defined as:

$$d_{\mathcal{H}}(p,q) = \frac{1}{\sqrt{c}} \operatorname{arccosh}\left(1 + \frac{2c\|p-q\|_2^2}{(1-c\|p\|_2^2)(1-c\|q\|_2^2)}\right) \tag{18}$$

$$d_{\mathcal{S}}(u,v) = \frac{1}{\sqrt{c}} \arccos\left(c\langle u, v\rangle\right) \tag{19}$$

## IV. Experimental Results and Analysis

### A. Datasets and Experimental Settings

We evaluate our framework on the GraspNet-1Billion dataset [30], which provides a large-scale benchmark for 6-DoF grasp pose estimation in cluttered scenes. The dataset contains 97,280 RGB-D images from 190 scenes, featuring 88 unique objects with diverse geometries and physical properties. To rigorously test generalization, the objects are partitioned into three splits: *Seen* (objects included in the training set), *Similar* (novel objects from categories present in training), and *Unknown* (novel objects from unseen categories). We adopt the standard evaluation metric, Average Precision (AP), which considers a grasp successful if it is physically stable and avoids collisions. AP is reported under different friction coefficients: mean AP (averaged over $\mu \in [0.2, 1.2]$), $\text{AP}_{0.8}$ (representing high-friction surfaces), and $\text{AP}_{0.4}$ (representing challenging low-friction surfaces).

For embodied interaction, we conduct instruction-driven grasping in a PyBullet simulation using a UR5 arm with a parallel-jaw gripper. Scenes contain 5-10 randomly placed objects viewed from three fixed cameras. A task is successful if the target object is lifted and held for 2 seconds without significant displacement of others. We report the success rate over 100 trials.

We train our model using Adam (initial learning rate $1 \times 10^{-3}$, decaying by 0.1 every 80 epochs) for 360 epochs with a batch size of 2. We sample 20,000 points per scene, applying random rotation, scaling, and jittering for augmentation, and generate $K = 15$ grasp candidates. Experiments are run on an Intel i9-10900K CPU and an NVIDIA RTX 3090 GPU.

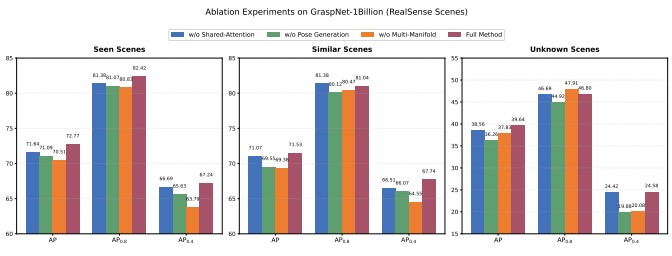


Fig. 4. Ablation Experiments on GraspNet-1Billion (RealSense Scenes)

### B. Ablation Experiments

We conduct ablation studies on GraspNet-1Billion (RealSense) to analyze each component, as summarized in Fig. 4.

TABLE I
ABLATION OF LOSS WEIGHTS ON GRASPNET

| Loss Weights | | Seen | | | Similar | | | Unknown | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_{hyp}$ | $\lambda_{sph}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ |
| 0.05 | 0.05 | **72.77** | **82.42** | 67.24 | **71.53** | 81.04 | **67.74** | **39.64** | 46.80 | **24.58** |
| 0.1 | 0.1 | 71.71 | 81.29 | 67.16 | 70.61 | 80.90 | 66.07 | 36.14 | 45.98 | 22.09 |
| 0.05 | 0.1 | 69.66 | 79.60 | 64.15 | 69.38 | 80.36 | 64.21 | 38.57 | 46.87 | 22.76 |
| 0.1 | 0.05 | 71.50 | 81.77 | 67.08 | 70.27 | 79.47 | 66.74 | 37.82 | 44.53 | 23.85 |
| 0 | 0.05 | 72.02 | 81.73 | 66.76 | 70.52 | 79.75 | 65.27 | 33.52 | 40.14 | 19.89 |
| 0.05 | 0 | 70.70 | 81.47 | 65.30 | 71.09 | 81.25 | 67.23 | 38.84 | **47.31** | 23.62 |
| 0.04 | 0.06 | 71.72 | 81.97 | **67.45** | 70.37 | 80.54 | 66.69 | 36.90 | 45.15 | 22.97 |
| 0.06 | 0.04 | 70.41 | 80.04 | 64.94 | 68.08 | 78.24 | 63.67 | 35.09 | 41.51 | 22.18 |

TABLE II
COMPARISON WITH SOTA ON GRASPNET

| Method | Seen | | | Similar | | | Unknown | | |
|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ |
| GraNet [31] | 41.48 | 49.84 | 33.86 | 35.29 | 43.15 | 26.89 | 11.57 | 14.31 | 5.24 |
| Scale-Bal. [14] | 58.95 | 68.18 | 54.88 | 52.97 | 63.24 | 46.99 | 22.63 | 28.53 | 12.00 |
| Real-to-Sim [32] | 64.71 | 76.09 | 58.00 | 56.93 | 66.21 | 52.78 | 26.31 | 32.70 | 15.63 |
| SGSIN [33] | 68.97 | 81.25 | 62.25 | 58.76 | 72.23 | 47.84 | 26.69 | 33.33 | 14.53 |
| FastGNet [34] | 70.42 | 82.19 | 64.95 | 60.65 | 73.43 | 52.18 | 25.96 | 32.47 | 13.98 |
| ZeroGrasp [35] | 70.53 | 82.28 | 64.26 | 62.51 | 74.26 | 54.97 | 26.46 | 33.13 | 15.11 |
| Mink-GraS. [36] | 71.20 | **82.84** | 65.97 | 61.47 | 74.44 | 53.06 | 25.06 | 31.31 | 13.52 |
| **Ours** | **72.77** | 82.42 | **67.24** | **71.53** | **81.04** | **67.74** | **39.64** | **46.80** | **24.58** |

TABLE III
PYBULLET SIMULATION RESULTS

| Method | Success Rate (%) | Num Actions |
|---|---|---|
| GraspSplats [37] | 58.0 | **2.05** |
| FoundationGrasp [38] | 73.7 | 4.70 |
| ThinkGrasp [39] | 74.5 | 4.11 |
| **Ours** | **81.3** | 4.04 |

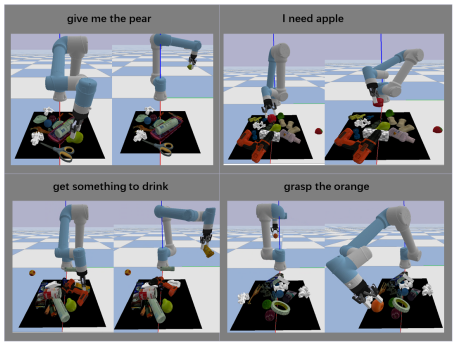


Fig. 5. Grasping visualization of instruction-driven tasks.

Removing the shared-attention mechanism drops AP on unknown objects (39.64 → 38.56), indicating that learnable bias vectors effectively promote semantic consistency across scales. Excluding utility-aware sampling and explicit geometric primitives causes the largest drop (39.64 → 36.26 AP), showing that purely implicit features miss fine-grained details, whereas explicit primitives provide crucial inductive biases. Finally, removing multi-manifold regularization reduces AP to 37.83, confirming that projecting features into hyperbolic and spherical spaces better captures hierarchical and directional structures than Euclidean embeddings.

Table I ablates the loss weights $\lambda_{hyp}$ and $\lambda_{sph}$. Setting either to zero degrades performance, confirming their complementarity. The best balance is $\lambda_{hyp} = \lambda_{sph} = 0.05$. Larger weights (e.g., 0.1) over-constrain the features, limiting flexibility. The hyperbolic branch mainly aids the *Unknown* split by capturing structural complexity, while the spherical branch improves directional stability under high friction ($AP_{0.8}$). Slight variations (0.04 and 0.06) show the model is stable and symmetric weighting yields optimal results.

### C. Comparison Experiments

Table II compares our method with state-of-the-art 6-DoF grasp detection approaches, where our framework achieves superior performance across all splits. Notably, it achieves 39.64 AP on the *Unknown* split, significantly outperforming SGSIN (26.69) and Mink-GraSNet (25.06). This highlights the generalization capability of multi-manifold priors for novel geometries. The *Seen* (72.77 AP) and *Similar* (71.53 AP) results also confirm that non-Euclidean embeddings provide a more discriminative feature space than standard Euclidean methods. Furthermore, our method shows exceptional robustness under low-friction ($\mu = 0.4$), reaching 24.58 AP on unknown objects (vs. ZeroGrasp's 15.11). This stability stems from our scene-aware generation module, which uses explicit geometric primitives to filter unstable candidates early.

For instruction-driven grasping, Table III shows our method achieves an 81.3% success rate in PyBullet simulations, outperforming FoundationGrasp (73.7%) and ThinkGrasp (74.5%). With a lower average of 4.04 actions per task, our model effectively grounds language into 3D geometry and generates high-quality initial candidates, reducing the need for iterative corrections in long-horizon tasks.

### D. Qualitative Analysis

Fig. 5 qualitatively illustrates that our framework can handle instruction-driven grasping in cluttered environments with objects of diverse shapes. The visualized examples are consistent with the quantitative results in Tables II and III, showing that the proposed method generates stable grasp poses for complex scenes without requiring additional heuristic post-processing.

## V. CONCLUSION

This paper introduces a multimodal 3D perception framework for robotic grasping, utilizing hyperbolic and spherical manifold priors to encode hierarchical structure and directional consistency. By integrating a shared-attention encoder and a scene-aware generation module, our method achieves physically plausible, instruction-aligned grasps. Experiments on GraspNet-1Billion demonstrate superior generalization, achieving 39.64 AP on unknown objects and an 81.3% success rate in PyBullet instruction-driven simulations. Future work will explore long-horizon tasks and multi-agent systems.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62571464, the Shenzhen Science and Technology Projects under Grant No. JCYJ20200109143035495, and the Natural Science Foundation of Fujian Province under Grant 2023J01003.

## REFERENCES

[1] J. K. Salisbury and B. Roth, "Kinematic and force analysis of articulated mechanical hands," *J*, 1983.

[2] C. Ferrari and J. Canny, "Planning optimal grasps," in *Proceedings., 1992 IEEE International Conference on Robotics and Automation*. IEEE, 1992, pp. 2290–2295.

[3] Z. Li and S. S. Sastry, "Task-oriented optimal grasping by multifingered robot hands," *IEEE Journal on Robotics and Automation*, vol. 4, no. 1, pp. 32–44, 2002.

[4] J. Bohg, M. Johnson-Roberson, B. León *et al.*, "Mind the gap-robotic grasping under incomplete observation," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 686–693.

[5] K. Hsiao, S. Chitta, M. Ciocarlie *et al.*, "Contact-reactive grasping of objects with partial shape information," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 1228–1235.

[6] A. T. Miller, S. Knoop, H. I. Christensen *et al.*, "Automatic grasp planning using shape primitives," in *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, vol. 2. IEEE, 2003, pp. 1824–1829.

[7] J. Mahler, J. Liang, S. Niyaz *et al.*, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Robotics: Science and Systems*, 2017.

[8] H. Liang, X. Ma, S. Li *et al.*, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3629–3635.

[9] A. Murali, A. Mousavian, C. Eppner *et al.*, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[10] C. Wang, H. S. Fang, M. Gou *et al.*, "Graspness discovery in clutters for fast and accurate grasp detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.

[11] Y. Li, T. Kong, R. Chu *et al.*, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3571–3578.

[12] Y. Chen, R. Xu, Y. Lin *et al.*, "A joint network for grasp detection conditioned on natural language commands," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4576–4582.

[13] G. Singh, S. Kalwar, M. F. Karim *et al.*, "Constrained 6-dof grasp generation on complex shapes for improved dual-arm manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7344–7350.

[14] H. Ma and D. Huang, "Towards scale balanced 6-dof grasp detection in cluttered scenes," in *Conference on robot learning*. PMLR, 2023, pp. 2004–2013.

[15] J. Liao, H. Zhang, H. Qian *et al.*, "Decision-making in robotic grasping with large language models," in *International Conference on Intelligent Robotics and Applications*. Singapore: Springer Nature Singapore, 2023, pp. 424–433.

[16] J. Xu, S. Jin, Y. Lei *et al.*, "Rt-grasp: Reasoning tuning robotic grasping via multi-modal large language model," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7323–7330.

[17] Y. Lu, Y. Fan, B. Deng *et al.*, "Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 976–983.

[18] M. Li, Q. Zhao, S. Lyu *et al.*, "Ovgnet: A unified visual-linguistic framework for open-vocabulary robotic grasping," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7507–7513.

[19] Q. Sun, H. Lin, Y. Fu *et al.*, "Language guided robotic grasping with fine-grained instructions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1319–1326.

[20] P. Xie, S. Chen, Y. Dai *et al.*, "Target-oriented object grasping via multimodal human guidance," in *European Conference on Computer Vision*. Cham: Springer, 2025, pp. 305–322.

[21] Z. Wu, S. Song, A. Khosla *et al.*, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[22] C. R. Qi, H. Su, K. Mo *et al.*, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[23] C. R. Qi, L. Yi, H. Su *et al.*, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[24] Y. Wang, Y. Sun, Z. Liu *et al.*, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[25] L. Wang, Y. Huang, Y. Hou *et al.*, "Graph attention convolution for point cloud semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 296–10 305.

[26] J. Lee, Y. Lee, J. Kim *et al.*, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 3744–3753.

[27] M. H. Guo, J. X. Cai, Z. N. Liu *et al.*, "Pct: Point cloud transformer," *Computational visual media*, vol. 7, pp. 187–199, 2021.

[28] H. Zhao, L. Jiang, J. Jia *et al.*, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268.

[29] H. Ma, M. Shi, B. Gao *et al.*, "Generalizing 6-dof grasp detection via domain prior knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024, pp. 18 102–18 111.

[30] H. S. Fang, C. Wang, M. Gou *et al.*, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[31] H. Wang, W. Niu, and C. Zhuang, "Granet: A multi-level graph network for 6-dof grasp pose generation in cluttered scenes," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 937–943.

[32] J.-F. Cai, Z. Chen, X.-M. Wu *et al.*, "Real-to-sim grasp: Rethinking the gap between simulation and real world in grasp detection," in *CoRL*, 2024, pp. 1109–1124.

[33] W. Wang, H. Zhu, and M. H. Ang Jr, "Sgsin: Simultaneous grasp and suction inference network via attention-based affordance learning," *IEEE Transactions on Industrial Electronics*, 2024.

[34] Z. Ding, A. Wang, M. Gao *et al.*, "Fastgnet: an efficient 6-dof grasp detection method with multi-attention mechanisms and point transformer network," *Measurement Science and Technology*, vol. 35, no. 4, p. 045020, 2024.

[35] S. Iwase, M. Z. Irshad, K. Liu *et al.*, "Zerograsp: Zero-shot shape reconstruction enabled robotic grasping," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 405–17 415.

[36] Y. Lyu, J. Li, A. Wang *et al.*, "Mink-grasnet: computationally efficient 6-dof grasp detection for large-scale ai-powered robotics," *The Journal of Supercomputing*, vol. 81, no. 11, pp. 1–23, 2025.

[37] M. Ji, R.-Z. Qiu, X. Zou *et al.*, "Graspsplats: Efficient manipulation with 3d feature splatting," in *CoRL*, 2024, pp. 1443–1460.

[38] C. Tang, D. Huang, W. Dong *et al.*, "Foundationgrasp: Generalizable task-oriented grasping with foundation models," *IEEE Transactions on Automation Science and Engineering*, 2025.

[39] Y. Qian, X. Zhu, O. Biza *et al.*, "Thinkgrasp: A vision-language system for strategic part grasping in clutter," in *CoRL*, 2024, pp. 3568–3586.