Decoupling Geometry from Semantics: Open-Vocabulary Grasping via SE(3) Diffusion

Yichen Xiao, Haoning Wu, Youyuan Tu, Xiaoping Wu†, Xiaoguang Niu†
School of Computer Science, Wuhan University, Wuhan, Hubei, China
{2023302111407, haoning.wu, yoyuan, wuxp, xgniu}@whu.edu.cn

***Abstract*—Open-vocabulary grasping requires learning physically valid actions on the non-Euclidean $\mathbf{SE(3)}$ manifold while aligning them with ambiguous, long-tailed semantic instructions. Existing approaches either bias grasp generation through end-to-end semantic conditioning, compromising geometric stability, or rely on task-specific semantic supervision that limits generalization.**

We propose a geometry–semantics decoupled framework that explicitly separates generative modeling from semantic decision-making. An $\mathbf{SE(3)}$-aware diffusion model first learns a geometry-only conditional distribution, generating diverse, physically valid grasp candidates from point clouds. A training-free vision–language pipeline (CLIP+SAM) then projects open-vocabulary 2D segmentations into 3D semantic target constraints. Finally, a post-hoc joint ranking strategy combines geometric quality and semantic relevance through a controllable utility function, enabling inference-time adjustment of risk preferences without re-sampling or re-training.

Extensive simulations show that our method achieves a 90.7% success rate on language-conditioned manipulation tasks, outperforms GraspNet in geometric metrics, and surpasses planning-based baselines such as Code as Policies and VoxPoser in semantic alignment.

***Index Terms*—$\mathrm{SE}(3)$ diffusion, open-vocabulary grasping, vision-language models, post-hoc ranking, zero-shot manipulation**

I. INTRODUCTION

Open-world robotic grasping requires reasoning over both geometric feasibility on the $\mathrm{SE}(3)$ manifold and open-vocabulary semantic alignment with natural language instructions. This poses a fundamental challenge: semantic signals are often noisy and weakly grounded in 3D, while grasp synthesis demands precise geometric reasoning.

Recent Vision-Language-Action (VLA) models [1], [2] condition action generation on language end-to-end but often sacrifice geometric precision in cluttered scenarios. Conversely, geometry-centered perception methods [3], [4] achieve strong geometric performance, while semantic grounding remains challenging in open-vocabulary settings [5], [6], limiting inference-time controllability. We argue that **geometry generation and semantic alignment should be decoupled**: learn an unconstrained $\mathrm{SE}(3)$ distribution, then apply semantic filtering post-hoc.

† Corresponding authors.

This work was supported in part by the Natural Science Foundation of China under Grant U25D9004 and the Key Research and Development Program of Hubei Province under Grant 2025BAB023.

We present a geometry-semantics decoupled framework with three components: **(1)** An $\mathrm{SE}(3)$-aware diffusion model generates diverse grasp candidates from point clouds. **(2)** A training-free pipeline projects open-vocabulary 2D segmentations (CLIP [7], SAM [8]) into 3D target regions. **(3)** Joint ranking fuses geometric stability and semantic relevance via a tunable utility function, enabling inference-time risk adjustment without re-sampling.

Contributions. (1) Formulating open-vocabulary grasping as geometry-only generation with post-hoc semantic ranking. **(2)** An $\mathrm{SE}(3)$-aware diffusion model capturing multi-modal grasp posteriors from point clouds. **(3)** Training-free 2D→3D semantic grounding via vision–language models. **(4)** Superior performance over geometry-only, cascade, and planning baselines (90.7% success).

II. RELATED WORK

Language-conditioned grasping. Early works focused on 2D planar grasping [9]. Recently, the focus has shifted to open-vocabulary 3D manipulation. Methods like LanPerAct [10] and AffordGrasp [6] leverage language and vision for task-oriented manipulation [11]. Unlike end-to-end approaches [1], which require massive training data, our method adopts a modular design similar to [10], utilizing VLM-guided ranking to explicitly align semantic intent with geometric candidates.

Open-vocabulary localization. Bridging 2D semantics to 3D space is critical. While some works build explicit 3D scene graphs [12], others project 2D foundation model features (SAM, CLIP) into 3D point clouds. We adopt the latter strategy for its efficiency, avoiding the heavy computation of neural fields [13] while maintaining sufficient precision for grasping.

Generative grasping. Diffusion models have revolutionized grasp generation [4], [14]. To handle the complex topology of rotation, recent works specifically apply diffusion on the $\mathrm{SE}(3)$ manifold [15], [16]. Advanced methods like EquiBot [17] further introduce equivariance to improve data efficiency, while recent works also study robustness and debiasing under imperfect supervision [18]. Closest to our work is UniDiffGrasp [5], which also combines VLMs with diffusion. However, as shown in Table I, we uniquely decouple geometry generation from semantic scoring. Unlike UniDiffGrasp which injects semantics during the expensive denoising process, our post-hoc ranking enables flexible, user-tunable risk preferences at inference time without re-generating candidates.

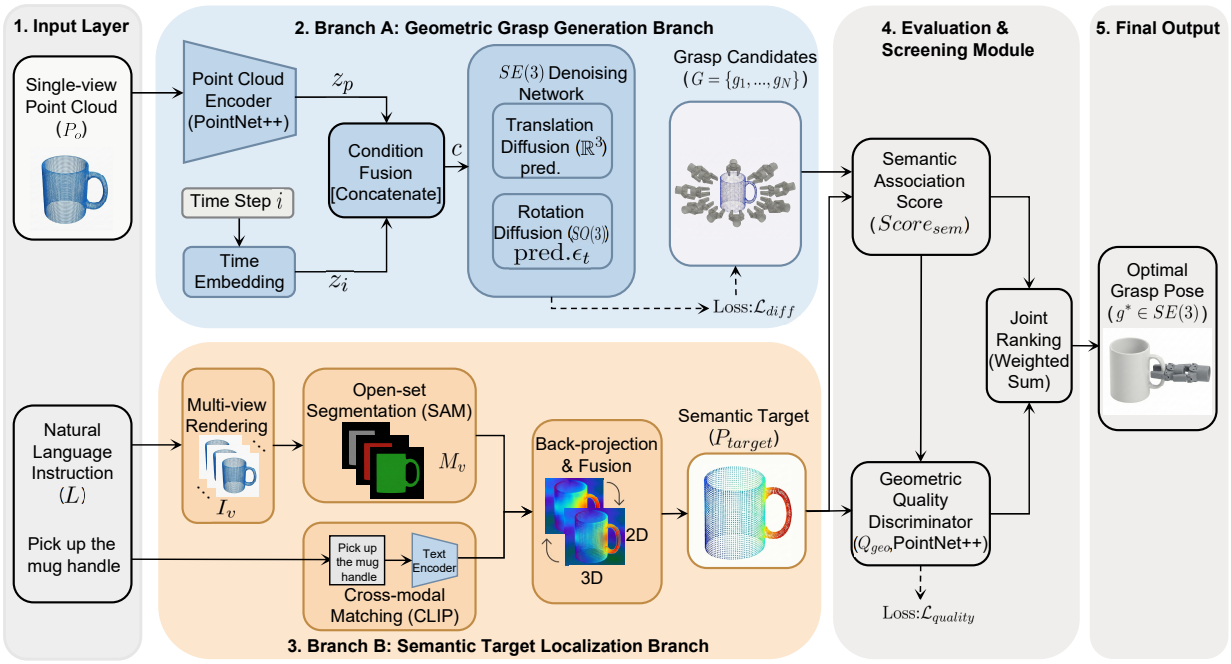

Fig. 1. Overview of the proposed open-vocabulary 6-DoF grasping framework. An SE(3) diffusion model generates grasp candidates (Branch A), while VLM+SAM localizes the target part for joint geometry–semantic ranking (Branch B).

TABLE I
COMPARISON WITH SOTA VLM-GRASPING METHODS

Method	Semantic Injection	Training	Inference
UniDiffGrasp [5]	Conditional Generation	VLM-Dependent	Fixed
AffordGrasp [6]	End-to-End	Heavy Supervision	Fixed
Ours	**Ranking Stage**	**Training-Free**	**Tunable**

III. METHOD

As shown in Fig. 1, our method decouples grasp generation into two parallel branches: **Geometric grasp candidate generation** adopts a diffusion model conditioned on the object point cloud to directly generate high-quality 6-DoF candidates in SE(3). **Semantic target region parsing** uses an LLM-based parser and VLM+SAM to localize the target region. At inference, we rank candidates via a joint scoring mechanism (Branch B).

A. SE(3) *Diffusion-Based Grasp Generation*

To address the challenge of generating high-quality and diverse grasp poses from a single-view point cloud, our geometric grasp module extends the core ideas of manifold diffusion. Instead of the traditional "sample-and-evaluate" discriminative paradigm, we adopt diffusion probabilistic models (DPMs) and formulate grasp generation as a conditional denoising process from noise to valid grasp poses. This enables learning complex multimodal grasp distributions and generating high-success 6-DoF grasp poses $\boldsymbol{G} = (\boldsymbol{t}, \boldsymbol{R})$ conditioned on local object geometry $\boldsymbol{P}_o$.

1) Manifold Diffusion Intuition: We represent a grasp pose as $\boldsymbol{G} = (\boldsymbol{t}, \boldsymbol{R})$ with $\boldsymbol{t} \in \mathbb{R}^3$ and $\boldsymbol{R} \in \mathrm{SO}(3)$, and model grasp generation directly on SE(3). To respect the manifold structure, we decouple translation and rotation and apply Euclidean diffusion for $\boldsymbol{t}$ and an SO(3) diffusion process for $\boldsymbol{R}$ [19]. This manifold-aware design is crucial for physical validity, as also highlighted in recent equivariant policy learning research [17].

2) SO(3) Rotation Diffusion: For rotation $\boldsymbol{R} \in \mathrm{SO}(3)$, to preserve manifold constraints, we follow [15] and use a diffusion framework based on isotropic Gaussian distributions on SO(3) (IGSO(3)). In the forward process, we progressively add noise on the rotation manifold via logarithmic mapping, ensuring $\boldsymbol{R}_i$ remains on the SO(3) manifold. Denoising is performed in the Lie algebra $\mathfrak{so}(3) \cong \mathbb{R}^3$.

3) Conditional SE(3) *Diffusion:* Combining the above translation and rotation diffusion models and introducing local object point cloud $\boldsymbol{P}_o$ as a condition, we obtain a complete conditional diffusion model operating on SE(3). A shared denoising network $\boldsymbol{\epsilon}_\theta$ jointly predicts translation and rotation noise. The network $\boldsymbol{\epsilon}_\theta(\boldsymbol{G}_i, i, \boldsymbol{c})$ takes the noisy grasp pose $\boldsymbol{G}_i = (\boldsymbol{t}_i, \boldsymbol{R}_i)$, the time step i, and the conditional context $\boldsymbol{c}$ encoded from point cloud $\boldsymbol{P}_o$ (using PointNet++ [20]), and outputs a 6D noise vector $[\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_R]$ where $\boldsymbol{\epsilon}_t \in \mathbb{R}^3$ and $\boldsymbol{\epsilon}_R \in \mathfrak{so}(3) \cong \mathbb{R}^3$. We train the model with a standard diffusion noise-prediction objective:

$$L(\theta) = L_t(\theta) + L_R(\theta) = \mathbb{E}_{i, \boldsymbol{G}_0, \boldsymbol{c}, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{G}_i, i, \boldsymbol{c}) \right\|^2 \right] \quad (1)$$

where $\boldsymbol{G}_0 \sim q(\boldsymbol{G}_0 | \boldsymbol{c})$ and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_R]$. At inference, we sample a small set of candidates from the learned conditional prior; in practice, we optionally apply lightweight collision-aware filtering/guidance to reduce obvious penetrations.

B. Open-Vocabulary Semantic Grounding

After an initial geometric understanding of the scene, the core challenge becomes how to accurately "understand" and "localize" the interaction target specified by a natural-language instruction L. While end-to-end Vision-Language-Action (VLA) models exist, our modular design leveraging CLIP and SAM avoids the need for massive grasping-specific finetuning and allows for explicit, interpretable geometric constraints. Specifically, we adopt a "3D→2D segmentation and matching, then back to 3D" route (Branch B in Fig. 1).

1) Multi-view Segmentation: We render multi-view 2D observations from the 3D point cloud and apply SAM [8] in a class-agnostic "segment everything" mode to obtain candidate masks per view. This yields a rich pool of regions (objects/parts) for subsequent open-vocabulary matching.

2) CLIP Semantic Matching: We introduce CLIP [7] to select the region most semantically matched to the user instruction L. We encode the target phrase and each masked region with CLIP, and calculate the cosine similarity:

$$s_{v,k} = f_{\text{text}} \cdot f_{\text{image},v,k}^{T} \tag{2}$$

To improve robustness, we select the best mask $m_v^* = \text{argmax}_{m_{v,k} \in M_v}(s_{v,k})$ within each view and record its score $s_v^* = \max_k(s_{v,k})$.

3) 3D Target Point Fusion: We lift each view's optimal mask m_v^* back to 3D: we traverse the original point cloud P and project points onto view v's image plane; points falling inside the mask form the view-specific target subset $P_{\text{target},v}$:

$$P_{\text{target},v} = \{p_j \in P \mid \Pi_{\text{2D}}(p_j, c_v) \in m_v^*\} \tag{3}$$

During fusion, we aggregate the matching confidence of each point p across visible views by taking the maximum score $S(p) = \max_{\{v|p \in P_{\text{target},v}\}}(s_v^*)$. Finally, we threshold $S(p)$ with τ_s to obtain the final target point set:

$$P_{\text{target}} = \{p \in P_{\text{union}} \mid S(p) > \tau_s\} \tag{4}$$

This subset $P_{\text{target}} \subset P$ serves as the semantic constraint input for subsequent grasp ranking.

C. Geometry–Semantics Joint Ranking

Given candidate grasps G from the diffusion prior and the semantic target points P_{target}, we rank grasps by jointly evaluating semantic alignment and geometric stability. **Unlike hard filtering strategies (e.g., GraspNet+CLIP) that impose binary decision boundaries, our soft semantic scoring preserves feasible candidates.**

1) Semantic Relevance: To quantify how well each candidate grasp g_i aligns with the user-specified target part, we design a positive semantic relevance scoring function $\text{Score}_{\text{sem}}(g_i)$ based on geometric overlap. **We adopt a contact-based formulation rather than centroid distance, as semantic grasp correctness depends on physical interaction regions.**

Concretely, for each candidate grasp g_i, we define the contact set $C(g_i)$ as scene points within distance $r_{\text{grip}} = 5\,\text{cm}$ of the gripper fingers after pose transformation:

$$C(g_i) = \{p \in P \mid \min_{q \in \mathcal{F}(g_i)} \|p - q\| < r_{\text{grip}}\} \tag{5}$$

where $\mathcal{F}(g_i)$ denotes the gripper finger surface points in pose g_i. We then compute semantic relevance as the proportion of contact points falling within distance $\delta = 2\,\text{cm}$ of the target region P_{target}:

$$\text{Score}_{\text{sem}}(g_i) = \frac{1}{|C(g_i)|} \sum_{p \in C(g_i)} I(\min_{p_t \in P_{\text{target}}} \|p - p_t\| < \delta) \tag{6}$$

where $I(\cdot)$ is an indicator function and $\text{Score}_{\text{sem}} \in [0, 1]$.

2) Geometric Quality: Although stage one diffusion generation already biases toward geometrically valid grasps, an independent geometric quality evaluator is still necessary to further ensure physical stability and real-world executability. We additionally train a lightweight grasp quality discriminator Q_{geo}. This discriminator is a neural network that takes a candidate grasp pose g_i and the surrounding local scene point cloud as input and outputs a scalar in $[0, 1]$, representing the geometric stability score:

$$\text{Score}_{\text{geo}}(g_i) = Q_{\text{geo}}(g_i, P) \tag{7}$$

Higher $\text{Score}_{\text{geo}}(g_i)$ indicates a more stable grasp.

3) Ranking and Selection: We compute a weighted fusion score:

$$\text{Score}_{\text{final}}(g_i) = w_{\text{sem}} \cdot \text{Score}_{\text{sem}}(g_i) + w_{\text{geo}} \cdot \text{Score}_{\text{geo}}(g_i). \tag{8}$$

where $w_{\text{sem}} + w_{\text{geo}} = 1$. This linear fusion strategy treats physical stability and semantic intent as orthogonal objectives. Unlike classifier-guided diffusion methods (e.g., UniDiffGrasp [5]) that require computationally expensive gradient calculations during generation, our **post-hoc ranking** approach is significantly more efficient. Furthermore, it decouples the risk preference from the generative model, allowing users to tune w_{sem} and w_{geo} at inference time to adapt to different safety requirements without re-training. We discard candidates with $\text{Score}_{\text{sem}} = 0$ and select the best remaining grasp:

$$g^* = \underset{g_i \in G | \text{Score}_{\text{sem}}(g_i) > 0}{\text{argmax}} (\text{Score}_{\text{final}}(g_i)) \tag{9}$$

D. Implementation Details

Data Construction. Following the standard evaluation protocol, we construct a simulation-based grasp dataset from 3D CAD models. We source object meshes from ShapeNet and generate grasp pose ground truth using GraspIt! with force-closure-based stability checks. We focus on a 10-category subset (CAT10) of common household/industrial objects. Each training sample consists of a partial single-view point cloud and a successful 6-DoF grasp pose. To reduce the gap between complete meshes and partial observations, we apply online rendering-based augmentations (random camera views, rotations, point dropout/cropping, and noise) to produce partial point clouds.

Training Setup. Our framework contains two trainable modules: an $\text{SE}(3)$-conditional diffusion grasp generator and the geometric quality discriminator Q_{geo}. The VLM and SAM used in stage two directly use pretrained weights and remain frozen. We downsample P to 4096 points and train for 150

epochs with `AdamW` (lr = 1×10^{-4}, cosine schedule) and standard augmentations. We train the diffusion generator with Eq. (1) and train the lightweight discriminator Q_{geo} with standard Binary Cross Entropy (BCE) loss.

IV. EXPERIMENTS

A. Experimental Setup

We follow the experimental setup established in standard benchmarks. Object meshes are sourced from ShapeNet; successful grasp poses are generated by GraspIt! and verified by force-closure/stability checks. Negative samples for the discriminator Q_{geo} are obtained by perturbing successful poses and verifying failures in MuJoCo. During evaluation, part-level ground truth is only used for offline statistics of semantic metrics such as SAR and is not used in inference.

1) Simulation Platform: We build a 6-DoF grasp simulation platform based on the MuJoCo physics engine, using a Franka Panda gripper. Evaluation procedure: import the object model, initialize the gripper pose according to the tested grasp, close the gripper and apply gravity; if the object is stably held beyond a threshold and does not slip, the grasp is considered successful. This subsection describes grasp-level evaluation (end-effector only); system-level end-to-end evaluation is described in Sec. IV-I.

B. Evaluation Metrics

We report grasp success rate (GSR), semantic accuracy rate (SAR) on physically successful grasps, and their harmonic mean (H-Mean). SAR measures the percentage of physically successful grasps whose contact points fall within the ground-truth semantic mask or within a distance threshold ($\delta = 2\,\text{cm}$) of the target part surface. We also report the classification accuracy of the geometric evaluator network.

C. Quantitative Comparison

We compare against two representative baselines:

- **6-DoF GraspNet** [21]: geometry-only 6-DoF grasp generation.
- **GraspNet + CLIP**: cascade baseline (generate first, then filter with CLIP).

We conduct large-scale tests on 10 object categories in the test set, with 100 independent grasp trials per category. Table II reports the quantitative results.

TABLE II
GRASP PERFORMANCE COMPARISON ON THE TEST SET

Method	GSR (%)	SAR (%)	H-Mean (%)
6-DoF GraspNet [21]	82.4	35.6	49.7
GraspNet + CLIP	76.8	84.2	80.3
Ours	**86.5**	**90.3**	**88.4**

All improvements significant at $p < 0.01$ (paired t-test, $n = 1000$).

Note: H-Mean is sensitive to low values and thus penalizes methods that prioritize only geometry or only semantics, more objectively reflecting overall multi-objective performance.

a) Result analysis: Table II shows that GraspNet has reasonable GSR but low SAR without semantic grounding. The cascade baseline improves SAR but reduces GSR due to hard filtering. Our method achieves the best overall trade-off (H-Mean 88.4%) by jointly optimizing geometry and semantics.

Note on UniDiffGrasp: We prioritize comparing our method against standard geometric baselines (GraspNet) and cascade architectures (GraspNet+CLIP) to isolate the efficacy of our proposed **decoupled ranking mechanism**. While recent works like UniDiffGrasp [5] also explore VLM-guided diffusion, they employ a classifier-guided paradigm that injects semantics during the denoising process. Due to significant differences in experimental protocols (e.g., custom datasets vs. standard ShapeNet subsets) and the lack of unified evaluation benchmarks for guided-diffusion grasping, a direct quantitative comparison is currently infeasible. However, conceptually, our **post-hoc ranking** offers superior flexibility in inference-time risk adjustment compared to their fixed conditional generation, as detailed in Table I.

D. Per-Category Results

We further break down performance across different geometric/semantic characteristics. Following our experimental protocol, we run 200 trials per category and report both the number of successful/semantic-correct grasps and the corresponding percentages. **The detailed statistics are reported in Table III.**

a) Analysis on Multi-modality: Notably, our diffusion-based method excels on asymmetric objects (e.g., Mugs, Hammers). Unlike unimodal regression or VAE baselines that often collapse to the mean pose (resulting in invalid grasps for asymmetric objects), the $\text{SE}(3)$ diffusion process effectively models the **complex multi-modal posterior distribution**. This allows the sampler to traverse the probabilistic manifold and locate valid modes within the specific semantic region (e.g., a mug handle) defined by the text.

E. Robustness Under Perception Disturbances

Real-world settings often involve occlusion and sparse point clouds. We stress-test robustness along two axes: (i) visible surface ratio under occlusion and (ii) point cloud density (downsampled from 1024 to 128 points). **These disturbances serve as a proxy for real-world sensor noise and occlusions.** As shown in Fig. 2, our method maintains high success rates even under severe downsampling, demonstrating strong potential for **Sim-to-Real transfer** where perfect perception is rarely available.

F. Candidate Distribution Visualization

To better explain why our method improves semantic accuracy, we visualize the spatial distribution of generated candidates on a representative part-centric object (mug). Compared with discriminative sampling, diffusion-guided candidates concentrate on the task-relevant functional part while maintaining local diversity.

TABLE III
PER-CATEGORY GRASP PERFORMANCE (200 TRIALS PER CATEGORY)

Category	6-DoF GraspNet		GraspNet+CLIP		Ours	
	GSR	SAR	GSR	SAR	**GSR**	**SAR**
Bottle	177 (88.4)	65 (32.5)	162 (81.2)	171 (85.4)	**181 (90.5)**	**182 (91.2)**
Bowl	173 (86.5)	77 (38.6)	157 (78.4)	176 (88.2)	**178 (89.2)**	**185 (92.5)**
Box	182 (91.2)	57 (28.4)	170 (85.0)	165 (82.5)	**185 (92.6)**	**177 (88.6)**
Mug	159 (79.5)	70 (35.2)	145 (72.6)	168 (84.1)	**171 (85.4)**	**187 (93.4)**
Hammer	154 (76.8)	62 (31.0)	139 (69.5)	161 (80.6)	**168 (84.2)**	**180 (89.8)**
Scissors	137 (68.4)	49 (24.5)	122 (61.2)	156 (78.2)	**157 (78.6)**	**173 (86.5)**
T-shaped	150 (75.2)	74 (36.8)	131 (65.4)	159 (79.5)	**167 (83.5)**	**180 (90.2)**
Average	165 (82.4)	71 (35.6)	154 (76.8)	168 (84.2)	**173 (86.5)**	**181 (90.3)**

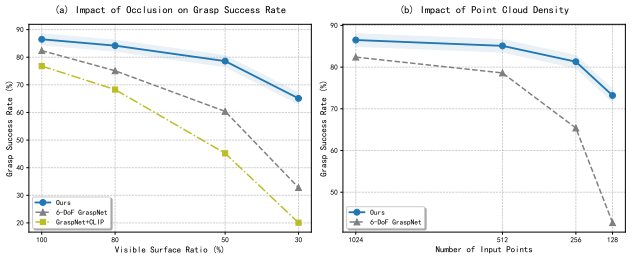

Fig. 2. Robustness under perception disturbances: (a) visible surface ratio vs grasp success; (b) point-cloud sparsity vs performance.

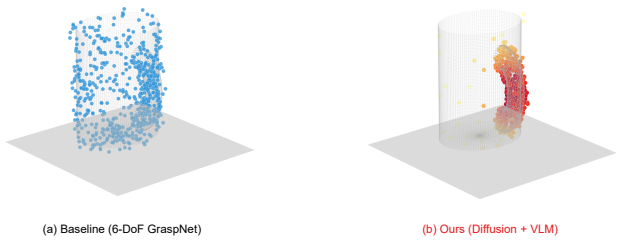

Fig. 3. Candidate grasp distribution visualization on a mug point cloud.

G. Inference Efficiency

We measure end-to-end latency on an Intel i9-12900K CPU and a single NVIDIA RTX 3090 GPU ($T = 25$, $K = 100$). The total inference time is **1204 ms**. The breakdown is: Instruction parsing (125.6 ms, 10.4%), Semantic grounding (315.4 ms, 26.2%), Pose generation (678.5 ms, 56.4%), and Ranking (84.5 ms, 7.0%). The generation stage dominates due to the iterative denoising process. Although 1.2s is slower than discriminative methods, it is well within the operational tolerance of service robots in **pick-and-place** scenarios. Typically, robot arm motion planning and execution span several seconds, during which the grasp generation can be performed in parallel or pre-computed during the approach phase. Furthermore, future work will investigate consistency distillation to compress the sampling steps (e.g., to < 100 ms), mitigating this latency.

H. Sampling Steps Ablation

Diffusion sampling steps T trade off accuracy and efficiency. Table IV evaluates different T values and reports GSR/SAR and denoising time.

I. End-to-End Language-Driven Manipulation Evaluation

We also evaluate an end-to-end language-driven pipeline where the instruction is parsed, the target is localized, and our grasp module provides the execution grasp. **Crucially, our method is used solely to provide the target execution grasp pose, while motion planning is handled by a standard planner (e.g., RRT-Connect) shared across all methods to ensure fair comparison.** We compare this end-to-end pipeline with two representative baselines: Code as Policies (CaP) [22], which directly generates code to call low-level action primitives; and VoxPoser [23], which encodes language constraints into a 3D value/cost field and plans actions via optimization. We test three task types: basic grasping, spatial reasoning, and semantically constrained grasping. Results are shown in Table V.

TABLE IV
IMPACT OF DIFFUSION STEPS T

T	GSR	SAR	Time
5	45.8	62.4	132.5
10	68.2	75.6	268.4
25	**85.4**	**90.3**	**685.2**
50	86.8	91.5	1342.6
100	87.2	92.0	2715.3

TABLE V
END-TO-END TASK SUCCESS (%)

Method	**Avg.**
CaP [22]	65.2
VoxPoser [23]	81.8
Ours	**90.7**

*Full breakdown in text.

Table V shows consistent gains across tasks, especially for semantic-constrained grasping. Results show that insufficient steps (e.g., $T < 10$) lead to performance drops. This is likely because the denoising trajectory on the non-linear $\mathrm{SE}(3)$ manifold requires sufficient iterations to converge from isotropic noise to a valid pose distribution; early truncation results in distorted rotation matrices and physical penetrations.

J. Qualitative Analysis

Fig. 4 demonstrates the adaptability of our method across diverse geometries. Whether for spherical, curved, or irregular industrial parts, our approach generates stable grasps that align with the object's topology.

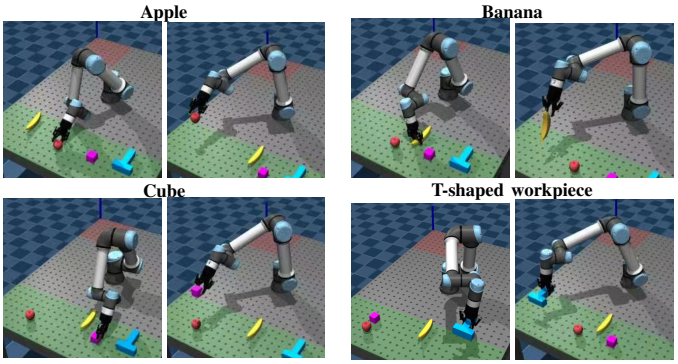

Fig. 4. Qualitative results. Left: Initial state; Right: Stable grasp.

K. Ablation Study

We ablate the optimization strategy and the evaluator-based screening.

1) Optimization Strategy: Table VI compares geometry-only, semantics-only, and joint ranking.

TABLE VI
ABLATION RESULTS OF GRASP OPTIMIZATION STRATEGIES

Strategy	**GSR (%)**	**SAR (%)**	**H-Mean (%)**
Baseline (no opt.)	65.2	42.5	51.5
Geometry-only	**88.4**	41.8	56.7
Semantics-only	58.6	**92.1**	71.6
Ours (Joint)	86.5	90.3	**88.4**

Joint ranking achieves the best balance between GSR and SAR.

2) Evaluator and Screening: In addition, for the proposed geometric evaluator network Q_{geo}, we test classification accuracy. The geometric evaluator achieves 83.56% classification accuracy; evaluator-based screening improves GSR from 65.2% to 86.5%.

V. CONCLUSION

We propose an open-vocabulary 6-DoF grasping framework based on an $\mathrm{SE}(3)$ diffusion prior and VLM-guided ranking. The framework generates a compact set of high-quality candidates via diffusion, grounds language targets using VLM+SAM to obtain 3D target points, and selects the final grasp via joint geometry–semantics scoring. Experiments demonstrate improved trade-offs over geometry-only and cascade baselines.

Limitations include the inherent latency of the iterative diffusion sampling process, which may impact high-frequency real-time control. **While our method decouples geometry and semantics effectively, the inference speed is currently bottlenecked by the iterative diffusion steps.** Future work will investigate consistency distillation [24] to accelerate inference, and explore replacing point clouds with 3D Gaussian Splatting [13], [25] for more robust open-vocabulary semantic representation.

REFERENCES

[1] M. J. Kim, K. Pertsch, S. Karamcheti *et al.*, "Openvla: An open-source vision-language-action model," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[2] H. Zhen *et al.*, "3d-vla: A 3d vision-language-action generative world model," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[3] H. Wu, K. Zhao, S. Wu, X. Wu, and X. Niu, "Aarr-net: An attention assistance feature fusion and model recursive recovery network for category-level 6d object pose estimation," in *Neural Information Processing*, ser. Lecture Notes in Computer Science. Springer, 2025, vol. 15292, pp. 402–416.

[4] B. Zuo, Z. Zhao, W. Sun, X. Yuan, Z. Yu, and Y. Wang, "GraspDiff: Grasping generation for hand-object interaction with multimodal guided diffusion," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 31, 2025.

[5] X. Guo, H. Hu, C. Song *et al.*, "Unidiffgrasp: A unified framework integrating vlm reasoning and vlm-guided part diffusion," in *arXiv preprint arXiv:2505.06832*, 2025.

[6] Y. Tang, S. Zhang, X. Hao *et al.*, "Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025.

[7] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[8] A. Kirillov, E. Mintun, N. Ravi *et al.*, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.

[9] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learn. (CoRL)*, 2022, pp. 894–906.

[10] H. Wu, S. Wu, Y. Tu, H. Zhou, S. Drew, and X. Niu, "Lanperact: A framework for language-driven perception and robotic manipulation," in *2025 IEEE Smart World Congress (SWC)*. IEEE, 2025, pp. 484–490.

[11] X. Li *et al.*, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[12] Q. Gu, A. Kuwajerwala, S. Morin *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.

[13] M. Ji, R.-Z. Qiu, X. Zou, and X. Wang, "Graspsplats: Efficient manipulation with 3d feature splatting," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[14] A. Murali, B. Sundaralingam *et al.*, "Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training," *arXiv preprint arXiv:2507.13097*, 2025.

[15] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 5923–5930.

[16] X. Zhu *et al.*, "Se(3)-equivariant diffusion policy in spherical fourier space," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2025.

[17] J. Yang, Z.-a. Cao, C. Deng *et al.*, "Equibot: Sim(3)-equivariant diffusion policy for generalizable and data efficient learning," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[18] J. Yang, L. Jing, Y. Xu, S. Wu, S. Drew, and X. Niu, "Label-expanded feature debiasing for single domain generalization," in *Pattern Recognition. ICPR International Workshops and Challenges*. Springer, 2024.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017.

[21] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2901–2910.

[22] J. Liang, W. Huang, F. Xia *et al.*, "Code as policies: Language model programs for embodied control," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 9493–9500.

[23] W. Huang, C. Wang, R. Zhang *et al.*, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *Proc. Conf. Robot Learn. (CoRL)*, 2023.

[24] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *International Conference on Machine Learning (ICML)*, 2023.

[25] O. Shorinwa *et al.*, "Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.