# SAGE-GRASP: AN END-TO-END TARGET-DRIVEN 6-DOF GRASPING FRAMEWORK VIA CROSS-MODAL SEMANTIC ADDRESSING AND GUMBEL-ROUTED EXPERTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Language-conditioned 6-DoF grasping in clutter requires identifying the referred object while generating physically feasible grasps. Existing task-agnostic detectors disregard semantic intent, whereas segment-then-grasp pipelines are susceptible to upstream segmentation errors and often discard useful contextual information. Thus, we propose SAGE-Grasp, which establishes a unified semantic-geometric representation that couples target understanding with full-scene physical reasoning. Cross-Modal Semantic Addressing (CMSA) and a Gumbel-routed mixture-of-experts architecture provide complementary mechanisms for semantic grounding and expert-routed grasp prediction, enabling the model to focus on the instructed target while retaining the context required for feasible grasping. SAGE-Grasp is jointly trained with text, visual, and zero queries, enabling target-aware semantic grounding while preserving task-agnostic geometric affordances. Experiments on a target-driven extension of GraspNet-1Billion demonstrate stronger target adherence than cascaded baselines and competitive scene-level grasping performance. These results indicate that target-conditioned grasping can be integrated into full-scene prediction while retaining task-agnostic scene-level capability.

## 1 INTRODUCTION

Robotic grasping in cluttered environments is a fundamental capability for robotic manipulation, enabling broad applications in real-world scenarios such as industrial assembly, order picking, and domestic service robotics (Hoang et al., 2022; Xu et al., 2023; Liu et al., 2024; Chisari et al., 2024). Recently, with the advent of large-scale real-world datasets such as GraspNet-1Billion (Fang et al., 2020), 6-DoF grasp detection has improved substantially. However, most state-of-the-art 6-DoF grasping models operate in a task-agnostic manner (Lu et al., 2022; Li & Cappelleri, 2024; Ma et al., 2024; Wang et al., 2025; Cheng et al., 2026). They are designed to perform indiscriminate bin-picking across the entire scene, aiming to grasp any available object. Instruction-guided manipulation requires a robot to identify and grasp a user-specified object in dense clutter. This target-driven setting differs from conventional scene-level grasp detection, which treats all graspable objects as valid targets.

Accurate target-driven grasping directly from raw, real-world sensory data is challenging due to severe object occlusions, partial geometries, and inherent sensor noise. Balancing precise semantic identification with robust geometric reasoning under these conditions remains a critical bottleneck. A commonly adopted solution is a cascade pipeline, often termed segment-then-grasp (Murali et al., 2020; Tziafas & Kasaei, 2024; Xie et al., 2025). This cascaded approach typically employs a 2D vision-language model to isolate the target point cloud, followed by a traditional 6-DoF network for grasp prediction. However, this disjointed pipeline not only makes the final prediction sensitive to upstream segmentation errors, but also discards surrounding scene geometry, thereby increasing the risk of collisions with surrounding objects during execution.

Fig. 1 contrasts the isolate-then-predict paradigm with SAGE-Grasp. We propose a unified end-to-end target-conditioned framework that predicts grasps directly from the full scene, where *end-*



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

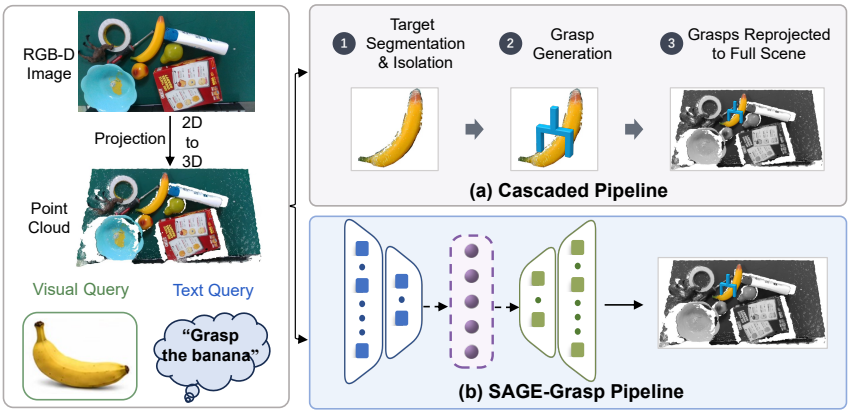



Figure 1: **Comparison of cascaded target-driven grasping and SAGE-Grasp.** Cascaded methods isolate the queried target before grasp generation, whereas SAGE-Grasp predicts target-conditioned grasps directly from the full scene using text or visual queries.

*to-end* denotes the unified grasp-prediction path after frozen SAM-CLIP semantic-prior extraction. Building on attention-based query–feature interaction (Vaswani et al., 2017; Tan & Bansal, 2019; Carion et al., 2020; Misra et al., 2021), CMSA performs query-conditioned point-wise semantic addressing in 3D space while preserving the surrounding geometry required for collision-aware grasp prediction. A Gumbel-routed MoE head (Jacobs et al., 1991; Shazeer et al., 2017) further adapts prediction to heterogeneous local geometry through expert routing. Joint training with text, visual, and zero queries supports flexible target conditioning while preserving task-agnostic grasping affordances, enabling both target-conditioned and scene-level inference.

The main contributions are as follows:

1. We propose SAGE-Grasp, a unified framework for target-conditioned 6-DoF grasping. By integrating semantic addressing into full-scene grasp prediction, it avoids hard target isolation associated with cascaded segment-then-grasp systems.
2. We introduce two complementary architectural components: CMSA performs fine-grained point-wise semantic addressing over sampled full-scene features, while a Gumbel-routed mixture-of-experts head adapts grasp prediction to diverse local geometries.
3. We establish a joint text, visual, and zero-query training paradigm. This approach enables multimodal intent grounding while retaining task-agnostic grasping affordances. Experiments on a target-driven extension of GraspNet-1Billion show improved target adherence and higher collision-free proposal rates under full-scene evaluation, while retaining competitive task-agnostic grasping performance.

## 2 Related Work

### 2.1 6-DoF Robotic Grasping in Clutter

Early robotic grasping methods largely relied on analytical formulations that require accurate 3D object models and explicit physical contact analysis (Ruan et al., 2020; Wang et al., 2024). With the emergence of large-scale grasping datasets such as GraspNet-1Billion (Fang et al., 2020), data-driven approaches have become dominant. PointNetGPD (Liang et al., 2019) and Contact-GraspNet (Sundermeyer et al., 2021) established effective point-cloud representations for grasp evaluation, while more recent models, including GSNet (Wang et al., 2021) and EconomicGrasp (Wu et al., 2024), improve dense grasp generation, efficiency, and geometric feature extraction in clutter. These methods nevertheless remain task-agnostic: they seek executable grasps throughout the scene without explicitly distinguishing the object requested by a human instruction, and therefore do not directly address target-conditioned manipulation.

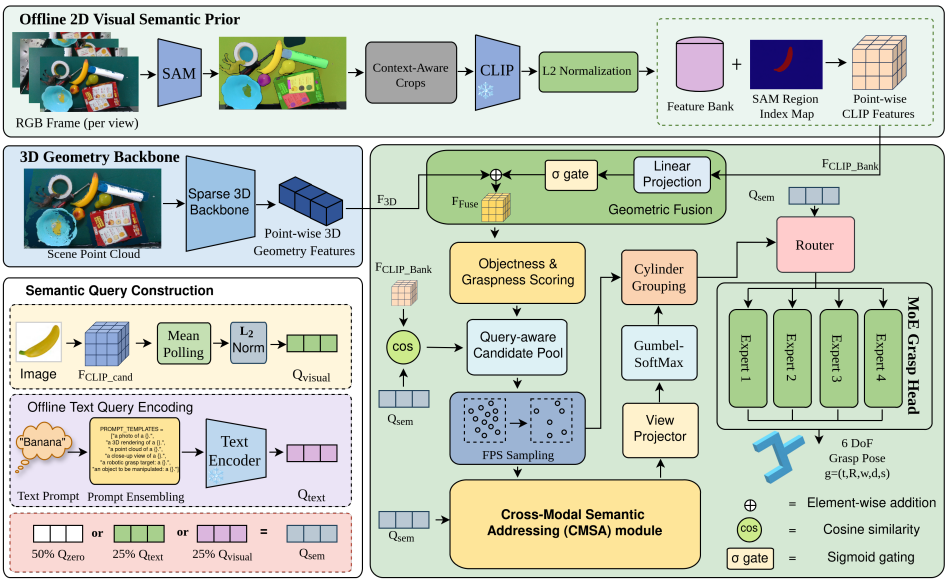



Figure 2: **Overall architecture of SAGE-Grasp.** Offline SAM-CLIP semantic priors are integrated with sparse 3D features to construct a query-aware candidate pool. CMSA refines target-relevant features, followed by Gumbel-based view selection and expert-routed 6-DoF grasp prediction.

### 2.2 Target-Driven and Language-Guided Grasping

While vision-language models have advanced 4-DoF manipulation (Li et al., 2024; Vuong et al., 2024; Qi et al., 2024) and simulated 6-DoF grasping (Nguyen et al., 2024), unified target-driven 6-DoF grasping on real-world point clouds of cluttered scenes remains less explored due to severe sensor noise and occlusion. To tackle real-world scenarios, cascaded segment-then-grasp pipelines (Lu et al., 2023; Jin et al., 2024; Song et al., 2025) commonly use foundation models (Radford et al., 2021; Kirillov et al., 2023) to isolate target point clouds before 6-DoF grasp synthesis. However, this cascaded design propagates upstream segmentation errors and discards global scene context. Instead, SAGE-Grasp performs target addressing within a learned full-scene predictor, preserving the spatial information used for collision-aware proposal generation.

### 2.3 Mixture of Experts in Vision

Mixture-of-Experts architectures (Jacobs et al., 1991; Shazeer et al., 2017) scale neural networks by activating input-conditioned subnetworks. They have achieved substantial success in natural language processing and 2D vision (Riquelme et al., 2021; Fedus et al., 2022), but remain less studied for 3D physical interaction. Target-driven grasping must accommodate large geometric variation across semantic categories, object poses, and local contact regions. In contrast to a monolithic MLP head, SAGE-Grasp uses Gumbel-based routing to enable input-dependent expert selection over local geometric features while maintaining differentiable training.

## 3 SAGE-Grasp

### 3.1 Overview

Given a point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ and semantic query $Q_{\text{sem}}$, the model predicts a sparse set of grasps

$$g = (t, R, w, d, s) \tag{1}$$

where $t \in \mathbb{R}^3$ is the gripper center, $R \in \mathrm{SO}(3)$ is its orientation, $w, d \in \mathbb{R}^+$ are the gripper width and depth, and $s \in [0, 1]$ is the confidence score.

As illustrated in Fig. 2, SAGE-Grasp combines cached SAM-CLIP semantic priors with sparse 3D geometry. Candidate sampling first retains high-graspness foreground points. For semantically valid

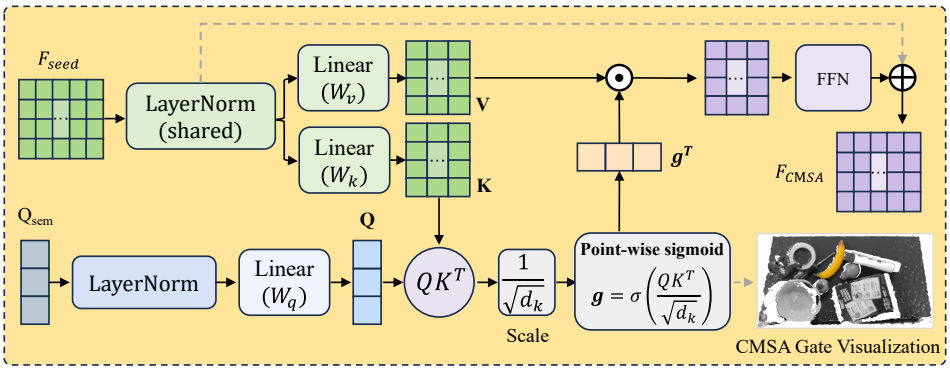



Figure 3: **CMSA architecture.** The semantic query and seed features are projected into query, key, and value representations. Point-wise sigmoid gates then modulate the value features, followed by FFN refinement and residual fusion.

queries, the candidate pool is additionally augmented with points having high cosine similarity between their cached CLIP features and $Q_{\text{sem}}$, followed by farthest-point sampling (FPS) to obtain $N_s$ seed points. CMSA then performs query-conditioned refinement of the sampled seed features. Gumbel-based view selection and local grouping further aggregate the surrounding full-scene geometry, and a semantic router dispatches the resulting local representations to the MoE head.

For zero-query samples, semantic candidate augmentation is disabled and candidate sampling relies on the task-agnostic objectness and graspness scores. CMSA reduces to the identity path, while the router receives a zero semantic vector such that expert routing is driven by the local geometric representation. The unified framework is jointly trained with text, visual, and zero queries.

### 3.2 Cross-Modal Semantic Addressing

After FPS, CMSA performs query-conditioned refinement of the sampled features $F_{\text{seed}}$ using the multimodal query $Q_{\text{sem}}$ (Fig. 3). We use pre-normalization for stable optimization, applying layer normalization (LN) to both inputs before projecting them into query ($Q$), key ($K$), and value ($V$) representations:

$$Q = \text{LN}_q(Q_{\text{sem}})W_Q \in \mathbb{R}^{1\times d_k} \tag{2}$$

$$\bar{F}_{\text{seed}} = \text{LN}_s(F_{\text{seed}}) \tag{3}$$

$$K = \bar{F}_{\text{seed}}W_K \in \mathbb{R}^{N_s\times d_k} \tag{4}$$

$$V = \bar{F}_{\text{seed}}W_V \in \mathbb{R}^{N_s\times D} \tag{5}$$

where $d_k$ denotes the query/key feature dimension. To estimate point-wise target relevance from background clutter, we compute an independent point-wise gate for each seed feature by evaluating the scaled query-key similarity:

$$\mathbf{g} = \sigma\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right) \in [0,1]^{1\times N_s} \tag{6}$$

The gate is broadcast over the feature dimension to modulate $V$, followed by FFN refinement and residual fusion:

$$Z = \bar{\mathbf{g}}^{\top} \odot V \in \mathbb{R}^{N_s\times D} \tag{7}$$

$$F_{\text{sem}} = Z + \text{FFN}(Z) \tag{8}$$

$$F_{\text{CMSA}} = mF_{\text{sem}} + (1-m)F_{\text{seed}} \tag{9}$$

where $m \in \{0,1\}$ is the semantic-valid indicator: $m = 1$ for text or visual queries and $m = 0$ for a zero query, and $\bar{\mathbf{g}} = m\mathbf{g} + (1-m)\mathbf{1}$. This formulation downweights seeds with low query similarity while retaining their coordinates and the neighboring geometry used by subsequent grouping operations.

### 3.3 GUMBEL-ROUTED MIXTURE OF EXPERTS GRASPING HEAD

Given the target-addressed features $F_{\mathrm{CMSA}}$, a view projector predicts approach-direction logits and selects one view per seed using Gumbel-Softmax:

$$R_v = \mathrm{ReLU}(\phi_{v1}(F_{\mathrm{CMSA}})), L_v = \phi_{v2}(R_v) \tag{10}$$

$$Y_{v,n} = \mathrm{GumbelSoftmax}(L_{v,n}; \tau, \text{ hard} = 1) \tag{11}$$

The residual is fused as $F_{\mathrm{view}} = F_{\mathrm{CMSA}} + R_v$, after which cylinder grouping along the selected view produces the local geometric representation $H_{\mathrm{local}}$.

The MoE router jointly conditions on local geometry and the normalized semantic query $Q_{\mathrm{sem}}$, which is set to zero for zero-query samples:

$$\mathbf{h}_n = r([H_{\mathrm{local},n}; Q_{\mathrm{sem}}]) \in \mathbb{R}^{N_e}, \tag{12}$$

where $h_{i,n}$ denotes the routing logit for expert $i$. The routing weights are obtained by

$$\pi_{i,n} = \frac{\exp((h_{i,n} + \epsilon_{i,n})/\tau)}{\sum_{j=1}^{N_e} \exp((h_{j,n} + \epsilon_{j,n})/\tau)}, \tag{13}$$

where $\epsilon_{i,n}$ is i.i.d. Gumbel noise. The final grasp prediction is

$$\hat{y}_n = \sum_{i=1}^{N_e} \pi_{i,n} E_i(H_{\mathrm{local},n}), \tag{14}$$

where $\hat{y}_n$ encodes grasp angle, depth, width, and confidence. Straight-through hard Gumbel-Softmax is used during training, whereas argmax selection is used at inference. Detailed routing and temperature settings are provided in Appendix B.

### 3.4 SEMANTIC QUERY CONSTRUCTION AND JOINT TRAINING

We construct semantic priors by associating SAM-derived object regions with frozen CLIP features (Fig. 2). Text queries are obtained from canonical descriptions of the 88 GraspNet object identities (Fang et al., 2020), while visual queries use CLIP image embeddings of target object regions. During training, visual regions are retrieved from an offline region bank; during inference, they are obtained from the current grasping scene. The resulting text or visual embedding is L2-normalized and used as $Q_{\mathrm{sem}}$.

To preserve task-agnostic grasping affordances, we jointly train with zero, text, and visual queries using probabilities of $0.5$, $0.25$, and $0.25$, respectively. Detailed semantic-prior construction and query protocols are provided in Appendix A.

### 3.5 TRAINING OBJECTIVE

We jointly optimize semantic gating, cross-modal alignment, MoE load balancing, and 6-DoF grasp prediction.

**Semantic Gate Supervision.** The semantic gate loss is computed exclusively for semantically valid samples. To prevent the larger number of background seeds from dominating the loss for small targets, we apply a balanced binary cross-entropy (BCE) loss where per-sample foreground and background weights are normalized to contribute equally. Specifically, the point-wise gate $\mathbf{g}$ is supervised by the aligned seed mask $M_{\mathrm{seed}}$:

$$\mathcal{L}_{\mathrm{att}} = \mathrm{BCE}_{\mathrm{bal}}(\mathbf{g}, M_{\mathrm{seed}}) \tag{15}$$

**Cross-Modal Contrastive Alignment.** The normalized gates pool the projected CMSA value features into a unified target representation. To bridge the modality gap, we apply a symmetric multi-positive InfoNCE loss $\mathcal{L}_{\mathrm{contra}}$ that aligns this visual representation with the corresponding normalized semantic query embeddings.

**MoE Load Balancing.** Let $\bar{P}$ and $\bar{W}$ be the batch-averaged soft routing probabilities and hard routing loads. We regularize both toward $U = (1/N_e, \ldots, 1/N_e)$:

$$\mathcal{L}_{\mathrm{bal}} = \mathrm{MSE}(\bar{P}, U) + \mathrm{MSE}(\bar{W}, U) \tag{16}$$

**Grasp Loss.** The network predicts objectness, graspness, view, angle, depth, width, and confidence. To ensure accurate geometric regression, invalid background seeds are explicitly excluded from the computation, while the confidence branch is supervised via class-weighted cross-entropy. The grasp regression loss is formulated as:

$$\mathcal{L}_{\text{grasp}} = \lambda_v \mathcal{L}_{\text{view}} + \lambda_a \mathcal{L}_{\text{angle}} + \lambda_d \mathcal{L}_{\text{depth}} + \lambda_w \mathcal{L}_{\text{width}} + \lambda_s \mathcal{L}_{\text{score}} \tag{17}$$

**Total Objective and Dynamic Curriculum.** The training objective is computed as a dynamically weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{att}} \mathcal{L}_{\text{att}} + \delta \mathcal{L}_{\text{bal}} + \beta(t) \mathcal{L}_{\text{contra}} + \gamma(t) \mathcal{L}_{\text{grasp}}. \tag{18}$$

where $\mathcal{L}_{\text{base}}$ denotes the objectness and graspness supervision inherited from the base grasp detector. $\lambda_{\text{att}} = 20$ and $\delta = 10$ are fixed weights for semantic-gate supervision and MoE load balancing. The time-dependent coefficients $\beta(t)$ and $\gamma(t)$ implement the contrastive-loss schedule and grasp-loss warm-up, respectively. Detailed optimization schedules are provided in Appendix B.

# 4 Experiments

## 4.1 Experimental Setup

**Dataset and Splits.** We use the standard Test-Seen, Test-Similar, and Test-Novel splits of GraspNet-1Billion (Fang et al., 2020). Target-conditioned evaluation is constructed over visible frame–object pairs whose object identities belong to the 88-object benchmark vocabulary.

**Implementation.** All models are trained and evaluated under a common preprocessing and physics-based evaluation protocol. Full training, preprocessing, and evaluation settings are provided in Appendix B. Cached inference efficiency and resource usage are reported in Appendix D.

## 4.2 Evaluation Metrics

We evaluate scene-level grasp quality and target-conditioned physics-based performance. For the target-conditioned query set $\mathcal{Q}$, let $\mathcal{G}_q = (g_{q,1}, \ldots, g_{q,K_q})$ denote the confidence-ranked evaluation candidates before full-scene collision removal. $o_q$ is the queried object, while $T(g, o_q)$, $C(g)$, and $F(g)$ indicate target assignment, collision-free status, and force closure, respectively. We define

$$V(g, o_q) = T(g, o_q) \wedge C(g) \wedge F(g). \tag{19}$$

**Target Assignment.** For each predicted grasp $g$, we assign it to the object whose transformed model point is nearest to the grasp center $t(g)$. Let $a(g)$ denote the resulting object identity. We define

$$T(g, o_q) = \mathbb{I}[a(g) = o_q]. \tag{20}$$

This assignment is performed before collision and force-closure evaluation and is shared by all target-conditioned physics-based metrics.

**Scene-Level Metrics.** We report the standard GraspNet mean AP, $\text{AP}_{0.4}$, and $\text{AP}_{0.8}$ over the benchmark's analytical friction thresholds.

**Hit Rate (HR@$k$).** HR@$k$ measures whether the top-$k$ pre-collision candidates contain at least one physically valid grasp on the queried object:

$$HR@k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}[\exists\, 1 \le i \le \min(k, K_q) : V(g_{q,i}, o_q)] \tag{21}$$

We report HR@1 and HR@5 to evaluate proposal quality without first removing colliding candidates from the ranked list.

**Semantic Grasp Success Rate (SGSR).** SGSR evaluates the highest-confidence collision-free candidate. Let $\mathcal{G}_q^{\text{cf}} = \{g \in \mathcal{G}_q \mid C(g) = 1\}$. When $\mathcal{G}_q^{\text{cf}} \neq \varnothing$, we define

$$g_q^* = \arg\max_{g \in \mathcal{G}_q^{\text{cf}}} s(g) \tag{22}$$

Table 1: **Target-conditioned physics-based evaluation (%).** The task-agnostic baselines are equipped with the same frozen SAM-CLIP retriever, while VL-Grasp is reported as an existing target-driven reference.

| Method | Query | Split | HR@1↑ | HR@5↑ | SGSR↑ | CF@1↑ |
|---|---|---|---|---|---|---|
| GraspNet Baseline (Fang et al., 2020) | Text | Seen | 6.45 | 15.42 | 14.53 | 17.75 |
| | | Similar | 5.80 | 15.33 | 14.63 | 15.73 |
| | | Novel | 0.97 | 3.18 | 2.62 | 12.58 |
| FGC-GraspNet (Lu et al., 2022) | Text | Seen | 5.20 | 9.81 | 8.47 | 18.31 |
| | | Similar | 4.10 | 8.00 | 6.54 | 14.68 |
| | | Novel | 0.72 | 1.90 | 1.33 | 11.60 |
| GraNet (Wang et al., 2023) | Text | Seen | 6.42 | 14.33 | 11.34 | 28.76 |
| | | Similar | 7.65 | 15.35 | 13.00 | 27.28 |
| | | Novel | 1.17 | 3.89 | 2.45 | 30.80 |
| EconomicGrasp (Wu et al., 2024) | Text | Seen | 12.75 | 23.10 | 19.21 | 30.87 |
| | | Similar | 11.01 | 21.18 | 17.53 | 27.21 |
| | | Novel | 1.98 | 4.65 | 2.99 | 18.10 |
| VL-Grasp (Lu et al., 2023) | Text | Seen | 7.62 | 14.94 | 12.74 | 19.36 |
| | | Similar | 1.81 | 3.88 | 3.34 | 12.57 |
| | | Novel | 0.37 | 0.96 | 0.67 | 8.21 |
| **SAGE-Grasp (Ours)** | Text | Seen | **37.04** | **61.00** | **43.51** | **72.17** |
| | | Similar | 14.78 | 29.22 | 19.31 | 58.47 |
| | | Novel | 4.24 | 9.99 | 5.40 | 50.78 |
| | Visual | Seen | 35.20 | 59.49 | 42.87 | 70.26 |
| | | Similar | **26.01** | **48.24** | **32.89** | **64.00** |
| | | Novel | **6.77** | **17.66** | **8.63** | **62.26** |

Queries with no collision-free candidate are counted as failures. Otherwise, success requires $g_q^*$ to target the queried object and satisfy force closure:

$$\mathrm{SGSR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}\left[\mathcal{G}_q^{\mathrm{cf}} \neq \varnothing\right] \mathbb{I}\left[T(g_q^*, o_q) \wedge F(g_q^*)\right] \tag{23}$$

**Collision-Free Rate at Rank 1 (CF@1):** CF@1 measures whether the top-ranked pre-collision candidate is collision-free, independently of target assignment and force closure:

$$\mathrm{CF@1} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}[K_q \geq 1 \wedge C(g_{q,1})] \tag{24}$$

### 4.3 Comparison with Existing Methods

Since most GraspNet-1Billion detectors are task-agnostic, we construct controlled cascaded target-driven variants of GraspNet (Fang et al., 2020), FGC-GraspNet (Lu et al., 2022), GraNet (Wang et al., 2023), and EconomicGrasp (Wu et al., 2024) using the same frozen SAM-CLIP retriever. We additionally report VL-Grasp (Lu et al., 2023) as an existing language-conditioned 6-DoF grasping reference. All methods are evaluated under the same target-conditioned physics-based evaluation protocol.

**Target-Conditioned Results.** Table 1 shows that SAGE-Grasp with text queries consistently outperforms both the controlled cascaded baselines and VL-Grasp across the Seen, Similar, and Novel splits. SAGE-Grasp achieves the highest HR@1, HR@5, SGSR, and CF@1 among all evaluated text-query methods. The improvement is particularly pronounced on Seen, where HR@1 increases from 12.75 to 37.04 and SGSR from 19.21 to 43.51 compared with the strongest competing results. The consistent gains across the three splits indicate that CMSA effectively associates the semantic query with target-relevant point-cloud features, enabling more accurate target-conditioned grasp prediction. The higher CF@1 further suggests that retaining contextual geometric information during target localization benefits collision-aware grasp generation.

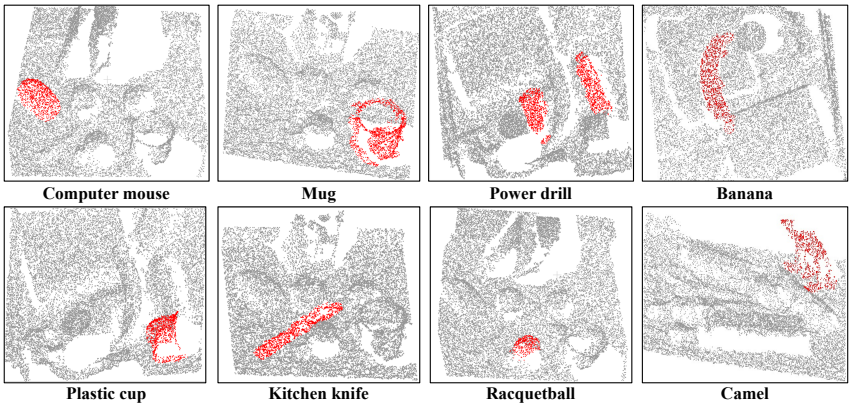



Figure 4: **Visualization of CMSA gate-response maps.**

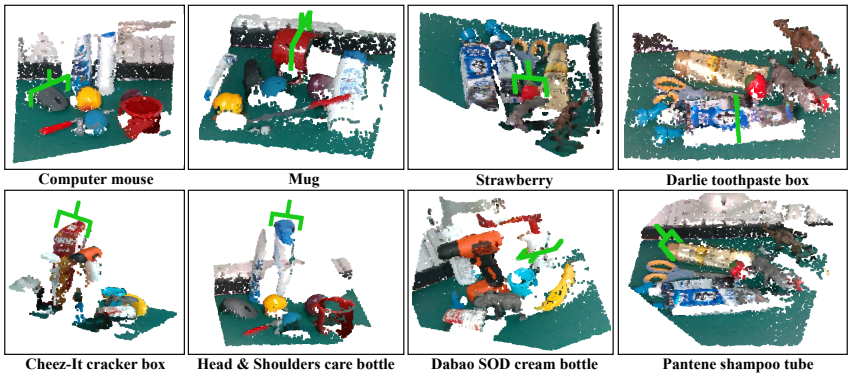



Figure 5: **Language-conditioned 6-DoF grasp predictions in clutter.**

The scene-derived visual query provides a complementary conditioning modality. It achieves performance comparable to the text query on Seen, while yielding higher target-conditioned metrics on Similar and Novel. This suggests that direct visual-semantic conditioning provides additional robustness to appearance variation. These results are reported as a modality analysis and are not included in the text-query comparison with the baseline methods.

**Analysis of Semantic Grounding and Grasping.** Fig. 4 visualizes the language-driven gate responses produced by CMSA. In the shown examples, high responses are concentrated on the queried objects, with limited activation on neighboring clutter. This behavior is observed across objects with different geometric characteristics, including complex shapes (e.g., mugs), slender objects (e.g., kitchen knives), and small objects (e.g., racquetballs), suggesting that CMSA provides consistent target-aware semantic refinement across these examples.

Fig. 5 further presents representative language-conditioned grasp predictions in cluttered scenes. The examples cover both category-level and fine-grained object names, as well as box-like, slender, and irregular object geometries. Across these cases, the predicted grasps remain localized on the queried targets despite nearby distractors, providing qualitative evidence that target-aware localization is retained by the downstream grasp predictor.

**Scene-Level Results.** To assess whether target-conditioned learning preserves task-agnostic grasping affordances, Table 2 evaluates SAGE-Grasp under zero-query scene-level grasping. SAGE-Grasp achieves AP values of 60.50, 51.99, and 23.16 on Seen, Similar, and Novel, exceeding the strongest listed baseline by 17.17, 12.01, and 8.26 points, respectively. Consistent gains in $\mathrm{AP}_{0.8}$ and $\mathrm{AP}_{0.4}$ further show that the joint model retains strong task-agnostic grasping across the three splits. Although the improvement margin is smaller on Novel, SAGE-Grasp still achieves the highest reported AP on this split. Additional scene-level qualitative predictions are provided in Appendix E.

Table 2: **Task-agnostic scene-level performance on the RealSense camera setting.**

| Method | Seen | | | Similar | | | Novel | | |
|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ |
| GPD (ten Pas et al., 2017) | 22.87 | 28.53 | 12.84 | 21.33 | 27.83 | 9.64 | 8.24 | 8.89 | 2.67 |
| GG-CNN (Morrison et al., 2018) | 15.48 | 21.84 | 10.25 | 13.26 | 18.37 | 4.62 | 5.52 | 5.93 | 1.86 |
| PointNetGPD (Liang et al., 2019) | 25.96 | 33.01 | 15.37 | 22.68 | 29.15 | 10.76 | 9.23 | 9.89 | 2.74 |
| S4G (Qin et al., 2020) | 25.71 | 31.19 | 20.84 | 18.45 | 23.14 | 12.86 | 9.04 | 11.23 | 3.61 |
| GraspNet Baseline (Fang et al., 2020) | 27.56 | 33.43 | 16.95 | 26.11 | 34.19 | 14.23 | 10.55 | 11.25 | 3.98 |
| TransGrasp (Liu et al., 2022) | 39.81 | 47.54 | 36.42 | 29.32 | 34.80 | 25.19 | 13.83 | 17.11 | 7.67 |
| GraNet (Wang et al., 2023) | 43.33 | 52.56 | 34.04 | 39.98 | 48.66 | 32.00 | 14.90 | 18.66 | 7.76 |
| **SAGE-Grasp (Ours)** | **60.50** | **71.78** | **52.80** | **51.99** | **64.30** | **40.42** | **23.16** | **29.00** | **11.39** |

Table 3: **Ablation studies.** (a) CMSA and the MoE head for target-conditioned grasping. (b) Zero-query joint training for task-agnostic scene-level grasping.

**(a) Target-conditioned grasping**

| Method | Seen | | | Similar | | | Novel | | |
|---|---|---|---|---|---|---|---|---|---|
| | HR@1 | HR@5 | SGSR | HR@1 | HR@5 | SGSR | HR@1 | HR@5 | SGSR |
| w/o CMSA | 36.55 | 59.17 | 42.80 | 12.79 | 24.84 | 16.94 | 3.52 | 8.41 | 4.88 |
| w/o MoE | 33.32 | 56.16 | 38.80 | 13.68 | 27.55 | 17.55 | 3.69 | 9.51 | 4.68 |
| **SAGE-Grasp** | **37.04** | **61.00** | **43.51** | **14.78** | **29.22** | **19.31** | **4.24** | **9.99** | **5.40** |

**(b) Task-agnostic scene-level grasping**

| Method | Seen | | | Similar | | | Novel | | |
|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ |
| w/o Joint Training | 2.91 | 3.48 | 0.90 | 2.05 | 2.49 | 0.72 | 2.42 | 2.85 | 0.89 |
| **SAGE-Grasp** | **60.50** | **71.78** | **52.80** | **51.99** | **64.30** | **40.42** | **23.16** | **29.00** | **11.39** |

### 4.4 Ablation Studies

Table 3 evaluates the three components according to their functional roles. Since CMSA and the MoE head directly participate in target-conditioned grasp prediction, they are evaluated under the target-conditioned setting, whereas zero-query joint training is designed to preserve task-agnostic scene-level grasping capability and is therefore evaluated using standard scene-level metrics. In Table 3(a), removing either CMSA or the MoE head consistently reduces HR@1, HR@5, and SGSR across all three splits. Removing CMSA causes the largest degradation on the Similar split, highlighting the benefit of point-wise semantic addressing under distribution shifts. The consistent performance drop without the MoE head further demonstrates the effectiveness of expert-routed prediction. Expert-routing distributions across the three splits are reported in Appendix C.

Table 3(b) shows that removing zero-query joint training causes a substantial collapse in scene-level grasping performance across all splits. This demonstrates that zero-query supervision is critical for retaining task-agnostic grasping capability during target-conditioned joint training.

## 5 Conclusion and Limitations

We presented SAGE-Grasp, a unified framework for target-conditioned 6-DoF grasping in clutter. By fusing semantic priors with 3D geometry via CMSA and Gumbel-routed MoE, our method supports target-conditioned grasp prediction across varied local geometries. Experiments demonstrate that SAGE-Grasp outperforms cascaded baselines in target-conditioned grasping and collision-free proposal generation, while zero-query training preserves strong scene-level grasping capability. The current method depends on offline SAM-CLIP features and has not yet been evaluated in closed-loop robot execution. Future work will investigate online semantic extraction, real-world execution, articulated objects, and dexterous manipulation.

## AI USE STATEMENT

Generative AI tools were used to assist with language polishing, translation, and manuscript organization. All scientific content, mathematical formulations, experimental design, implementation, experimental results, citations, and conclusions were reviewed and verified by the authors. AI-generated suggestions were used only under human supervision, and the authors take full responsibility for the final content of the manuscript.

## REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the main paper and appendices. Sections 3.1–3.5 specify the model architecture, semantic conditioning mechanisms, routing formulation, and training objectives, while Sec. 4.2 defines the evaluation metrics. Appendix A describes semantic-prior construction and the text and visual query protocols. Appendix B reports the hardware, preprocessing settings, optimization schedules, loss weights, random seed, evaluation configuration, and query-level aggregation protocol. Additional routing statistics, cached inference measurements, and qualitative results are provided in Appendices C–E, respectively. Each training configuration was run once, as explicitly stated in Appendix B.

## REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229, Cham, 2020. Springer. doi: 10.1007/978-3-030-58452-8_13.

Yaofeng Cheng, Fusheng Zha, Wei Guo, Pengfei Wang, Chao Zeng, Lining Sun, and Chenguang Yang. PCF-Grasp: Converting point completion to geometry feature to enhance 6-DoF grasp. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(1):617–628, 2026. doi: 10.1109/TSMC.2025.3628946.

Eugenio Chisari, Nick Heppert, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. CenterGrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-DoF grasp estimation. *IEEE Robotics and Automation Letters*, 9(6):5094–5101, 2024. doi: 10.1109/LRA.2024.3388850.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. GraspNet-1Billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453, 2020. doi: 10.1109/CVPR42600.2020.01146.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Dinh-Cuong Hoang, Johannes A. Stork, and Todor Stoyanov. Context-aware grasp generation in cluttered scenes. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 1492–1498. IEEE, 2022. doi: 10.1109/ICRA46639.2022.9811371.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Shiyu Jin, Jinxuan Xu, Yutian Lei, and Liangjun Zhang. Reasoning grasping via multimodal large language model. *arXiv preprint arXiv:2402.06798*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00371.

Juncheng Li and David J. Cappelleri. Sim-Grasp: Learning 6-DoF grasp policies for cluttered environments using a synthetic benchmark. *IEEE Robotics and Automation Letters*, 9(9):7645–7652, 2024. doi: 10.1109/LRA.2024.3430712.

Meng Li, Qi Zhao, Shuchang Lyu, Chunlei Wang, Yujing Ma, Guangliang Cheng, and Chenguang Yang. OVGNet: A unified visual-linguistic framework for open-vocabulary robotic grasping. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7507–7513. IEEE, 2024. doi: 10.1109/IROS58592.2024.10802654.

Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. PointNetGPD: Detecting grasp configurations from point sets. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3629–3635. IEEE, 2019. doi: 10.1109/ICRA.2019.8794435.

Yushi Liu, Alexander Qualmann, Zehao Yu, Miroslav Gabriel, Philipp Schillinger, Markus Spies, Ngo Anh Vien, and Andreas Geiger. Efficient end-to-end detection of 6-DoF grasps for robotic bin picking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5427–5433. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611417.

Zhixuan Liu, Zibo Chen, Shangjin Xie, and Wei-Shi Zheng. TransGrasp: A multi-scale hierarchical point transformer for 7-DoF grasp detection. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1533–1539. IEEE, 2022. doi: 10.1109/ICRA46639.2022.9812001.

Yuhao Lu, Beixing Deng, Zhenyu Wang, Peiyuan Zhi, Yali Li, and Shengjin Wang. Hybrid physical metric for 6-DoF grasp pose detection. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8238–8244. IEEE, 2022. doi: 10.1109/ICRA46639.2022.9811961.

Yuhao Lu, Yixuan Fan, Beixing Deng, Fangfu Liu, Yali Li, and Shengjin Wang. 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)*, pp. 976–983. IEEE, 2023. doi: 10.1109/IROS55552.2023.10341379.

Haoxiang Ma, Modi Shi, Boyang Gao, and Di Huang. Generalizing 6-DoF grasp detection via domain prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18102–18111, 2024.

Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2906–2917. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00290.

Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. In *Proceedings of Robotics: Science and Systems XIV*, Pittsburgh, Pennsylvania, 2018. doi: 10.15607/RSS.2018.XIV.021.

Adithyavairavan Murali, Arsalan Mousavian, Clemens Eppner, Chris Paxton, and Dieter Fox. 6-DoF grasping for target-driven object manipulation in clutter. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6232–6238. IEEE, 2020. doi: 10.1109/ICRA40945.2020.9197318.

Toan Nguyen, Minh Nhat Vu, Baoru Huang, An Vuong, Quan Vuong, Ngan Le, Thieu Vo, and Anh Nguyen. Language-driven 6-dof grasp detection using negative prompt guidance. In *Computer Vision – ECCV 2024*, volume 15077 of *Lecture Notes in Computer Science*, pp. 363–381. Springer, 2024. doi: 10.1007/978-3-031-72655-2_21.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. ShapeLLM: Universal 3D object understanding for embodied interaction. In *Computer Vision – ECCV 2024*, volume 15101 of *Lecture Notes in Computer Science*, pp. 214–238, Cham, 2024. Springer. doi: 10.1007/978-3-031-72775-7_13.

Yuzhe Qin, Rui Chen, Hao Zhu, Meng Song, Jing Xu, and Hao Su. S4G: Amodal single-view single-shot SE(3) grasp detection in cluttered scenes. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 53–65. PMLR, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8583–8595. Curran Associates, Inc., 2021.

Jian Ruan, Houde Liu, Anshun Xue, Xueqian Wang, and Bin Liang. Grasp quality evaluation network for surface-to-surface contacts in point clouds. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pp. 1467–1472. IEEE, 2020. doi: 10.1109/CASE48305.2020.9216808.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Yaoxian Song, Penglei Sun, Piaopiao Jin, Yi Ren, Yu Zheng, Zhixu Li, Xiaowen Chu, Yue Zhang, Tiefeng Li, and Jason Gu. Learning 6-DoF fine-grained grasp detection based on part affordance grounding. *IEEE Transactions on Automation Science and Engineering*, 22:15200–15214, 2025. doi: 10.1109/TASE.2025.3566461.

Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13438–13444. IEEE, 2021. doi: 10.1109/ICRA48506.2021.9561877.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514.

Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13–14):1455–1473, 2017. doi: 10.1177/0278364917735594.

Georgios Tziafas and Hamidreza Kasaei. Towards open-world grasping with large vision-language models. *arXiv preprint arXiv:2406.18722*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

An Dinh Vuong, Minh Nhat Vu, Baoru Huang, Nghia Nguyen, Hieu Le, Thieu Vo, and Anh Nguyen. Language-driven grasp detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17902–17912. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01695.

An-Lan Wang, Nuo Chen, Kun-Yu Lin, Yuan-Ming Li, and Wei-Shi Zheng. Task-oriented 6-DoF grasp pose detection in clutters. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5692–5698. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11128749.

Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15964–15973. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01566.

Haowen Wang, Wanhao Niu, and Chungang 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)*, pp. 937–943. IEEE, 2023. doi: 10.1109/IROS55552.2023.10341549.

Zixi Wang, Zeyi Liu, Nicolas Ouporov, and Shuran Song. ContactHandover: Contact-guided robot-to-human object handover. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9916–9923. IEEE, 2024. doi: 10.1109/IROS58592.2024.10801777.

Xiao-Ming Wu, Jia-Feng Cai, Jian-Jian Jiang, Dian Zheng, Yi-Lin Wei, and Wei-Shi Zheng. An economic framework for 6-DoF grasp detection. In *Computer Vision – ECCV 2024*, volume 15085 of *Lecture Notes in Computer Science*, pp. 357–375, Cham, 2024. Springer. doi: 10.1007/978-3-031-73383-3_21.

Pengwei Xie, Siang Chen, Yixiang Dai, Dingchang Hu, Kaiqin Yang, and Guijin Wang. Target-oriented object grasping via multimodal human guidance. In *Computer Vision – ECCV 2024 Workshops*, volume 15634 of *Lecture Notes in Computer Science*, pp. 305–322. Springer, 2025. doi: 10.1007/978-3-031-92591-7_19.

Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, and He Wang. UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4737–4746, 2023.

# A Query Construction and Protocol

**Semantic-Prior Construction.** For each RGB observation, SAM is used to generate 2D instance regions. Masks smaller than 800 pixels are discarded. For each retained region, we construct a context-aware crop from its bounding box. Instead of hard-masking the background, the luminance of background pixels is attenuated by 80% to preserve local contextual cues. Each crop is encoded using the frozen CLIP image encoder and associated with the corresponding 3D points to form the cached semantic-prior bank used by SAGE-Grasp.

**Text Query.** We adopt a controlled object-referential text-query protocol. Specifically, each of the 88 GraspNet object identities is associated with a canonical visible description containing its object category and, when useful for disambiguation, observable attributes such as color, size, shape, or brand identity. Examples include "yellow banana," "small blue plastic cup," and "Darlie toothpaste box."

For each object description, we instantiate six contextual templates: "a photo of a $\{\}$," "a 3D rendering of a $\{\}$," "a point cloud of a $\{\}$," "a close-up view of a $\{\}$," "a robotic grasp target: a $\{\}$," and "an object to be manipulated: a $\{\}$." The resulting prompts are encoded independently by the frozen CLIP text encoder. Their embeddings are L2-normalized, averaged, and normalized again to produce one 512-dimensional query representation for each object identity. We additionally examine pairwise cosine similarities among the 88 query embeddings.

This protocol evaluates text-conditioned object reference under a fixed benchmark vocabulary rather than unrestricted natural-language instruction following. The current experiments do not systematically cover unseen paraphrases, compositional descriptions, spatial relations, ambiguous instructions, or disambiguation between multiple instances of the same category.

**Visual Query.** For visual-query training, we construct an offline bank of SAM-derived object regions. For each semantic training sample, the region corresponding to the target object is retrieved and encoded by the frozen CLIP image encoder. The resulting feature is L2-normalized and used as the visual semantic query.

At inference, the visual-query region is obtained from the current grasping scene using SAM and encoded by the same frozen CLIP image encoder. The resulting embedding replaces the text embedding as $Q_{\mathrm{sem}}$; subsequent candidate sampling, CMSA refinement, view selection, expert routing, and grasp prediction remain unchanged. This setting therefore evaluates scene-derived visual conditioning rather than unrestricted external reference-image querying.

**Joint Query Sampling.** For each training sample, a zero query is used with probability $0.5$. Among the remaining semantic-query samples, text and visual conditioning are selected with equal probability, resulting in overall probabilities of $0.25$ for text and $0.25$ for visual queries.

# B Experimental Setup and Reproducibility

**Training and Hardware.** All experiments were conducted on a Linux workstation equipped with an Intel Core i9-14900K CPU, 128 GB of system memory, and two NVIDIA RTX 4090D GPUs. Model training, inference, and evaluation were implemented using PyTorch 2.5.1.

Each SAGE-Grasp configuration, including the ablation variants, was trained once for 40 epochs, and each reported result was obtained from a single evaluation run of the corresponding checkpoint. For target-conditioned evaluation, each query instance corresponds to a visible object in a particular test frame whose identity belongs to the 88-object query vocabulary. The same object identity may therefore contribute multiple query instances across different scenes and camera viewpoints. HR@1, HR@5, SGSR, and CF@1 are averaged over all frame–object query instances satisfying the visibility and vocabulary criteria in each split, rather than over the 88 object identities. If candidate sampling yields no seed associated with the queried target, the query remains in the evaluation set and is counted as a failure. Scene-level SAGE-Grasp results were obtained from a single evaluation run over the complete official test split, and the cascaded baselines in the target-conditioned comparison were evaluated once under the same evaluation protocol. Reported results were not averaged over multiple independent training runs or random seeds.

Table 4: **Expert-routing distribution over target seeds (%).**

| Split | Expert 1 | Expert 2 | Expert 3 | Expert 4 |
|---|---|---|---|---|
| Seen | 10.09 | 42.82 | 44.49 | 2.60 |
| Similar | 16.05 | 27.90 | 52.87 | 3.18 |
| Novel | 13.15 | 27.97 | 56.40 | 2.48 |

Table 5: **Cached inference efficiency of SAGE-Grasp-Text.**

| Metric | Value |
|---|---|
| Latency per scene | 6.62 s |
| Latency per query | 662.21 ms |
| Throughput | 1.51 queries/s |
| Peak GPU memory | 6.47 GB |
| Peak host memory | 3.97 GB |
| Average GPU utilization | 34.2% |
| Average GPU power | 168.5 W |
| Checkpoint size | 224.88 MB |

**Preprocessing and Evaluation Configuration.** Each scene is voxelized using a 0.005 m voxel size and sampled to 20,000 points, from which 1,024 seed points are retained. During evaluation, we use a score threshold of 0.05, NMS thresholds of 0.03 m and $15^\circ$, full-scene collision checking with a 0.01 m margin, a friction coefficient of 0.4, a maximum gripper width of 0.10 m, a collision-model voxel size of 0.008 m, and at most 50 candidates per query. The random seed is fixed to 42. We use $N_e = 4$ and $N_s = 1024$ in all experiments.

**Optimization and Curriculum Details.** We use AdamW with an initial learning rate of $10^{-3}$, a minimum learning rate of $10^{-5}$, weight decay of $10^{-4}$, and gradient clipping at 5.0. The learning rate is linearly warmed up for the first two epochs and then follows cosine decay. The loss weights for objectness, graspness, view, angle, depth, score, and width are 1, 10, 20, 1, 1, 1, and 2, respectively. The semantic-gate and MoE-balancing losses use fixed weights of 20 and 10, respectively. The contrastive loss has a weight of 0.5 when active and is enabled from the fifth training epoch. The InfoNCE temperature is $0.07$.

For both view selection and expert routing, the Gumbel-Softmax temperature is exponentially annealed from $1.0$ to $0.2$ over the first 30 epochs. Straight-through hard Gumbel-Softmax routing is used during training, and argmax routing is used at inference.

## C Additional Ablation Analysis

Table 4 reports the expert-routing distribution over target seeds. The routing is non-uniform, with Experts 2 and 3 receiving most assignments. Notably, their relative usage changes from 42.82%/44.49% on Seen to 27.97%/56.40% on Novel, while the other experts remain active. This redistribution across data splits indicates that the router adjusts expert allocation according to the input distribution rather than collapsing to a fixed routing pattern. Since routing frequency alone does not reveal the functional role of each expert, we interpret these results as evidence of data-dependent routing rather than explicit expert specialization.

## D Cached Inference Efficiency and Resource Usage

Table 5 reports the cached inference efficiency of SAGE-Grasp-Text when the SAM–CLIP semantic priors and their 2D–3D associations are precomputed. The evaluation covers 10 scenes and 100 target queries.

Under this cached setting, SAGE-Grasp-Text requires 6.62 s per scene and 662.21 ms per query, corresponding to a throughput of 1.51 queries/s. Peak GPU and host memory usage are 6.47 GB and

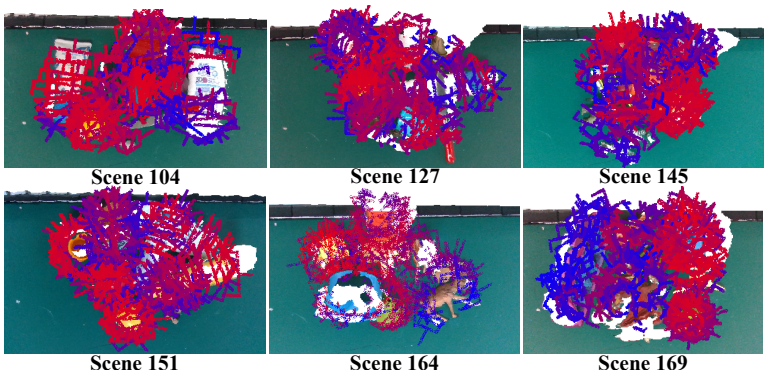



Figure 6: **Scene-level grasp predictions.** Colors encode confidence from blue (low) to red (high).

3.97 GB, respectively, while the average GPU utilization and power draw are 34.2% and 168.5 W. The trained checkpoint occupies 224.88 MB.

These measurements characterize the cached inference setting used in our benchmark evaluation. They exclude the one-time costs of SAM mask generation, CLIP feature extraction, 2D–3D association, and Feature Bank construction. Once these priors are available, each target query is processed through the unified SAGE-Grasp prediction path without target-specific segmentation or isolated-object grasp inference. A complete online deployment would additionally incur semantic-prior construction costs, which are not measured in the current implementation.

## E Additional Qualitative Results

Fig. 6 presents additional scene-level grasp predictions on representative GraspNet-1Billion test scenes. Under zero-query inference, predicted grasps are distributed across multiple objects and local contact regions in clutter, rather than concentrating on a single semantic target. The predictions also span different grasp orientations and object geometries across the shown scenes, qualitatively complementing the task-agnostic scene-level results in Table 2. Grasp colors encode prediction confidence from blue (low) to red (high).