Grasp the Invisibility by Vision-Language guided Active View Planning

Yitian Shi*, Di Wen*, Edgar Welte, Kunyu Peng, Rainer Stiefelhagen, Rania Rayyes

***Abstract*— We propose a new framework, VISO-Grasp, that handles grasping in environments with extreme occlusion by integrating vision-language understanding. Our approach harnesses the spatial reasoning capabilities of large Foundation Models (FMs) to actively guide viewpoint selection and build a dynamic, instance-focused map of object relationships. This evolving representation improves grasp reliability under limited visibility and enables informed Next-Best-View (NBV) planning and sequential grasp execution when direct access is obstructed. To further enhance robustness, we propose a multi-view fusion strategy driven by uncertainty, which adaptively refines grasp confidence and directional estimates in real-time. The Video is under: https://www.youtube.com/watch?v=HsJCMzc-Zas**

I. Introduction

Robotic grasping in unstructured, cluttered environments remains a significant challenge, particularly for executing target-oriented grasps under partial or complete occlusions [1]. Humans instinctively overcome occlusion during targeted searches by adjusting their viewpoints and intuitive reasoning about spatial relationships to infer potential target locations. As inspired by this, we propose VISO-Grasp, a Vision-language Informed Spatial Object-centric grasping framework that integrates off-the-shelf Foundation Models (FMs) [2]–[4] with active vision and occlusion-aware $\mathrm{SE}(3)$ grasp planning, which leverages the inherent prior of VLMs to achieve human-like decision-making. VISO-Grasp introduces evolving spatial reasoning that continuously integrates spatial and occlusion relationships. By incorporating an online grasp fusion mechanism, our approach dynamically refines target visibility and substantially improves grasp efficiency in heavy occlusions.

II. Methodology

We consider a robotic manipulation system operating in an unknown, cluttered environment containing a set of *objects*. Among these, a distinct *target object* is subject to significant occlusions, causing partial or complete invisibility.

A. System Overview

Fig. 2 illustrates the fundamental components of VISO-Grasp, which comprises: i) Adaptive Multi-view Open-Vocabulary 3D Object Detector (AMOV3D); ii) Target-guided View Planner (TGV-Planner); iii) Real-Time Uncertainty-guided Multi-view Grasp Fusion (RT-UMGF).

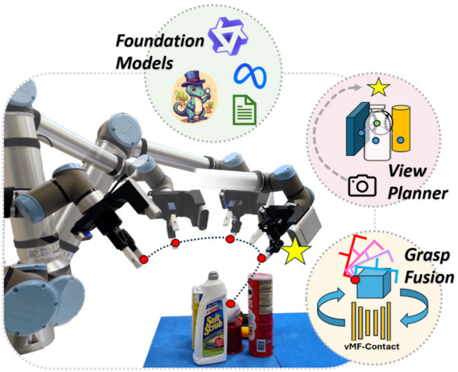

Fig. 1: VISO-Grasp, a unified system integrating Foundation Models (FMs) into target-aware active view planning and uncertainty-driven real-time 6-DoF grasp fusion.

B. Adaptive Multi-View Open-Vocabulary 3D Object Detection (AMOV3D)

In general, the AMOV3D module leverages FMs to achieve robust 3D object detection and segmentation in cluttered environments. Given an input image and a prompt-specified target object, a VLM generates structured descriptions for each detected object, prioritizing the identification of the target to mitigate the risk of occlusion-induced omission. Each object instance is described using its label and three attributes, which serve as a text prompt for open-vocabulary object detection [3], [4]. A historical object list, which stores all objects currently perceived within the scene, is maintained and updated at each new viewpoint to refine detected object attributes and resolve occlusions. This list persists throughout the planning process, dynamically evolving as new observations improve object descriptions and scene understanding. Each viewpoint captured in the history provides additional observations of the scene. During Next-Best-View (NBV) planning, AMOV3D updates the detected object set $\mathcal{O}'$ by integrating new observations and verifying consistency with prior data. By continuously refining $\mathcal{O}'$ and selecting viewpoints that maximize the visibility of the target object, AMOV3D improves the accuracy of object detection and occlusion reasoning, thereby enhancing the spatial reasoning accuracy of the TGV-Planner.

If the target object is not detected in the current view, the system initiates an occlusion reasoning process to infer potential obstructing objects. We employ in-context learning within the VLM to analyze the scene's spatial configuration and identify potential occluders based on structured object descriptions and the current view image. To ensure accurate inference without additional computational cost, we enforce

Karlsruhe Institute of Technology, Karlsruhe, Germany. Email: yitian.shi@kit.edu. This work is sponsored by the DFG SFB-1574 Circular Factory project and Baden-Württemberg Ministry of Science, Research and the Arts within InnovationCampus Future Mobility.

*Equal contribution

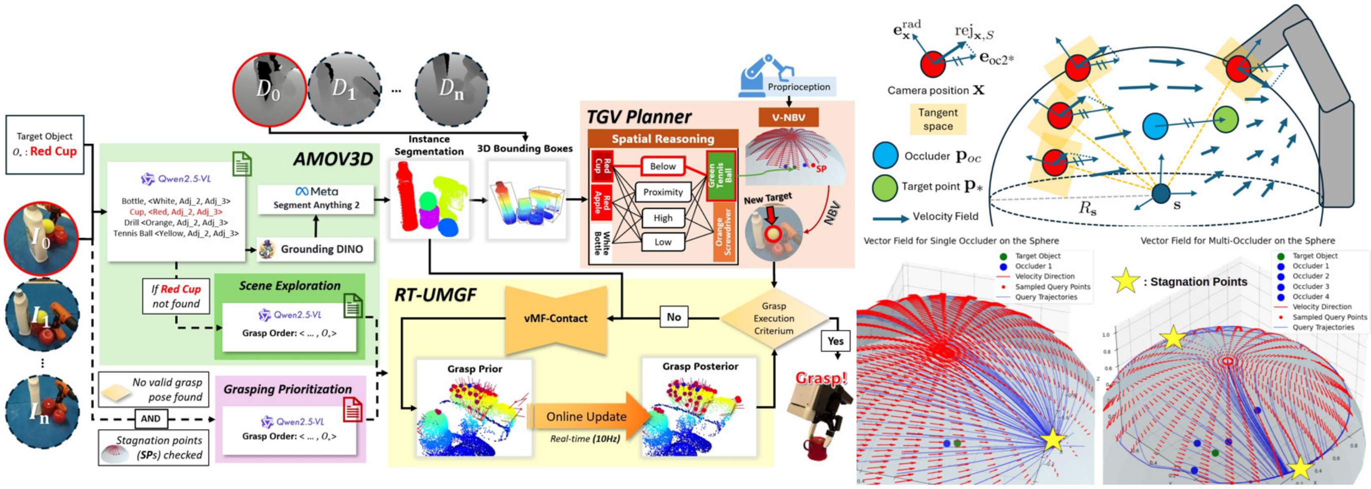

Fig. 2: Overview of VISO-Grasp (left) and principle for TGV-Planner (right).

a Mixture-of-Reasoning-Experts (MoRE) [5] paradigm, guiding the VLM to concurrently assess spatial relations, material properties, and geometric constraints via structured prompting. The outputs are aggregated using voting to enhance occlusion identification robustness.

C. Target-Guided View Planner (TGV-Planner)

The TGV-Planner performs spatial reasoning to determine whether the target object can be directly grasped or if an occluder must be removed first. If occlusions prevent direct grasping, the system prioritizes occluder removal. However, when direct removal is infeasible or does not sufficiently expose the target, the *Velocity-field-based NBV (V-NBV)* module optimizes the view planning. *V-NBV* is applied to either enhance the visibility of for grasp execution or to improve occluder removal efficiency by selecting a more informative camera perspective. The general principle to construct V-NBV is illustrated in Fig. 2 (right), and relevant details are explained in [6].

D. Real-Time Uncertainty-guided Multi-view Grasp Fusion (RT-UMGF)

In general, RT-UMGF continuously refines grasp predictions by fusing multi-view observations. It applies Bayesian updates using a von Mises-Fisher (vMF) distribution, ensuring stable grasp selection in cluttered environments. We employ the vMF-Contact [7], which is adapted to real-time (10Hz) inference through pose-centric uncertainty modeling and buffered online Bayesian fusion. In runtime, assume that the raw point cloud is captured in the time frame t, a set of contact grasps [8] is inferred as $G_t = \{\mathbf{g}_n^t = (\mathbf{c}, \mu_{\mathbf{c}}, \kappa_{\mathbf{c}}, \Delta_{\mathbf{c}}, w_{\mathbf{c}}, q_{\mathbf{c}})_n^t \mid n \in N\}$ that parameterizes: 1) The queried contact points $\mathbf{c} \in \mathbf{R}^3$; 2) Baseline vector distributions: $p(\mathbf{b}|\mathbf{c}) = \mathrm{vMF}(\mathbf{b}|\mu_{\mathbf{c}}, \kappa_{\mathbf{c}}) = Z(\kappa_{\mathbf{c}})\exp(\kappa_{\mathbf{c}}\mu_{\mathbf{c}}{}^{\top}\mathbf{b})$, with μ_c the mean direction of the baseline and κ_c as the directional precision. $Z(\kappa_{\mathbf{c}})$ is the normalization factor; 3) The quantized approach vector, represented by $\Delta_{\mathbf{c}} = \{\delta_{\mathbf{c}}^k\}_{k=0,\ldots,K}$ represents discrete categorical bin scores that define a direction constrained to lie on a plane perpendicular to a given baseline. 4) The grasp width $w_{\mathbf{c}}$ and 5) The contact quality score $q_{\mathbf{c}}$.

The grasp fusion process is designed as follows: For contact point positions $\mathbf{c}_j^t$, we adopt the same regime as [9] utilizing weighted sum by grasp quality q: $\mathbf{c}_i^t = \frac{q_i^{t-1}\mathbf{c}_i^{t-1}+\sum_j(q_j\mathbf{c}_j^t)}{q_i^{t-1}+\sum_j(q_j)}$ with grouped element indices $j \in J$. For baselines and approach vectors, the conjugate prior of vMF baseline distribution is initialized as: $\mu_{\mathbf{c}} \sim \mathrm{vMF}(\mu_{\mathbf{c}}^0, \kappa_{\mathbf{c}}^0)$ for $t = 0$. For the Bayesian inference in time frame t, we may update the posterior distribution following the rule of the exponential family [10] by:

$$\mu_{\mathbf{c}}^t = \frac{\kappa_{\mathbf{c}}^{t-1}\mu_{\mathbf{c}}^{t-1}+\sum_j\kappa_{\mathbf{c}_j}^t\mu_{\mathbf{c}_j}^t}{\kappa_{\mathbf{c}}^{t-1}+\sum_j\kappa_{\mathbf{c}_j}^t}, \kappa_{\mathbf{c}}^t = \kappa_{\mathbf{c}}^{t-1}+\sum_j\kappa_{\mathbf{c}_j}^t. \quad (1)$$

The approach categories are updated by $\delta_{\mathbf{c}}^{t,i} = \delta_{\mathbf{c}}^{t-1,i} + \sum_j \delta_{\mathbf{c}_j}^{t,i}$. Here $\kappa_{\mathbf{c}}^t$ represents the precision on the observed mean likelihood $\mu_{\mathbf{c}}$. We refer interesting readers for the theoretical background to [6], [7].

III. Conclusion

We develop VISO-Grasp, a novel vision-language-informed system for target-oriented grasping in highly unstructured environments including entire invisibility. By integrating a Vision-Language Model (VLM) with object-centric View planning and real-time uncertainty-driven grasp fusion, our system enhances scene understanding and improves grasp success through continuous velocity fields and semantic spatial reasoning for adaptive grasping in occluded environments with complete invisibility. VISO-Grasp leverages robust multi-view aggregation to refine grasp selection by integrating uncertain grasp hypotheses, ensuring superior stability and accuracy. More details to our experiments are documented under [6].

References

[1] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic *et al.*, "Deep learning approaches to grasp synthesis: A review," *TRO*, vol. 39, no. 5, 2023.

[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, “Qwen2. 5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.

[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.

[4] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in *ECCV*. Springer, 2024.

[5] C. Si, W. Shi, C. Zhao, L. Zettlemoyer, and J. Boyd-Graber, “Getting more out of mixture of language model reasoning experts,” *arXiv preprint arXiv:2305.14628*, 2023.

[6] Y. Shi, D. Wen, G. Chen, E. Welte, S. Liu, K. Peng, R. Stiefelhagen, and R. Rayyes, “Viso-grasp: Vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility,” *arXiv preprint arXiv:2503.12609*, 2025.

[7] Y. Shi, E. Welte, M. Gilles, and R. Rayyes, “vmf-contact: Uncertainty-aware evidential learning for probabilistic contact-grasp in noisy clutter,” *arXiv preprint arXiv:2411.03591*, 2024.

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

[9] J. Zhang, N. Gireesh, J. Wang, X. Fang, C. Xu, W. Chen, L. Dai, and H. Wang, “Gamma: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion,” in *ICRA*, 2024.

[10] B. Charpentier, O. Borchert, D. Zügner, S. Geisler, and S. Günnemann, “Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions,” in *ICLR*, 2022.