Semantically Safe Robot Manipulation: From Semantic Scene Understanding to Motion Safeguards

Lukas Brunke, Yanni Zhang, Ralf Römer, Jack Naimer, Nikola Staykov, Siqi Zhou, and Angela P. Schoellig

***Abstract*— Ensuring safe robot interactions in human environments requires adhering to common-sense safety (e.g., preventing spilling of water by keeping a cup straight). While safety in robotics is extensively studied, semantic understanding is rarely considered. We propose a semantic safety filter that certifies robot actions against semantically defined and geometric constraints. Our approach builds a 3D semantic map from perception inputs and uses large language models to infer unsafe conditions, which are enforced via control barrier certification. We validate our framework in teleoperated and learned manipulation tasks, demonstrating its effectiveness in real-world scenarios beyond traditional collision avoidance.**

I. INTRODUCTION

Safety is a key issue in robotics [1], [2]. In the control theory literature, safety is achieved through set invariance (i.e., to prevent a system from leaving a safe set) [2]. Various safety filters with this goal have been developed, which can be applied to unsafe control inputs and turn them into safe inputs [1], [3]. Existing safety filters based on control barrier function (CBF) [4] can provide theoretical safety guarantees. Still, they assume the safety constraints are given and typically restricted to geometrically defined constraints (e.g., collisions). In contrast, robots must adhere to *semantic constraints* that reflect "common sense" (see Fig. 1) to operate safely around humans. For example, a manipulator carrying a cup of water should avoid moving over electronic devices to prevent spills and limit rotation to avoid pouring. Such semantic constraints are not necessarily "visible," but are critical for real-world applications.

II. PROBLEM STATEMENT

In this work, we consider a manipulator transporting objects using teleoperation or a motion policy. These policies can be unsafe, so we aim to design a safety filter to ensure safe operation, adhering to semantic constraints (e.g., spatial, behavioral, and pose-based) and geometric constraints (e.g., environment-collision avoidance). We also assume that the environment is only perceived through RGB-D images and their associated camera poses.

III. METHODOLOGY

To ground our safety filter in the real world, we construct open-vocabulary 3D object-level point clouds of the environment similar to [5], [6]. We identify three types of semantic safety: *(i)* unsafe spatial relationships (e.g., "don't move a candle below a balloon"), *(ii)* behavioral constraints (e.g., "slow down when holding a knife"), and *(iii)* pose constraints (e.g., "keep a cup upright to avoid spilling"). These constraints are object- and scene-specific, making manual specification tedious. Therefore, we automate the synthesis using large language models (LLM) [7].

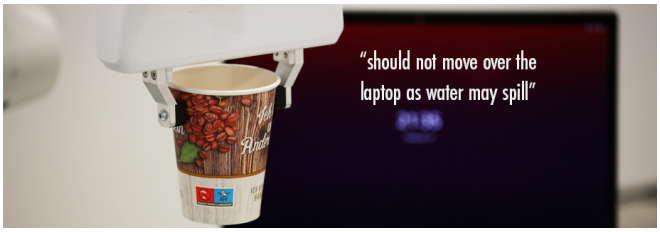

Fig. 1: We propose a semantic safety filter framework that leverages semantic scene understanding and contextual reasoning capabilities of large language models to certify robot motions with "common sense" constraints. A video of the full experimental results can be found at https://tiny.cc/semantic-manipulation and on our website https://utiasdsl.github.io/semantic-manipulation/.

We design a prompt for the LLM, which consists of multiple in-context examples and a final request as the true query. For each object in the scene, the requests contain the following components: *(i)* a high-level description of the scene, *(ii)* the object the robot is manipulating, and *(iii)* the object itself. Using these requests, we determine three sets of semantic constraints. First, the set of unsafe spatial relationships is $\mathcal{S}_\text{r}(o) = \{(l_i, r_i)\}_{i=1}^{N_\text{r}}$, where o is the manipulated object (e.g., `cup of water`), l_i is an object in the scene (e.g., `laptop`, `book`, etc.), r_i is an unsafe spatial relationship (e.g., `above`, `under`, etc.), and N_r is the number of unsafe spatial relationships. Second, the set of unsafe behaviors is $\mathcal{S}_\text{b}(o) = \{(l_i, b_i)\}_{i=1}^{N_\text{b}}$, where b_i indicates `caution` or `no caution` and N_b is the number of unsafe behaviours. Finally, the pose-based constraint set is $\mathcal{S}_\text{T}(o) = \{T\}$, where T specifies the end effector orientation constraint (`constrained rotation` or `free rotation`). The set of semantic constraints $\mathcal{S}(o)$ is the union of all the semantic constraints above.

Our semantic safety filter is designed based on CBFs [4] using $\mathcal{S}(o)$. We denote the joint positions by $\boldsymbol{q} \in \mathbb{R}^n$ (here $n = 7$) and assume joint velocity control $\dot{\boldsymbol{q}}$ [8].

1) Spatial Relationship Constraints: The semantic constraint sets are parameterized as the 0-super-level sets of continuously differentiable functions $\boldsymbol{h}_\text{sem}$. For each pair (l_i, r_i) in $\mathcal{S}_\text{r}(o)$, based on the point cloud $\boldsymbol{p}_i$ of the object l_i and the undesirable spatial relationship r_i, we fit a

The authors are with the Learning Systems and Robotics Lab and the Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80333 Munich, Germany. This work has been supported by the Robotics Institute Germany, funded by BMBF grant 16ME0997K, and the EU Horizon 2024 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101155035 (SSDM). Email: `firstname.lastname@tum.de`

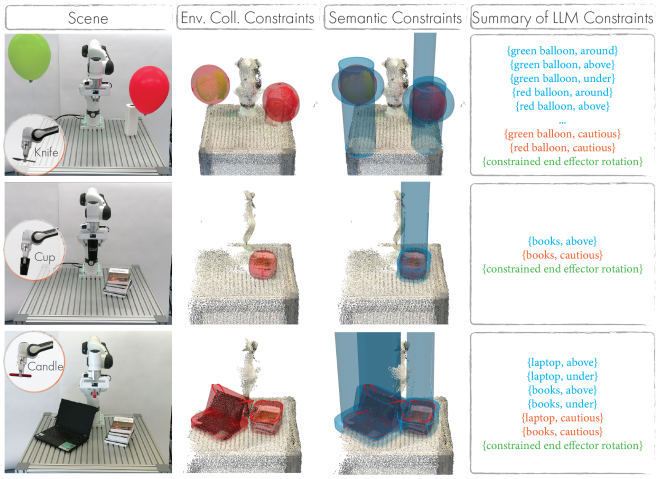

Fig. 2: Examples of the environment collision and semantic constraints enforced by our proposed semantic safety filter. For each scene, environment collision constraints are generated based on the point clouds of individual objects while the semantic constraints are synthesized based on the point clouds and labels of individual objects as well as the semantic safety conditions from the LLM. The semantic safety conditions are further categorized into spatial relationship constraints (blue text), behavioural constraints (orange text), and end effector pose constraints (green text).

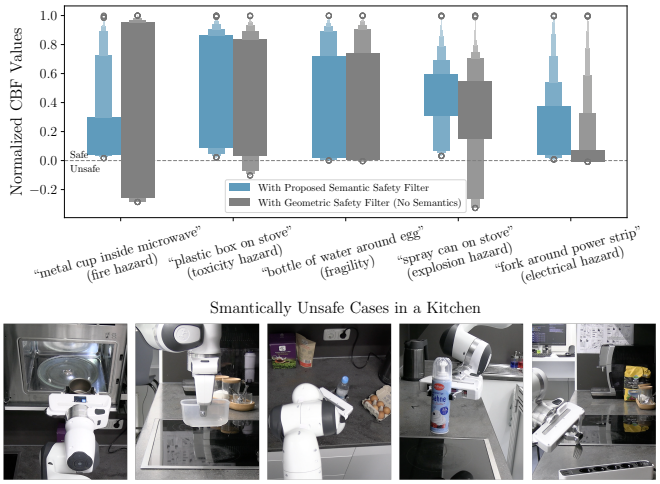

Fig. 3: A comparison of normalized CBF values for applying the proposed semantic safety filter *(top, blue plots)* versus the typical geometric safety filter *(top, grey plots)* to diffusion policies across five different scenarios *(bottom)*. The proposed semantic safety filter effectively addresses common sense constraints of different types, ranging from the considerations for fragile items to the prevention of fire and electrical hazards.

superquadric $h_{\text{sem},i}$ [9] to capture the set of points which the robot end effector should not enter. To account for the spatial relationship `above`, we extend the superquadric in the positive z-direction. We define similar superquadrics for relationships such as `under` and `around` (see Fig. 2).

2) Behavioral Constraints: The behavioral constraints are implemented using constraints on the time derivative of the CBF, i.e., the control invariance condition $\dot{\boldsymbol{h}}_{\text{sem}}(\boldsymbol{q}, \dot{\boldsymbol{q}}) \geq -\boldsymbol{\alpha}_{\text{sem}}(\boldsymbol{h}_{\text{sem}}(\boldsymbol{q}); \mathcal{S}_{\text{b}}(o); \mathcal{S}_{\text{r}}(o))$ [4]. Intuitively, the condition bounds how fast the robot system is allowed to approach the semantic safety boundary through the design of $\boldsymbol{\alpha}_{\text{sem}}$ and ensures that the constraints defined by $\boldsymbol{h}_{\text{sem}}$ are always satisfied. In particular, we design $\boldsymbol{\alpha}_{\text{sem}}$ to adhere to behavioral semantic constraints b_j from $\mathcal{S}_{\text{b}}(o)$ such that the system approaches the safe set boundary of the object with label l_j more slowly and exhibits the desired level of caution.

3) Pose Constraints: The pose constraint is active if $\mathcal{S}_{\text{T}}(o) = \{\texttt{constrained rotation}\}$. We use a softened pose constraint through the objective $\boldsymbol{w}_{\text{rot}}(\mathcal{S}_{\text{T}}(o))^{\text{T}} \boldsymbol{L}_{\text{rot}}(\boldsymbol{q}, \dot{\boldsymbol{q}})$, with $\boldsymbol{w}_{\text{rot}} > 0$ if $T = \texttt{constrained rotation}$ and $\boldsymbol{w}_{\text{rot}} = 0$ otherwise. The cost $\boldsymbol{L}_{\text{rot}}$ determines the difference between the predicted orientation at the next timestep and the desired orientation of the manipulator's end effector.

Given the semantic constraints $\mathcal{C}_{\text{sem}}$ and the set $\mathcal{S}$, our goal is to modify potentially unsafe end effector velocity commands sent by a human operator or from a motion policy. We convert the command to desired joint velocities $\dot{\boldsymbol{q}}_{\text{cmd}}$, and the semantic safety filter computes a certified input $\dot{\boldsymbol{q}}_{\text{cert}}$ that best matches the desired joint velocity $\dot{\boldsymbol{q}}_{\text{cmd}}$ while ensuring semantic and geometric constraint satisfaction:

$$\begin{aligned} \dot{\boldsymbol{q}}_{\text{cert}} = \underset{\dot{\boldsymbol{q}} \in \mathbb{U}}{\text{argmin}} \quad & \|\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}_{\text{cmd}}\|_2^2 + \boldsymbol{w}_{\text{rot}}(\mathcal{S}_{\text{T}}(o))^{\text{T}} \boldsymbol{L}_{\text{rot}}(\boldsymbol{q}, \dot{\boldsymbol{q}}) \\ \text{s.t.} \quad & \dot{\boldsymbol{h}}_{\text{sem}}(\boldsymbol{q}, \dot{\boldsymbol{q}}; \mathcal{S}_{\text{r}}(o)) \geq -\boldsymbol{\alpha}_{\text{sem}}(\boldsymbol{h}_{\text{sem}}(\boldsymbol{q}); \mathcal{S}_{\text{b}}(o)) \\ & \dot{\boldsymbol{h}}_{\text{geo}}(\boldsymbol{q}, \dot{\boldsymbol{q}}) \geq -\boldsymbol{\alpha}_{\text{geo}}(\boldsymbol{h}_{\text{geo}}(\boldsymbol{q})), \end{aligned}$$

where we added environment, collision-avoidance, joint angle, and velocity constraints through additional CBFs $\boldsymbol{h}_{\text{geo}}(\boldsymbol{q})$ and compact polyhedral input constraints $\mathbb{U}$. The above optimization problem is convex and can be efficiently solved online.

IV. EXPERIMENTS

In our real-world experiments, a robotic manipulator is deployed with our proposed semantic safety filter in closed-loop to prevent potentially unsafe commands from a learned motion policy. To demonstrate the applicability of our proposed filter, we conducted experiments in a real-world kitchen environment and trained diffusion policies for five different transportation tasks involving various semantically unsafe constraints. These constraints include handling fragile items and preventing fire and electrical hazards. Clips of this set of experiments are included in the supplementary video. Fig. 3 compares the normalized CBFs for our proposed semantic safety filter and a nominal geometric safety filter that does not account for semantic constraints. The plot shows that the proposed semantic safety filter successfully prevents unsafe actions, such as placing a metal cup inside a microwave or putting a pressurized spray can on a stove. This set of experiments highlights the generalizability of our proposed approach and its applicability in real-world settings.

V. CONCLUSION

We propose a semantic safety filter framework that combines scene understanding and LLM's reasoning capabilities with CBF-based safe control. This framework ensures adherence to "common sense" constraints not visible in 3D maps, while guaranteeing collision-free motion and robot-specific safety. Demonstrated in real-world tasks, our work emphasizes the importance of semantic understanding for achieving human-like safety beyond collision avoidance.

REFERENCES

[1] K.-C. Hsu, H. Hu, and J. F. Fisac, "The safety filter: A unified view of safety-critical control in autonomous systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2023.

[2] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[3] K. P. Wabersich, A. J. Taylor, J. J. Choi, K. Sreenath, C. J. Tomlin, A. D. Ames, and M. N. Zeilinger, "Data-driven safety filters: Hamilton-Jacobi reachability, control barrier functions, and predictive methods for uncertain systems," *IEEE Control Systems Magazine*, vol. 43, no. 5, pp. 137–177, 2023.

[4] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *Proc. of the European Control Conf. (ECC)*, 2019, pp. 3420–3431.

[5] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, *et al.*, "ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[6] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "OpenMask3D: Open-Vocabulary 3D Instance Segmentation," in *Proc. of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," in *Proc. of the Advances in Neural Info. Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[8] A. Singletary, W. Guffey, T. G. Molnar, R. Sinnet, and A. D. Ames, "Safety-critical manipulation for collision-free food preparation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 954–10 961, 2022.

[9] W. Liu, Y. Wu, S. Ruan, and G. S. Chirikjian, "Robust and accurate superquadric recovery: A probabilistic approach," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2676–2685.