From Semantic Understanding to Geometric Features: Using Foundation Models for Novel Robotic Tasks

Technion - Autonomous system program

Nizan Mashall, Erez Karpas, Miriam Zacksenhouse

Goal | Household robotics

Setup of the near future:

- Pre-trained with fundamental capabilities (walking, grasping)
- Integrated sensing capabilities (e.g., cameras)
- **Perform new tasks autonomously**

Household Robot

Novel Tasks | Learning-based methods

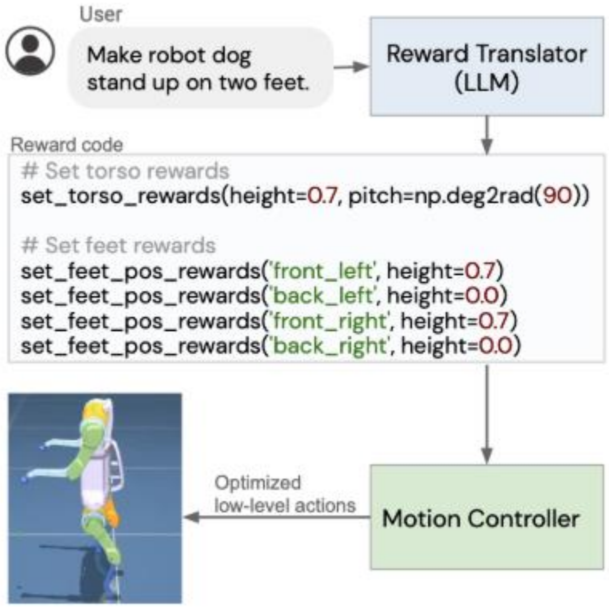
- **Require extensive training through:**
  - Demonstration videos
  - Teleoperation records

- **Have limited ability to generalize across:**
  - New tasks
  - Different tools
  - Various environments

Novel Tasks | Model-Based Planning

Model-based methods encounter two main challenges in automating novel tasks:

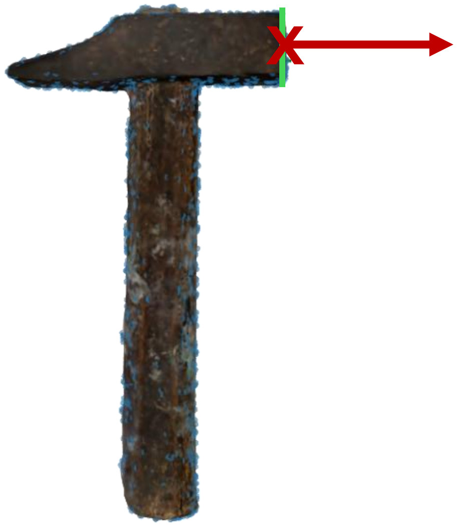
- Automating plan generation
- Automating reference frame assignment


User
Make robot dog stand up on two feet.
Reward Translator (LLM)
Reward code
Optimized low-level actions
Motion Controller

```
# Set torso rewards
set_torso_rewards(height=0.7, pitch=np.deg2rad(90))

# Set feet rewards
set_feet_pos_rewards('front_left', height=0.7)
set_feet_pos_rewards('back_left', height=0.0)
set_feet_pos_rewards('front_right', height=0.7)
set_feet_pos_rewards('back_right', height=0.0)
```

Yu, Wenhao, et al. "Language to rewards for robotic skill synthesis." *arXiv preprint arXiv:2306.08647* (2023).



Novel Tasks | Model-Based Planning Example

Input(Point, Direction):

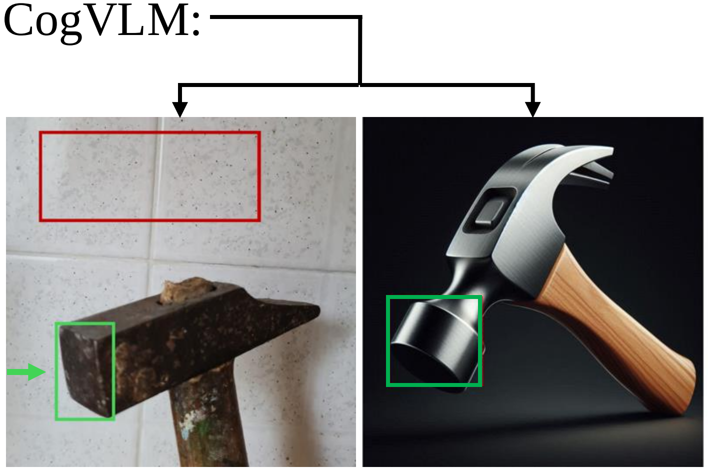
- Colinear
- Coincident

We designed a fully automated system for 3D geometric feature detection and reference frame assignment

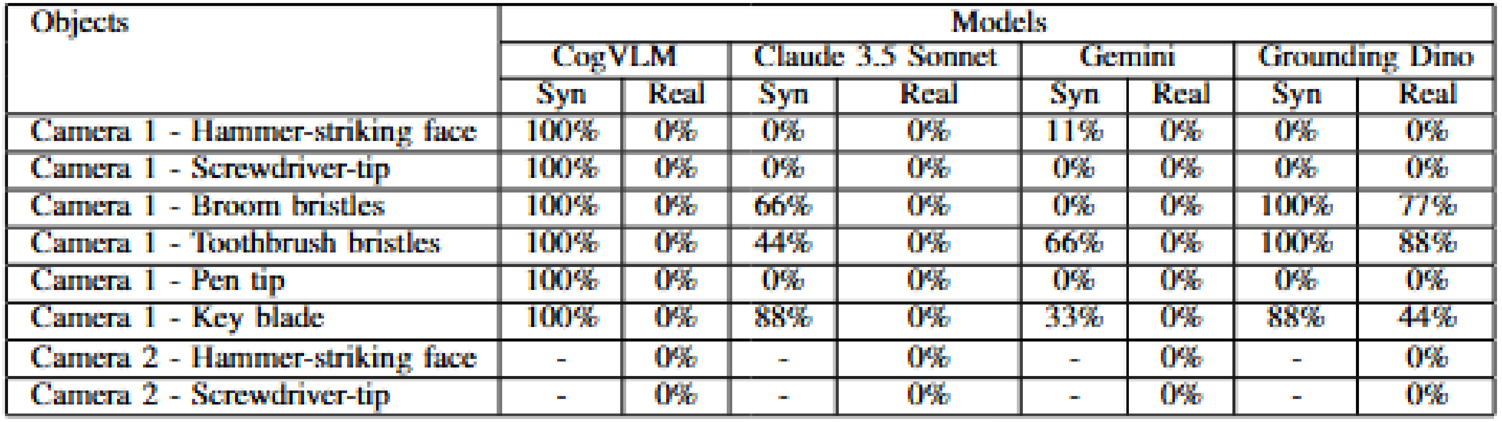
Key Insights | VLM Limitations

- VLMs can perform 2D affordance grounding
- However, there is a dramatic performance gap between real and synthetic images

User: “Using your visual understanding capabilities, locate the hammer's main impact surface.”
CogVLM:



ground truth



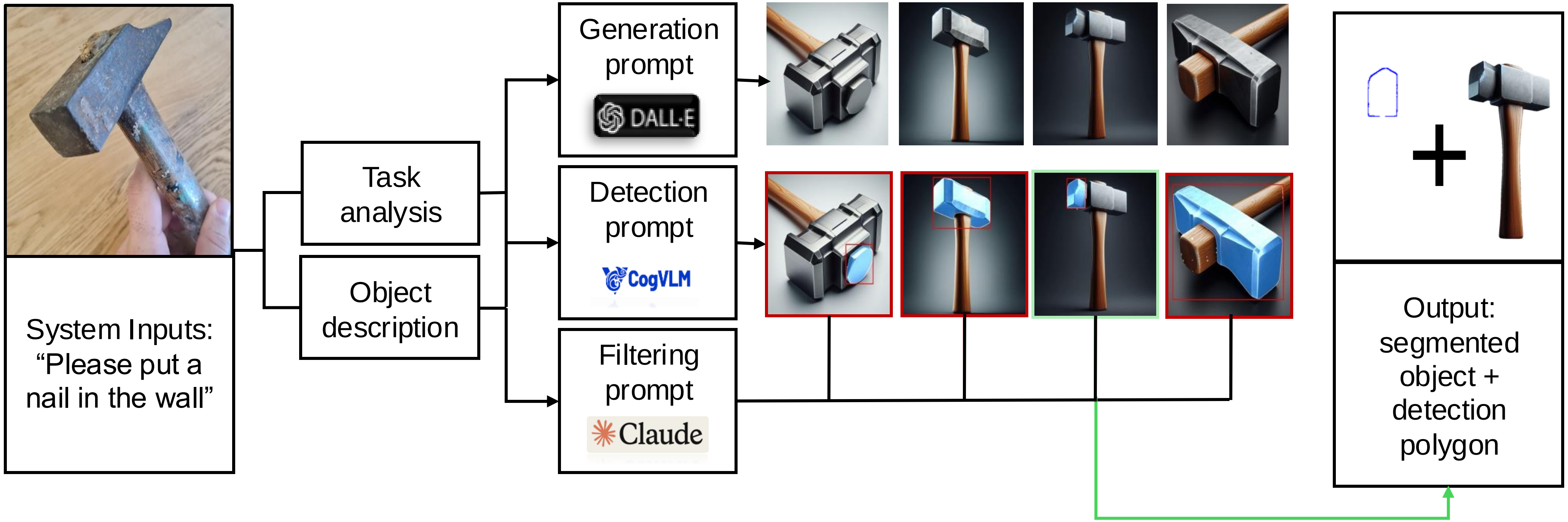
Objects	Models							
	CogVLM		Claude 3.5 Sonnet		Gemini		Grounding Dino	
	Syn	Real	Syn	Real	Syn	Real	Syn	Real
Camera 1 - Hammer-striking face	100%	0%	0%	0%	11%	0%	0%	0%
Camera 1 - Screwdriver-tip	100%	0%	0%	0%	0%	0%	0%	0%
Camera 1 - Broom bristles	100%	0%	66%	0%	0%	0%	100%	77%
Camera 1 - Toothbrush bristles	100%	0%	44%	0%	66%	0%	100%	88%
Camera 1 - Pen tip	100%	0%	0%	0%	0%	0%	0%	0%
Camera 1 - Key blade	100%	0%	88%	0%	33%	0%	88%	44%
Camera 2 - Hammer-striking face	-	0%	-	0%	-	0%	-	0%
Camera 2 - Screwdriver-tip	-	0%	-	0%	-	0%	-	0%

Key Insights | Common Hand Tool Similarity

- Hand tools share common geometric patterns across different variants
- This similarity enables geometric feature detection to be transferred across variants

Method | Task Analysis and Detection over Digital Twin

Detecting the geometric feature in digital twin



System Inputs: “Please put a nail in the wall”

Task analysis

Object description

Generation prompt

DALL·E

Detection prompt

CogVLM

Filtering prompt

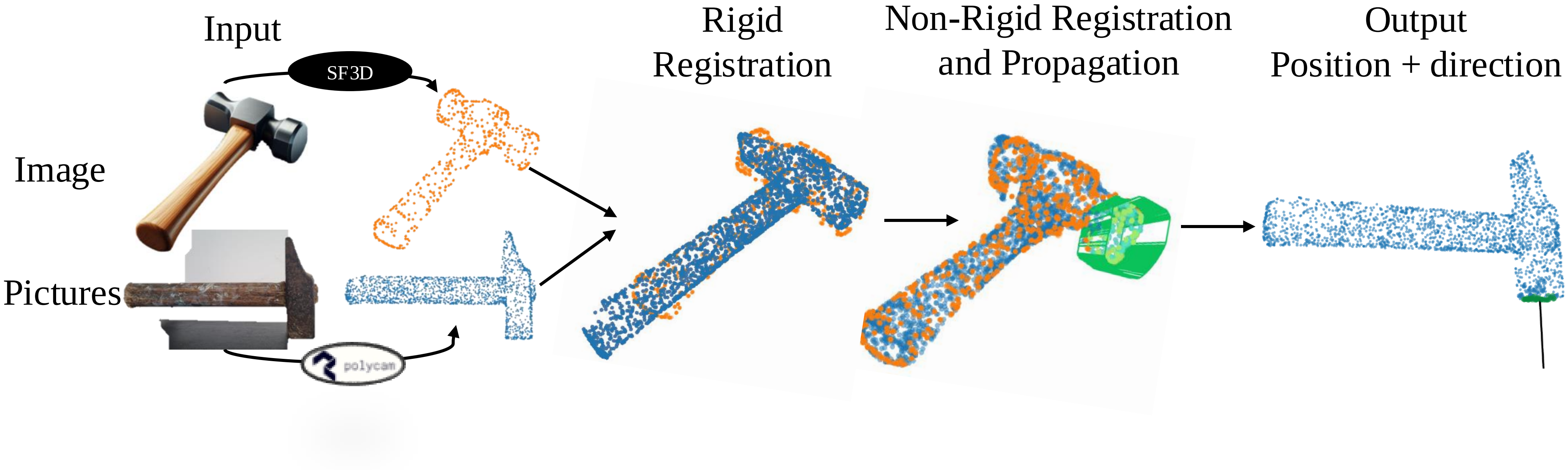
Claude

+

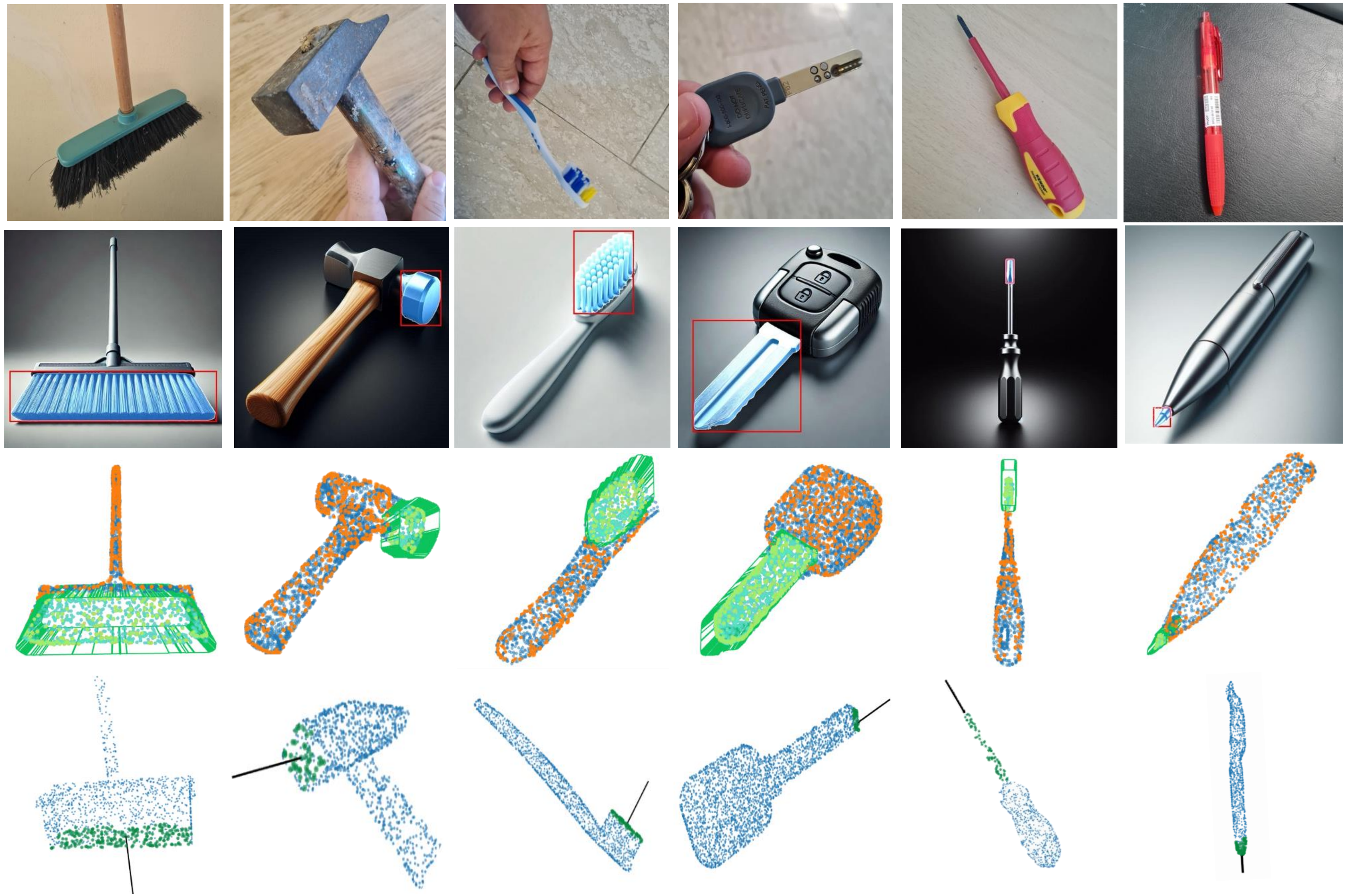
Output: segmented object + detection polygon

Method | Registration of Digital Twin with Real Tool

Transferring the detected geometric feature to the real object

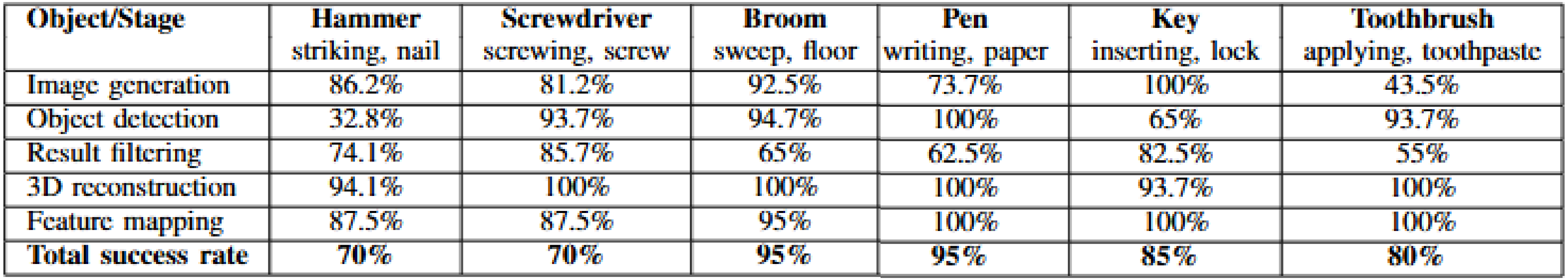


Method | Results



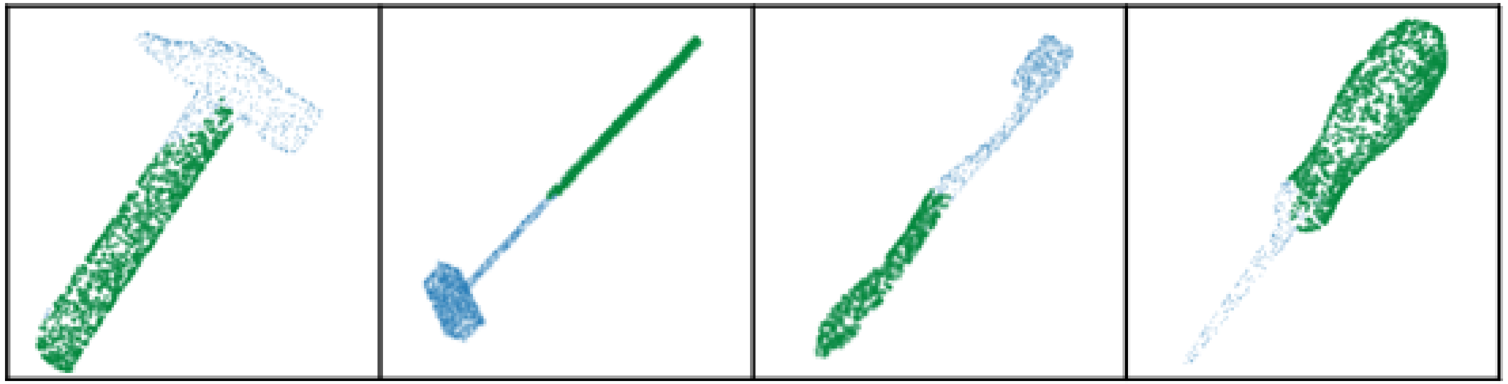
Method | Evaluation

Summary table of the results



Object/Stage	Hammer striking, nail	Screwdriver screwing, screw	Broom sweep, floor	Pen writing, paper	Key inserting, lock	Toothbrush applying, toothpaste
Image generation	86.2%	81.2%	92.5%	73.7%	100%	43.5%
Object detection	32.8%	93.7%	94.7%	100%	65%	93.7%
Result filtering	74.1%	85.7%	65%	62.5%	82.5%	55%
3D reconstruction	94.1%	100%	100%	100%	93.7%	100%
Feature mapping	87.5%	87.5%	95%	100%	100%	100%
Total success rate	**70%**	**70%**	**95%**	**95%**	**85%**	**80%**

Additional Usage



Main Contributions

- Zero-shot 3D detection
- Generalization across diverse objects and tasks
- No training or demonstration required

From Semantic Understanding to Geometric Features: Using Foundation Models for Novel Robotic Tasks

Thank You!

Questions?

Nizan Mashall

Erez Karpas

Miriam Zacksenhouse