

GRIM: Task-Oriented Grasping with Conditioning on Generative Examples

Shailesh¹, Alok Raj¹, Nayan Kumar¹, Priya Shukla²,
Andrew Melnik³, Michael Beetz³, Gora Chand Nandi²

¹IIT Dhanbad, India

²IIT Allahabad, India

³University of Bremen, Germany

Abstract

Task-Oriented Grasping (TOG) requires a robot to select grasps that are not only stable but also functionally appropriate for a given task. This presents a significant challenge, demanding a nuanced understanding of task semantics, object affordances, and functional constraints. Existing learning-based approaches often struggle with generalization due to the scarcity of large-scale, task-annotated grasp datasets. To overcome these limitations, we introduce GRIM (Grasp Re-alignment via Iterative Matching), a novel, training-free framework for TOG. GRIM operates on a retrieve, align, and transfer paradigm. It first queries a memory of object-task examples, built from diverse sources including generative AI, web images, and human demonstrations. Given a new scene object, GRIM retrieves a semantically similar example and aligns its 3D geometry to the scene object using a robust coarse-to-fine strategy. This alignment is guided by a combination of geometric cues and a semantic similarity score over dense DINO features. Finally, the task-oriented grasp from the memory instance is transferred to the scene object and refined against a set of geometrically stable grasps to ensure task compatibility and physical feasibility. By eschewing task-specific training, GRIM demonstrates strong generalization, achieving state-of-the-art performance on benchmark datasets with only a small number of conditioning examples.

Code, Dataset & Appendix — <https://grim-tog.github.io/>

Introduction

The ability for robots to physically interact with the world is fundamental to their utility. While grasp synthesis has made significant strides in achieving geometric stability, true manipulation intelligence lies in selecting grasps that are functionally suitable for a specific goal. This problem, known as Task-Oriented Grasping (TOG), moves beyond the question of “Can I pick this up?” to “How should I pick this up to complete task X?”. For example, a hammer must be grasped by its handle to be used for hammering, not by its head. This requires a deep understanding of object affordances, task semantics, and the functional constraints they impose (2020). A primary bottleneck for progress in TOG is the data-scarcity problem. Supervised learning methods (Murali

et al. 2020; Tang et al. 2023a) are powerful but depend on large, manually annotated datasets that specify which grasps are suitable for which tasks. Creating such datasets is labor-intensive and scales poorly, limiting the ability of these models to generalize to novel objects and tasks not seen during training.

To address these challenges, we propose **GRIM** (Grasp Re-alignment via Iterative Matching), a novel **training-free** framework that leverages the power of pre-trained foundation models in a retrieve-align-transfer pipeline (Kuang et al. 2024; Di Palo and Johns 2024). Our approach follows the pipeline shown in Figure 1. Since GRIM learns grasping from generated video and the underlying video generation models are too computationally heavy for real-time execution, we create and successfully leverage memory option for grasp retrieval. Instead of training a model on a fixed dataset, GRIM builds a dynamic, evergreen memory (Figure 2) of object-task interactions from diverse and easily accessible sources: synthetic data from generative models, in-the-wild images from the web, and on-demand human demonstrations. This diverse memory provides a rich source of functional priors, inspired by the cognitive concept of the world as an external memory (Melnik et al. 2018).

When faced with a new object and task, GRIM’s workflow is as follows (Figure 3):

1. **Retrieve:** It queries its memory to find the most relevant prior experience, using a joint similarity metric that considers both the visual appearance of the object (via DINO embeddings (Oquab et al. 2024)) and the semantics of the task description (via CLIP embeddings (Radford et al. 2021)).
2. **Align:** It robustly aligns the 3D point cloud of the retrieved memory object with the scene object. This is a key contribution, employing a coarse-to-fine strategy that first uses PCA-reduced DINO features for a semantically-aware coarse alignment, followed by a precise ICP (Besl and McKay 1992) refinement.
3. **Transfer & Refine:** The task-specific grasp pose from the memory instance is transferred to the aligned scene object. This transferred pose then serves as a powerful prior to select and refine the best grasp from a set of pre-computed, geometrically stable candidates for the scene object.

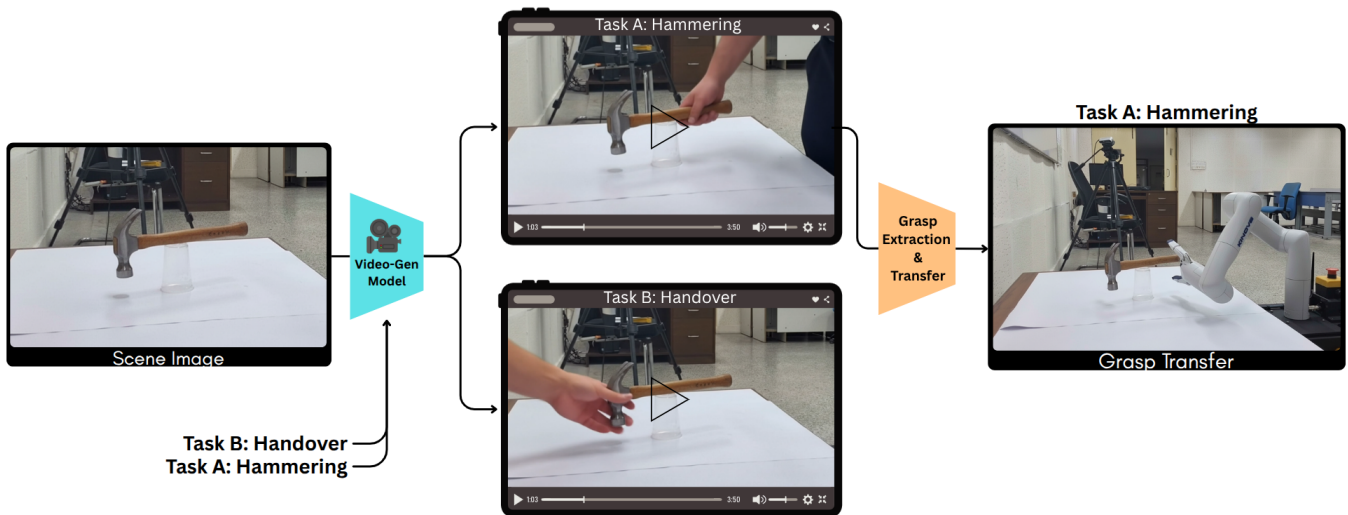


Figure 1: The GRIM framework for task-oriented grasp synthesis. From a single scene image, the Video-Gen model generates task-specific video examples, such as hammering (Task A) and handover (Task B). Grasps are extracted from these generated videos and then transferred to a robotic arm to execute the specified task in the real world as shown for hammering (Task A).

Our main contributions are:

- A flexible and scalable memory construction pipeline that integrates object-task experiences from diverse sources, including a novel application of generative AI, circumventing the need for manually annotated datasets.
- A novel 3D alignment strategy that prioritizes semantic correspondence over geometric shape. By matching dense DINO features, our method works effectively even with sparse, partial point clouds where traditional geometry-based alignment techniques often fail.
- A complete training-free framework that shows generalization to both novel objects and novel tasks, validated through extensive experiments and real-world robot demonstrations.

Related Work

Task-Oriented Grasping (TOG) research has evolved from analytical methods to data-driven techniques, with a recent shift towards leveraging large-scale pre-trained models.

Data-Driven Approaches

Early data-driven methods learned direct mappings from object classes and tasks to grasps (Dang and Allen 2012; Liu, Daruna, and Chernova 2019). However, these approaches often lacked semantic understanding and struggled to generalize (Tang et al. 2023a). To inject semantic knowledge, subsequent works utilized knowledge bases (KBs) and probabilistic logic (Song et al. 2010; Huang et al. 2022; Liu et al. 2023; Ardón et al. 2019; Zese et al. 2014), but these systems often require significant engineering to construct and scale the KBs.

The release of the TaskGrasp dataset by Murali et al. (2020) was a significant step, enabling methods like GCN-Grasp which uses a Graph Convolutional Network. How-

ever, such methods are inherently limited by the contents of their training data and knowledge graph, struggling to generalize to concepts unseen during training. More recent works like GraspGPT (Tang et al. 2023a) and GraspMolmo (Deshpande et al. 2025) leverage Large Language Models (LLMs) and Vision-Language Models (VLMs) to incorporate open-world knowledge, improving generalization (Mikami et al. 2024). Nevertheless, these models still rely on a foundational training phase on task-specific datasets (Tang et al. 2023b; Jin et al. 2024; Nguyen et al. 2023), inheriting the associated data acquisition bottleneck.

GRIM diverges fundamentally from these paradigms. It is entirely training-free, obviating the need for task-specific grasp annotations. By dynamically building a memory from heterogeneous data, it directly tackles the data scarcity and annotation challenges that constrain prior methods.

Training-Free and Retrieval-Based Approaches

The advent of powerful foundation models has spurred the development of training-free TOG methods. Many approaches use LLMs or VLMs to provide semantic guidance, mapping a language command to a region on an object where a grasp should be executed (Rashid et al. 2023; Mirjalili et al. 2024; Li et al. 2024). While these methods avoid training, they typically only provide coarse spatial priors (e.g., “grasp the handle”), not directly executable 6D grasp poses.

Closer to our work are retrieval-based methods. RTA-Grasp (Dong et al. 2024) also proposes a training-free approach using a memory of human demonstrations. It retrieves a relevant video and uses 2D feature matching to transfer a grasp point. While effective, its reliance on 2D matching can be ambiguous and less robust to viewpoint changes. RoboABC (Ju et al. 2024) uses CLIP to retrieve contact points but struggles to determine the full 6D grasp

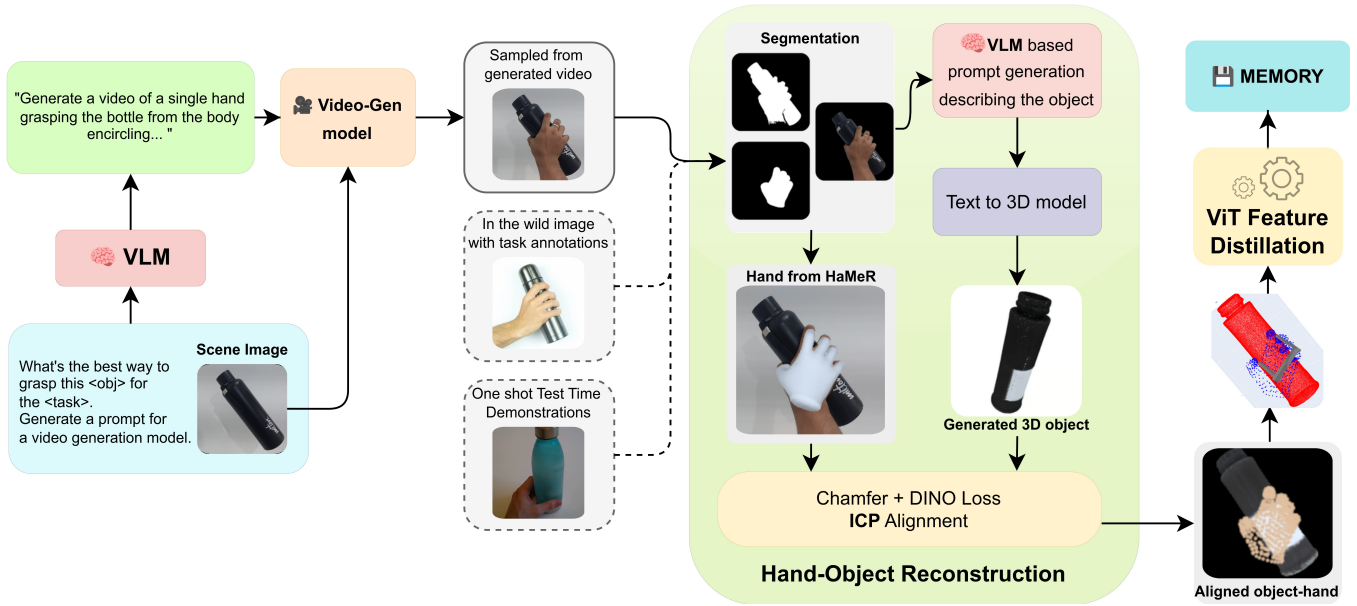


Figure 2: Our memory creation pipeline. A diverse set of inputs (AI-generated video frames, web images, human demonstrations) are processed by a hand-object reconstruction module (Wu et al. 2024). This yields an object mesh and a corresponding task-oriented grasp pose. We enrich the object mesh with dense DINO features to create a feature mesh, which is stored in memory alongside the task label and grasp pose.

pose, particularly the crucial grasp orientation.

GRIM builds upon the strengths of retrieval but makes several key improvements. Our retrieval is guided by a joint 3D visual (DINO (Oquab et al. 2024)) and task-semantic (CLIP (Radford et al. 2021)) similarity. Crucially, we introduce a robust, semantically-aware 3D alignment strategy that aligns entire object point clouds, not just 2D features (Di Palo and Johns 2024). This allows for a more precise transfer of the full 6D grasp pose, which is then further refined against the scene object’s specific geometry. This holistic process addresses both “where” and “how” to grasp with high precision and adaptability, without the limitations of pre-defined datasets or explicit training.

Methodology

We introduce GRIM (Grasp Re-alignment via Iterative Matching), a training-free framework for TOG. Our approach follows a retrieve-align-transfer pipeline, detailed below.

Memory Creation

To generalize to novel scenes, we construct a memory \mathcal{M} of object-task experiences from diverse data sources. Each instance in \mathcal{M} is a tuple (F_M, G_t, T, O) , containing the object’s feature mesh F_M , a 6D task-oriented grasp pose G_t , the corresponding task description T , and the object name O .

The pipeline to create a single memory instance (Figure 2) begins with an image or video frame I_{HO} depicting a functional grasp. We use a hand-object reconstruction model (Wu et al. 2024) to extract the object mesh and hand mesh.

We then derive a 6D parallel-jaw gripper pose G_t from the hand mesh. This conversion is done by first identifying the centroids of hand segments: the thumb, the combined index and middle fingers, and the palm. The gripper’s center (translation) is defined as the midpoint between the centroid of the thumb and the combined centroid of the opposing fingers. The vector connecting these centroids establishes the closing direction, and the palm’s centroid provides a reference point to determine the approach vector.

To create the feature mesh F_M , we sample points from the object mesh and compute a dense DINOv2 feature vector for each point, effectively creating a semantic descriptor field on the object’s surface, similar to Wang et al. (2023) and PS et al. (2024).

While GRIM primarily learns from AI-generated videos, our pipeline is flexible and can incorporate additional data sources as well:

AI-Generated Videos: To create a scalable and diverse data source, we leverage generative AI (Melnik et al. 2024). For an object and task from a source like TaskGrasp (Murali et al. 2020), we prompt a VLM (Gemini Pro) to generate a detailed textual description of a video showing the correct grasp. This description, along with a starting image frame, is then used as a prompt for a video generation model such as VEO2 (Google 2025) to generate a short video. We sample a frame from this video to serve as I_{HO} . This process allows for cheap, large-scale creation of functionally-grounded grasp data.

In-the-Wild Web Images: We use images scraped from the web that show human grasping actions. For each image, we use a VLM to generate a plausible task description T .

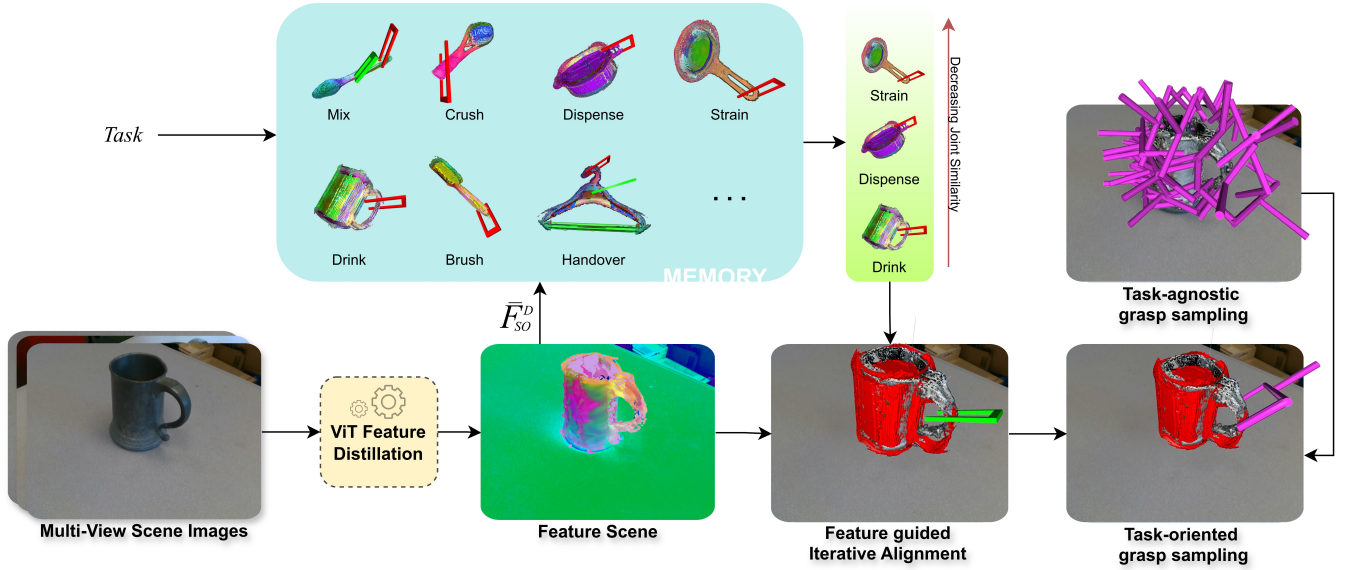


Figure 3: The GRIM pipeline for a given scene object and task. **(1) Retrieval:** The system queries its memory using joint visual and task similarity to find the best matching prior experience (a cup for the task ‘drink’). **(2) Alignment:** The retrieved memory object (red point cloud) is aligned with the scene object (grey point cloud) using our feature-guided iterative alignment. The colors on the objects represent PCA-reduced DINO features, showing semantic correspondence. **(3) Transfer & Refine:** The grasp from the memory object is transferred to the scene object and used to select the best among a set of task-agnostic, stable grasp candidates (cluster of purple grasps), resulting in the final task-oriented grasp (single purple grasp).

Test-Time Expert Demonstrations: Our framework supports lifelong learning. If the robot fails on a task, a human can provide a single-image demonstration, which is seamlessly processed and added to the memory \mathcal{M} , improving future performance on similar tasks (Malato et al. 2024).

Memory Retrieval

Given a novel scene containing a target object (represented by its point cloud P_{SO} with per-point DINO features F_{SO}^D) and a task command T_S , we retrieve the most relevant memory instance.

First, we compute a global visual descriptor \bar{F}_{SO}^D for the scene object by averaging its per-point DINO features. We encode the task command T_S into a text embedding E_{T_S} using CLIP’s text encoder.

For each memory instance $i \in \mathcal{M}$ with its global object descriptor $\bar{F}_{MO,i}^D$ and task embedding $E_{T_{M,i}}$, we compute a joint similarity score:

$$S_{\text{joint}}(i) = \alpha \cdot \text{sim}_{\text{cos}}(\bar{F}_{SO}^D, \bar{F}_{MO,i}^D) + (1 - \alpha) \cdot \text{sim}_{\text{cos}}(E_{T_S}, E_{T_{M,i}}) \quad (1)$$

where $\text{sim}_{\text{cos}}(\cdot, \cdot)$ is the cosine similarity and α is a hyperparameter balancing visual and task similarity (we use $\alpha = 0.5$). The memory instance (F_M^*, G_t^*, T^*, O^*) with the highest S_{joint} is selected for the next stage.

Semantic 3D Alignment

After retrieving a memory object (source point cloud P_{MO} , DINO features F_{MO}^D), we must align it to the scene object

(target point cloud P_{SO} , features F_{SO}^D). A purely geometric alignment like standard ICP would fail if the objects have different shapes (e.g., aligning a metal spatula to a plastic one). We therefore propose a coarse-to-fine alignment strategy guided by semantic features.

Coarse Alignment: To reduce the dimensionality and noise of the DINO features, we apply PCA, projecting both F_{MO}^D and F_{SO}^D into a lower 4-dimensional space ($D_{\text{PCA}} = 4$). We then perform a grid search over a discretized set of initial rotations to find a promising coarse alignment. Specifically, we sample 8 steps for each of the three Euler angles (roll, pitch, yaw), resulting in $8^3 = 512$ candidate rotations $\{R_i\}$. For each candidate, we compute a transformation $T_{\text{init},i}$ that aligns the point cloud centroids and applies the rotation. The quality of this initial transformation is evaluated using a joint feature-geometric score. For each point in the transformed source cloud, we find its $K = 3$ nearest neighbors in the target cloud and compute a cost based on a weighted sum of the squared Euclidean distance ($w_g = 10$) and the feature dissimilarity ($w_f = 100$). By heavily weighting the feature component, we prioritize finding a semantically meaningful alignment over a purely geometric one. The top 10 transformations with the lowest cost are selected as candidates for the fine refinement stage.

Fine Refinement: The best coarse alignment is then used to initialize the Iterative Closest Point (ICP) algorithm. This standard ICP step refines the alignment to be geometrically precise. This two-step process, where semantics guide the initial guess and geometry refines it, allows for robust alignment even between objects that are semantically similar but

geometrically distinct. The final output is a transformation T_{final} that maps points from the memory object’s coordinate frame to the scene object’s frame.

Grasp Transfer and Refinement

With the alignment T_{final} , we transfer the task-oriented grasp G_M from memory to the scene object: $G_S = T_{\text{final}} \cdot G_M$. However, due to small alignment errors or geometric differences, G_S may not be perfectly stable or executable.

To find an optimal, executable pose, we follow a sample-and-refine strategy inspired by Dong et al. (2024). First, we use a task-agnostic grasp sampler, AnyGrasp (Fang et al. 2023), to generate a set of N geometrically stable grasp candidates $\{G_{A,i}\}_{i=1}^N$ on the scene object, each with a geometric quality score $S_{\text{geo},i}$.

We then re-rank these candidates based on their compatibility with our transferred task-oriented grasp $G_S = (R_S, \mathbf{t}_S)$. We define a task-compatibility score $S_{\text{task},i}$ for each candidate grasp $G_{A,i} = (R_{A,i}, \mathbf{t}_{A,i})$:

$$S_{\text{task},i} = \underbrace{(\mathbf{v}_{\text{target}} \cdot \mathbf{v}_{A,i})}_{\text{Orientation Sim.}} + \underbrace{\exp\left(-\frac{\|\mathbf{t}_{A,i} - \mathbf{t}_S\|^2}{2\sigma^2}\right)}_{\text{Position Sim.}} \quad (2)$$

where $\mathbf{v}_{\text{target}}$ and $\mathbf{v}_{A,i}$ are the approach vectors of the grasps (e.g., the z-axis of the gripper frame), and σ is a bandwidth parameter (set to 0.02m). This score rewards candidates that are close in both position and orientation to the transferred task-centric pose.

The final score for each candidate is a weighted sum of its task compatibility and geometric quality:

$$S_i = w_{\text{task}} S_{\text{task},i} + w_{\text{geo}} S_{\text{geo},i} \quad (3)$$

We heavily prioritize task-compatibility by setting $w_{\text{task}} = 0.95$ and $w_{\text{geo}} = 0.05$, as AnyGrasp already ensures candidates have high geometric quality. The grasp candidate G_A^* with the highest final score S_i is selected for execution.

Experiments and Results

We conduct extensive experiments to evaluate GRIM’s performance, focusing on its ability to generalize to novel objects and tasks.

Experimental Setup

Baselines: We compare GRIM against three representative baselines:

- **Random:** A task-agnostic baseline that randomly selects a geometrically stable grasp from the candidates provided by AnyGrasp.
- **RTAGrasp (Dong et al. 2024):** A state-of-the-art training-free method that uses 2D feature matching to transfer grasps from a video memory.
- **GraspMolmo (Deshpande et al. 2025):** A state-of-the-art learning-based VLM, which was fine-tuned on a mixture of its primary synthetic dataset (PRISM, 379k examples) and a portion of the TaskGrasp.

Dataset: We evaluate all methods on the TaskGrasp dataset (Murali et al. 2020), which provides object point clouds and annotated task-oriented grasps. To rigorously test generalization, we use two challenging splits:

- **Held-out Objects:** The memory contains no objects of the same category as the test object.
- **Held-out Tasks:** The memory contains no examples of the task being performed, even if it has seen the object category before.

Memory: Our memory for GRIM and RTAGrasp is constructed from a combination of 180 AI-generated video frames, 15 web images, and 15 human demonstrations, totaling 210 instances. This small size highlights the data efficiency of our approach. To ensure a fair comparison with RTAGrasp, we build its memory from the same source images and derive its required 2D grasp points from our 6D poses.

Evaluation Metric: Following standard practice, we evaluate the methods on their ability to identify the correct task-oriented grasps from a set of proposals. We use the 25 annotated grasps for each object instance in TaskGrasp as candidates. A predicted grasp is considered correct if it is one of the positive examples for the given task. We report the Mean Average Precision (mAP) over all object-task pairs. Since, these 25 grasp poses doesn’t have any geometric quality score, we set $w_{\text{geo}} = 0.0$ for final score calculation.

Quantitative Results

GRIM’s effectiveness and data efficiency are demonstrated in our quantitative evaluations (Table 2). On the full TaskGrasp dataset, GRIM achieves a Mean Average Precision (mAP) of 0.67. This result not only surpasses the state-of-the-art training-free method, RTAGrasp (0.58), but also, remarkably, outperforms GraspMolmo (0.62). This comparison is particularly significant: GraspMolmo is a powerful VLM trained on a massive dataset of 379,000 synthetic task-oriented grasp examples, whereas GRIM’s memory contains only 210 instances from heterogeneous, un-curated sources. This result strongly validates our central thesis: by effectively retrieving and re-aligning functional priors from a small but diverse memory, it is possible to achieve superior generalization without relying on vast, expensive, and potentially biased training datasets. Additionally, Table 1 shows a more granular insight for a few object categories from the dataset.

Furthermore, GRIM’s performance advantage is most pronounced in the challenging generalization splits. On held-out objects and tasks, GRIM’s mAP degrades by only 3%, whereas RTAGrasp’s performance drops by over 10%. This underscores the robustness of our 3D semantic alignment strategy, which successfully transfers functional knowledge even without direct categorical or task precedents—a scenario where 2D feature matching proves less effective.

To understand why our method works well, we tested it without its key parts in an ablation study (Table 3). The results clearly show that semantic alignment is the most

Method	Novel Instances								
	Paint roller	Brush	Tongs	Strainer	Frying Pan	Fork	Mortar	Ice Scrapper	Pizza Cutter
Random	0.30	0.66	0.23	0.24	0.32	0.26	0.31	0.60	0.50
RTAGrasp	0.39	0.93	0.28	0.55	0.42	0.35	0.37	0.91	0.57
GraspMolmo	0.56	0.73	0.55	0.44	0.46	0.53	0.66	0.65	0.58
GRIM(Ours)	0.89	0.90	0.58	0.58	0.60	0.76	0.72	0.71	0.92

Table 1: Per-category Average Precision on novel object instances from the TaskGrasp dataset.

critical component. Without it (GRIM w/o Semantic Alignment), performance drops to 0.50 mAP, which is nearly as poor as the random baseline. This confirms that using features is crucial for aligning objects for a task, especially when their shapes differ. The grasp refinement step is also important. Without it (GRIM w/o Grasp Refinement), performance falls to 0.59 mAP. This means the transferred grasp is a good starting point for the task, but it must be fine-tuned to the scene object’s geometry to be successful. In summary, both components are vital: semantic alignment provides the correct functional idea, and refinement makes that idea physically work.

A qualitative analysis further illuminates the behavior of the semantic alignment module, particularly its failure modes. Its performance is intrinsically linked to the fidelity of the input point cloud. In scenarios with severe sensor noise or extreme sparsity, the process of establishing dense feature correspondences can break down. This corrupts geometric priors like the centroid and leads to a flawed coarse alignment from which the local ICP refinement cannot recover. The final transferred grasp is consequently misplaced and functionally irrelevant. This underscores a key dependency: while GRIM is robust to partial views, its ability to reason functionally is contingent on receiving a partial point cloud of sufficient quality to support the crucial semantic alignment stage.

Method	All Data	Held-out Obj.	Held-out Tasks
Random	0.49	0.41	0.43
RTAGrasp	0.58	0.52	0.51
GraspMolmo	0.62	0.57	0.55
GRIM (Ours)	0.67	0.65	0.64

Table 2: Mean Average Precision (mAP) on the TaskGrasp dataset. GRIM consistently outperforms all baselines, with particularly strong performance on the held-out splits, demonstrating superior generalization.

Real-World Robot Validation

To demonstrate the practical applicability of GRIM, we deployed it on a Kinova Gen3 Lite manipulator. The scene is captured by two RGB-D cameras. We used the same 210-instance memory from our simulation experiments, containing no instances of the test objects. We evaluated

Configuration	mAP (All Data)
<i>Ablations:</i>	
GRIM w/o Semantic Alignment	0.50
GRIM w/o Grasp Refinement	0.59
GRIM (Full Model)	0.67

Table 3: Ablation study of GRIM’s key components. Results are reported as Mean Average Precision (mAP) on the full TaskGrasp dataset, demonstrating the critical role of both semantic alignment and grasp refinement.

GRIM on 5 novel objects with associated tasks: a mallet (task: ‘hammering’), a kettle (task: ‘pour’), a spray bottle (task: ‘spray’), an aerosol-can (task: ‘spray’), and a spoon (task: ‘scoop’). For each object-task pair, we performed 10 trials. GRIM achieved a high success rate, successfully executing the task-oriented grasp in 39 out of 50 trials. Failures were not due to flawed grasp selection but were instead traced to perception errors; specifically, noise in point cloud reconstruction and calibration inaccuracies were able to disrupt the subsequent 3D alignment stage. Figure 4 shows qualitative examples of successful executions.

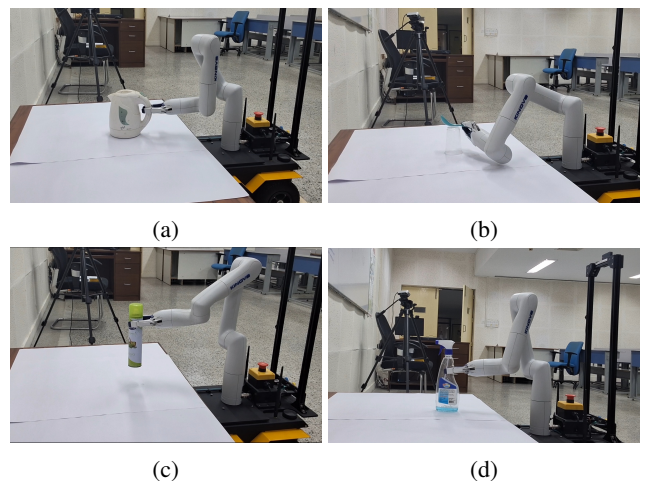


Figure 4: **Real-world deployment of GRIM with novel objects.** The system correctly plans and executes task-oriented grasps. The Kinova Gen3 Lite robot successfully executing the planned grasp.

Limitations

As a training-free framework reliant on upstream pre-trained models such as Gemini-Pro, Veo 2, Genie, SAM, and hand-object reconstruction models, it is susceptible to hallucinations, low-quality outputs, or biases inherited from these models, potentially affecting grasp accuracy in edge cases. Additionally, while online inference is efficient at 10 seconds, the offline memory creation incurs a one-time cost of 7 minutes per item, which may limit scalability for very large memories. Future work could address these by incorporating robustness checks and expanding real-world benchmarks.

Conclusion

We have presented GRIM, a training-free framework for task-oriented grasping that demonstrates remarkable generalization capabilities by retrieving and re-aligning functional priors from a diverse memory. Our key innovation is a robust 3D alignment process guided by semantic features, which allows for effective knowledge transfer between objects that are functionally similar but geometrically different. By leveraging generative models and other readily available data sources, GRIM circumvents the data bottleneck that plagues traditional supervised methods. Our extensive experiments show that GRIM significantly outperforms existing training-free and learning-based approaches, particularly in its ability to handle novel objects and tasks.

Future work could explore incorporating explicit geometric reasoning, perhaps through the generation of digital twins (Melnik et al. 2025), to further refine the alignment and grasp transfer process. Nevertheless, GRIM represents a significant step towards building more general, adaptable, and data-efficient robotic manipulation systems.

Acknowledgments

The research reported in this paper has been (partially) supported by IHFC-TIH of Department of Science and Technology, Govt. of India, project #GP/2021/HRI/002 and the German Research Foundation (DFG), as part of the Collaborative Research Center (Sonderforschungsbereich) 1320, Project-ID 329551904, “EASE – Everyday Activity Science and Engineering”, University of Bremen (<http://www.ease-crc.org/>).

References

Ardón, P.; Pairet, É.; Petrick, R. P. A.; Ramamoorthy, S.; and Lohan, K. S. 2019. Learning Grasp Affordance Reasoning Through Semantic Relations. *IEEE Robotics and Automation Letters*, 4: 4571–4578.

Besl, P. J.; and McKay, N. D. 1992. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14: 239–256.

Dang, H.; and Allen, P. K. 2012. Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1311–1317.

Deshpande, A.; Deng, Y.; Ray, A.; Salvador, J.; Han, W.; Duan, J.; Zeng, K.-H.; Zhu, Y.; Krishna, R.; and Hendrix, R. 2025. GraspMolmo: Generalizable Task-Oriented Grasping via Large-Scale Synthetic Data Generation. *arXiv:2505.13441*.

Di Palo, N.; and Johns, E. 2024. DINOBot: Robot Manipulation via Retrieval and Alignment with Vision Foundation Models. *arXiv preprint arXiv:2402.13181*.

Dong, W.; Huang, D.; Liu, J.; Tang, C.; and Zhang, H. 2024. RTAGrasp: Learning Task-Oriented Grasping from Human Videos via Retrieval, Transfer, and Alignment. *arXiv preprint arXiv:2409.16033*.

Fang, H.-S.; Wang, C.; Fang, H.; Gou, M.; Liu, J.; Yan, H.; Liu, W.; Xie, Y.; and Lu, C. 2023. AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. *arXiv:2212.08333*.

Google. 2025. Veo 2. via Google AI Studio. Accessed in May 2025. Model available at <https://aistudio.google.com/>.

Huang, W.; Xia, F.; Xiao, T.; Chan, H.; Liang, J.; Florence, P.; Zeng, A.; Tompson, J.; Mordatch, I.; Chebotar, Y.; Sermanet, P.; Brown, N.; Jackson, T.; Luu, L.; Levine, S.; Hausman, K.; and Ichter, B. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *arXiv preprint arXiv:2207.05608*.

Jin, S.; Xu, J.; Lei, Y.; and Zhang, L. 2024. Reasoning Grasping via Multimodal Large Language Model. *ArXiv*, abs/2402.06798.

Ju, Y.; Hu, K.; Zhang, G.; Zhang, G.; Jiang, M.; and Xu, H. 2024. Robo-ABC: Affordance Generalization Beyond Categories via Semantic Correspondence for Robot Manipulation. *arXiv preprint arXiv:2401.07487*.

Kuang, Y.; Ye, J.; Geng, H.; Mao, J.; Deng, C.; Guibas, L.; Wang, H.; and Wang, Y. 2024. RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation. *arXiv preprint arXiv:2407.04689*.

Li, S.; Bhagat, S.; Campbell, J.; Xie, Y.; Kim, W.; Sycara, K.; and Stepputtis, S. 2024. ShapeGrasp: Zero-Shot Task-Oriented Grasping with Large Language Models through Geometric Decomposition. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10527–10534.

Liu, W.; Daruna, A.; Patel, M.; Ramachandruni, K.; and Chernova, S. 2023. A survey of Semantic Reasoning frameworks for robotic systems. *Robotics and Autonomous Systems*, 159: 104294.

Liu, W.; Daruna, A. A.; and Chernova, S. 2019. CAGE: Context-Aware Grasping Engine. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2550–2556.

Malato, F.; Leopold, F.; Melnik, A.; and Hautamäki, V. 2024. Zero-shot imitation policy via search in demonstration dataset. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7590–7594. IEEE.

Melnik, A.; Alt, B.; Nguyen, G.; Wilkowski, A.; Wu, Q.; Harms, S.; Rhodin, H.; Savva, M.; Beetz, M.; et al. 2025.

- Digital Twin Generation from Visual Data: A Survey. *arXiv preprint arXiv:2504.13159*.
- Melnik, A.; Ljubljanc, M.; Lu, C.; Yan, Q.; Ren, W.; and Ritter, H. 2024. Video Diffusion Models: A Survey. *Transactions on Machine Learning Research*.
- Melnik, A.; Schüler, F.; Rothkopf, C. A.; and König, P. 2018. The world as an external memory: The price of saccades in a sensorimotor task. *Frontiers in behavioral neuroscience*, 12: 253.
- Mikami, Y.; Melnik, A.; Miura, J.; and Hautamäki, V. 2024. Natural Language as Policies: Reasoning for Coordinate-Level Embodied Control with LLMs. *arXiv preprint arXiv:2403.13801*.
- Mirjalili, R.; Krawez, M.; Silenzi, S.; Blei, Y.; and Burgard, W. 2024. Lan-grasp: Using Large Language Models for Semantic Object Grasping. *arXiv:2310.05239*.
- Murali, A.; Liu, W.; Marino, K.; Chernova, S.; and Gupta, A. 2020. Same Object, Different Grasps: Data and Semantic Knowledge for Task-Oriented Grasping. In *Conference on Robot Learning*.
- Nguyen, T.; Vu, M. N.; Huang, B.; Vo, T. V.; Truong, V.; Le, N.; Vo, T. D.; Le, B.; and Nguyen, A. 2023. Language-Conditioned Affordance-Pose Detection in 3D Point Clouds. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 3071–3078.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *arXiv:2304.07193*.
- PS, A.; Melnik, A.; Nandi, G. C.; et al. 2024. SplatR: Experience Goal Visual Rearrangement with 3D Gaussian Splatting and Dense Feature Matching. *arXiv preprint arXiv:2411.14322*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020*.
- Rashid, A.; Sharma, S.; Kim, C. M.; Kerr, J.; Chen, L. Y.; Kanazawa, A.; and Goldberg, K. 2023. Language Embedded Radiance Fields for Zero-Shot Task-Oriented Grasping. In *7th Annual Conference on Robot Learning*.
- Song, D.; Huebner, K.; Kyrki, V.; and Kragic, D. 2010. Learning task constraints for robot grasping using graphical models. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1579–1585.
- Tang, C.; Huang, D.; Ge, W.; Liu, W.; and Zhang, H. 2023a. GraspGPT: Leveraging Semantic Knowledge from a Large Language Model for Task-Oriented Grasping. *arXiv preprint arXiv:2307.13204*.
- Tang, C.; Huang, D.; Meng, L.; Liu, W.; and Zhang, H. 2023b. Task-Oriented Grasp Prediction with Visual-Language Inputs. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4881–4888.
- Wang, Y.; Zhang, M.; Li, Z.; Kelestemur, T.; Driggs-Campbell, K.; Wu, J.; Fei-Fei, L.; and Li, Y. 2023. D³Fields: Dynamic 3D Descriptor Fields for Zero-Shot Generalizable Rearrangement. *arXiv preprint arXiv:2309.16118*.
- Wu, J.; Pavlakos, G.; Gkioxari, G.; and Malik, J. 2024. Reconstructing Hand-Held Objects in 3D from Images and Videos. *arXiv preprint arXiv:2404.06507*.
- Zese, R.; Bellodi, E.; Lamma, E.; Riguzzi, F.; and Aguiari, F. 2014. Semantics and Inference for Probabilistic Description Logics. In Bobillo, F.; Carvalho, R. N.; Costa, P. C.; d’Amato, C.; Fanizzi, N.; Laskey, K. B.; Laskey, K. J.; Lukasiewicz, T.; Nickles, M.; and Pool, M., eds., *Uncertainty Reasoning for the Semantic Web III*, 79–99. Cham: Springer International Publishing. ISBN 978-3-319-13413-0.