

GraspCoT: Integrating Physical Property Reasoning for 6-DoF Grasping under Flexible Language Instructions

Xiaomeng Chu¹, Jiajun Deng^{2*}, Guoliang You¹, Wei Liu¹, Xingchen Li¹, Jianmin Ji¹, Yanyong Zhang^{1*}

¹ University of Science and Technology of China, ² The University of Adelaide

Abstract

Flexible instruction-guided 6-DoF grasping is a significant yet challenging task for real-world robotic systems. Existing methods utilize the contextual understanding capabilities of the large language models (LLMs) to establish mappings between expressions and targets, allowing robots to comprehend users' intentions in the instructions. However, the LLM's knowledge about objects' physical properties remains underexplored despite its tight relevance to grasping. In this work, we propose GraspCoT, a 6-DoF grasp detection framework that integrates a Chain-of-Thought (CoT) reasoning mechanism oriented to physical properties, guided by auxiliary question-answering (QA) tasks. Particularly, we design a set of QA templates to enable hierarchical reasoning that includes three stages: target parsing, physical property analysis, and grasp action selection. Moreover, GraspCoT presents a unified multimodal LLM architecture, which encodes multi-view observations of 3D scenes into 3D-aware visual tokens, and then jointly embeds these visual tokens with CoT-derived textual tokens within LLMs to generate grasp pose predictions. Furthermore, we present IntentGrasp, a large-scale benchmark that fills the gap in public datasets for multi-object grasp detection under diverse and indirect verbal commands. Extensive experiments on IntentGrasp demonstrate the superiority of our method, with additional validation in real-world robotic applications confirming its practicality. The code is available at <https://github.com/cxmomo/GraspCoT>.

1. Introduction

The language-guided 6-DoF grasp detection task tackles the problem of robotic grasping that interprets human instructions and 3D environments [1, 27, 52]. This capability holds critical importance for domestic assistance, healthcare, and industrial automation. However, it presents greater com-

*Corresponding authors.

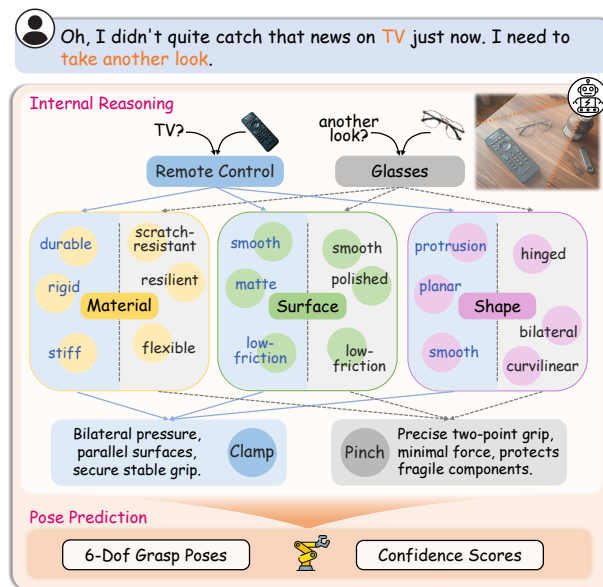


Figure 1. Example of GraspCoT. Facing a rewatching TV request, GraspCoT employs CoT reasoning, including target parsing, physical property analysis, and action selection, to interpret the scene and predict 6-DoF grasp poses with confidence scores for critical objects, namely remote control and glasses.

plexity than its language-agnostic robotic grasping counterparts [14, 16, 24, 35, 36], mainly due to the cross-modal alignment between textual and visual representations. Consequently, existing methods [19, 29, 37] use a text encoder [6, 17, 34] to map textual and visual features into a unified latent space, enabling cross-modal correspondence for instruction-guided grasping.

With the emergence of LLMs and their multi-modal variants (MLLMs), there has been a surge of research interest in integrating semantic knowledge and common sense into object grasping. Early works leverage LLMs mainly for semantic understanding in task-oriented grasping. For example, GraspGPT [38] uses LLMs to describe objects and task-relevant information, enabling interaction with novel object classes, while SemGrasp [18] fine-tunes an MLLM to gen-

erate human-like grasps. These models primarily handle explicit instructions, such as “Give me the remote control”, where the target object and action are clearly stated. A more recent trend focuses on exploring the reasoning ability of LLMs to interpret implicit grasping instructions. The methods [13, 32] following this trend infer the intended action from indirect cues, such as recognizing the need to fetch a remote control from “I want to watch TV”. Such practices makes human-computer interaction more natural by understanding users’ intention. Nevertheless, merely understanding the semantics or users’ intentions does not fully leverage the common knowledge of LLMs to improve grasp detection. The physical properties of objects, essential for real-world manipulation, remain underexplored. Beyond geometric similarity, effective grasping requires reasoning about functional and physical attributes—for instance, a fragile vase demands precise neck grasping to prevent breakage, whereas a sturdy mug allows more flexible handling based on material strength and texture.

By consolidating this idea, we explore a new paradigm that can hierarchically integrate objects’ physics-centric analyses into grasping by leveraging LLMs’ remarkable reasoning capabilities. Besides, the previous methods generally assume that the grasping task only involves a single object, we give up such an assumption to make the grasping algorithm more applicable. Notably, in this work, we use “flexible instruction” to indicate the expression that does not explicitly specify the target object or even object quantity.

Formally, we introduce **GraspCoT**, a 6-DoF grasp detection framework featuring a physical-property-oriented chain-of-thought (CoT) [45] reasoning mechanism via auxiliary question-answering (QA) tasks. Specifically, we develop QA templates for a three-stage CoT pipeline covering target parsing, physical property analysis, and grasp action selection. QA texts and flexible language instructions are encoded as textual tokens. Fig. 1 provides an example of the CoT reasoning process of GraspCoT. Departing from conventional pipelines that decouple multimodal understanding into isolated modules, our framework integrates visual-textual comprehension directly into 6-DoF grasp pose prediction, inspired by breakthroughs in 3D MLLMs [2, 4, 50]. This unified architecture not only enhances cross-modal feature alignment but also reduces system complexity. On the visual side, we construct multi-view observations of the 3D scene by projecting colored point clouds into complementary RGB images and depths. The images are processed into 2D patches via CLIP, enriched with depth-derived 3D position embeddings, and back-projected into geometrically grounded 3D tokens [33, 55, 56]. Finally, these 3D tokens, along with textual tokens mentioned above, undergo deep alignment within LLMs to predict grasp poses with confidence scores for the target objects.

Besides, there is no public benchmark for evaluating

grasp detection under flexible instructions, especially for grasping multiple target objects. To bridge this gap, we propose IntentGrasp, a novel benchmark adapted from the Grasp-Anything-6D dataset [30]. For each 3D scene, we employ structured prompt templates processed by Llama-3 [11, 39] to generate linguistically diverse requirements that identify target objects through contextual clues rather than explicit identifiers. This benchmark challenges robotic systems to resolve referential ambiguity by joint visual perception with textual semantics, thereby comprehensively assessing their cross-modal cognitive abilities.

We evaluate GraspCoT on the IntentGrasp benchmark, demonstrating the superiority of our proposed GraspCoT for both pose estimation accuracy and collision avoidance rate. Besides, we deploy the proposed GraspCoT on a Kinova Gen3 robotic arm. This extended experiment further validates its effectiveness in real-world robotic applications.

In summary, our main contributions are as follows:

- We propose a physical-property-oriented CoT reasoning paradigm that bridges 3D scene understanding and grasp detection via auxiliary QA tasks.
- We introduce a unified grasp detection framework, GraspCoT, that deeply embeds LLMs for the joint optimization of textual-visual feature alignment via dual-task learning.
- We propose IntentGrasp, the first benchmark supporting multi-target grasping under flexible instructions.
- We demonstrate the superior performance of GraspCoT by comprehensive evaluations on the IntentGrasp benchmark and real-world deployment on the robot arm.

2. Related Work

Robotic Grasping. The emergence of grasping datasets [5, 7, 8, 25] lay the groundwork for data-driven grasping, while later works extend these foundations to diverse and task-aware scenarios. 6-DOF GraspNet [26] employs variational autoencoders to sample diverse poses from partial observations, and AnyGrasp [9] introduces temporal consistency for dynamic environments. Recognizing the need for hardware generalization, UniGrasp [36] pioneers gripper-agnostic frameworks through kinematic latent encoding, while optimization-centric paradigms like SE(3)-DiffusionFields [41] reformulate grasping as gradient-based pose-trajectory co-optimization. Recent efforts balance these directions: Wang et al. [42] accelerates inference via geometric heuristics, and Wang et al. [43] enhances physical interaction through auxiliary task learning. Collectively, these works demonstrate a shift from isolated geometric analysis to hybrid systems that embed functional adaptability and computational efficiency within unified frameworks.

Language-guided Grasping. Recent advances in vision-language models have revolutionized language-guided robotic grasping by enabling semantic alignment between

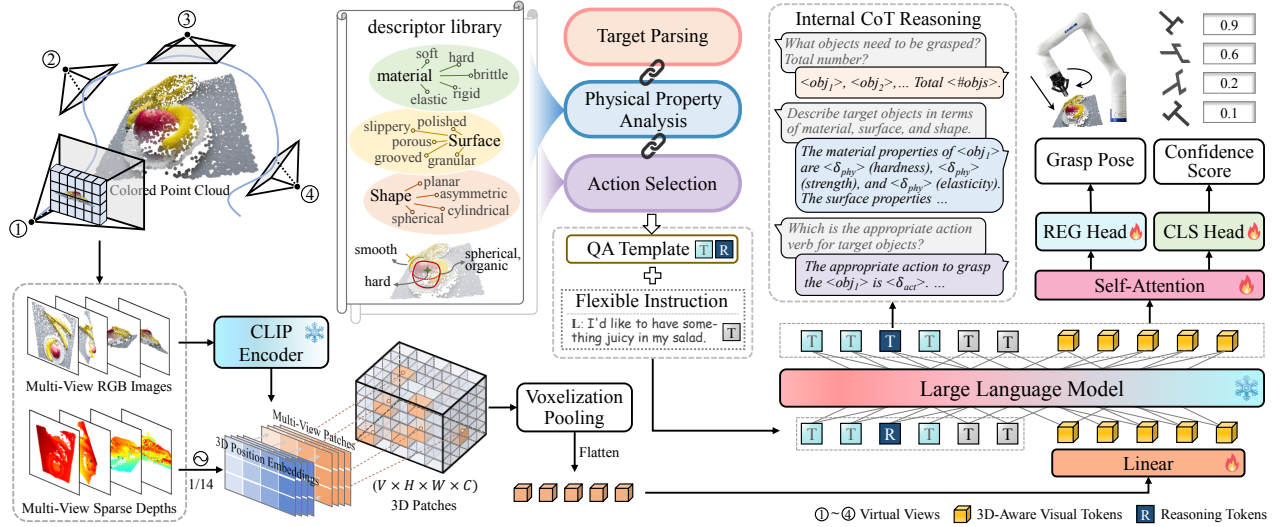


Figure 2. **Overview of GraspCoT.** Given a colored point cloud, we project it into multi-view RGB-D images. CLIP-encoded RGB patches and depth-derived 3D positional embeddings are fused and back-projected into 3D patches, which are voxel-pooled and linearly projected into 3D-aware visual tokens. We propose a three-stage Chain-of-Thought (CoT) framework based on fill-in-the-blank question-answering (QA) templates. This framework consists of three stages: target parsing, physical property analysis, and action selection, all enhanced with descriptor libraries. In this setup, the blank fields within the templates are represented as reasoning tokens, which the large language model (LLM) resolves into physical or action descriptors in order to generate answers. The fixed textual components of the templates serve as contextual QA tokens. After performing reasoning in the LLM, the autoregressive generated output tokens are all fed into a self-attention layer, followed by parallel grasp pose regression and confidence prediction.

textual instructions and visual perception. Pioneering works [19, 23, 28–30, 37, 46, 51] leverage a text backbone [6, 34] to embed textual features, where some explore diffusion models for 3D manipulation—3DAPNet [29] extends grasp pose prediction paradigm to open-vocabulary affordance-pose joint learning, while LGrasp6D [30] introduces the Grasp-Anything-6D dataset and incorporates the negative prompt guidance for 6-DoF grasp detection. OVGNet [19] proposes a unified visual-linguistic framework combining CLIP-style cross-modal alignment with grasp pose prediction, enabling generalization to unseen categories without category-specific retraining. The emergence of (M)LLMs [17, 22, 39] further expands the visual and textual interpretation capabilities in manipulation tasks [18, 31, 38, 40]. ThinkGrasp [32] introduces a language-guided, stepwise clutter removal strategy powered by GPT-4o’s contextual inferring, enabling robotic grasping in occluded environments through goal-oriented obstacle displacement and few-step target recovery. RT-Grasp [49] proposes a framework that unlocks LLMs’ capacity for numerical prediction in robotics by embedding structured inferring phases into training, while Jin et al. [13] leverage an end-to-end multimodal LLM framework to bridge implicit human intentions to grasp poses, enabling object/part-level grasping under indirect instructions. Recent efforts combine visual grounding with attribute prediction of required objects using multimodal foundation

models. ShapeGrasp [20] pioneers zero-shot task-oriented grasping through geometric decomposition of objects into convex shapes represented as attribute-rich graphs, combined with LLM-driven semantic reasoning to infer part-task utility. DeliGrasp [48] proposes LLM-driven object properties prediction for adaptive grasping, converting semantic object descriptions into grasp policies via inferred mass/friction/spring parameters. Instead of requiring LLMs to directly predict precise physical parameters—a task challenging current LLMs’ capabilities—our method encodes physical properties into structured descriptors, explicitly guiding grasp pose prediction while avoiding error-prone low-level physics estimation.

3. Method

In this section, we formally define the task of flexible-instruction-guided 6-DoF grasp detection, which requires jointly interpreting visual scene understanding and flexible language to determine optimal grasp configurations. Next, we present the framework, detailing its core components: the physical-property-oriented Chain-of-Thought (CoT) design and grasp decoding with the auxiliary question-answering (QA) task via large language models (LLMs). Finally, we outline the construction of the IntentGrasp benchmark, designed to evaluate grasping under linguistically flexible instructions.

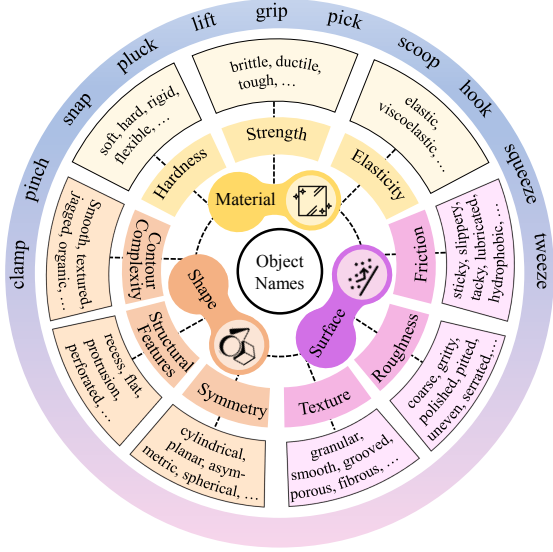


Figure 3. Hierarchical CoT Reasoning: target parsing (inner ring), multi-aspect physical property analysis (material/surface/shape in middle three rings), and grasp action selection represented by the action descriptor libraries (outer ring).

3.1. Preliminary

The flexible-instruction-guided 6-DoF grasp pose detection takes as input a colored point cloud $\mathcal{P} = \{(p_i, c_i)\}_{i=1}^N$, where $p_i \in \mathbb{R}^3$ represents the 3D coordinates and $c_i \in \mathbb{R}^3$ denotes the RGB color of each point, along with a flexible language instruction t . The output consists of CoT-relevant questions and answers $\mathcal{Q} = \{(q_k, a_k)\}_{k=1}^M$ that analyze the categories, physical properties, and grasp action of the target object(s), followed by the generation of multiple 6-DoF grasp poses $g \in SE(3)$ for the identified target(s). Formally, this process Φ is defined as:

$$\Phi : (\mathcal{P}, t) \mapsto (\mathcal{Q}_{\text{CoT}}, \mathcal{G}), \quad (1)$$

where \mathcal{G} represents the grasp pose set. Each $g_j = (R, T, w) \in SE(3)$ defines a 6-DoF gripper configuration composed of the orientation $R \in SO(3)$, the position $T \in \mathbb{R}^3$, and the suitable gripper width $w \in \mathbb{R}^+$.

3.2. Model Architecture

With growing advances in 3D multi-modal LLMs (MLLMs) [3, 10, 12, 53], recent breakthroughs have been achieved in aligning 3D geometric representations with linguistic instructions. As shown in Fig. 2, our architecture operates on colored point cloud inputs through a dual-branch processing pipeline. In the visual processing branch, we construct multiple visual viewpoints for projection to generate multi-view sparse RGB-D representations. These projections undergo CLIP-based feature encoding, which is subsequently combined with 3D position embeddings from depth data. After back-projection, voxel pooling, and linear

projection, we transform the reconstructed 3D patches into 3D-aware visual tokens. Simultaneously, the textual branch employs a hierarchical CoT reasoning framework comprising target parsing, physical property analysis, and action selection. To guide a robust reasoning process, we establish open-world descriptor repositories and develop semi-supervised QA templates that produce supervised QA tokens along with unsupervised reasoning tokens for descriptor instantiation. The autoregressive tokens produced by the LLM’s reasoning stage are first passed into a self-attention layer, then jointly utilized for parallel predictions of grasp poses and their confidence scores.

3.3. Physical-Property-Oriented CoT Design

To bridge scene understanding and grasp poses, we design fill-in-the-blank QA templates and implement a three-stage CoT reasoning pipeline that sequentially grounds target categories, infers physical properties, and maps grasp actions. Structured QA templates are designed as follows to guide the reasoning process for grasp detection:

- Q1: Target Parsing** - Which objects need to be grasped?
A1: [$\langle \text{obj}_{j_1} \rangle$, $\langle \text{obj}_{j_2} \rangle$, ...]. Total $\langle \# \text{obj}_j \rangle$.
- Q2: Physical Property Analysis** (e.g., material) - For each target object, analyze its material from three aspects: hardness [options: soft, hard, rigid, flexible, etc.], strength [options: brittle, ductile, tough, etc.], and elasticity [options: elastic, viscoelastic, etc.].
A2: The material properties of $\langle \text{obj}_{j_k} \rangle$ are $\langle \delta_{\text{phy}} \rangle$ (hardness), $\langle \delta_{\text{phy}} \rangle$ (strength), and $\langle \delta_{\text{phy}} \rangle$ (elasticity).
- Q3: Grasp Action Selection** - For each target object, select an appropriate grasp action [options: clamp, pinch, snap, pluck, lift, grip, etc.].
A3: The appropriate verb to grasp the $\langle \text{obj}_{j_k} \rangle$ is $\langle \delta_{\text{act}} \rangle$.

First, the target parsing stage resolves referential ambiguities in an answering template with keyword slots, where target categories are explicitly supervised. Next, the pipeline takes advantage of the common sense and reasoning ability of MLLMs to infer physical properties, which are hierarchically categorized into three primary groups: material, surface, and shape¹. Each property is further subdivided into three fine-grained characteristics described by appropriate physical descriptors δ_{phy} in open-world libraries, as illustrated in Fig. 3. For example, a glass cup’s material properties are characterized as $\langle \text{rigid} \rangle$ in hardness, $\langle \text{brittle} \rangle$ in strength, and $\langle \text{inflexible} \rangle$ in elasticity. Similarly, its surface attributes are described as $\langle \text{smooth} \rangle$ in texture, $\langle \text{polished} \rangle$ in roughness, and $\langle \text{slippery} \rangle$ in friction. This structured classification enables a comprehensive representation of physical properties for the prediction of grasp poses. Finally, the action se-

¹While “shape” is geometrically defined, our analysis targets inferring mass distribution and optimal grasping contact forces, thus integrating it into the broader category of physical properties for unified reasoning.

lection stage requires LLMs to select reasonable action descriptors δ_{act} (e.g., “pinch” for fragile objects, “clamp” for textured surfaces) from a predefined verb library, dynamically linking these choices to geometric parameters such as gripper orientation and contact force. Notably, the physical/action descriptor tokens remain unsupervised, leveraging LLMs’ extensive corpora knowledge to bypass manual annotation of exhaustive object-specific properties.

3.4. Auxiliary QA Task and Grasp Decoding

The synergistic integration of 3D-aware visual tokens and flexible-instruction tokens enhances the model’s comprehension of 3D environments and interactive intent. As mentioned before, GraspCoT unifies grasp-related QA tasks and grasp pose prediction through joint training, enabling deep fusion of task-relevant token representations within the pre-trained LLM architecture. We implement answer templates for three question categories in Sec. 3.3. Crucially, the physical properties of targets and grasp actions remain unspecified descriptors that require no explicit annotation. Instead, we introduce special “reasoning tokens”, which are resolved into physical or action descriptors by the LLM, to represent these latent variables. During training, these tokens receive masked supervision with zero loss weighting.

The core output of GraspCoT consists of multiple 6-DoF grasp poses with corresponding confidence scores, which inspired us to adopt a query-based detection decoding approach. Specifically, we first process all output tokens from the LLM through a self-attention layer, then feed them into two separate heads: a regression head for grasp pose estimation and a classification head for confidence prediction. During training, we employ the Hungarian matching algorithm to align predicted grasp poses with ground truth annotations and compute L1 loss L_{Reg} for pose refinement. Simultaneously, we apply focal loss [21] L_{Cls} to optimize the confidence scores of the predicted poses. Combined with the cross-entropy loss L_{QA} for the QA task, the total loss of the model integrates three components as follows:

$$L = L_{QA} + L_{Reg} + L_{Cls}. \quad (2)$$

3.5. IntentGrasp Benchmark Construction

Most existing language-guided grasping benchmarks utilize explicit instructions such as “Bring me a bottle of water” or “Grab the black remote control”, which explicitly specify target objects by name and salient visual attributes. However, real-world human-robot interaction demands support for implicit requests. For instance, a user stating “I’m thirsty” or “I want to watch TV”, where the robot should infer the need for water or the remote control. Equipped with visual scene understanding, the robot can further engage in contextual reasoning: upon detecting “glasses” in the environment while addressing a “watch TV” request, it might

Please generate **3 to 5** linguistically flexible instructions to grasp **one or more** objects from “Human” based on the given **scene description**, following the rules below:

- [R1]** The names of the targets should **not** be explicitly mentioned.
- [R2]** If multiple targets need to be grasped, the instructions should establish a reasonable and logical connection between the targets.
- [R3]** Each instruction should be authentic, concise, and independent.

Example:

Scene description: “Black **car keys**, a ceramic **mug**, a red **notebook**, and a pair of **sunglasses** on the table.”

Output: {Instruction: “It’s **sunny** outside and I want to **go for a drive**.”,
Target Objects: [car keys, sunglasses]};
{Instruction: “It’s **scorching** out here-I’m **parched!**”,
Target Objects: [mug]};
...

Figure 4. Prompt used for flexible instruction generation.

proactively retrieve them for the user. We define such commands as flexible instructions, where one or more target objects are required but neither explicitly named nor uniquely defined. The robot must interpret the user’s intent through contextual cues within a plausible scenario, potentially retrieving one or multiple contextually relevant targets.

To address the lack of public benchmarks supporting multi-target grasp detection under such flexible instructions, we develop IntentGrasp, a benchmark containing one million (1M) 3D colored point cloud scenes with approximately 3M objects, sourced from the Grasp-Anything-6D dataset [30], which provides scene descriptions, explicit instructions, and grasp pose annotations. For each scene, we generate 3-5 flexible instructions using the open-source LLM Llama3-70B [39] guided by a customized prompt template. As shown in Fig. 4, the prompt directs the LLM to create plausible requests that implicitly demand the grasping of one or multiple objects without naming targets. For example, in a scene containing “black car keys, a ceramic mug, a red notebook, and sunglasses on the table”, the Grasp-Anything-6D dataset provides explicit instructions like “Grasp the black car keys”, whereas our method generates contextual statements such as “It’s sunny outside and I want to go for a drive”, implicitly requiring the robot to retrieve both car keys and sunglasses. IntentGrasp fills the critical gap in flexible-instruction-driven multi-target grasping benchmarks. Its generation design can be further integrated with detection/segmentation models [15, 44] to extend to any other grasping dataset.

4. Experiments

In this section, we validate the effectiveness of GraspCoT using flexible instructions on both the IntentGrasp benchmark and real-world experimental platforms.

4.1. Main Results on IntentGrasp Benchmark

Setup. To evaluate the effectiveness of our grasp pose prediction method, we employ three metrics: coverage

Methods	CR@0.4↑	CR@0.3↑	CR@0.2↑	EMD↓	EW-CFR↑
3DAPNet [29]	0.9249	0.7378	0.2904	0.3066	0.3521
LGrasp6D [30]	0.9294	0.7567	0.3349	0.2935	0.3847
GraspCoT (ours) w/o CoT	<u>0.9468</u>	<u>0.7941</u>	<u>0.4164</u>	<u>0.2878</u>	<u>0.4173</u>
GraspCoT (ours)	0.9715	0.8797	0.5587	0.2520	0.4229

Table 1. Results on IntentGrasp benchmark under flexible language instructions. “CR@ θ ” denotes the success rate when predicted poses are within a threshold θ of ground truth poses.

rate (CR) [26], earth mover’s distance (EMD) [41], and collision-free rate (CFR) [54] with appropriate adaptations. CR quantifies the proportion of ground-truth grasp poses that are sufficiently covered by predictions. Specifically, it measures the percentage of ground-truth poses where at least one predicted pose lies within a distance threshold θ (Euclidean distance in SE(3) space). We evaluate CR under three increasingly stringent thresholds: $\theta = 0.4, 0.3$, and 0.2 . EMD measures how holistically the predicted poses match the spatial distribution of ground-truth annotations in SE(3) space, where lower values indicate better distributional consistency. CFR evaluates the physical feasibility of predicted grasp poses by measuring their ability to avoid collisions with the target object and environment. However, a critical limitation arises when directly using the CFR metric: erroneous grasp poses far from the target object may artificially achieve higher scores because they avoid contact with environmental point clouds entirely. To jointly evaluate geometric consistency and physical feasibility where higher values indicate safer grasps, we propose an EMD-Weighted CFR (EW-CFR) metric defined as follows:

$$\text{EW-CFR} = \frac{1}{N} \sum_{i=1}^N (1 - \mathcal{C}(\mathcal{P}, \mathcal{M}(g_i))) \cdot \left(1 - \frac{\text{EMD}(g_i, \hat{G})}{R}\right), \quad (3)$$

where R normalizes the $\text{EMD}(g_i, \hat{G})$. The function \mathcal{C} denotes a binary collision check between the point cloud \mathcal{P} and the gripper model $\mathcal{M}(g_i)$.

The IntentGrasp dataset is partitioned following conventional practice, with 80% of samples dedicated to training and the remaining 20% preserved as the evaluation set.

Baseline. In comparative experiments, we benchmark our method against two recent language-guided 6-DoF grasp detection models: 3DAPNet [29] and LGrasp6D [30]. To align with scenarios under flexible instructions, we replace their original explicit grasp command inputs with flexible instructions while preserving their default training protocols. Specifically, we eliminate 3DAPNet’s affordance prediction module to better align with our task requirements. For consistency across all models, the valid grasp pose’s positions (x, y, z) and orientations along x- and y-axes are limited to the range $[-1, 1]$, and z-axis orientations are bounded within $[0, \pi]$. To ensure fairness, we uniformly sample 600 valid grasp poses of each scene generated by 3DAPNet and LGrasp6D for validation. Since each grasp pose in Grasp-

Material	Surface	Shape	Action	CR@0.3↑	CR@0.2↑	EMD↓
				0.7941	0.4164	0.2878
✓				0.8507	0.5091	0.2698
	✓			0.8334	0.4818	0.2738
		✓		0.8306	0.4716	0.2742
✓	✓			0.8609	0.5177	0.2653
✓	✓	✓		0.8717	0.5386	0.2572
✓	✓	✓	✓	0.8797	0.5587	0.2520

Table 2. Ablation study on physical property analysis and grasp action selection in CoT Reasoning.

CoT is assigned a confidence score, we select the top 600 highest-confidence poses as valid candidates.

Grasp Label Pruning. Wu [47] et al. demonstrate that excessively dense supervision incurs label ambiguity, particularly caused by the diversity of grasp orientations, as a critical bottleneck hindering model convergence. Our analysis reveals that certain objects in the Grasp-Anything-6D dataset contain over 700 annotated grasp poses. This overabundance of positive samples not only exacerbates the aforementioned issues but also severely challenges Hungarian matching during training. Consequently, we implemented a pruning strategy in IntentGrasp to reduce similar grasp labels, especially redundant rotation-similar annotations, per object to 100 instances. The detailed pruning method is provided in the supplementary material.

Training. Our models are trained using the Adam optimizer with a batch size of 8 per GPU over 4 epochs, initialized with LLaVA-3D [56] pre-trained weights from the HuggingFace repository. The learning rate followed a cosine annealing schedule, starting at 1×10^{-4} . All experiments are executed on 8×24 GB NVIDIA RTX3090 GPUs.

Quantitative Results. In Tab. 1, our method presents superior performance across all metrics on the IntentGrasp benchmark under flexible language instructions. Notably, GraspCoT attains a CR of 0.5587 at the strictest threshold ($\theta=0.2$), outperforming 3DAPNet and LGrasp6D by a large margin of 0.2683 and 0.2238, respectively. This substantial improvement highlights our method’s ability to generate geometrically precise grasp poses even under high-precision requirements. The progressive performance gap across $\theta=0.4, 0.3$, and 0.2 further demonstrates that our ap-

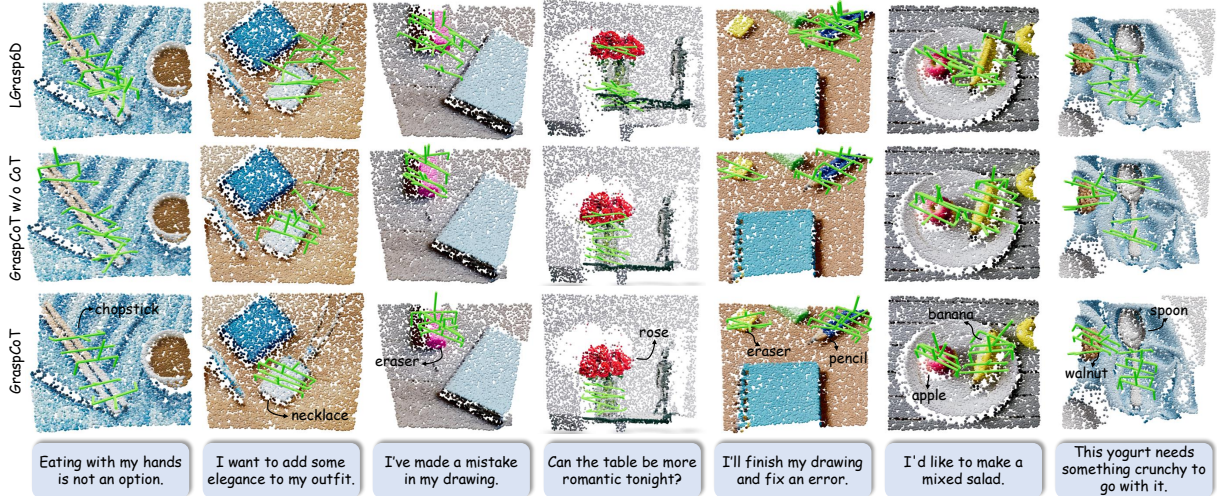


Figure 5. Visualization of 6-DoF grasp detection results based on flexible instructions.

#Targets	CoT	CR@0.4 \uparrow	CR@0.3 \uparrow	CR@0.2 \uparrow	EMD \downarrow
Single		0.9465	0.7896	0.4108	0.2810
Multiple		0.9492	0.7977	0.4183	0.3106
Single	\checkmark	0.9723	0.8840	0.5590	0.2453
Multiple	\checkmark	0.9680	0.8754	0.5571	0.2756

Table 3. Grasp detection results of GraspCoT vs. the w/o CoT version in single- and multi-target scenarios.

proach maintains effectiveness as task difficulty escalates. The EMD of GraspCoT reflects a 14.1% relative reduction compared to LGrasp6D, indicating superior alignment between the predicted and ground-truth pose distributions. For our proposed EW-CFR metric, GraspCoT achieves a score of 0.4229, surpassing 3DAPNet and LGrasp6D by 20.1% and 9.9%, respectively. The w/o CoT version of GraspCoT confirms the critical role of our CoT reasoning pipeline: removing CoT results in a 25.4% decline in CR@0.2 and a 9.1% increase in EMD, respectively.

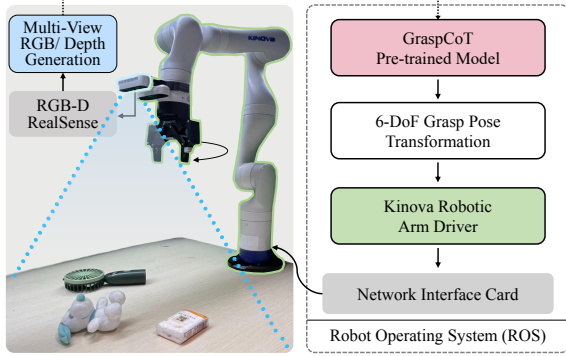
Qualitative Results. In Fig. 5, we present the comparative visualization results of GraspCoT for 6-DoF grasp detection under flexible instructions, covering both single- and multi-target scenarios. Our method exhibits more precise and contextually reasonable predictions in target parsing and grasp positioning. For example, GraspCoT correctly identifies grasp regions such as the body of a rose vase or the handle of a spoon. Notably, GraspCoT and the w/o CoT version both effectively avoid target omission in multi-target tasks, demonstrating its enhanced comprehension of both linguistic instructions and 3D scene contexts.

4.2. Ablation Studies

Physical-Property-Oriented CoT Reasoning. As shown in Tab. 1, CoT reasoning significantly enhances GraspCoT’s

performance, validating its necessity. Tab. 2 quantifies the contributions of material, surface, and shape property analysis within the CoT pipeline, alongside grasp action selection. Material property analysis notably elevates performance, improving CR@0.3 and CR@0.2 by 7.1% and 22.3% respectively over the w/o CoT version, while reducing EMD by 6.3%, demonstrating its critical role in refining pose prediction. Integrating surface or shape property sorely also boosts performance, achieving a 15.7% and 13.3% CR@0.2 gain, respectively. This finding reveals that even single-aspect physical property integration exerts a significant directional influence on LLMs’ decoding of grasp-relevant hidden states. Cumulative integration of all three physical analyses achieves the second-best pose estimation quality, reaching an EMD of 0.2572. Grasp action selection further enhances precision, resulting in a 3.7% increase in CR@0.2 and a 2.0% reduction in EMD.

Single- and Multi-Target Grasping. Our default evaluation protocol randomly selects one flexible instruction from 3–5 variants per scene, encompassing both single- and multi-target grasping scenarios. We then detail the respective performance of GraspCoT and its w/o CoT version in these two scenarios. In Tab. 3, the w/o CoT version shows marginally superior CR metrics for multi-target grasping compared to single-target cases, albeit at the cost of 10.5% higher EMD. When integrating CoT reasoning, significant improvements emerge: single-target performance achieves 12.0% improvement in CR@0.3, 36.1% improvement in CR@0.2, a 12.7% decrease in EMD, while multi-target grasping attains 9.7% improvement in CR@0.3, 33.2% improvement in CR@0.2, and 11.3% decrease in EMD. These results demonstrate that our CoT mechanism delivers stronger optimization for single-target scenarios because physical property analysis targets the object itself.



(a) Real-world experimental setup.



(b) Example of the grasping task.

Figure 6. An application example on a Kinova Gen3 robotic arm. The video demonstration is included in supplementary materials.

Number of Virtual Views. As presented in Tab. 4, we evaluate the impact of varying numbers of complementary virtual views, up to five. In ablation studies with no fewer than four views, performance using k (<4) views reveals that $CR@0.2$ experiences the most significant fluctuations: a 3.70% relative increase when using 3 versus 2 views, followed by a substantial 12.6% leap at 4 views. Beyond four views, performance plateaus indicate limited enhancement. We therefore adopt four complementary views as the default configuration, optimally balancing precision gains against computational overhead.

4.3. Real-World Experiment

Setup. We deploy a Kinova Gen3 robotic arm with an Intel RealSense D435i depth camera, as shown in Fig. 6. The system detects 6-DoF grasp poses from RGB-D data, maps them to end-effector poses via hand-eye calibration, and executes trajectories autonomously. Our ROS-based platform (NVIDIA RTX 3090) processes real-time camera data and controls the robot via a network interface. We categorize the test scenarios into explicit and flexible instructions, further dividing the latter into single-target grasping and multi-target sequential grasping. For each setting, we conduct 35 repeated grasping experiments, with the success rate as the

#Views	CR@0.4 \uparrow	CR@0.3 \uparrow	CR@0.2 \uparrow	EMD \downarrow	EW-CFR \uparrow
2	0.9637	0.8485	0.4785	0.2600	0.3984
3	0.9666	0.8566	0.4962	0.2569	0.4103
4	0.9715	0.8797	0.5587	0.2520	0.4229
5	0.9700	0.8811	0.5581	0.2517	0.4239

Table 4. Performance analysis of virtual view quantity.

Methods	Explicit	Flexible	
		Single	Multiple
w/o CoT	51.4%	48.6%	44.0%
GraspCoT	54.2%	54.2%	46.7%

Table 5. Success rate (%) in real-world experiments.

evaluation metric. Partial successes (e.g., grasping one of two required targets) are counted as individual successes, with the rate defined as successful grasps / total targets.

Results. As shown in Tab. 5, when handling explicit grasping instructions, CoT-enhanced GraspCoT achieved a 54.2% success rate, outperforming its non-CoT counterpart by 2.8%. For flexible instructions, the CoT version maintained stable performance with a 5.6% advantage over the non-CoT variant, further validating its advantages in optimizing the grasp pose targeted at the object itself. Notably, in multi-target grasping scenarios where avoiding omissions of targets with weaker implicit relevance remains challenging, the CoT implementation improved success rates from 44.0% to 46.7%, which is a 2.7% enhancement, verifying its effectiveness in handling complex requirements.

5. Conclusion and Discussion

This work presents a novel 6-DoF grasp detection framework that bridges the gap between flexible instructions and robotic grasping. By integrating a physical-property-oriented CoT mechanism with QA-guided hierarchical reasoning, our method aligns target parsing, physical property analysis, and grasp action selection with grasp detection. The proposed multimodal LLM architecture, augmented by 3D-aware visual-textual token fusion, achieves state-of-the-art performance on a newly established benchmark IntentGrasp for diverse verbal-command grasping. Real-world validations further underscore its practical applicability. While our method demonstrates robust performance under flexible instructions, some limitations persist. For instance, grasp predictions for slender objects (e.g., chopsticks) occasionally deviate from mass centers despite leveraging shape descriptors, suggesting the need to model latent physical properties beyond geometric cues. Additionally, future directions include extending the framework to sequential video inputs for dynamic scenes, such as grasping moving targets, which would require video-based grasping benchmarks to advance real-world adaptability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62332016) and the Key Research Program of Frontier Sciences, CAS (No. ZDBS-LY-JSC001).

References

- [1] Shehan Caldera, Alexander Rassau, and Douglas Chai. Review of deep learning methods in robotic grasp detection. *Multimodal Technologies and Interaction*, 2(3):57, 2018. 1
- [2] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024. 2
- [3] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *CoRR*, abs/2405.10370, 2024. 4
- [4] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian D. Reid. 3d-llava: Towards generalist 3d llms with omni superpoint transformer. *CoRR*, abs/2501.01163, 2025. 2
- [5] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018. 2
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1, 3
- [7] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021. 2
- [8] 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*, pages 11444–11453, 2020. 2
- [9] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023. 2
- [10] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *CoRR*, abs/2403.11401, 2024. 4
- [11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2
- [12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 4
- [13] Shiyu Jin, Jinxuan Xu, Yutian Lei, and Liangjun Zhang. Reasoning grasping via multimodal large language model. *Conference on Robot Learning*, 2024. 2, 3
- [14] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018. 1
- [15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5
- [16] Quoc V Le, David Kamm, Arda F Kara, and Andrew Y Ng. Learning to grasp objects with multiple contact points. In *2010 IEEE International Conference on Robotics and Automation*, pages 5062–5069. IEEE, 2010. 1
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 3
- [18] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. In *European Conference on Computer Vision*, pages 109–127. Springer, 2024. 1, 3
- [19] 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)*, pages 7507–7513. IEEE, 2024. 1, 3
- [20] Samuel Li, Sarthak Bhagat, Joseph Campbell, Yaqi Xie, Woojun Kim, Katia P. Sycara, and Simon Stepputtis. Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, pages 10527–10534. IEEE, 2024. 3
- [21] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007, 2017. 5
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [23] Teli Ma, Zifan Wang, Jiaming Zhou, Mengmeng Wang, and Junwei Liang. GLOVER: generalizable open-vocabulary

- affordance reasoning for task-oriented grasping. *CoRR*, abs/2411.12286, 2024. 3
- [24] Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. *arXiv preprint arXiv:1804.05172*, 2018. 1
- [25] Douglas Morrison, Peter Corke, and Jürgen Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters*, 5(3):4368–4375, 2020. 2
- [26] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2901–2910, 2019. 2, 6
- [27] Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, et al. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, 39(5):3994–4015, 2023. 1
- [28] Nghia Nguyen, Minh Nhat Vu, Baoru Huang, An Vuong, Ngan Le, Thieu Vo, and Anh Nguyen. Lightweight language-driven grasp detection using conditional consistency model. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13719–13725. IEEE, 2024. 3
- [29] Toan Nguyen, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le, Thieu Vo, Bac Le, and Anh Nguyen. Language-conditioned affordance-pose detection in 3d point clouds. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3071–3078. IEEE, 2024. 1, 3, 6
- [30] 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 *European Conference on Computer Vision*, pages 363–381. Springer, 2024. 2, 3, 5, 6
- [31] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7587–7597, 2024. 3
- [32] Yaoyao Qian, Xupeng Zhu, Ondrej Biza, Shuo Jiang, Linfeng Zhao, Haojie Huang, Yu Qi, and Robert Platt. Thinkgrasp: A vision-language system for strategic part grasping in clutter. *CoRR*, abs/2407.11298, 2024. 2, 3
- [33] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 2
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [35] Philipp Schmidt, Nikolaus Vahrenkamp, Mirko Wächter, and Tamim Asfour. Grasping of unknown objects using deep convolutional neural networks based on depth images. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6831–6838. IEEE, 2018. 1
- [36] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020. 1, 2
- [37] Qiang Sun, Haitao Lin, Ying Fu, Yanwei Fu, and Xiangyang Xue. Language guided robotic grasping with fine-grained instructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1319–1326. IEEE, 2023. 1, 3
- [38] Chao Tang, Dehao Huang, Wenqi Ge, Weiyu Liu, and Hong Zhang. Graspopt: Leveraging semantic knowledge from a large language model for task-oriented grasping. *IEEE Robotics Autom. Lett.*, 8(11):7551–7558, 2023. 1, 3
- [39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 2, 3, 5
- [40] Georgios Tzifas and Hamidreza Kasaei. Towards open-world grasping with large vision-language models. *arXiv preprint arXiv:2406.18722*, 2024. 3
- [41] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5923–5930. IEEE, 2023. 2, 6
- [42] 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*, pages 15964–15973, 2021. 2
- [43] Lirui Wang, Yu Xiang, Wei Yang, Arsalan Mousavian, and Dieter Fox. Goal-auxiliary actor-critic for 6d robotic grasping with point clouds. In *Conference on Robot Learning*, pages 70–80. PMLR, 2022. 2
- [44] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023. 5
- [45] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [46] Yi-Lin Wei, Jian-Jian Jiang, Chengyi Xing, Xian-Tuo Tan, Xiao-Ming Wu, Hao Li, Mark Cutkosky, and Wei-Shi Zheng. Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems*, 37:46881–46907, 2024. 3

- [47] 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 *European Conference on Computer Vision*, pages 357–375. Springer, 2024. 6
- [48] William Xie, Maria Valentini, Jensen Laverling, and Nikolaus Correll. Deligrasp: Inferring object properties with llms for adaptive grasp policies. In *8th Annual Conference on Robot Learning*, 2024. 3
- [49] Jinxuan Xu, Shiyu Jin, Yutian Lei, Yuqian Zhang, and Liangjun Zhang. Rt-grasp: Reasoning tuning robotic grasping via multi-modal large language model. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7323–7330. IEEE, 2024. 3
- [50] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024. 2
- [51] Houjian Yu, Mingen Li, Alireza Rezazadeh, Yang Yang, and Changhyun Choi. A parameter-efficient tuning framework for language-guided object grounding and robot grasping. *CoRR*, abs/2409.19457, 2024. 3
- [52] Hanbo Zhang, Jian Tang, Shiguang Sun, and Xuguang Lan. Robotic grasping from classical to modern: A survey. *arXiv preprint arXiv:2202.03631*, 2022. 1
- [53] Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3d-zero: An agent for zero-shot 3d understanding. In *European Conference on Computer Vision*, pages 186–202. Springer, 2024. 4
- [54] Binglei Zhao, Hanbo Zhang, Xuguang Lan, Haoyu Wang, Zhiqiang Tian, and Nanning Zheng. Regnet: Region-based grasp network for end-to-end grasp detection in point clouds. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13474–13480. IEEE, 2021. 6
- [55] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8995–9006, 2025. 2
- [56] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering llms with 3d-awareness. *CoRR*, abs/2409.18125, 2024. 2, 6