A Generic Framework for Damage-Free Grasping of Delicate Produce Using LLMs and Volume Estimation

Ziye Zhang[b,1], Xiaoyu Xia[b,1], Yuhao Jin[a,b], Qizhong Gao[a,b], Lin Qiao[a,b], Jinglei Chen[b], Yong Yue[b], ShanLiang Yao[c], Xiaohui Zhu[a,*]

[a] *Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK*
[b] *School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*
[c] *School of Information Engineering, YanCheng Institute of Technology, Yancheng 224051, China*

Abstract

Despite notable progress in agricultural robotics, achieving stable, adaptive, and damage-free grasping of irregular, fragile, and diverse produce remains a key unsolved challenge. Traditional approaches rely on geometric heuristics or fixed-force strategies, which cannot accurately model or adapt to the complex physical properties of various produce, often causing instability or damage. To the best of our knowledge, this paper proposes the first adaptive damage-free grasping framework that integrates large language models (LLMs) with multimodal perception to produce. We develop an intelligent framework combining semantic understanding from visual images, geometric awareness from 3D point clouds, and LLMs commonsense knowledge to infer the minimal stable grasping force, enhancing safety and adaptability. To enhance accuracy and robustness, we introduce a point-cloud-based volume estimation module that directly leverages spatial geometry, thereby minimizing reliance on LLMs reasoning and significantly improving perception quality and estimation precision. Moreover, we design a hybrid multi-agent collaboration mechanism that efficiently coordinates perception, reasoning, and control agents, improving reasoning and execution in complex environments. Experiments on a custom dataset of 25 typical produce categories show our framework significantly outperforms state-of-the-art baselines in both volume estimation accuracy and damage-free grasp success rate, demonstrating excellent stability, robustness, and truly damage-free capability. This work validates the effectiveness of combining LLMs and multimodal perception in agricultural robotics and suggests a promising direction for future research toward intelligent and reliable grasping. Our framework introduction is at https://wyattzzzz.github.io/DamageFreeGrasp

Keywords: Multimodal Perception, Large Language Model, Agricultural Robot, Damage-free Grasping, Commonsense Physical Reasoning

1. Introduction

Autonomous robotic grasping is critical for agricultural tasks such as produce harvesting and pick-and-place operations (Zhang et al., 2020; Jin and Han, 2024). In recent years, deep learning-based grasp detection models, such as GGCNN (Ghasemzadeh et al., 2023), GR-ConvNet (Kumra et al., 2020), and VMGNet (Jin et al., 2024), can efficiently generate candidate grasping poses from visual or point cloud data, providing essential support for stable and robust autonomous grasping (Jin et al., 2025). However, existing grasping strategies still lack sufficient force adaptability, hindering the robot's ability to dynamically adjust the applied force during the grasping process (Li et al., 2024). First, existing methods lack a generalizable physical reasoning mechanism, making it difficult to adapt to delicate produce with diverse shapes and textures (Zhang et al., 2020; Huang et al., 2025). As a result, the grasping force cannot be precisely controlled, often leading to grasp failure or produce damage (Zheng et al., 2021).

To address such challenges, researchers have begun incorporating multimodal information, such as visual, tactile, and semantic data, to enhance the robustness and intelligence of grasping strategies (Li et al., 2023b, 2025). Among these, large language models (LLMs) have gained increasing attention for their potential in robotic grasping tasks (Rashid et al., 2023). LLMs can perform commonsense physical reasoning to infer key physical properties of produce, including volume, density, mass, friction coefficient with the gripper, and elastic modulus (Xie et al., 2024). This capability enables force-controlled grippers to adjust grasping force more precisely. Consequently, the integration of LLMs holds promise for equipping robots with greater adaptability and generalization when interacting with previously unseen produce (Xu et al., 2024). However, despite the physical reasoning capabilities of LLMs, relying solely on LLMs may lead to limited generalization and reduced accuracy in physical inference, particularly when dealing with non-standard produce shapes (Chu et al., 2025; Xie et al., 2024). To mitigate this limitation, point cloud data is incorporated to assist in volume estimation and correct inference bias, thereby reducing dependence on LLMs and improving the accuracy of grasp parameter prediction. This also enhances the framework's robustness and generalization to diverse produce morphologies.

In real-world agricultural environments, which are often complex and task-diverse, a single model may struggle to complete the full loop from perception to decision-making (Jin et al., 2025; Hu et al., 2025a). Therefore, this work further introduces a multi-agent collaboration mechanism that systematically integrates semantic reasoning and geometric estimation modules into a cohesive workflow designed for damage-free grasping. This mechanism significantly enhances the efficiency of multi-source information fusion and grasp parame-

*Corresponding author.
Email Address: Xiaohui.Zhu@xjtlu.edu.cn
[1]These authors contributed equally to this work.

ter decision-making in complex environments, offering an effective solution for achieving stable grasping of diverse agricultural produce.

Our contributions are summarized as follows:

(1) To the best of our knowledge, this study is the first to propose a grasp reasoning framework that integrates visual, semantic, and point cloud information to intelligently infer the minimal stable grasping force required for delicate produce, significantly improving safety and adaptability in soft grasping.
(2) By introducing point cloud-based volume estimation of produce, the framework reduces reliance on LLMs commonsense reasoning ability, thereby improving the accuracy of grasp parameter estimation and enhancing the framework's robustness and generalization across diverse produce morphologies.
(3) A multi-agent collaboration system is incorporated into the damage-free grasping of delicate produce, effectively improving the coordination efficiency and general applicability of the framework in complex perception and parameter reasoning tasks.

The structure of this paper is organized as follows: In the Section 2, we review the literature on the application of LLMs in robotic grasping, as well as grasping techniques for fragile produce. The Section 3 provides a detailed explanation of our proposed framework. The Section 4 presents our experimental setup and results. Finally, Section 5 summarizes the paper and discusses the limitations of our work along with potential future directions.

2. Related Work

Recent research has explored a variety of approaches to achieving intelligent, stable, and damage-free grasping of delicate produce, propelled by advances in LLMs and multimodal perception technologies. These developments have opened up new opportunities for integrating high-level semantic reasoning with low-level force control, improving the adaptability and safety of soft grasping systems. This section reviews related work from two perspectives: LLMs in robotic grasping and delicate grasping techniques.

2.1. Large Language Models in Grasping

In recent years, the application of LLMs in robotic grasping and manipulation tasks has garnered increasing attention. Benefiting from pretraining on large-scale corpora, LLMs not only possess strong capabilities in natural language understanding and generation but also exhibit a degree of commonsense physical reasoning (Wang et al., 2023). This enables them to infer implicit properties of produce, including volume, mass, and grasp stability, based on semantic labels. Building on this idea, RT-Grasp (Xu et al., 2024) proposes a "Reasoning Tuning" approach, which incorporates a structured reasoning stage during

training, allowing multimodal LLMs to generate numerical predictions, such as grasp pose, grounded in semantic context. This significantly enhances control accuracy in complex tasks. Similarly, SayCan (Ahn et al., 2022) combines language comprehension with physical constraint reasoning to guide robots through step-by-step execution of intricate tasks. However, these methods often exhibit instability when dealing with irregularly shaped produce with complex internal structures. This is primarily due to the lack of common semantic labels and insufficient training data related to such produce, leading to unreliable physical inference by LLMs.

To address these limitations, recent research has explored architectural innovations to enhance the synergy between multimodal perception and language understanding, which are critical for high-level reasoning in robotic systems. For example, Point-BERT (Yu et al., 2022) and PointCLIP (Zhang et al., 2022) employ Transformer-based architectures to encode point cloud data and align it with textual modalities, improving cross-modal semantic consistency. ULIP (Xue et al., 2023) further extends this idea by learning unified representations of images, point clouds, and text for open-vocabulary 3D understanding. Similarly, LLMs such as GPT-4o and QWen-Plus adopt a vision encoder coupled with a language model, enabling them to jointly reason over visual and textual modalities. These models exhibit remarkable generative and reasoning abilities in complex visual-textual scenarios, offering valuable insights for robotic decision-making. Collectively, these advancements highlight the potential of integrating semantic reasoning and geometric perception to enable more intelligent and context-aware decision-making in robotic grasping systems.

Existing approaches typically use LLMs to directly estimate produce volume or mass based on semantic priors and commonsense knowledge. However, these methods predominantly infer from semantic labels and are limited in their ability to perceive the precise geometric structure of individual instances. This often results in high estimation errors or unstable predictions, particularly when dealing with irregular produce that exhibits diverse and complex shapes.

2.2. Grasping Delicate Produce

In agricultural robotic grasping, the irregular shapes, fragile skins, and diverse varieties of produce present significant challenges, motivating extensive research on perception, strategies, and force control. In the domain of visual perception, a method (Chen et al., 2024) integrates 3D object detection with feedforward control, enabling precise grasping of thin-skinned produce. This effectively reduces damage during the grasping process. To further improve the stability of grasp decision-making, a three-finger soft gripper with integrated force feedback and slip detection was designed for apple harvesting (Chen et al., 2022), which significantly reduces bruising and slippage in field experiments. Considering the difficulties of detecting and grasping small produce, as well as their susceptibility to damage, another study (Visentin et al., 2023) proposed a soft gripper that combines RGB visual feedback and tactile sensing. This design not only improved the detection accuracy and grasp success rate for small pro-

duce but also minimized potential damage as much as possible. More recently, a soft gripper equipped with tactile sensing and slip detection was developed (Liu et al., 2024), demonstrating strong adaptability to delicate produce such as grapes and tomatoes in experimental settings. Despite these advances, current approaches often rely on limited visual or geometric cues, making it difficult to assess the underlying physical properties of produce, which are critical for achieving stable, damage-free grasping.

To address this challenge, researchers have developed methods that explicitly incorporate texture and firmness perception to assist grasp force modulation and enable selective, damage-free harvesting. For instance, an integrated vision-tactile sensor was developed to estimate produce texture and firmness during the picking process, thereby assisting in grasp force modulation (Ma et al., 2024). Furthermore, a soft gripper combined with vision-tactile modalities was proposed to implement a non-invasive firmness estimation mechanism (Lin et al., 2023), which was validated to be effective for handling a wide range of produce categories. In addition, a compact tandem-actuated gripper, which is capable of adaptively switching between suction and compliant mechanical grasping modes based on environmental context, enables selective and damage-free harvesting even under partially occluded conditions (Velasquez et al., 2024). These studies underscore the critical role of integrating texture and firmness perception with adaptive control strategies to achieve more intelligent, reliable, and context-aware damage-free harvesting in diverse agricultural environments.

Existing approaches in agricultural robotic grasping often rely on complex soft gripper designs that integrate multiple sensing modalities. While these designs enhance adaptability and reduce produce damage, they frequently require intricate mechanical structures and tight sensor integration, which increases system complexity, fabrication cost, and maintenance demands. On the perception side, traditional models for grasp planning and control are typically large and resource-intensive, relying on heavyweight neural architectures that limit real-time deployment on embedded or mobile robotic platforms.

3. Methodology

This section outlines the proposed framework for damage-free grasping of delicate produce. It is structured into four subsections, starting with a high-level overview of the framework and followed by detailed explanations of each module.

3.1. Framework Overview

This study introduces an innovative framework that combines 3D point cloud geometric perception with LLMs for the task of damage-free grasping of delicate produce. By combining visual perception and language-based reasoning, the framework significantly enhances adaptability and accuracy in the grasping process, as illustrated in Fig. 1. The overall workflow consists of three

main modules: (1) a 3D perception module, (2) a physical property reasoning module, and (3) a grasp strategy generation and execution module. In the 3D perception module, the system performs geometric shape perception using RGB images and depth maps, extracting produce 3D geometric information to support subsequent grasping decisions. The physical property reasoning module then combines semantic labels with geometric information, leveraging LLMs to infer physical properties such as density and friction coefficients, providing accurate parameters for force-controlled grasping. Finally, the grasp strategy generation and execution module formulates the grasp strategy based on the reasoning results and executes it with a force-controlled gripper to ensure stable, damage-free grasping. The seamless integration of these modules enables the system to perform stable grasping for delicate produce while effectively preventing damage, marking the first attempt to apply the combination of 3D geometric perception and LLMs in agricultural robotics.

In the 3D perception module, the system receives RGB images and depth maps of the produce. Combined with a semantic label set including apples, bananas, grapes, and others, multimodal feature fusion and semantic-guided querying are performed using the multimodal fusion encoder and Language-guided Query Selection (LGQS) modules. The fused features are then processed by the Multimodal Decoder, which generates 2D coordinates of the target produce, enabling precise localization and semantic label alignment. Next, the identified produce mask, along with its depth map, is processed through Planar Surface Extraction and Non-Planar Segmentation, extracting 3D contour information and forming a complete point cloud and shape representation of the produce. A coarse-grained volume estimation of produce is then obtained, which will be input to the grasp execution module.

In the physical property reasoning module, the system utilizes the semantic label extracted from the 3D perception module, along with information about the gripper material, to drive three LLMs-powered agents called FricAgent, AdjustAgent, and DensAgent, which respectively infer the friction coefficient, volume correction factor, and density estimate of the produce.

In the grasping strategy generation and execution module, the module determines the minimum stable grasping force using the Contact Force Estimator, based on the coarse volume and inferred physical attributes of the produce, thereby ensuring safe manipulation without damage. Simultaneously, the 2D coordinates and depth map of the produce are processed by the Grasp Pose Generator, an extension of VMGNet (Jin et al., 2024), to predict candidate grasp poses. These poses are jointly optimized with the estimated minimum force and then transmitted to the robotic arm controller for execution. The grasping action is carried out by a force-controlled gripper, which maintains a stable hold while preventing excessive compression or surface damage, ultimately achieving reliable and damage-free grasping.

By combining the commonsense reasoning capabilities of language models with the geometric perception capabilities of point cloud and image data, this framework offers an adaptable and generalizable solution for damage-free grasping across a wide variety of produce types in flexible agricultural robotics.

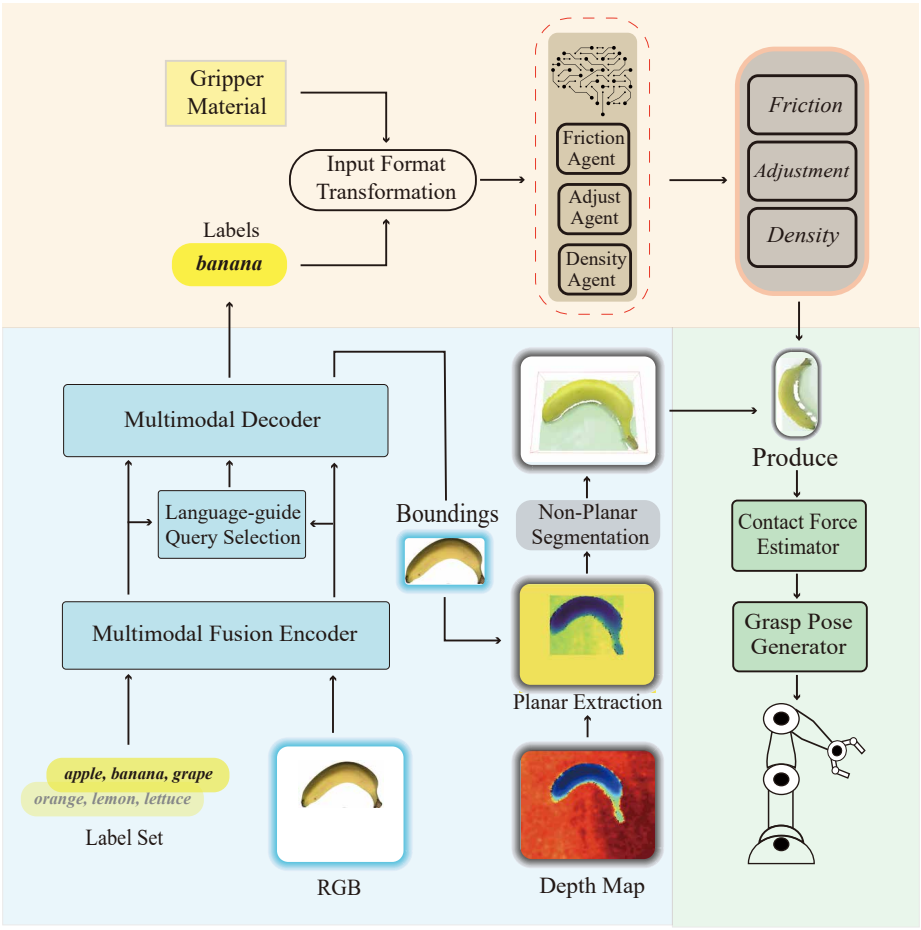

Fig. 1. Architecture overview of the proposed framework, composed of three main modules, (1) 3D perception module in blue, (2) LLM-based commonsense physical reasoning in orange, (3) grasp execution in green.

3.2. 3D Perception Module

Accurate 3D perception is essential for reliable produce localization and manipulation in unstructured agricultural environments. To this end, the perception module integrates both semantic and geometric cues to achieve robust detection and volume estimation. As illustrated in the following subsections, we first employ a visual-language-based detection pipeline to localize target produce using RGB images and category prompts. Subsequently, we introduce a depth-enhanced, geometry-aware estimation process that leverages masked point clouds and commonsense reasoning to infer volumetric properties of the localized produce.

3.2.1. Multimodal Instance Detection

Grounding DINO is an open-vocabulary object detection model derived from the DETR family, enabling flexible detection guided by textual prompts. Its architecture comprises four primary components: an Image Backbone, a Text Backbone, a Feature Enhancer that employs a dual-stream Transformer to fuse and enhance visual and textual features, and the core modules, LGQS and Cross-Modality Decoder.

As shown in Fig. 2, to facilitate effective multimodal interaction, our framework introduces a multimodal fusion encoder. The core function of the multimodal fusion encoder is to fuse image and text features. Initially, text and image features are processed through a feature extractor, after which they enter the cross-modality attention layer to enhance the interaction between the two modalities. Following that, self-attention and deformable self-attention layers are applied to further optimize the feature alignment, ensuring that the model can extract and integrate effective information from different modalities, providing strong feature representations for downstream tasks. This process ensures efficient multimodal information fusion, allowing the model to make accurate judgments in complex scenarios.

The multimodal query-based decoder then takes the aligned features and generates 2D coordinates of the target produce, enabling precise localization. The decoder further optimizes the aligned features and assists in semantic label alignment, ensuring the accuracy and reliability of the feature representation. In addition, the decoder is not just a feature processor; it also integrates semantic and geometric information, providing more accurate input for downstream tasks, such as grasp pose estimation. With this powerful encoder-decoder architecture, our framework significantly enhances grasp prediction for irregular produce categories, ensuring both accuracy and stability in grasp pose estimation.

Following multimodal feature fusion, the LGQS module plays a pivotal role in generating semantically aligned queries for object detection. Such module generates semantically guided Cross-Modality Queries by combining image features with language prompts. These queries are then passed to the decoder to predict the object Bounding Box. This design enables the model to align semantics and localize regions based on arbitrary natural language descriptions,

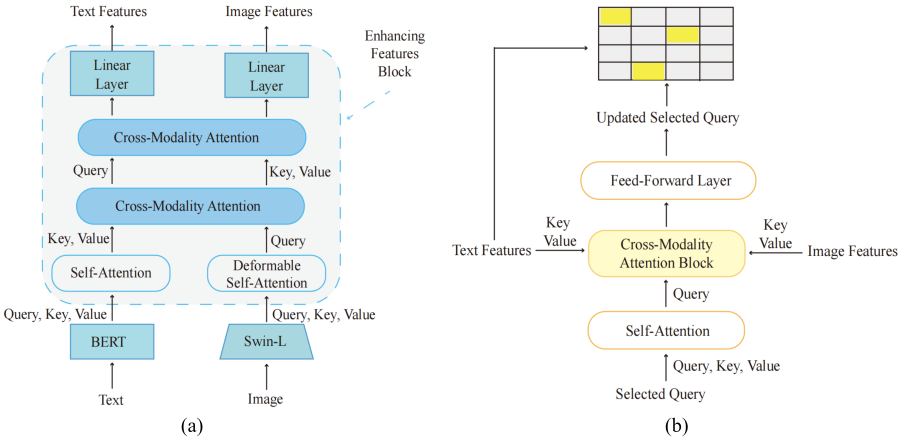

Fig. 2. Illustration of the encoder and decoder. (a) shows the Multimodal Feature Fusion Encoder, and (b) shows the multimodal query-based decoder.

equipping it with zero-shot detection capabilities (Hu et al., 2025b).

While the encoder-decoder modules are responsible for feature alignment and spatial localization, the LGQS module plays a critical role in bridging semantics with vision. Its internal computation process is illustrated in Fig. 3. The module takes as input the image features and text features produced by the respective backbones and computes a cross-modal similarity matrix, where each element reflects the semantic affinity between an image token and a text token. To identify the most semantically relevant regions in the image, the highest relevance score of each image token with respect to the entire text prompt is calculated, producing a relevance vector. Based on this vector, the image tokens with the highest semantic scores, corresponding to the most text-aligned visual regions, are selected and used as input queries for the cross-modality decoder.

In our system, we inherit the core architecture of Grounding DINO and adopt its multimodal fusion encoder and LGQS module to produce detection based on RGB images and a predefined label set. The LGQS module computes a similarity matrix between image and text features and selects the most relevant image tokens by identifying the maximum semantic response. The index selection process can be formally expressed as follows:

$$\mathbf{I}_{N_q} = \mathrm{Top}_{N_q}\left(\mathrm{Max}^{(-1)}\left(\mathbf{X}_I\mathbf{X}_T^{\top}\right)\right). \tag{1}$$

where Top picks the most relevant image tokens, N_q denotes the number of LGQS module output indices fixed at 900, $\mathrm{Max}^{(-1)}$ extracts each image token's maximum relevance to the text, $\mathbf{X}_I \in \mathbb{R}^{N_I \times d}$ is the image features, $\mathbf{X}_T \in \mathbb{R}^{N_T \times d}$ is the text features, N_I and N_T denote the number of image and text tokens, and d is the shared embedding dimension.

To improve the system's flexibility and adaptability, a user-defined produce list mechanism is introduced, enabling dynamic adjustment of detection cat-

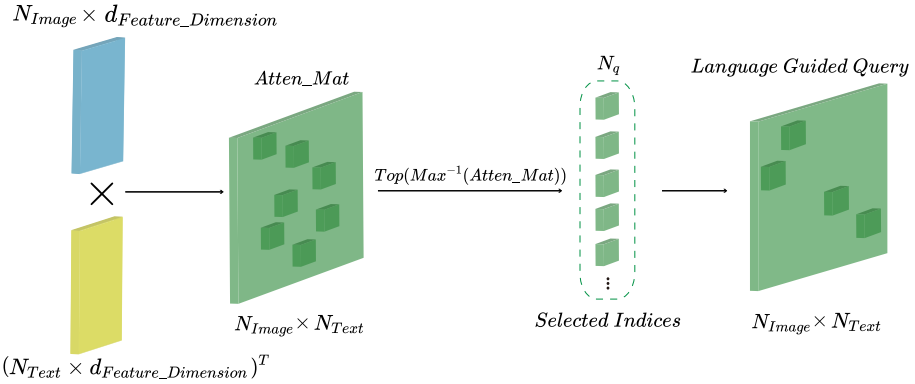

Fig. 3. Brief schematic of the LGQS module, where image and text features are combined to form an attention matrix, and the image token indices with the highest semantical score are selected to generate language-guided queries.

egories in response to specific grasping requirements. For example, produce categories such as blueberries or spinach can be dynamically added, and irrelevant categories can be filtered out based on specific task requirements. This design improves the system's practical applicability and robustness in dynamic agricultural environments.

3.2.2. Geometry-aware Volume Estimation

To improve volume estimation accuracy and enhance system adaptability, this study introduces a multi-stage estimation pipeline that combines visual perception with LLM-based commonsense reasoning. As shown in Fig. 4, the pipeline first acquires RGB images and the corresponding 2D coordinators of the target produce. Based on these coordinators, a 2D masked image is generated and aligned with the depth map to isolate the region of interest, which is then transformed into point cloud data using the camera intrinsic parameter matrix. Planar Segmentation removes background surfaces, and the remaining points undergo Point Cloud Partitioning and Outlier Removal before Bounding Box Extraction provides a coarse geometric volume estimate. The original depth map is progressively refined through masking, point cloud filtering, and geometric fitting, effectively isolating the target produce, removing background interference, and constructing an enclosing bounding box. While this geometric enclosure often leads to overestimation due to redundant space, these visual results clearly illustrate how geometric information is extracted and utilized in the volume estimation process.

Regarding prompt design and further data processing, we observed that using more concrete and descriptive natural language expressions, for example "a piece of lettuce" outperformed abstract labels like "lettuce" by providing clearer semantic cues. This improvement enabled precise localization of target produce within cluttered scenes and laid a more robust visual groundwork for

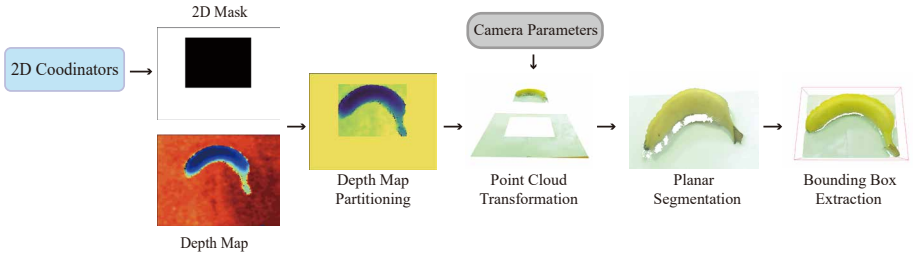

Fig. 4. Multi-stage pipeline for commonsense reasoning, combining 2D masks generation and depth maps converted to point clouds, followed by filtering and bounding box extraction.

downstream grasp parameter reasoning. In light of Grounding DINO's image-text pair input format, we predefine a set of common produce categories as a default token pool for category matching during inference. With this design, the module requires only a single RGB image to automatically perform produce detection, without the need for additional textual input.

Once the target region is localized, its 2D mask is aligned with the corresponding depth map to isolate the produce's depth profile. This masked depth data is then converted into a partial point cloud using the camera's intrinsic parameters. Subsequent processing includes planar segmentation to remove background surfaces, Euclidean clustering to separate the target from nearby objects, and statistical outlier removal to reduce noise. The cleaned point cloud is then fitted with an oriented bounding box to obtain coarse geometric measurements, which serve as the basis for initial volume estimation.

Together, these two modules provide a complete 3D perception pipeline that enables both robust instance detection and geometry-aware volume estimation, forming the basis for downstream grasping and manipulation.

3.3. LLM-Based Reasoning Module

The module first receives the semantic label of the target produce from the 3D perception module, such as banana, together with the gripper material specification. These inputs are processed by an Input Format Transformation stage, which re-formats the information for reasoning by a set of LLMs-driven agents, including FrictionAgent, AdjustAgent, and DensityAgent.

Without directly accessing point cloud data, the AdjustAgent leverages the semantic label to infer a canonical geometric primitive, such as an ellipsoid, cone, or cylinder-based on shape priors learned during pretraining. From this inferred geometry, the agent estimates the occupancy ratio within the bounding box, yielding a volume correction factor, which is used to refine the coarse volume estimate resulting in a more accurate produce volume. DensityAgent and FrictionAgent infer the produce density based on semantic knowledge and prior learning and estimate the friction coefficient between the produce surface and the gripper contact material, respectively.

Finally, the module further leverages LLMs to infer the produce's density and friction coefficient. These parameters, combined with Coulomb's law of friction, are used to calculate the minimum grasping force required for stable manipulation. First, the formula used to calculate the mass of tacet produce is as follows:

$$m = \rho \cdot k \cdot V \tag{2}$$

where ρ is the produce's density, k is the volume correction factor, and V is the estimated coarse volume. Then, the grasp force calculation formula is as follows:

$$F_{\text{grasp}} = \mu \cdot m \cdot g \tag{3}$$

where F_{grasp} is the required grasping force, μ is the friction coefficient and g is the gravitational acceleration valued as 9.81 m/s^2. The estimated force is finally passed as a control parameter to the force-controlled gripper module.

By incorporating high-level semantic reasoning and physical property inference, the proposed framework substantially improves the accuracy of volume estimation and grasping force prediction. This enables the force-controlled gripper to execute stable, damage-free manipulation of complex, irregular, or deformable produce.

3.4. Grasping Execution Module

VMGNet is a grasp representation framework that generates 2D planar grasping predictions from RGB or RGB-D images. Its core architecture comprises the Visual State Space (VSS) and the Fusion Bridge Module (FBM), which together enable efficient grasp pose inference and planning. The VSS component compresses image information to reduce computational complexity, while the FBM enhances the model's ability to capture multi-scale features, thereby improving the accuracy and robustness of grasp point generation (Jin et al., 2024).

Although VMGNet exhibits strong grasp representation capabilities under 2D image inputs, its original design primarily targets structurally stable general objects. This makes it less adaptable to agricultural scenarios, where produce often exhibits soft textures, irregular morphologies, and unstructured features. To enhance VMGNet's performance in damage-free agricultural grasping tasks, we implement targeted optimizations to its input pipeline and grasping strategy.

In the case of VMGNet, the Region of Interest (ROI) is directly defined as a graspable region and parameterized using a grasp rectangle representation, expressed as follows:

$$ROI = (x, y, \theta, h, w) \tag{4}$$

where (x, y) denotes the center coordinates of the grasp rectangle, θ represents the grasp angle relative to the horizontal axis, and h, w correspond to the height and width of the proposed grasp. These parameters are regressed directly from fused multi-scale features via the decoder, without requiring explicit cross-modal alignment or segmentation.

We introduce the LGQS module from Grounding DINO to extract the semantic label of the target produce and generate a precise 2D mask for localizing its boundary within the image. This semantic-guided mechanism ensures that the detection region is both accurate and complete, mitigating grasp errors caused by false detections or background interference. We then align the detected produce region with its corresponding depth map to construct a fused image input focused on the target produce, significantly improving the quality of grasp candidate regions and reducing interference from non-target areas. To demonstrate the practical effectiveness of VMGNet in real-world agricultural scenarios, we visualize the predicted grasp poses for various delicate produce. As shown in Fig. 5, the VMGNet framework accurately identifies ROI for grasping, as depicted by the bounding boxes on the produce. These bounding boxes guide the Contact Force Estimator to predict stable grasp forces by ensuring that the contact force is appropriate for each object's physical characteristics, thereby preventing damage during manipulation.

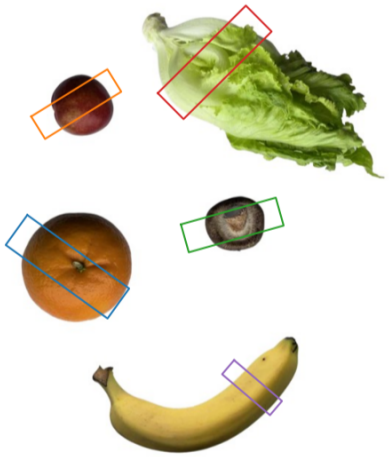

Fig. 5. Graspable regions predicted by VMGNet for various agricultural produce for minimal manipulation damage.

Considering the sensitivity of force control in handling delicate produce, we design a Contact Force Estimator based on LLM-inferred physical attributes. This module utilizes key parameters, such as density, friction coefficient, and a volume correction factor inferred through language model reasoning, to estimate the minimum stable contact force required for damage-free grasping. The estimated force guides the grasp planning module in determining whether a grasp action complies with damage-free constraints, thereby preventing produce damage caused by excessive gripping force or inaccurate object understanding.

4. Experiment

This experiment aims to evaluate the proposed multimodal damage-free grasping framework in terms of LLMs reasoning-based performance and grasp force control, with a focus on comparing different estimation pathways and assessing the effectiveness of LLMs in physical property reasoning and grasp feasibility.

4.1. Experimental Setup and Data Collection

We implement the proposed framework following the architecture described in Section 3, which integrates a 3D perception module, an LLM-based physical reasoning module, and a grasp execution module. We conduct all experiments using four NVIDIA RTX 4090 GPUs with each 24GB of memory. The implementation is based on PyTorch 2.0.1 and Python 3.8.20. All text embedding processing and tokenization steps are performed using BERT. The training and evaluation are carried out on Ubuntu 20.04.

To enhance the detection of agricultural produce categories insufficiently represented in the pretrained model, we fine-tuned Grounding DINO on an augmented dataset combining Objects365 V1 (Shao et al., 2019), OpenImages V6 (Kuznetsova et al., 2020), and GoldG(Kebe et al., 2021). RGB and depth data are captured using an Intel RealSense D435 camera, producing RGB-D images and point clouds for 3D bounding box construction and mass estimation. For each produce sample, high-precision electronic scales are used to obtain ground-truth mass measurements.

The experimental dataset consists of 25 typical produce categories, covering berries, fruits, mushrooms, leafy vegetables, hollow produce, and part of the produce structure. These samples exhibit diverse shapes, densities, surface textures, and structural stiffness, reflecting the challenges of real-world agricultural grasping. Representative examples are shown in Fig. 6, and the full category list is provided in the Appendix.

4.2. Evaluation Metrics

To comprehensively evaluate the effectiveness of the proposed multimodal damage-free grasping framework, we design three experimental tasks: comparative reasoning performance analysis across different LLMs, mass estimation accuracy assessment, and grasp force feasibility validation, each with well-defined quantitative metrics.

4.2.1. Reasoning-based Tasks Metrics

To ensure consistency and comparability across reasoning-based tasks, we adopt a unified set of evaluation metrics for both the mass estimation task and the LLMs reasoning capability comparison task. Specifically, we assess model performance along two dimensions: accuracy and stability.

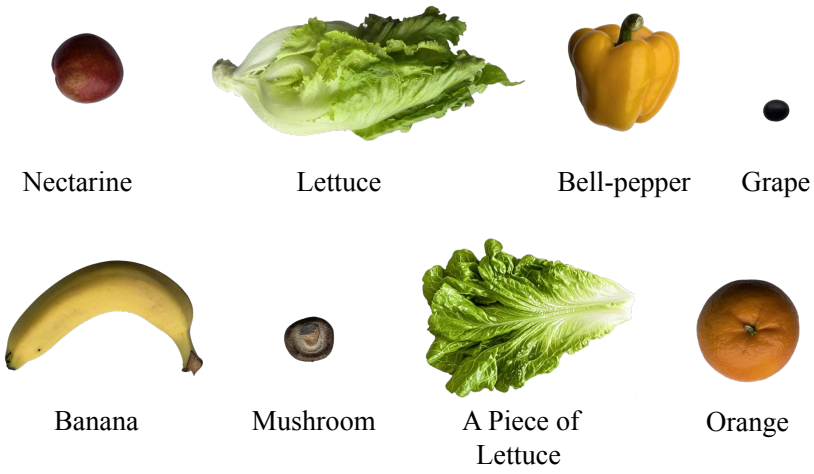

Fig. 6. Representative samples from the experimental dataset, covering diverse produce categories with varying shapes, textures, and physical properties.

Accuracy is quantified using the Mean Absolute Percentage Error (MAPE) as follows:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right|\right) \times 100\% \tag{5}$$

where $\hat{y}_i$ is the predicted value of the i-th sample, y_i is the ground truth value of the i-th sample, and N represents the total number of samples. Lower MAPE indicates higher estimation accuracy. Stability of predictions across samples is measured by the Standard Deviation (STD) as follows:

$$STD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{6}$$

where x_i is the i-th observed value, $\bar{x}$ is the sample mean, N is the total number of samples. Smaller STD reflects more consistent predictions.

The ground-truth mass of each sample is measured using a high-precision electronic scale, serving as the baseline for evaluation. Estimation accuracy is quantified by the MAPE, where lower values indicate higher accuracy. To ensure numerical stability, a minimum denominator threshold is applied in the MAPE computation to avoid division by very small ground-truth values. Prediction stability is assessed by the STD of the errors.

4.2.2. Grasp Feasibility Evaluation Metrics

To assess the accuracy of predicted grasping forces in a safety-aware manner, we adopt a tolerance-based hit rate metric. For each produce category c, we denote the reference *success force* of sample i as $F^{(i)}$. Based on empirical

fragility considerations, an asymmetric tolerance interval is defined as

$$\left[L^{(i)}, U^{(i)}\right] = \left[(1-\delta_c^-)F^{(i)},\ (1+\delta_c^+)F^{(i)}\right], \tag{7}$$

where δ_c^- and δ_c^+ represent the category-specific lower and upper tolerance ratios, respectively. Given a calibrated model prediction $\hat{F}^{(i)}$, a hit indicator is defined as

$$h^{(i)} = \begin{cases} 1, & \text{if } L^{(i)} \le \hat{F}^{(i)} \le U^{(i)}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The **hit rate** for model m within category c is then computed as the empirical average:

$$\text{HitRate}_{m,c} = \frac{1}{N_c} \sum_{i=1}^{N_c} h^{(i)}, \tag{9}$$

where N_c denotes the number of samples in category c.This metric reflects the proportion of predictions that fall into the physically safe and effective grasping force interval. It complements continuous error metrics (e.g., MAPE) by directly quantifying the likelihood that a model's predictions are simultaneously non-damaging and sufficient for successful grasping.

Together, these metrics provide a comprehensive and task-relevant evaluation of both the physical property reasoning and the downstream force-controlled grasp execution capabilities of the proposed system, reflecting its practicality and reliability in real-world agricultural environments.

4.3. Perception-Fusion Mass Estimation

To systematically evaluate the contribution of geometric perception to mass estimation, we compare two paradigms: (i) a **reasoning-only** setting, where the LLMs directly infer the mass of the target produce in kilograms from its semantic label without any geometric perception, and (ii) our proposed **perception-fusion** framework, which explicitly incorporates 3D information. In the latter, the module first derives a 3D bounding box and corresponding volumetric estimate from RGB-D observations, then applies an LLM-inferred shape correction factor to refine the raw volume, and subsequently combines the corrected volume with an LLM-inferred density to obtain the final mass prediction. This design enables a clear isolation of the benefits brought by integrating geometric structure into the reasoning process.

Among the models assessed, such as Deepseek-r1, Deepseek-v3, qwen-plus, QWen-2.5 (14B), QWen-max, ChatGPT-4o, and Claude-3.5-Sonnet, the 2.5 (14B) model represents the smallest publicly released system available at the time of study. While its limited parameter count constrains baseline reasoning capability, we mitigate this by performing domain-specific fine-tuning on the Camel dataset (Li et al., 2023a). This specialized adaptation allows the model to acquire inductive biases tailored for volumetric and density reasoning, forming the backbone of our final framework. The results highlight not only the value of geometric integration but also the potential of compact LLMs

when reinforced through carefully curated training, with experimental results across different LLMs summarized in Table 1 and further details provided in the Appendix.

Table 1
Average MAPE and STD in produce mass estimation using reasoning-only vs. perception-fusion methods, with model parameter sizes (B), on a custom dataset of seven produce categories.

Models	**Reasoning-only**		**Perception-fusion**		**Parameter (B)**
	MAPE	**STD**	**MAPE**	**STD**	
ChatGPT4o	792.842	1022.937	88.349	95.469	220
Claude-3.5-Sonnet	142.843	2126.330	81.729	99.335	$\geq$ 175
Deepseek R1	312.601	284.372	190.326	220.491	671
Deepseek V3	170.083	125.146	82.123	69.441	671
QWen-max	301.782	249.868	81.2889	61.204	$\geq$ 72
QWen-plus	227.963	234.358	83.822	82.937	$\geq$72
Ours	376.804	340.810	45.809	23.590	14

In the task of produce mass estimation, our experimental results systematically demonstrate the pronounced advantages of incorporating point cloud assistance. First, in terms of overall error levels, the introduction of geometric information from point clouds significantly improves both prediction accuracy (MAPE) and stability (STD) across almost all evaluated models. For instance, ChatGPT4o under the reasoning-only setting exhibits an average MAPE as high as 792.84% with an STD exceeding 1000, indicating severe instability when relying solely on semantic inference across diverse produce categories. In contrast, under the perception-fusion setting, its MAPE drops to 88.35% and STD to 95.47, a reduction of nearly 90%, clearly demonstrating that three-dimensional geometric cues effectively compensate for the inherent limitations of purely semantic reasoning. Similarly, the MAPE of Claude-3.5-Sonnet decreases from 142.84% to 81.73%, highlighting that even large-scale models benefit substantially from the integration of geometric perception.

Second, the advantages of point cloud integration are even more pronounced for medium- and small-scale models. For example, Qwen-plus and Qwen-max under the reasoning-only paradigm exhibit average errors in the 200%-300% range, but these are reduced to 83.82% and 81.29%, respectively, under the perception-fusion setting, corresponding to relative improvements of more than 60%. This demonstrates that geometric structural information serves as a highly effective complement for models with limited parameter scales, enabling them to achieve performance levels comparable to or even surpassing those of much larger models. Of particular note is our 14B-parameter model (Ours), which achieves the lowest overall error with the lowest MAPE(45.81%) and the smallest STD (23.59) under the perception-fusion setting. This performance not only far exceeds that of its reasoning-only counterpart, which leads to relatively

high MAPE(3776.804%) and STD(340.810), but also substantially outperforms larger models such as ChatGPT4o and Claude-3.5-Sonnet. These findings underscore that through the synergy of geometric perception and semantic reasoning, even medium-scale models can deliver performance superior to state-of-the-art large-scale models in complex produce mass estimation tasks.

More importantly, from an agricultural application perspective, different produce categories exhibit substantial variations in geometric morphology and physical properties, and the point cloud-assisted method demonstrates strong stability across these heterogeneous classes. For instance, in categories such as leafy vegetables and parts of produce structures, reasoning-only methods often suffer from large fluctuations in estimation errors, whereas the perception-fusion strategy effectively reduces both bias and variance, thereby substantially lowering the risk of damage in non-destructive grasping. For fragile produce such as berries and mushrooms, point cloud perception not only improves the precision of volumetric estimation but also provides more reliable priors to guide subsequent grasping strategies. This cross-category robustness carries direct significance for automated harvesting, sorting, and packaging in agricultural production. Detailed per-category statistics supporting these observations are provided in the Appendix.

4.4. Reasoning Performance Comparison Across LLMs

In damage-free grasping tasks, accurate force control strategies critically depend on reliable estimation of the target object's mass and frictional properties. As mass is not directly measurable during inference, commonsense reasoning informed by visual cues and semantic labels becomes essential for approximating physical properties. Therefore, this section focuses on comparing the performance of different LLMs in commonsense reasoning and evaluating their ability to support grasp control strategies.

This study selects seven state-of-the-art LLMs, including GPT-4o, Qwen-plus, Qwen-max, Qwen-2.5 with fine-tuned (Ours), Deepseek-r1, Deepseek-v3, and Claude-3.5-Sonnet. Each model receives standardized semantic labels of produce as input and outputs both density and volume correction factors. These values are combined with volume estimates obtained from RGB-D point cloud-derived 3D bounding boxes, enabling mass calculation. Based on the inferred mass and friction coefficients, the system applies Coulomb's friction model to calculate the theoretical minimum grasping force. This force is then applied using a force-controlled gripper equipped with closed-loop control capabilities. The system records whether the grasp is successful and whether any deformation or damage occurred.

In parallel, the MAPE is used to quantitatively compare the mass estimation accuracy of the seven LLMs across seven representative produce structural categories. The results are summarized in Table 2.

Among all evaluated models under the perception-fusion setting, Ours achieves the lowest overall average MAPE of 45.81%, substantially outperforming larger-scale models such as ChatGPT-4o (88.35%) and Claude-3.5-Sonnet (81.73%).

Table 2
Comparison of MAPE across LLMs on seven representative produce categories under point cloud-assisted mass estimation.

	MAPE						
	Deepseek		Qwen		ChatGPT	Claude	Ours
	r1	v3	max	plus	4o	3.5-Sonnet	
1	46.889	**36.958**	41.558	69.246	50.002	45.408	69.393
2	47.579	**42.513**	55.664	63.831	47.563	49.745	52.192
3	825.177	**21.704**	49.333	31.094	34.266	37.068	22.008
4	35.135	36.565	42.739	42.289	36.876	**35.1**	47.676
5	335.094	136.478	180.718	135.478	**56.999**	171.998	78.484
6	**12.236**	250.664	144.071	233.544	383.41	201.499	41.597
7	30.171	49.982	54.939	11.271	9.324	31.286	**9.311**
All	190.326	82.123	81.289	83.822	88.349	81.729	**45.809**

Note: 1 stands for Berries; 2 for Round Fruits; 3 for Elongated Fruits; 4 for Elliptical Fruits; 5 for Leafy Vegetables; 6 for Part of Produce Structure; and 7 for Mushroom Family.

This demonstrates its strong generalization capability and robustness, highlighting its suitability as a compact yet reliable inference backbone for multi-category produce grasping systems. In contrast, Deepseek-r1 exhibits the weakest generalization with an average MAPE of 190.33%, indicating limited adaptability across structurally diverse produce.

At the category level, Ours consistently delivers leading performance, achieving the lowest MAPE in Mushroom Family (9.31%) and Part of Produce Structure (41.60%), both of which are characterized by fragility and structural complexity. Deepseek-v3 shows relative strength in Elongated Fruits (21.70%), suggesting an ability to capture density-to-volume relationships in slender geometries. However, the Deepseek series also reveals clear bottlenecks, for example, Deepseek-r1 reached 47.58% in Round Fruits and 335.09% in Leafy Vegetables, underscoring its instability in categories that are either density-variable or structurally loose.

Other models demonstrate intermediate patterns. ChatGPT-4o exhibits balanced performance across most categories, reflecting a trade-off between semantic understanding and reasoning ability. Claude-3.5-Sonnet achieves strong results in Mushroom Family (31.29%), but its performance fluctuated significantly in other categories, such as Leafy Vegetables (171.99%), indicating dependence on semantic clarity and training data relevance.

Overall, these results underscore the importance of integrating geometric perception with language-based reasoning in agricultural mass estimation. The cross-category variations in MAPE highlight that not all state-of-the-art LLMs generalize equally well. Notably, the fine-tuned compact model (Ours) demonstrates the greatest potential as a reliable backbone for damage-free robotic grasping, particularly in scenarios involving diverse, fragile, and structurally ambiguous produce.

4.5. Comparison with Baseline Approaches

To further evaluate the practical adaptability and control feasibility in real-world grasping scenarios, a grasp force validation experiment is conducted between Ours and DeliGrasp (Xie et al., 2024). Evaluation metrics include the hit rate, which measures the proportion of predictions falling within asymmetric safety tolerances, and the MAPE, which quantifies continuous estimation error. Together, these results provide insight into how different baseline methods balance accuracy and safety compliance in damage-free grasping, where a higher hit rate and lower MAPE indicate a better trade-off between the two.

A calibration factor serves as a lightweight linear adjustment that rescales the raw force estimates produced by each model. Different categories of produce often exhibit systematic overestimation or underestimation due to variations in density, geometry, or fragility, and a single global model may not capture these nuances. By learning a multiplicative factor s for each model-category pair, we align predictions with the physical reference scale, thereby reducing systematic bias while preserving relative variations across samples. Experimental results are illustrated in Table 3.

Table 3
Performance comparison of DeliGrasp (Xie et al., 2024) and Ours across produce categories, reporting hit rate, MAPE, and calibration factor.

Categories	DeliGrasp (Xie et al., 2024)			Ours		
	Hit Rate	MAPE	Factor	Hit Rate	MAPE	Factor
1	0.667	0.147	0.756	**0.8**	**0.136**	0.78
2	0.6	0.354	0.661	**0.8**	**0.352**	0.7
3	0.667	0.311	0.673	**1**	**0.206**	0.585
4	0.375	**0.251**	0.554	**0.667**	0.378	0.98
5	**0.375**	**0.412**	0.56	0.25	0.76	0.85
6	**0.889**	**0.165**	0.708	0.667	0.242	0.6
7	0.667	**0.424**	0.743	0.667	0.659	0.89
All	0.606	**0.295**	-	**0.693**	0.39	-

Note: 1 stands for Berries; 2 for Round Fruits; 3 for Elongated Fruits; 4 for Elliptical Fruits; 5 for Leafy Vegetables; 6 for Part of Produce Structure; and 7 for Mushroom Family.

At the aggregate level, Ours achieves a superior balance between accuracy and safety compliance. The results show that its average hit rate reaches 0.693, whereas DeliGrasp attains only 0.606, demonstrating a more consistent ability to generate grasping force estimates that remain within the asymmetric safety tolerances. Although the MAPE is slightly higher at 0.390 compared with 0.295 for DeliGrasp, the improvement in hit rate highlights the stronger robustness of Ours when applied to fragile produce. This indicates that in real-world scenarios, our framework is more effective at preventing damage caused by excessive force while still maintaining acceptable predictive precision.

A category-wise analysis further emphasizes these advantages. In highly fragile categories such as Berries and Round Fruits, Ours attains hit rates of 0.8 in both cases, exceeding the 0.667 and 0.6 achieved by DeliGrasp, while also delivering lower estimation errors. In Elongated Fruits, Outs achieves a perfect hit rate of 1.0, in contrast to only two-thirds for DeliGrasp, and simultaneously yields smaller errors. These results demonstrate that our framework generalizes more effectively across produce with diverse geometries and structural properties, providing accurate estimates while preserving the structural integrity of delicate produce.

The analysis of calibration factors provides additional evidence of robustness. In several categories, the scaling factors learned by Ours remain close to unity, with values such as 0.78 for Berries and 0.98 for Part of Produce Structure, indicating minimal systematic bias between predicted and reference values. In contrast, DeliGrasp exhibits much stronger deviations, with factors of only 0.554 for Part of Produce Structure and 0.56 for Leafy Vegetables, suggesting larger inherent biases. The reduced reliance of Ours on calibration adjustments confirms the intrinsic stability of its raw predictions and enhances interpretability, as the outputs remain well aligned with the physical reference scale.

Beyond the numerical results, a qualitative comparison of grasping outcomes further underscores the effectiveness of our framework in achieving damage-free manipulation. Fig. 7 provides a side-by-side illustration of grasping instances produced by our method and DeliGrasp. In the top row, grasps executed by our framework demonstrate stable force regulation that firmly secures the objects while preserving their original structural integrity. The absence of deformation or surface damage highlights the capability of our method to deliver reliable control even for fragile and irregularly shaped produce. In contrast, the bottom row depicts grasps from DeliGrasp, where uneven or excessive forces often result in visible deformation, slippage, or local crushing. These shortcomings are particularly evident in delicate categories such as leafy vegetables and elongated fruits, where excessive compression compromises the objects' usability and market value.

The visual evidence aligns closely with the quantitative analysis reported in Table 3. Categories such as berries, round fruits, and elongated fruits show both higher hit rates and lower estimation errors for our method, which directly translates into the more consistent and damage-free grasps visible in Fig. 7. DeliGrasp, despite achieving a lower average MAPE, fails to guarantee force regulation within the asymmetric safety tolerances, which manifests as the physical deformations shown in the figure. This illustrates that a lower error metric alone is insufficient for safe grasping, and that the ability to maintain forces within the safe operational range is more critical for real-world applications.

Taken together, these results demonstrate that our method offers not only quantitative improvements but also practical advantages in terms of adaptability, robustness, and safety compliance. By achieving accurate force predictions that directly translate into stable and damage-free manipulation, our framework provides a stronger foundation for damage-free grasping in real-world agricultural and industrial environments.

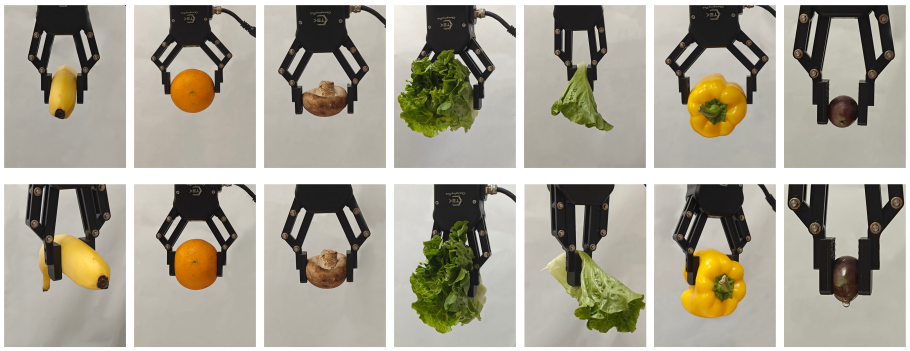

Fig. 7. Comparison of grasping performance. The top row shows successful grasping by Ours, while the bottom row displays results from DeliGrasp (Xie et al., 2024), which may lead to deformation or even damage.

5. Conclusion

We propose a novel grasping framework, which builds upon the planar grasping framework by incorporating semantic enhancement and commonsense physical reasoning to improve adaptability and robustness across diverse produce categories and morphologies. The framework integrates the LGQS module from Grounding DINO, enabling deep alignment between natural language inputs and visual features. This enables the framework not only to localize potential grasp regions but also to identify the produce category and key attributes, thereby supporting semantically driven grasp strategy generation. Building on semantic identification, further leverages the extensive knowledge base and reasoning capability of LLMs. By interpreting class-specific semantic descriptions of produce, the framework can autonomously infer implicit physical attributes such as density and friction coefficient, which are then used to estimate appropriate grasping forces. This semantic-physical reasoning mechanism provides a robust cognitive foundation for executing damage-free grasping tasks. In addition, this framework utilizes point cloud data of the target produce, combined with the heuristic estimation ability of LLMs, to improve volume estimation and grasp region identification for irregularly shaped produce such as density-varying mushrooms, curved bananas, or leafy greens with internal voids. This enhances generalization and adaptability in handling complex produce geometries.

Despite the promising results achieved in damage-free grasping of delicate produce, the framework still faces several limitations. First, LLMs exhibit instability in density inference for structurally complex produce such as dragon-fruit and avocados. Second, the current volume correction mechanism relies on static factors rather than a learned model, limiting its accuracy. Future work will focus on improving system performance by incorporating visual context cues to enhance the physical reasoning capabilities of LLMs and introducing a learnable volume correction network to improve the geometric estimation accuracy for diverse produce types.

CRediT authorship contribution statement

Ziye Zhang: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft & Review & Editing, Visualization. **Xiaoyu Xia**: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft & Review & Editing, Visualization. **Yuhao Jin**: Conceptualization, Methodology, Resources, Writing - Review & Editing. **Qizhong Gao**: Visualization, Writing - Review & Editing. **Lin Qiao**: Resources, Formal analysis. **Jinglei Chen**: Formal analysis, Data Curation. **Yong Yue**: Supervision, Funding acquisition, Validation. **Xiaohui Zhu**:

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the technical support and device contribution to Xi'an Jiaotong-Liverpool University (XJTLU).

Funding

This research was funded by Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004), the XJTLU Artificial Intelligence (AI) University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU, and Suzhou Industrial Park (SIP) AI innovation platform (YZCXPT2022103).

Appendix A. Produce sample with corresponding category

In Table A1, produce items are classified by category and grasp-related characteristics, providing a structured basis for analyzing grasp difficulty and estimation challenges.

Appendix B. Complete Experimental Data

This appendix provides the complete per-category results of mass estimation across seven representative produce categories using seven different LLMs. The Table B1 and Table B2 report MAPE and STD under both reasoning-only and perception-fusion settings, complementing the aggregated results presented in the main text.

Table A1
Produce classification and their grasping-related characteristics.

Produce	**Categories**	**Characteristic**
grape, cherry, berry, strawberry, cherry-tomato	Berries	Small volume; Smooth surface; Risky of slippage
apple, tomato, nectarine	Round Fruits	Spherical; high density; Easy to predict
banana, cucumber	Elongated Fruits	Elongated and curved; Difficult to shape fit
lemon, avocado, orange, dragon-fruit, kiwi-fruit	Elliptical Fruits	Rough surface; Medium volume; Easy to deformation
bell-pepper	Hollow Produce	Hard to estimate
spinach, lettuce, Chinese-cabbage, baby-cabbage	Leafy Vegetables	Lightweight; Large volume; Loosely structured
watermelon-flesh, broccoli, a-piece-of-lettuce	Part of Produce Structure	Difficult to quantify size; Incomparable qualitatively
button-mushroom, shiitake-mushroom, King-Oyster-mushroom	Mushroom Family	Volume - Mass imbalance

Table B1

Complete per-category results of produce mass estimation across seven LLMs and seven representative categories. Reported are MAPE and STD under reasoning-only and perception-fusion paradigms.

Models	Categories	Reasoning-only		Perception-fusion	
		MAPE	STD	MAPE	STD
Deepseek R1	1	53.17	21.057	46.889	28.739
	2	35.543	5.009	47.579	12.827
	3	389.258	525.284	825.177	1140
	4	30.089	17.029	35.135	24.269
	5	462.955	703.512	335.094	309.018
	6	943.952	451.066	12.236	9.623
	7	273.237	267.648	30.171	18.964
	All	312.601	284.372	190.326	220.491
Deepseek V3	1	88.557	8.506	36.958	31.825
	2	359.015	578.433	42.513	29.127
	3	28.802	3.157	21.704	8.158
	4	24.862	21.976	36.565	15.733
	5	178.343	123.152	136.478	185.692
	6	448.224	106.216	250.664	209.68
	7	62.779	34.581	49.982	5.87
	All	170.083	125.146	82.123	69.441
QWen-max	1	85.084	12.012	41.558	34.592
	2	16.279	19.613	55.664	19.738
	3	7.972	5.74	49.333	0.146
	4	50.747	41.75	42.739	16.948
	5	1255.166	1053.88	180.718	145.58
	6	485.95	390.708	144.071	191.636
	7	211.276	225.374	54.939	19.789
	All	301.782	249.868	81.2889	61.204
QWen-plus	1	43.68	46.944	69.246	38.27
	2	35.717	44.879	63.831	9.007
	3	36.53	39.434	31.094	9.8
	4	38.085	19.375	42.289	19.308
	5	552.95	512.414	135.478	164.507
	6	830.379	916.582	233.544	322.436
	7	58.4	60.881	11.271	17.232
	All	227.963	234.358	83.822	82.937
ChatGPT4o	1	65.144	21.403	50.002	24.634
	2	36.828	10.438	47.563	22.433
	3	5.686	2.867	34.266	26.382
	4	51.503	32.251	36.876	13.642
	5	394.81	621.336	56.999	56.794
	6	4947.936	6463.047	383.41	521.184
	7	47.986	9.22	9.324	3.214
	All	792.842	1022.937	88.349	95.469

Note: 1 stands for Berries; 2 for Round Fruits; 3 for Elongated Fruits; 4 for Elliptical Fruits; 5 for Leafy Vegetables; 6 for Part of Produce Structure; and 7 for Mushroom Family.

Table B2
Complete per-category results of produce mass estimation across seven LLMs and seven representative categories. Reported are MAPE and STD under reasoning-only and perception-fusion paradigms (continue).

Models	**Categories**	**Reasoning-only**		**Perception-fusion**	
		MAPE	STD	MAPE	STD
Claude-3.5-Sonnet	1	137.061	196.967	45.408	23.365
	2	59.42	49.95	49.745	15.541
	3	159.26	38.671	37.068	4.514
	4	77.929	81.73	35.1	12.788
	5	237.538	10186.482	171.998	450.291
	6	265.742	4324.687	201.499	173.815
	7	62.948	5.825	31.286	15.029
	All	142.843	2126.330	81.729	99.335
Ours	1	79.356	28.154	69.393	34.377
	2	34.704	40.799	52.192	17.355
	3	12.624	5.555	22.008	2.33
	4	63.54	50.884	47.676	19.281
	5	1099.925	1021.552	78.484	51.089
	6	861.828	1171.669	41.597	37.315
	7	485.651	67.059	9.311	3.385
	All	376.804	340.810	45.809	23.590

Note: 1 stands for Berries; 2 for Round Fruits; 3 for Elongated Fruits; 4 for Elliptical Fruits; 5 for Leafy Vegetables; 6 for Part of Produce Structure; and 7 for Mushroom Family.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al., 2022. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 doi:10.48550/arXiv.2204.01691.

Chen, K., Li, T., Yan, T., Xie, F., Feng, Q., Zhu, Q., Zhao, C., 2022. A soft gripper design for apple harvesting with force feedback and fruit slip detection. Agriculture 12, 1802. doi:10.3390/agriculture12111802.

Chen, X., Sun, Y., Zhang, Q., Dai, X., Tian, S., Guo, Y., 2024. Three-dimension object detection and forward-looking control strategy for non-destructive grasp of thin-skinned fruits. Applied Soft Computing 150, 111082. doi:10.1016/j.asoc.2023.111082.

Chu, X., Deng, J., You, G., Liu, W., Li, X., Ji, J., Zhang, Y., 2025. Graspcot: Integrating physical property reasoning for 6-dof grasping under flexible language instructions. arXiv preprint arXiv:2503.16013 doi:10.48550/arXiv.2503.16013.

Ghasemzadeh, P., Hempel, M., Wang, H., Sharif, H., 2023. Ggcnn: An efficiency-maximizing gated graph convolutional neural network architecture for automatic modulation identification. IEEE Transactions on Wireless Communications 22, 6033–6047. doi:10.1109/TWC.2023.3239311.

Hu, F., Wu, F., Gu, H., Abbas, G., Alanazi, M.D., Othmen, S., Wang, J., Zhang, T., 2025a. Transforming agriculture with advanced robotic decision systems via deep recurrent learning. Expert Systems with Applications 259, 125123. doi:10.1016/j.eswa.2024.125123.

Hu, Z., Gao, K., Zhang, X., Yang, Z., Cai, M., Zhu, Z., Li, W., 2025b. Efficient grounding dino: Efficient cross-modality fusion and efficient label assignment for visual grounding in remote sensing. IEEE Transactions on Geoscience and Remote Sensing doi:10.1109/TGRS.2025.3536015.

Huang, H., Wang, R., Huang, F., Chen, J., 2025. Analysis and realization of a self-adaptive grasper grasping for non-destructive picking of fruits and vegetables. Computers and Electronics in Agriculture 232, 110119. doi:10.1016/j.compag.2025.110119.

Jin, T., Han, X., 2024. Robotic arms in precision agriculture: A comprehensive review of the technologies, applications, challenges, and future prospects. Computers and Electronics in Agriculture 221, 108938. doi:10.1016/j.compag.2024.108938.

Jin, Y., Gao, Q., Zhu, X., Yue, Y., Lim, E.G., Chen, Y., Wong, P., Chu, Y., 2024. Vmgnet: A low computational complexity robotic grasping network based on vmamba with multi-scale feature fusion. arXiv preprint arXiv:2411.12520 doi:10.48550/arXiv.2411.12520.

Jin, Y., Xia, X., Gao, Q., Yue, Y., Lim, E.G., Wong, P., Ding, W., Zhu, X., 2025. Deep learning in produce perception of harvesting robots: A comprehensive review. Applied Soft Computing , 112971doi:10.1016/j.asoc.2025.112971.

Kebe, G.Y., Higgins, P., Jenkins, P., Darvish, K., Sachdeva, R., Barron, R., Winder, J., Engel, D., Raff, E., Ferraro, F., et al., 2021. A spoken language dataset of descriptions for speech-based grounded language learning. Advances in neural information processing systems URL: https://openreview.net/forum?id=Yx9jT3fkBaD.

Kumra, S., Joshi, S., Sahin, F., 2020. Antipodal robotic grasping using generative residual convolutional neural network, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 9626–9633. doi:10.1109/IROS45743.2020.9340777.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision 128, 1956–1981. URL: https://storage.googleapis.com/openimages/web/index.html.

Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D., Ghanem, B., 2023a. Camel: Communicative agents for "mind" exploration of large scale language model society. doi:10.48550/arXiv.2303.17760, arXiv:2303.17760.

Li, S., Yu, H., Ding, W., Liu, H., Ye, L., Xia, C., Wang, X., Zhang, X.P., 2023b. Visual–tactile fusion for transparent object grasping in complex backgrounds. IEEE Transactions on Robotics 39, 3838–3856. doi:10.1109/TRO.2023.3286071.

Li, T., Yan, Y., Yu, C., An, J., Wang, Y., Chen, G., 2024. A comprehensive review of robot intelligent grasping based on tactile perception. Robotics and Computer-Integrated Manufacturing 90, 102792. doi:10.1016/j.rcim.2024.102792.

Li, Z., Liu, J., Li, Z., Dong, Z., Teng, T., Ou, Y., Caldwell, D., Chen, F., 2025. Language-guided dexterous functional grasping by llm generated grasp functionality and synergy for humanoid manipulation. IEEE Transactions on Automation Science and Engineering doi:10.1109/TASE.2024.3524426.

Lin, J., Hu, Q., Xia, J., Zhao, L., Du, X., Li, S., Chen, Y., Wang, X., 2023. Non-destructive fruit firmness evaluation using a soft gripper and vision-based tactile sensing. Computers and Electronics in Agriculture 214, 108256. doi:10.1016/j.compag.2023.108256.

Liu, Y., Zhang, J., Lou, Y., Zhang, B., Zhou, J., Chen, J., 2024. Soft bionic gripper with tactile sensing and slip detection for damage-free grasping of fragile fruits and vegetables. Computers and Electronics in Agriculture 220, 108904. doi:10.1016/j.compag.2024.108904.

Ma, C., Ying, Y., Xie, L., 2024. Visuo-tactile sensor development and its application for non-destructive measurement of peach firmness. Computers and Electronics in Agriculture 218, 108709. doi:10.1016/j.compag.2024.108709.

Rashid, A., Sharma, S., Kim, C.M., Kerr, J., Chen, L.Y., Kanazawa, A., Goldberg, K., 2023. Language embedded radiance fields for zero-shot task-oriented grasping, in: 7th Annual Conference on Robot Learning. URL: https://openreview.net/forum?id=k-Fg8JDQmc.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J., 2019. Objects365: A large-scale, high-quality dataset for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8430–8439. URL: www.objects365.org.

Velasquez, A., Grimm, C., Davidson, J.R., 2024. Compact robotic gripper with tandem actuation for selective fruit harvesting. arXiv preprint arXiv:2408.06674 doi:10.48550/arXiv.2408.06674.

Visentin, F., Castellini, F., Muradore, R., 2023. A soft, sensorized gripper for delicate harvesting of small fruits. Computers and Electronics in Agriculture 213, 108202. doi:10.1016/j.compag.2023.108202.

Wang, Y.R., Duan, J., Fox, D., Srinivasa, S., 2023. Newton: Are large language models capable of physical reasoning? arXiv preprint arXiv:2310.07018 doi:10.48550/arXiv.2310.07018.

Xie, W., Valentini, M., Lavering, J., Correll, N., 2024. Deligrasp: Inferring object properties with llms for adaptive grasp policies. arXiv preprint arXiv:2403.07832 doi:10.48550/arXiv.2403.07832.

Xu, J., Jin, S., Lei, Y., Zhang, Y., Zhang, L., 2024. Rt-grasp: Reasoning tuning robotic grasping via multi-modal large language model, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 7323–7330. doi:10.1109/IROS58592.2024.10801718.

Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S., 2023. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1179–1189. doi:10.1109/CVPR52729.2023.00120.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J., 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 19313–19322. doi:10.1109/CVPR52688.2022.01871.

Zhang, B., Xie, Y., Zhou, J., Wang, K., Zhang, Z., 2020. State-of-the-art robotic grippers, grasping and control strategies, as well as their applications in agricultural robots: A review. Computers and Electronics in Agriculture 177, 105694. doi:10.1016/j.compag.2020.105694.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H., 2022. Pointclip: Point cloud understanding by clip, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8552–8562. doi:10.48550/arXiv.2112.02413.

Zheng, W., Xie, Y., Zhang, B., Zhou, J., Zhang, J., 2021. Dexterous robotic grasping of delicate fruits aided with a multi-sensory e-glove and manual grasping analysis for damage-free manipulation. Computers and Electronics in Agriculture 190, 106472. doi:10.1016/j.compag.2021.106472.