Journal of Field Robotics

WILEY

RESEARCH ARTICLE

SimLiquid: A Simulation-Based Liquid Perception Pipeline for Robot Liquid Manipulation

Yan Huang[1] | Jiawei Zhang[1] | Ran Yu[1] | Shoujie Li[1] | Wenbo Ding[1,2]

[1]Shenzhen Ubiquitous Data Enabling Key Lab, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China | [2]RISC-V International Open Source Laboratory, Shenzhen, China

Correspondence: Shoujie Li (lsj20@mails.tsinghua.edu.cn) | Wenbo Ding (ding.wenbo@sz.tsinghua.edu.cn)

Received: 1 December 2024 | **Revised:** 6 February 2025 | **Accepted:** 15 March 2025

Funding: This study was supported by the National Key R&D Program of China under Grant (No. 2024YFB3816000), Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Shenzhen Science and Technology Program (JCYJ20220530143013030), Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197), Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005), Shenzhen Higher Education Stable Support Program (WDZC20231129093657002), and Meituan.

Keywords: liquid estimation | LLM | Sim2Real | transparent object

ABSTRACT

Transparent liquid volume estimation is crucial for robot manipulation tasks, such as pouring. However, estimating the volume of transparent liquids is a challenging problem. Most existing methods primarily focus on data collection in the real world, and the sensors are fixed to the robot body for liquid volume estimation. These approaches limit both the timeliness of the research process and the flexibility of perception. In this paper, we present SimLiquid20k, a high-fidelity synthetic data set for liquid volume estimation, and propose a YOLO-based multi-task network trained on fully synthetic data for estimating the volume of transparent liquids. Extensive experiments demonstrate that our method can effectively transfer from simulation to the real world. In scenarios involving changes in background, viewpoint, and container variations, our approach achieves an average error of 5% in real-world volume estimation. In addition, our work conducts two application experiments integrating with GPT-4, showcasing the potential of our method in service robotics. The accompanying videos and supporting Information are available at https://simliquid.github.io/.

1 | Introduction

The ability to estimate the state of a liquid within a container is essential for domestic service robots to handle liquid safely and agilely. When pouring, they must assess the amount of liquid to stop at the right time. In drink-serving task, they need to identify which containers are unfilled. However, liquid is difficult to perceive, especially in transparent containers. Most liquids are textureless and refractive, which makes them challenging to detect using either RGB or depth-based methods.

To estimate the liquid state, several previous works utilized contact-based measurement methods, such as force-torque sensor, accelerometer, and tactile sensor by Matl et al. (2019), Chen et al. (2016), and Huang et al. (2022). These methods showed promising accuracy in estimating liquid physical properties, while they cannot operate without contacting with liquid containers. With the advancement of deep learning algorithms, some works turn to noncontact learning-based approaches for liquid sensing, such as liquid segmentation and volume estimation. These methods are typically data-driven, so a sufficiently diverse data set is necessary for algorithm training. However, liquids are deformable and can be contained in various shapes of containers, making it relatively challenging to construct a large scale data set with diversity in both liquid volume and container types.

For the liquid data set collection, previous works have adopted various methods. Schenck and Fox (2017b) built a system for

liquid auto-annotation, which utilized a RGBD camera calibrated with a thermal camera to generate pixel level annotation for heated liquids. Some projects built annotation frameworks with robotic arms, vision sensors, audio sensors and digital scales to collect synchronized multi-modal data for training liquid perception models by Wilson et al. (2019); Liang et al. (2020). Work by Narasimhan et al. (2022) trained a GAN model to transform colored liquid to transparent liquid, for auto mask annotation of transparent liquid. Despite these approaches having achieved remarkable performance, they still have limitations in generalization ability due to the limited-scale data collected from the lab scenarios. To solve this issue, we propose a synthetic data set generation pipeline, which is able to build a large-scale liquid data at low cost for training data-driven models with strong generalization ability. Our pipeline is built upon Blender, with Cycles engine supporting physically based rendering (PBR). Before data set generation, the process involves importing container CAD models into Blender and extracting their inner shells, which are then modeled to liquids with different liquid PBR (Physically Based Rendering) materials. A splitting plane is arranged for each container in the simulation and it randomly moves along the z-axis of the container to generate random volume liquids. We developed a data generation framework based on BlenderProc, proposed by Denninger and Sundermeyer (2022), which renders hundreds of photo-realistic liquid-container per hour. Moreover, the framework provides multi-modal accurate annotations, including segmentaions, poses and bounding boxes for both liquids and containers, along with liquid volume annotations at milliliter-level precision. Based on the pipe, we built a large scale data set, SimLiquid20k, containing 20 kinds of containers and four common liquids.

We developed a YOLO-based model trained on the SimLiquid20k to demonstrate its Sim2Real performance. Our model takes RGB or RGB-D images as input and outputs multiple prediction, including class labels, bounding boxes, segmentation masks, and liquid volume. To enable end-to-end training and inference of the multi-task model, we implemented a hybrid loss function specifically tailored for liquid state estimation. We evaluated the model's performance through comprehensive benchmarking in both simulated and real-world environments. Evaluation on the synthetic validation data set shows that the model accurately predicts liquid volumes using both RGB and RGB-D input modalities. The real-world benchmark demonstrates that our model estimates the liquid volume across different containers with a mean error of less than 5%. Furthermore, our model is robust across various camera views. To prove the practical value of our approach, we conducted pouring experiments with the assistance of the LLM. Results indicate that incorporating our liquid estimation model substantially improved the LLM's decision-making robustness in the liquid manipulation task.

To summarize, the contributions of this paper are:

1. We propose a synthetic data set generation pipeline for liquid state estimation, constructing a large-scale data set across various containers, with high-fidelity images, precise ground-truth annotations, and extensive domain randomization across illumination conditions, backgrounds, camera viewpoints, and liquid volumes.
2. We design a novel neural network that simultaneously addresses multiple tasks, including container detection, liquid segmentation, and volume estimation, enabling accurate liquid state estimation in complex environments.
3. We conduct comprehensive benchmarking in simulation and real-world settings, demonstrating the robust sim-to-real transfer ability of our data set and network. Furthermore, we integrate the network with a large language LLM for pouring manipulation, showcasing the practical applicability of our approach in domestic service robotics.

2 | Related Work

2.1 | Transparent Object Perception

In the field of robotic perception, transparent object sensing is a challenging task due to the complex optical properties of transparent objects, such as refraction and reflection. The textures of these objects are highly influenced by various environmental factors, leading to significant variations. Furthermore, transparent objects are non-Lambertian, meaning their light paths do not conform to the geometric assumptions made by classical stereo vision algorithms. This characteristic makes it difficult for standard 3D sensors to accurately estimate the depth of transparent objects, often resulting in noise or distortion.

Previous works, such as those by Xie et al. (2020) and Liao et al. (2020), have focused on the perception of transparent containers and developed methods for tasks like transparent object segmentation. Similarly, Albrecht and Marsland (2013) as well as Dai et al. (2023) have worked on depth estimation, while Liu et al. (2020) have contributed to keypoint detection. However, collecting real-world data set for transparent object perception and manipulation tasks is a time-consuming and labor-intensive process. With the advancement of computer graphics simulation tools, many studies have adopted synthetic data set to support related tasks. Examples include ClearGrasp by Sajjan et al. (2020), a synthetic data set for depth completion, DexNeRF by Ichnowski et al. (2021), a synthetic data set for transparent object detection and localization, and the Omniverse data set proposed by Zhu et al. (2021), a large-scale synthetic data set containing 60K images generated using the NVIDIA PhysX engine. Recent work by Yu et al. (2024) further advances this field by proposing a method for depth restoration of hand-held transparent objects, specifically targeting human-to-robot handover scenarios. This highlights the growing importance of synthetic data set in enabling robust and reliable robotic manipulation of transparent objects. Their approach leverages synthetic data to train models that can accurately reconstruct depth information for transparent objects, even in dynamic and interactive settings. This highlights the growing importance of synthetic data set in enabling robust and reliable robotic manipulation of transparent objects. Such synthetic or simulated data set significantly save time and financial resources. In particular, when transparent objects contain contents like liquid, high-quality synthetic datasets play a crucial role in bridging the gap between simulation and real-world scenarios.

2.2 | Transparent Liquid Perception

Compared to transparent object perception, sensing transparent liquid is often more challenging due to the lack of fixed shapes and geometric characteristics. Work by Kennedy et al. (2017) shows that colored liquid typically have distinct boundaries that facilitate their segmentation from the background or other objects. However, transparent liquid lack such clear boundaries, making its perception significantly more difficult. Additionally, the instability of liquid flow often leads to more complex optical effects for transparent liquid compared to transparent objects.

One approach to sense transparent liquid is the use of thermal imaging, as demonstrated by Schenck and Fox (2017a, 2017b), to obtain liquid labels. However, frequently heating liquids in real-world scenarios is a cumbersome and impractical process. Another method involves using depth sensors to acquire transparent liquid labels, as explored by Do and Burgard (2019), as well as Dong et al. (2019), but the depth data for transparent objects and liquid often suffer from significant distortion and noise. In some experiments, Kennedy et al. (2019) employed real-time weighing of containers filled with liquids to determine the liquid weight. This approach necessitates additional experimental equipment, making the data collection and experimental processes less convenient.

2.3 | Volume Estimation in Robotics

When a robot grasps and pours a container filled with liquid, understanding the volume of liquid inside the container is often crucial for subsequent operations. To perceive liquid volume, previous studies, such as Zhu et al. (2022), have attempted to estimate the amount of liquid poured out using force sensors embedded in robotic arms. Other approaches, like those proposed by Brandi et al. (2014), estimate liquid volume based on motion and CAD models. Additionally, some methods, including Zhu et al. (2022), utilize multi-modal fusion techniques for liquid volume estimation. Additionally, thermal imaging has been employed to accurately detect the height and volume of hot water.

However, these approaches often rely on nonvisual sensors or additional information and typically place sensors at positions that directly interact with the liquid container, such as the gripper of a robotic arm. This setup imposes constraints on both visual and sensory capabilities. In contrast, humans can roughly estimate liquid volume using only visual perception. Thus, we hypothesize that liquid volume can be estimated using only RGB or RGB-D images as input. Furthermore, most prior studies collected data in real-world environments, making transparent liquid estimation tasks time-consuming and labor-intensive.

To address these challenges, we generated a synthetic data set for transparent objects and liquids using image simulation software. We carefully calibrated parameters such as lighting and background to minimize the gap between the synthetic data set and real-world scenarios. In our volume estimation task, the camera is mounted with flexible positioning rather than being fixed relative to the robotic arm, allowing degrees of freedom comparable to human visual mobility.

3 | Methodology

3.1 | Data Set Generation and Annotation

Training a liquid volume estimation model requires a large data set, which is challenging for manual annotation methods. To address this challenge, we generate a sizable synthetic liquid estimation data set, SimLiquid20k, using BlenderProc. SimLiquid20k is a high domain randomization data set comprising 20,000 synthetic images in a variety of environmental conditions.

As shown in Figure 1, SimLiquid20k contains 20 kinds of containers and four types of liquids: clean water, milk, orange juice, and red wine. These non-transparent liquids were added to the model to improve its perception of the liquid within the cups.

To ensure diversity in lighting and background, 773 HDRI panoramic photos are randomly applied to each scene as part of the data set generation procedure to guarantee backdrop and illumination variation. Cameras are placed around the target items in each scene using a shell-based sampling

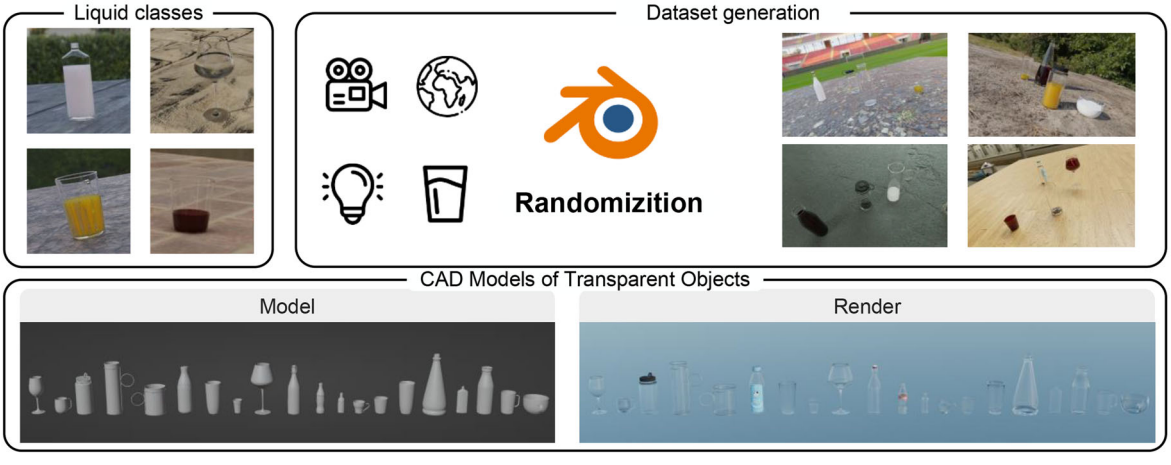

FIGURE 1 | Illustration of data set generation. [Color figure can be viewed at wileyonlinelibrary.com]

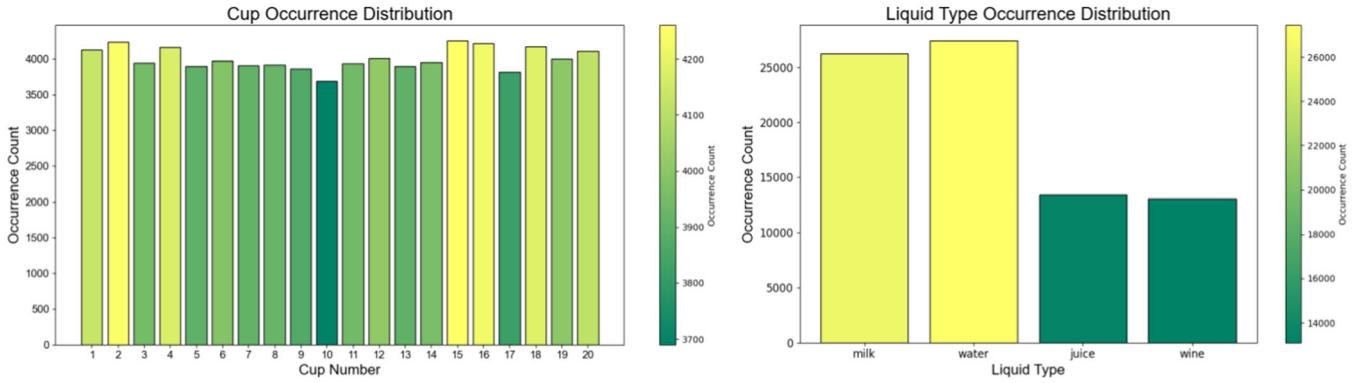

FIGURE 2 | Illustration of the training data distribution. [Color figure can be viewed at wileyonlinelibrary.com]

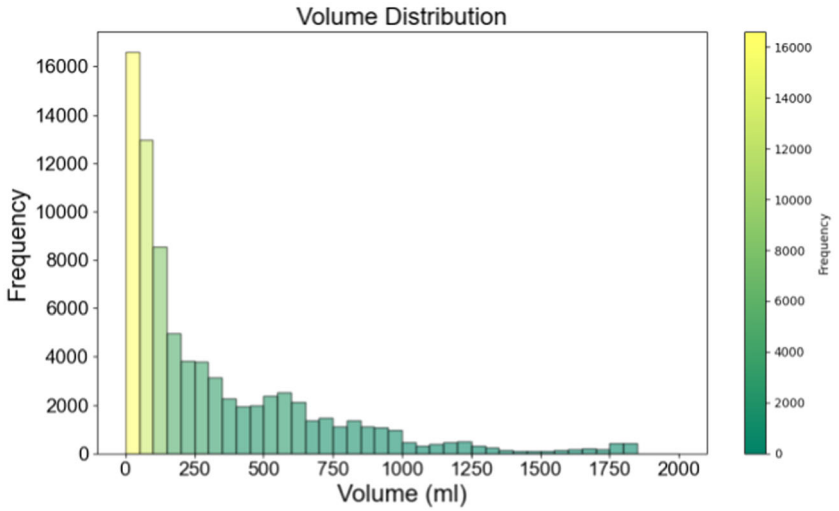

FIGURE 3 | Illustration of the liquid volume distribution in SimLiquid20k. [Color figure can be viewed at wileyonlinelibrary.com]

distribution. To simulate naturalistic viewing angles and add variability to the data set, slight rotations and perturbations are applied to the cameras. Liquid volumes are sampled within predefined ranges specific to each cup. The variation in volume is achieved by dynamically adjusting the height of the liquid's surface plane.

In addition to the generation of synthetic images, we also collected bounding box annotations, segmentation masks, volume labels, and depth images during the data set creation process, ensuring a comprehensive set of ground truth data to support the training and evaluation of a wide range of tasks. To further enhance the data set's applicability, the pose of each container was also recorded. While this study does not specifically address grasping tasks, the inclusion of pose data would facilitate future research in areas such as robotic manipulation and pose estimation.

To enhance the generalization capability of SimLiquid20k, we randomly combined 20 prepared containers and four types of liquids to construct our data set, as illustrated in Figure 2. We ensured that different containers and liquids appeared with nearly uniform frequencies, aiming to comprehensively represent the relevant features of both containers and liquids. Additionally, due to the volume constraints of each container, a maximum capacity was predefined for each container, and the liquid volumes were randomly sampled within this range. The distribution of liquid volumes in our data set is shown in Figure 3. Due to the inclusion of containers with varying sizes, the distribution of liquid volumes exhibits a decreasing trend as the volume increases.

3.2 | Data Processing

The RGB(-D) images generated from Blenderproc have a resolution of 960×600. To standardize the data dimensions, we rescaled the original images to 640×640 while preserving the aspect ratio, padding the remaining areas with plain background. Accurately constraining volume predictions to a minimal error margin using visual input alone presents a considerable challenge. Although the impact of resolution on volume estimation is not explicitly addressed in this study, we acknowledge it as an important factor that warrants further investigation.

To eliminate invalid depth values, we applied a clipping operation to the collected depth maps, restricting the maximum depth to 50 cm. Subsequently, both RGB and depth

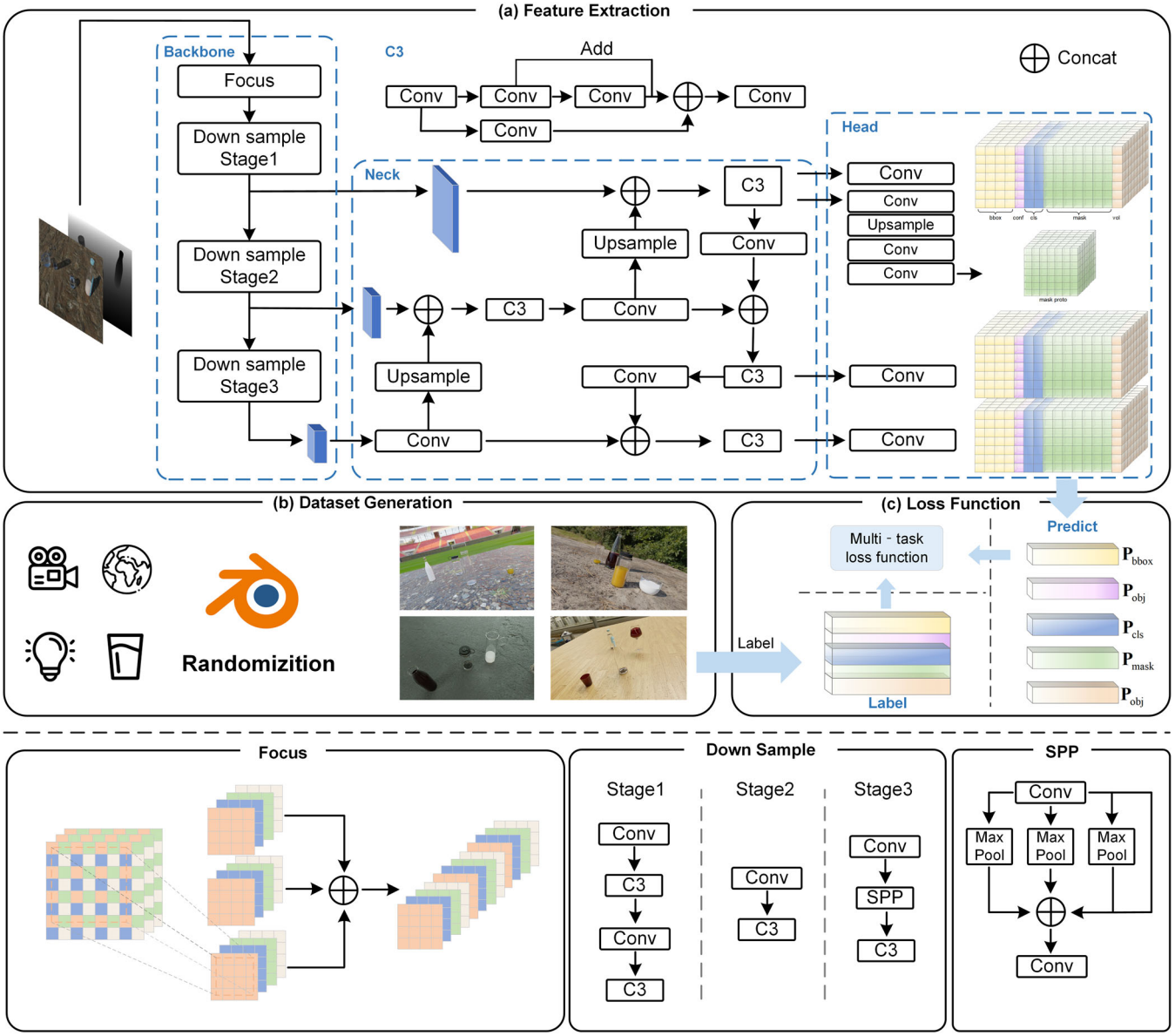

FIGURE 4 | Illustration of our working pipeline. (a) Multi-task features are extracted from the RGB and depth images. Segment head predicts the output of yolo-based model. (b) SimLiquid20k data set is generated by Blenderproc with randomization. (c) Multi-task loss function enables simultaneous output of the bounding box, mask, and volume of the target. [Color figure can be viewed at wileyonlinelibrary.com]

images were normalized and concatenated at the input stage. Depth maps acquired from simulation environments and those captured by real-world depth sensors often exhibit substantial discrepancies, which complicates direct transferability. However, during training within the simulated environment, no additional scaling was applied to either the depth or RGB images. In our input design, we assume that both types of images provide an equivalent amount of information.

3.3 | Liquid Estimation

From the perspective of robotic perception, our goal is to design a function $f(o)$ that can predict the volume of liquid within a container manipulated by the robot based on observational data collected from sensors in the current environment. This function is implemented as a YOLO-based deep neural network as shown in Figure 4, leveraging its end-to-end learning capabilities to efficiently process sensory inputs and generate accurate volume estimations.

In liquid perception tasks, we consider depth information to be crucial. For monocular tasks, depth is an indispensable component for reconstructing 3D scene information from images. Therefore, in our model design, we opted for two input modes: RGB and RGB-D.

Our network architecture consists of three main components: a backbone for down-sampling, a neck for feature fusion, and a detection head based on the fused features. As shown in Figure 4, the detection head outputs feature vectors at three different scales. From these scale-specific feature vectors, we extract predictions for bounding boxes, class labels (cls),

confidence scores (conf), segmentation masks (mask), and liquid volume (vol).

The corresponding losses are computed based on these predictions, and we integrate the weighted losses using $\lambda_d L_d$ into a total loss function for multi-task learning, defined as $L_{\text{total}} = \sum_{d \in D} \lambda_d L_d$, where L_{total} enables the network to jointly learn multiple tasks and accurately estimate the liquid volume within the container.

3.3.1 | End-to-End Prediction

3.3.1.1 Design of the network. In our network, the RGB(-D) images first pass through the Focus module when entering the YOLOv5 Backbone. The Focus module slices the image according to a predefined stride and concatenates these slices into a new feature map, effectively condensing the high-resolution spatial information from the original image. The RGB(-D) image is encoded by three hierarchical down-sampling modules, producing three multi-scale feature representations.

Subsequently, these feature tensors are processed through up-sampling or down-sampling operations to achieve feature fusion. Finally, the network outputs prediction tensors at three different scales. To compute the error between the predicted results and the ground-truth labels, we design a hybrid loss function combining mean-squared-error (MSE) and binary cross-entropy (BCE). This hybrid loss is well-suited for multi-task learning, enabling the network to accurately estimate the volume of the liquid.

3.3.1.2 Parameter Settings. We conducted our network training tasks on an NVIDIA A6000 GPU. To accelerate the training process and achieve better convergence, we utilized pre-trained weights from an object classification model trained on ImageNet as the initialization for our network. The batch size was set to 16, and the training was performed for 100 epochs. The initial learning rate was set to 0.01, and we employed a cosine learning rate scheduler to dynamically adjust the learning rate during the training process.

3.3.2 | Multi-Task Learning

Our network supports multi-task learning, enabling it to simultaneously output the bounding box, mask, and volume of the target. In a specific liquid detection scenario, the liquid volume is often strongly correlated with the mask of the liquid and the mask of the container. For a fixed container volume, the liquid volume can be estimated using the mask. By sharing feature vectors across the three task outputs, we want to enhance the performance of the volume estimation task. In our network, we employ multi-task learning techniques by combining MSE and BCE losses to optimize the shared representation and improve overall performance.

The total loss function is expressed as:

$$\mathcal{L}_{\text{total}} = \sum_{t \in D} \lambda_d L_d, \tag{1}$$

where $D = \{\text{box}, \text{obj}, \text{cls}, \text{seg}, \text{vol}\}$ represents the set of all loss components. These include the bounding box regression loss ($\mathcal{L}_{\text{box}}$), objectness loss ($\mathcal{L}_{\text{obj}}$), classification loss ($\mathcal{L}_{\text{cls}}$), mask segmentation loss ($\mathcal{L}_{\text{seg}}$), and volume regression loss ($\mathcal{L}_{\text{vol}}$). The objectness loss ($\mathcal{L}_{\text{obj}}$) refers to the confidence score that indicates the likelihood of an object being present in a predicted bounding box. In other words, it is a measure of how confident the model is that a given region contains an object of interest. In this formulation, λ_d represents the weight assigned to each loss component, and L_d is the loss value for each component $t \in D$.

Our loss function primarily utilizes two types of loss formulations: BCE and MSE. These are used to compute the different components of the total loss.

The objectness loss, classification loss, and segmentation mask loss can be expressed as:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{N} \sum_{i=1}^{N} \text{BCE}(p_{\text{d},i}, t_{\text{d},i}), \tag{2}$$

where $\mathcal{L}_{\text{obj}}$, $\mathcal{L}_{\text{cls}}$, and $\mathcal{L}_{\text{seg}}$ follow this form, with respective $p_{\text{d},i}$ and $t_{\text{d},i}$ representing objectness, class labels, and segmentation masks.

The volume loss, $\mathcal{L}_{\text{vol}}$, is calculated as the MSE between the predicted volume p_{vol} and the target volume t_{vol}:

$$\mathcal{L}_{\text{vol}} = \frac{1}{N} \sum_{i=1}^{N} (p_{\text{vol},i} - t_{\text{vol},i})^2, \tag{3}$$

where N is the number of non-zero target volumes, $p_{\text{vol},i}$ and $t_{\text{vol},i}$ represent the predicted and true volumes for the i-th target, respectively. Only non-zero target volumes are included in the computation to avoid introducing bias from empty objects.

4 | Experiments

To evaluate the performance of our YOLO-based liquid volume estimation network, we conducted experiments in both simulated and real environments. First, we tested the network's volume estimation capability in the simulation environment. Next, we transferred the trained network to real-world scenarios to assess its sim-to-real transferability using the SimLiquid20k data set. Finally, we demonstrated several applications integrated with LLM based on liquid volume estimation to showcase the considerable practical value of our method.

4.1 | Simulation Experiments

4.1.1 | Validation Data Set

Consistent with the training data set, we also used BlenderProc to generate our test data set. The test data set consists of 730 images, and the container and liquid volume distributions are shown in Figure 5. Since data with volumes greater than 1000 mL make up a small proportion of the data set and are

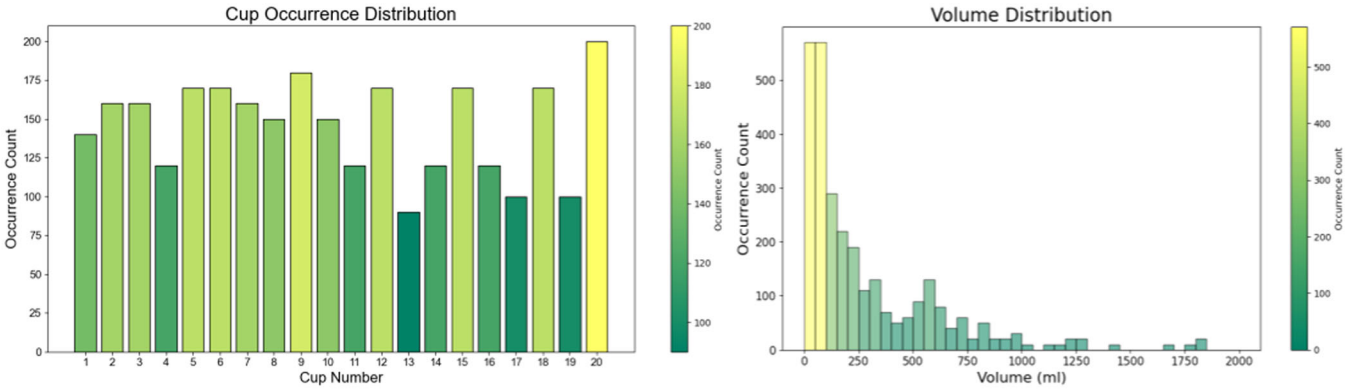

FIGURE 5 | Illustration of validation data distribution. [Color figure can be viewed at wileyonlinelibrary.com]

rarely involved in most operational tasks that require handling excessively large liquid volumes, we only analyzed the data with volumes below 1000 mL in the subsequent tests.

4.1.2 | Performance Across Different Inputs

To investigate the role of depth in the volume estimation task, we applied various data processing techniques. Since depth sensors often suffer from significant distortion when capturing depth information of transparent objects, we specifically processed the depth data for these objects. We set the depth values in the transparent object mask region to zero, thereby creating a lossy simulated depth map. In real-world scenarios, depth completion techniques are commonly used to restore depth information for transparent objects. To simulate this, we employed TransCG by Fang et al. (2022) for depth completion on the lossy simulated depth map, mimicking the depth map obtained after completion in a real-world environment.

The depth-to-3D conversion process is mathematically modeled as follows: Let D be the depth image, a matrix of size $H \times W$, where each element $D_{i,j}$ represents the depth (distance in meters) of the pixel at row i and column j. Using the camera's intrinsic parameters f_x, f_y, c_x, and c_y, the camera intrinsic matrix $\mathbf{K}$ is given by:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

To project a pixel (i, j) from 2D image coordinates to 3D space, we use the inverse of the intrinsic matrix $\mathbf{K}^{-1}$ and the depth value $D_{i,j}$. The 3D coordinates (X, Y, Z) of the pixel (i, j) are computed by:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} D_{i,j}, \tag{5}$$

where $D_{i,j}$ is the depth value at pixel (i, j), and (X, Y, Z) are the 3D coordinates in the camera frame. This mathematical relationship helps in reconstructing the spatial structure of the scene. Additionally, we recorded the camera intrinsic parameters in the simulation environment, allowing us to reconstruct spatially scaled images from the RGB and depth images.

Based on the recorded relative error shown in Table 1, depth estimation plays a critical role in improving the accuracy of liquid volume prediction. The relative error is calculated as:

$$\mathcal{E}_{\text{vol}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|p_{\text{vol},i} - t_{\text{vol},i}|}{t_{\text{vol},i}}, \tag{6}$$

where N is the total number of samples in specific range. This computation allows for a consistent measure of prediction accuracy across different liquid volumes.

The inclusion of depth information, as provided by RGB-D methods, significantly enhances the volume estimation compared to RGB alone. In any liquid volume range between 50 and 1000 mL, the average relative error of the RGB-D method with depth data is consistently lower than that of RGB alone, suggesting that depth information helps in distinguishing liquid from background, leading to more accurate volume predictions.

Among the RGB-D variants, the RGB-D(reconstruct) method, which focuses on reconstructing depth with higher fidelity, consistently performs better than RGB-D(lossy). This indicates that reconstructing depth with more accuracy contributes to improved liquid volume prediction, while lossy depth representations introduce additional errors.

However, the inclusion of depth scaling (RGB-D(scale)) introduces a challenge. While scaling depth to real-world units may seem advantageous, it results in a performance drop in comparison to RGB-D. This drop in performance with scaled depth data may indicate that depth scaling introduces additional noise, which negatively affects the prediction. The effectiveness of depth scaling likely depends on the accuracy of the depth recovery process, which may require more advanced methods, such as 3D networks, to handle the spatial complexity of scaled depth maps. Without such advanced techniques, the use of scaled depth could hinder the overall performance of volume prediction.

TABLE 1 | Experimental results on relative errors for different inputs in different liquid volume ranges.

Volume Range	RGB	RGB-D	RGB-D(Lossy)	RGB-D(Reconstruct)	RGB-D(Scale)
50–100 mL	0.28	0.23	1.48	0.29	0.31
100–150 mL	0.22	0.17	1.06	0.26	0.25
150–200 mL	0.19	0.10	0.47	0.25	0.21
200–250 mL	0.15	0.11	0.50	0.27	0.20
250–300 mL	0.15	0.13	0.33	0.24	0.20
300–350 mL	0.12	0.11	0.47	0.27	0.20
350–400 mL	0.08	0.08	0.39	0.19	0.12
400–450 mL	0.06	0.04	0.25	0.10	0.06
450–500 mL	0.08	0.09	0.32	0.21	0.15
500–550 mL	0.13	0.10	0.30	0.21	0.15
550–600 mL	0.07	0.07	0.23	0.17	0.12
600–650 mL	0.11	0.08	0.24	0.19	0.14
650–700 mL	0.06	0.06	0.29	0.14	0.09
700–750 ml	0.08	0.06	0.40	0.19	0.08
750–800 mL	0.08	0.14	0.29	0.14	0.16
800–850 mL	0.08	0.06	0.27	0.17	0.09
850–900 mL	0.03	0.02	0.14	0.07	0.02
900–950 mL	0.07	0.05	0.64	0.17	0.08
950–1000 mL	0.03	0.04	0.19	0.11	0.07

4.2 | Real Experiments

4.2.1 | Hardware Setup

The hardware setup for the real-world experiments is shown in Figure 6. A RealSense D415i camera was mount near the experiment platform to capture images with a resolution of 640×480. We use a soft gripper as the robotic gripper, which is fixed to the end-effector of the UR3 robotic arm. The experiment platform was covered with a black curtain to minimize the impact of external light sources and surrounding scenes on the experimental results.

In our experiments, due to the limited size of the platform, large containers could not be fully captured by the camera. As a result, all containers used in the real-world experiments had a volume of 300 mL or less. Based on our synthetic data set, we selected three containers with sizes and shapes similar to those in the data set, though each container differed in shape. To accurately measure the required liquid volume, we used a measuring cup with a clear scale as the reference for liquid volume.

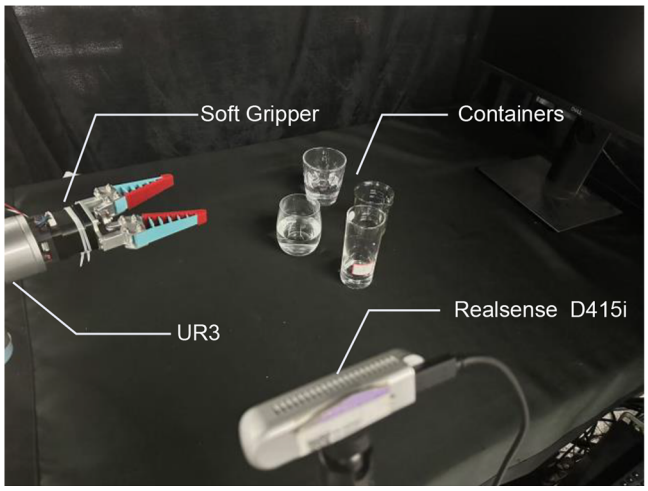

FIGURE 6 | Hardware setup of real experiments. [Color figure can be viewed at wileyonlinelibrary.com]

4.2.2 | Liquid Estimation

In this experiment, due to constraints imposed by the physical setup, such as the limited size of the experimental platform and the distance between the camera and the containers, we selected three transparent containers with volumes not exceeding 300 mL. These containers were chosen based on their similarity to the shapes and sizes of containers in a simulated data set, though it is important to note that the containers in the real-world experiment were not identical in shape.

The focus of this experiment was not to evaluate volume estimation for very low liquid levels because at such low volumes, the liquid surface becomes difficult to distinguish from the bottom of the container. The transparent bottom of the container often introduces complex optical refractions, making it challenging to clearly differentiate between the liquid surface and the container boundary. As a result, we chose to start

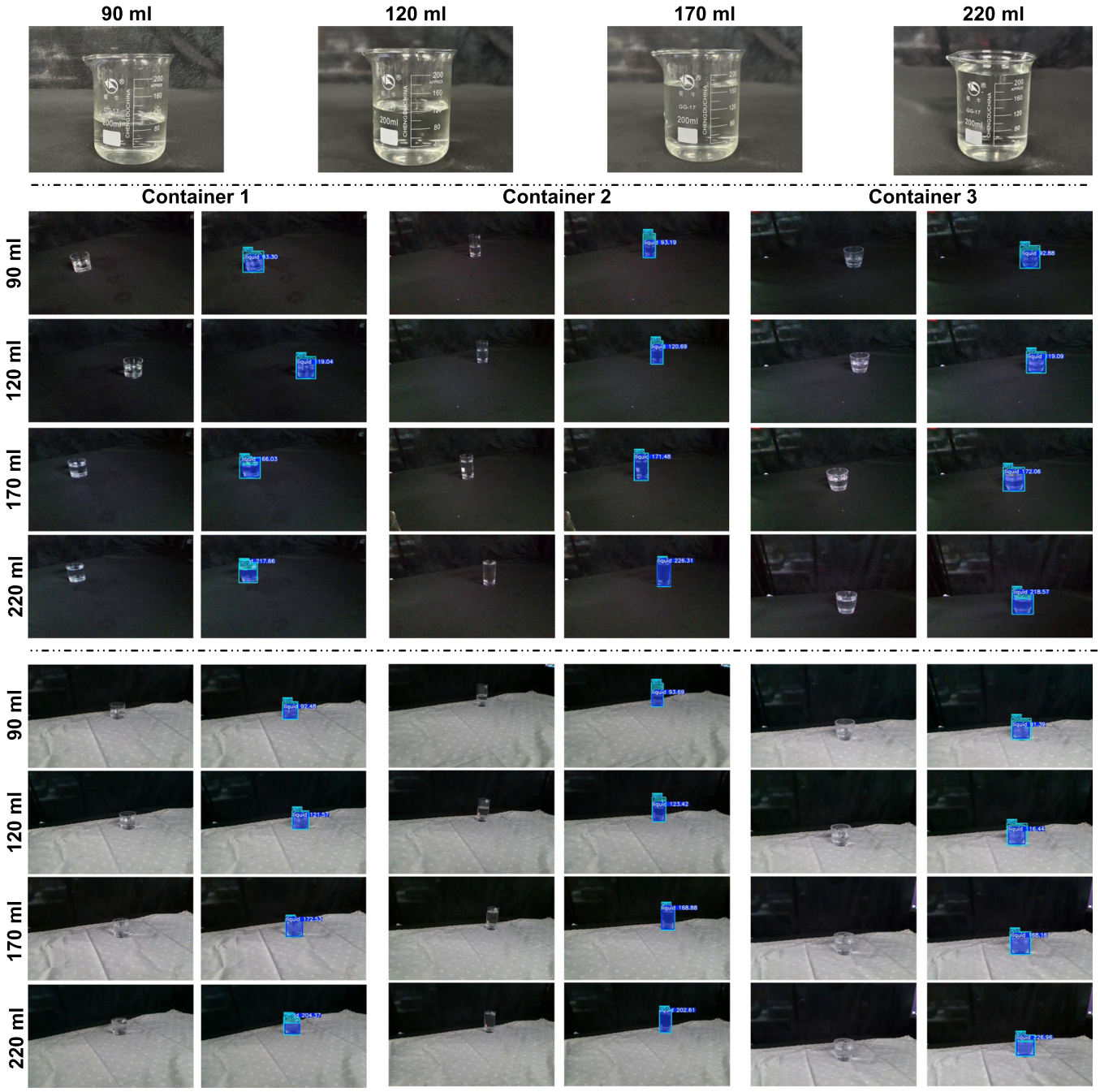

FIGURE 7 | Transparent liquid volume estimation across different backgrounds, containers, and liquid volumes. In our experiments, we selected three containers with slightly different shapes and conducted real-world tests under two types of backgrounds: a plain background and a complex background. We evaluated the system's performance at four different liquid volumes: 90, 120, 170, and 220 mL. [Color figure can be viewed at wileyonlinelibrary.com]

volume measurements from 90 mL and continued to assess volumes at 120, 170, and 220 mL.

This experiment is a model-based task. As illustrated in Figure 7, we selected three containers with similar sizes but slightly different shapes, and conducted liquid volume prediction accuracy tests under both plain and complex backgrounds.

In real-world scenarios, to mitigate temporal jitter during the prediction process, we employed a Kalman filter to obtain stable outputs. During testing, we recorded the prediction results over a period of time, randomly selecting 10 frames from this sequence. The final prediction was obtained by averaging the results of these 10 frames. The corresponding prediction results are presented in Table 2.

The experimental results indicate that our model provides accurate volume predictions with relatively small errors across different containers and volume ranges. The error margin typically falls within a reasonable range, with most errors under 5%. For instance, in the 220 mL range, the relative errors for all three containers are below 5%, with Container 1 having the

smallest error of only 0.76%. These results demonstrate the effectiveness of our volume estimation approach in real-world scenarios.

4.2.3 | Multi-View Estimation

To evaluate the robustness of the liquid volume estimation model under varying camera perspectives, we conducted an experiment where the liquid volume was fixed at 120 mL. The camera's position relative to the container was systematically adjusted within a specified range to simulate different viewpoints.

The results, as shown in Figure 8, demonstrate that the model maintained stable performance across different camera perspectives, with the estimated liquid volume falling within the range of 114.58–122.94 mL. This indicates that the model is robust to variations in camera angle, which is essential for real-world applications where the camera's position may not be fixed.

4.2.4 | Application

In application-focused experiments, we combine liquid volume estimation with LLM to enable task-specific action inference for robotic manipulation. Upon receiving the estimated volume range, we submit the information to the LLM, which analyzes the liquid volume range alongside predefined user requirements. The LLM is tasked with reasoning and inferring the appropriate actions needed to interact with the liquid, ultimately generating commands for the robotic arm.

To ensure the accuracy of the predicted liquid volume and to mitigate potential negative effects from depth inaccuracies, we positioned a calibrated camera directly above the experimental setup. This camera placement minimizes the impact of optical distortions caused by the transparent containers.

As shown in Figure 9, we designed two experiments involving interaction with the LLM. The first task involves combining two 100 mL volumes of water, while the second task involves the robotic arm performing a coffee brewing task. In the first task, two 100 mL cups of liquid are placed on the surface. Using our network, the LLM is provided with the target volume of 200 mL. The LLM then determines that the two liquids can be combined to achieve the desired volume and instructs a robotic arm to perform the task. In the second task, the LLM is informed that 100 mL of water needs to be poured into a cup containing coffee bean. The LLM then drives the robotic arm to carry out the pouring task, stopping once the estimated liquid volume reaches 100 mL.

TABLE 2 | Experimental results of predicting the average of ten data in different containers in different liquid volumes.

Volume range	Container1	Container2	Container3
90 mL	99.32 mL	98.26 mL	91.11 mL
120 mL	122.79 mL	118.04 mL	116.21 mL
170 mL	177.73 mL	172.38 mL	178.42 mL
220 mL	218.34 mL	227.13 mL	230.45 mL

4.2.5 | Limitations

Volume estimation is a task that involves spatial scaling, where depth information or prior knowledge about the scale of surrounding objects plays a crucial role. However, depth information obtained from simulation software exhibits a significant style difference compared to depth data captured from real-world depth sensors. This disparity limits the sim-to-real transferability of transparent object perception and liquid volume estimation tasks. Bridging the gap between simulated and real-world depth data is an important research area for improving liquid volume estimation.

In the tasks deployed in this paper, our approach is a model-based learning method, which requires a sufficiently large data set of containers with similar shapes and sizes to complete the task. However, when faced with untrained containers that have similar shapes but significantly different sizes, the performance tends to be suboptimal. We believe that providing the network with a fixed size reference or spatial scale perception, and leveraging prior knowledge to estimate the container's volume as well as the liquid volume, is a crucial component for achieving generalized liquid volume estimation across various containers.

During the pouring process or when the liquid surface is unstable, the liquid volume estimation often becomes inaccurate. This is because the liquid's dynamic nature and surface fluctuations introduce additional challenges that our current approach struggles to address effectively.

5 | Conclusions

Transparent objects and liquids possess complex optical properties, and liquid volume perception has always been a significant challenge. In this paper, we propose an end-to-end, model-based approach that can predict the liquid volume in a container using monocular visual information. Acquiring liquid

FIGURE 8 | Liquid volume estimation results for different viewpoints in real experiments. [Color figure can be viewed at wileyonlinelibrary.com]

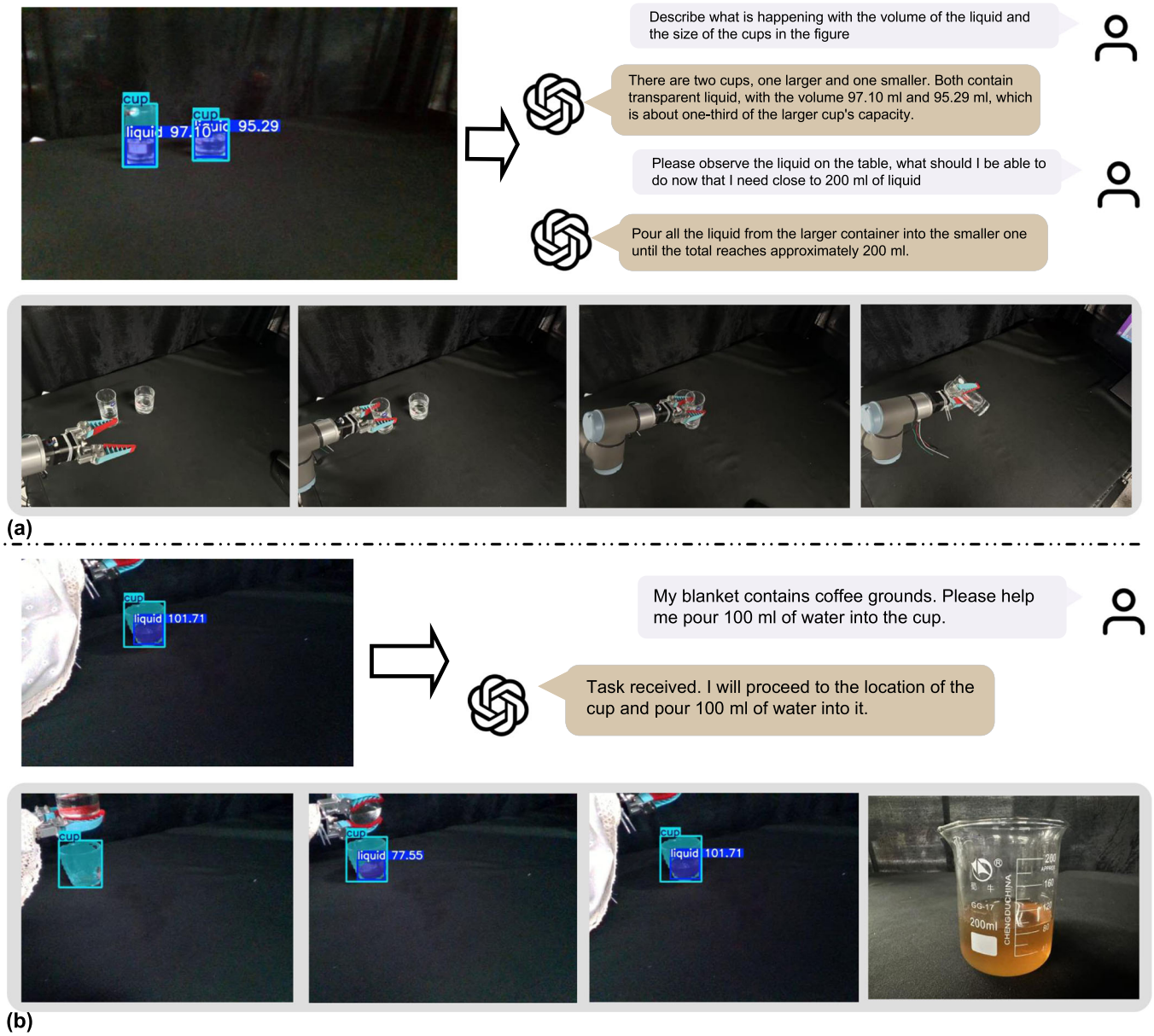

FIGURE 9 | Interact with LLM to accomplish robot manipulation tasks. (a) Illustrates the step-by-step process where the LLM receives the identified volumes of two cups of 100mL water, directing the robot to perform the action of pouring the water together. (b) Demonstrates the interaction between the LLM and the robot for brewing 100 mL of coffee. [Color figure can be viewed at wileyonlinelibrary.com]

volume data is often a challenging task, and unlike other methods, our approach is based on a synthetic data set. We created the SimLiquid20k data set using 20 different container models and liquid materials, and the network trained on this data set can transfer well to real-world scenarios. To explore the roles of RGB and depth information in the task, we tested different input structures on the synthetic data set. We also conducted evaluations in real-world scenarios for tasks involving liquid volumes between 50 and 200 mL, where the average error was approximately 5%, which is acceptable.

Our approach has some limitations that can be addressed in future work. First, our method is model-based and requires pre-modeling containers with similar shapes and sizes. Additionally, the system's performance declines when the liquid is in motion, and unstable liquid volume estimates occur during pouring. For future work, we hope to enable the network to learn prior knowledge, eliminating the need for container modeling. We also aim to explore solutions for stable volume estimation in situations where the liquid surface is unstable.

Acknowledgments

This study was supported by the National Key R&D Program of China under Grant (No. 2024YFB3816000), Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Shenzhen Science and Technology Program (JCYJ20220530143013030), Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197), Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005), Shenzhen Higher Education Stable Support Program (WDZC20231129093657002), and Meituan.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions. The data that supports the findings of this study are available in the Supporting Information of this article.

References

Albrecht, S., and S. Marsland. 2013. "Seeing the Unseen: Simple Reconstruction of Transparent Objects From Point Cloud Data." In Robotics: Science and Systems, Vol. 3, 1–6.

Brandi, S., O. Kroemer, and J. Peters. 2014. "Generalizing Pouring Actions Between Objects Using Warped Parameters." In *2014 IEEE-RAS International Conference on Humanoid Robots*, 616–621. IEEE.

Chen, C. L., J. O. Snyder, and P. J. Ramadge. 2016. "Learning to Identify Container Contents Through Tactile Vibration Signatures." In *2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, 43–48. IEEE.

Dai, Q., Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang. 2023. "Graspnerf: Multiview-Based 6-DOF Grasp Detection for Transparent and Specular Objects Using Generalizable NERF." In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 1757–1763. IEEE.

Denninger, M., and M. Sundermeyer. 2022. "Blenderproc 2.0: A Procedural Pipeline for Photorealistic 3d Scene Generation." *arXiv preprint arXiv:2206.03028*.

Do, C., and W. Burgard. 2019. "Accurate Pouring With an Autonomous Robot Using an RGB-D Camera." In *Intelligent Autonomous Systems 15: Proceedings of the 15th International Conference IAS-15*, 210–221. Springer.

Dong, C., M. Takizawa, S. Kudoh, and T. Suehiro. 2019. "Precision Pouring Into Unknown Containers by Service Robots." In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5875–5882. IEEE.

Fang, H., H.-S. Fang, S. Xu, and C. Lu. 2022. "Transcg: A Large-Scale Real-World Dataset for Transparent Object Depth Completion and a Grasping Baseline." *IEEE Robotics and Automation Letters* 7, no. 3: 7383–7390.

Huang, H.-J., X. Guo, and W. Yuan. 2022. "Understanding Dynamic Tactile Sensing for Liquid Property Estimation." *arXiv preprint arXiv:2205.08771*.

Ichnowski, J., Y. Avigal, J. Kerr, and K. Goldberg. 2021. "DEX-NERF: Using a Neural Radiance Field to Grasp Transparent Objects." *arXiv preprint arXiv:2110.14217*.

Kennedy, M., K. Queen, D. Thakur, K. Daniilidis, and V. Kumar. 2017. "Precise Dispensing of Liquids Using Visual Feedback." In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1260–1266. IEEE.

Kennedy, M., K. Schmeckpeper, D. Thakur, C. Jiang, V. Kumar, and K. Daniilidis. 2019. "Autonomous Precision Pouring From Unknown Containers." *IEEE Robotics and Automation Letters* 4, no. 3: 2317–2324.

Liang, H., C. Zhou, S. Li, et al. 2020. "Robust Robotic Pouring Using Audition and Haptics." In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10880–10887. IEEE.

Liao, J., Y. Fu, Q. Yan, and C. Xiao. 2020. "Transparent Object Segmentation From Casually Captured Videos." *Computer Animation and Virtual Worlds* 31, no. 4–5: e1950.

Liu, X., R. Jonschkowski, A. Angelova, and K. Konolige. 2020. " Keypose: Multi-View 3d Labeling and Keypoint Estimation for Transparent Objects." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11602–11610.

Matl, C., R. Matthew, and R. Bajcsy. 2019. "Haptic Perception of Liquids Enclosed in Containers." In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7142–7149. IEEE.

Narasimhan, G., K. Zhang, B. Eisner, X. Lin, and D. Held. 2022. "Self-Supervised Transparent Liquid Segmentation for Robotic Pouring." In *2022 International Conference on Robotics and Automation (ICRA)*, 4555–4561. IEEE.

Sajjan, S., M. Moore, M. Pan, et al. 2020. "Clear Grasp: 3d Shape Estimation of Transparent Objects for Manipulation." In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 3634–3642. IEEE.

Schenck, C., and D. Fox. 2017a. "Towards Learning to Perceive and Reason About Liquids." In *2016 International Symposium on Experimental Robotics*, 488–501. Springer.

Schenck, C., and D. Fox. 2017b. "Visual Closed-Loop Control for Pouring Liquids." In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2629–2636. IEEE.

Wilson, J., A. Sterling, and M. C. Lin. 2019. "Analyzing Liquid Pouring Sequences via Audio-visual Neural Networks." In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7702–7709. IEEE.

Xie, E., W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo. 2020. "Segmenting Transparent Objects in the Wild." In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII 16, 696–711. Springer.

Yu, R., H. Yu, S. Li, H. Yan, Z. Song, and W. Ding. 2024. Depth Restoration of Hand-Held Transparent Objects for Human-to-Robot Handover. https://arxiv.org/abs/2408.14997.

Zhu, F., R. Jia, L. Yang, et al. 2022. "Visual-Tactile Sensing for Real-time Liquid Volume Estimation In Grasping." In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 12542–12549. IEEE.

Zhu, L., A. Mousavian, Y. Xiang, et al. 2021. RGB-D Local Implicit Function for Depth Completion of Transparent Objects. https://arxiv.org/abs/2104.00622.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.