Intent-Handover: Grounding Language in Human-Usage Regions for Trustworthy Robot-to-Human Handovers

Hanxin Zhang[1,2], Abdulqader Dhafer[1,2], Hongbiao Dong[3], Zhou Daniel Hao[1,2]

***Abstract*— Spoken instructions in robot-to-human handovers may specify either an object ("the cup") or an intended use ("pour water"); in both cases, successful handover requires the robot to infer the target object and the region remaining available for the human to hold. If the robot grasps that hold region, the object could become awkward to receive and immediately use, potentially reducing perceived competence and trust; if the gripper approaches too close to the receiving hand during delivery, perceived safety may also suffer. We present Intent-Handover, which grounds unconstrained speech and visual scene context into explicit grasp and delivery constraints. Given a spoken instruction and a scene observation, a vision–language model identifies the target object and the intended human-usage region. A grasp optimization module then selects a feasible grasp keeping this region accessible while enforcing clearance from the predicted receiving hand. During execution, the robot tracks upper-body key points to estimate the user's receiving pose and places the handover at an ergonomically feasible location. In a within-subjects ablation study (n=30), human-usage region awareness increases perceived trust, hand-gripper collision avoidance increases perceived safety, and interaction comfort is highest when both are enabled. Website and code: https://robot-future.github.io/intent-handover/.**

I. Introduction

Robot-to-human (R2H) object handover is a fundamental capability in human-robot collaboration [1], [2]. Existing approaches attempt to build trustworthy handovers by improving grasp stability [3], delivery safety [4], and user comfort [5], but have not fully considered human's post-handover usage intent. Instructions such as "pour water" implicitly constrain the grasp region: the cup handle may be better left for the human to grasp. A robot may grasp the object stably and deliver it successfully, but the human still needs to readjust their grasp posture before using it. Prior studies suggest that such readjustment can reduce perceived competence and degrade trust [6], [7]. This raises a first concern: the robot may occupy the region the human intends to grasp.

A second concern arises during delivery: the robot end-effector may approach too close to the receiver's hand. Insufficient clearance between the two can potentially reduce perceived safety and comfort [8], [5]. These two concerns are separable: human-usage region awareness shapes perceived competence and trust, while hand-gripper collision avoidance shapes perceived safety and comfort. Trustworthy R2H handover therefore requires satisfying both constraints simultaneously.

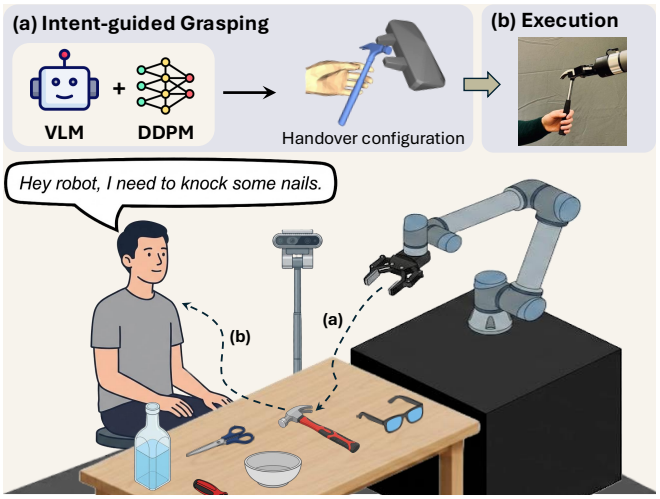

Fig. 1: **Intent-Guided Robot-to-Human Object Handover.** (a) *Intent-Guided Grasping*: A Vision–Language Model and diffusion model infer user intent and human grasp pose from speech and RGB inputs. The system then optimizes the robot grasp configuration. (b) *Execution*: The robot aligns the object with the receiving hand for ergonomic delivery.

To address both concerns, the robot needs to consider which region the human intends to grasp and how the human is likely to grasp it. Current handover systems can identify the target object from human verbal and visual cues [9], [10], [11], but may not reason about intended human grasp region. Systems that account for grasp region typically rely on static affordance labels [12] or demonstration-derived contact maps [13], which are difficult to infer from natural language. In addition, requiring explicit object specification [9] or auxiliary sensing such as wearable devices [14], [15] may limit interaction naturalness. Therefore, the core technical challenge remains: computing an optimal robot grasp from a spoken instruction and an RGB scene image that avoids the human-usage region and prevents hand-gripper collision.

In this paper, we present **Intent-Handover**, an R2H handover system that grounds spoken instructions into explicit spatial constraints and jointly optimizes for human-usage region awareness and hand-gripper collision avoidance. As shown in Fig. 1, the system operates in two phases. In the intent-guided grasping phase (Sec. III-A), a vision–language pipeline maps the spoken instruction and RGB scene image to a structured intent identifying the target object

[1]Authors are with DANiLab, University of Leicester, Leicester, UK {hz273,aamd2,d.hao}@leicester.ac.uk.
[2]Authors are with the School of Computing and Mathematical Sciences, University of Leicester, Leicester, UK.
[3]The author is with the School of Metallurgy and Materials, University of Birmingham, Birmingham, UK. h.dong.1@bham.ac.uk.

and its intended usage region. A grasp optimization module then selects a grasp that avoids the inferred grasp region and minimizes proximity to the predicted receiving hand. During execution, the robot estimates the user's receiving pose from skeletal key points and delivers the object at an ergonomically appropriate location.

We conduct a within-subjects ablation user study with 30 participants, independently disabling human-usage region awareness and hand-gripper collision avoidance, to isolate the effect of each constraint. The study provides empirical evidence for two design hypotheses: (**H1**) users trust the robot more when it delivers the object in a grasp that is easy for them to use, and (**H2**) users feel less safe and less comfortable when the robot does not account for hand-gripper collision avoidance during delivery.

The contributions of this work are as follows.

1) **Intent grounding.** A pipeline from speech and RGB scene image that predicts the human's natural grasp pose for the intended object, jointly resolving the target object and its usage region from unconstrained spoken instructions.
2) **Grasp optimization from inferred intent.** A grasp selection method that enforces human-usage region awareness and hand-gripper collision avoidance, with both constraints derived from natural language rather than predefined labels or demonstrations.
3) **Empirically validated design principles.** A within-subjects ablation user study demonstrating that (1) human-usage region awareness enhances perceived trust, (2) hand-gripper collision avoidance enhances perceived safety, and (3) comfortable handover requires both constraints simultaneously.

II. Related Work

Robot-to-Human Object Handovers. R2H handover requires the robot to infer the receiver's intent and deliver the object safely, comfortably, and in a functionally appropriate manner [2], [6], anticipating the shared grasp pose, usage intent, and handover timing [1]. Intent has been inferred from language [16], gestures [17], or eye gaze [18]; compared to physical modalities, natural language conveys intent at a higher level of semantic abstraction, allowing users to express intended uses directly. In this work, we infer handover intent from language and visual scene context using vision–language models (VLMs), jointly enforcing human-usage region awareness and hand-gripper collision avoidance.

Handover Intent Identification. Human receivers typically communicate handover intent by extending their hands toward the robot [19], [20], [21]; some systems instead infer intent from verbal instructions using speech recognition to identify the target object [9], [10]. These systems typically rely on external sensing devices [14], [15] and require users to name the target object explicitly, without inferring the intended usage region. Vision–language models (VLMs) infer object identity and usage intent without auxiliary hardware. Recent approaches ground handover intent via structured prompting [16], conversational reasoning [22], or predefined skills [11].

However, they identify the target object only, and do not infer which part of the object the user intends to grasp. We introduce a hierarchical prompting strategy that produces structured intent outputs encoding the target object and the region the user will grasp, providing a language prior for predicting the human's natural grasp pose to constrain robot grasp selection.

Grasp Pose Prediction. Most R2H handover systems predict a shared grasp pose to guide handover execution [4]. Meng et al. [23], [5] predict a receiving hand pose using GANHand [24] and select a grasp opposite the hand direction for fast delivery, but do not consider object affordance. Lehotsky et al. [12] target large stable regions identified by AffNet-DR [25] to improve grasp usability, but rely on static labels predefined per object. ContactHandover [13] reranks candidate grasps via a hand–object contact map to avoid unintended hand contact, but does not consider which region the user needs for the intended task. None of these methods jointly enforces human-usage region awareness and hand-gripper collision avoidance as constraints derived from the user's spoken intent. We derive both constraints from natural language rather than from predefined labels or observed demonstrations.

III. Methodology

The Intent-Handover consists of two key stages: intent-guided grasping phase and execution phase, as shown in Fig. 1. During the intent-guided grasping phase, the system first translates user instructions and scene images into textual handover intent, then predicts the corresponding receiving hand pose, and optimizes a robotic grasp. During the execution phase, the system first grasps the object, then localizes the receiver's hand and aligns it with the predicted pose, ultimately completing the handover.

A. *Intent-Guided Grasping Phase*

The intent-guided grasping phase aims to generate a grasp configuration that aligns with the receiver's intended use of the object while minimizing interference with the human hand. As illustrated in Fig. 2, given a spoken instruction and an RGB observation of the tabletop scene, the system first infers the textual handover intent, then predicts a corresponding receiver hand pose together with a set of candidate robot grasps. A final score is computed for each candidate grasp by combining human-usage region awareness and hand-gripper collision avoidance criteria. The grasp with the highest score is selected for execution.

*1) **Intent Identification**:* We translate multimodal user input into structured handover intent using a hierarchical prompting framework.

The process begins when Whisper [26] transcribes the spoken instruction into text T. An Intel RealSense Depth Camera D435i captures the scene image $I \in \mathbb{R}^{H\times W\times C}$. FastVLM [27] parses the scene image I to produce a visual description of the objects. The text-image pair (T, I) is then

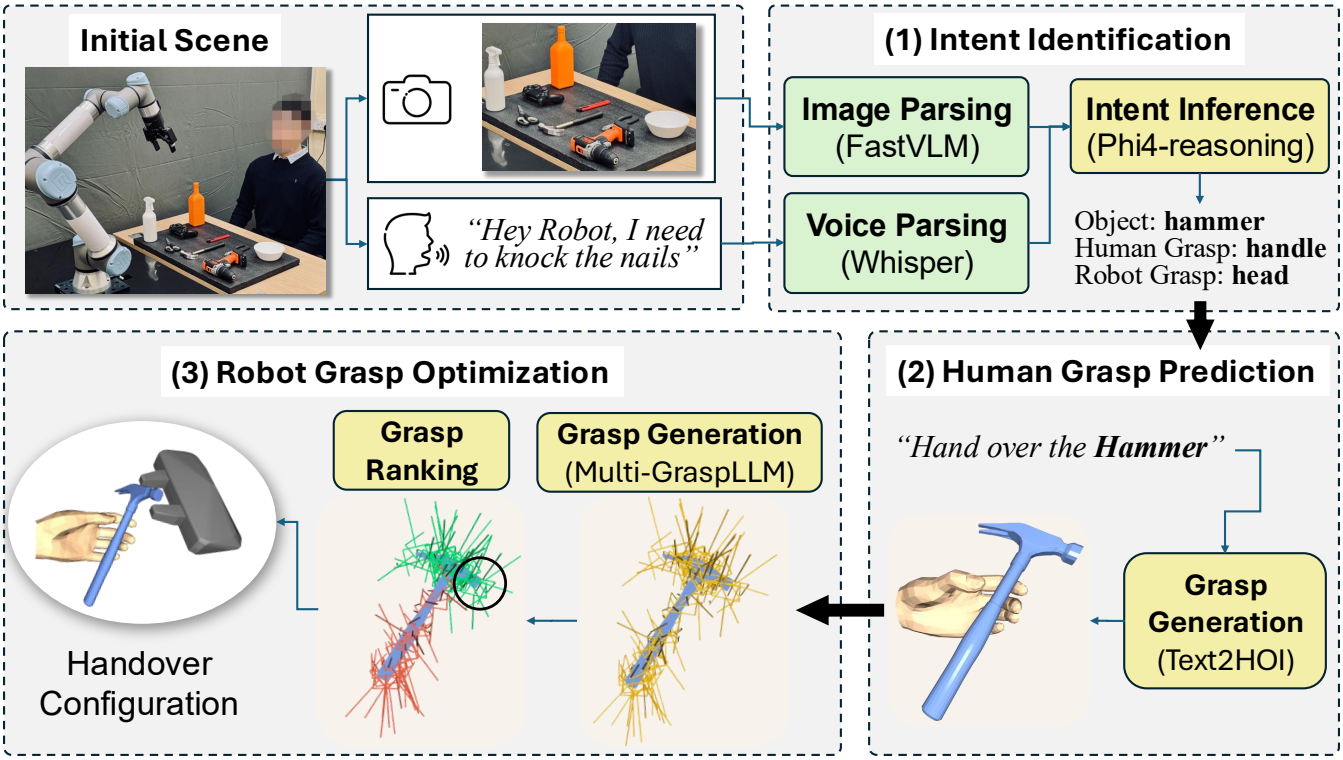

Fig. 2: **Intent-Guided Grasping Phase.** Given a natural language instruction and a scene image, the system proceeds in three steps. (1) A vision-language model (VLM, Phi4-Reasoning) identifies the target object and infers the intended usage region from the instruction and scene image. (2) A diffusion model (Text2HOI) conditioned on language predicts the expected pose of the receiving hand. (3) Candidate robot grasp poses are evaluated based on the predicted hand pose and the intended usage region; the optimal grasp is selected for execution.

processed by Phi4-Reasoning [28], which refines the raw instruction into a structured intent representation T'.

The processed intent T' encodes the predicted object category $\hat{o}$ and a textual usage representation $\hat{u}$ that specifies which object regions are compatible with the intended use, following a fixed template filled by the VLM (e.g., "*I want to drill a hole*" $\rightarrow$ "*the robot grasps the **drill head**, the human grasps the **handle***").

*2) **Human Grasp Prediction**:* To predict the human receiving pose, we train Text2HOI [29], a language-conditioned diffusion model. Given structured intent T' and object point cloud $\mathbf{x}_{obj}$, it generates a plausible interaction using a transformer-based denoiser conditioned on language embeddings and a geometric contact prior.

The output of the model is the predicted MANO parameter vector $\mathbf{x}_{hand}$ of the receiving hand, which defines its global pose and articulated joint configuration in 3D space. We train the interaction generation components of Text2HOI on a handover dataset constructed by integrating data from H2O [30], GRAB [31], and ARCTIC [32] . To construct the training dataset, we manually curated 400 grasp-related trajectories from these datasets, retaining only frames where the hand is in contact with the object and discarding approach and post-use frames. Each retained sequence was manually annotated with a structured intent prompt T'.

During training, the model learns to reconstruct $\mathbf{x}_{hand}$ from $(T', \mathbf{x}_{obj})$, capturing the distribution of natural receiving poses. We use $\mathbf{T}_{ab} \in SE(3)$ to denote the rigid transform that maps coordinates from frame b to frame a. Finally, the predicted pose is transformed to the robot base frame via $\mathbf{T}_{rh} = \mathbf{T}_{rw}\mathbf{T}_{wo}\mathbf{T}_{oh}$, where $\mathbf{T}_{oh}$, $\mathbf{T}_{wo}$, and $\mathbf{T}_{rw}$ denote the hand-to-object, object-to-world, and world-to-robot transforms, respectively.

*3) **Robot Grasp Optimization**:* For each object o in our dataset, we use the grasp annotations provided by Multi-GraspLLM [33] as the candidate grasp set $\mathcal{G}_o = \{g_i\}_{i=1}^{N}$, where each g_i is a 6-DoF gripper pose defined in the object coordinate frame. Let $\mathcal{S}_o$ denote the object surface point set and let $\mathcal{R}_o \subset \mathcal{S}_o$ denote the human-usage region specified by the structured intent T'.

Human-usage Region Awareness. For each candidate grasp g_i, we verify geometric feasibility and ensure the grasp lies outside the human-usage region. Let $w(g_i)$ denote the object width measured along the gripper closing direction under g_i, and let $\mathbf{x}_{int}(g_i) \in \mathcal{S}_o$ denote the object surface point intersected by the gripper approach axis. A grasp g_i is considered valid if it satisfies

$$w(g_i) \leq w_{\max} \ \wedge \ \mathbf{x}_{int}(g_i) \notin \mathcal{R}_o,$$

where $w_{\max}$ is the maximum gripper aperture.

Hand-Gripper Collision Avoidance. Let $\mathbf{v}_h \in \mathbb{R}^3$ denote the wrist-to-middle-finger direction of the predicted receiving

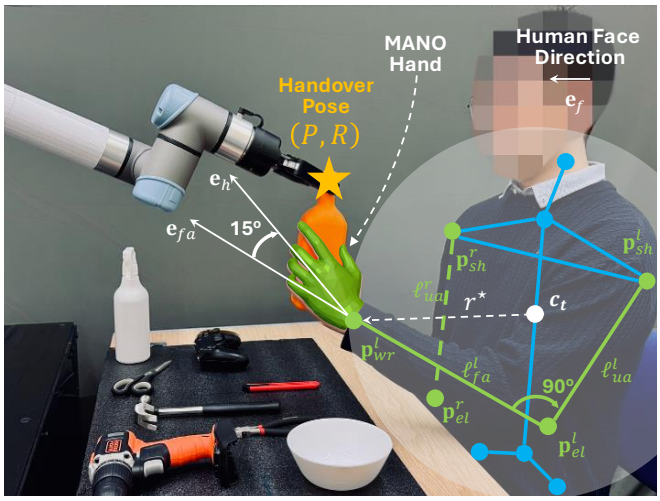

Fig. 3: **Execution Phase: Handover Pose Estimation.** The handover position P and orientation R are derived from skeletal keypoints to ensure comfortable reachability. See Sec. III-B for details.

hand, and let $\mathbf{v}_{g_i} \in \mathbb{R}^3$ denote the approach direction of g_i. Let $\mathbf{p}_h$ and $\mathbf{p}_{g_i}$ denote the centers of the hand and gripper, respectively. We define the avoidance cost

$$A(g_i) = \frac{\mathbf{v}_{g_i} \cdot \mathbf{v}_h}{\|\mathbf{v}_{g_i}\|\|\mathbf{v}_h\|} - \|\mathbf{p}_{g_i} - \mathbf{p}_h\|.$$

The first term is minimized when the gripper and hand approach from opposite directions ($\cos\theta \to -1$), maximizing spatial separation during handover. The second term favors grasps whose center is far from the predicted hand center.

Final selection. The optimal grasp minimizes the avoidance cost among all feasible candidates:

$$g^\star = \arg\min_{g_i \in \mathcal{G}_o} A(g_i) \quad \text{s.t.} \quad w(g_i) \le w_{\max},\ \mathbf{x}_{int}(g_i) \notin \mathcal{R}_o.$$

The selected grasp $g^\star$ is passed to the execution module.

B. *Execution Phase*

As illustrated in Fig. 3, the robot estimates the handover position and orientation using an RGB perception setup with depth estimation. An Intel RealSense D435i camera is mounted at a height of $2.30\,\mathrm{m}$ above the floor and positioned $1.90\,\mathrm{m}$ away from the calibrated world-frame origin in the horizontal plane, at a $45°$ diagonal offset from the table front edge. Its optical axis is oriented toward the center of the workspace at a $45°$ horizontal angle. All observations are transformed into a unified world frame via extrinsic calibration, and 3D skeletal key points are extracted using MediaPipe [34].

*1) **Handover Position**:* The handover position is defined to ensure reachability and arm comfort for a seated user. Following ergonomic principles [35] and prior handover studies [13], we model the human reachable workspace as a sphere [36], [37] centered at an upper-body reference frame.

Let $\mathbf{p}^l_{sh}, \mathbf{p}^r_{sh} \in \mathbb{R}^3$ denote the left and right shoulder positions. The shoulder midpoint is

$$\mathbf{c}_{sh} = \frac{\mathbf{p}^l_{sh} + \mathbf{p}^r_{sh}}{2}.$$

Let z_{desk} denote the desk height. To account for seated tabletop interaction, the torso reference center is defined as

$$\mathbf{c}_t = \left[(\mathbf{c}_{sh})_x, (\mathbf{c}_{sh})_y, \frac{(\mathbf{c}_{sh})_z + z_{\text{desk}}}{2}\right].$$

Let $\mathbf{p}_{el}, \mathbf{p}_{wr}$ denote the elbow and wrist positions, and define the upper-arm and forearm lengths as

$$\ell_{ua} = \|\mathbf{p}_{el} - \mathbf{c}_{sh}\|, \quad \ell_{fa} = \|\mathbf{p}_{wr} - \mathbf{p}_{el}\|.$$

To avoid extreme extension, we select a mid-range posture with an elbow flexion of approximately $90°$, which is regarded as a neutral configuration in ergonomics [38]. The corresponding comfortable reach radius is

$$r^\star = \sqrt{\ell^2_{ua} + \ell^2_{fa}}.$$

Let $\mathbf{e}_f$ denote the unit horizontal facing direction of the human, computed as the horizontal component of the vector perpendicular to the left-right shoulder axis $(\mathbf{p}^r_{sh} - \mathbf{p}^l_{sh})$, pointing forward. The delivery position is defined as

$$P = \mathbf{c}_t + r^\star \mathbf{e}_f.$$

The predicted handover configuration specifies a fixed object-to-end-effector transform $\mathbf{T}_{eo}$. Therefore, the delivery position P uniquely determines the target end-effector position.

*2) **Handover Orientation**:* The user is assumed to receive the object at delivery position P. The hand orientation is determined from the forearm direction estimated from upper-limb keypoints.

Let $\mathbf{e}_{fa}$ denote the unit forearm direction from elbow to wrist. Natural grasping involves a slight wrist extension rather than strict collinearity with the forearm axis; we adopt $15°$ as it minimizes musculoskeletal cost while remaining within the comfortable wrist extension range [39],

$$\mathbf{e}_h = R_{ext}(15°)\,\mathbf{e}_{fa},$$

where $R_{ext}(15°)$ denotes a rotation of $15°$ in the wrist flexion–extension plane.

Let $\mathbf{e}_{mano}$ denote the canonical wrist-to-middle-finger axis of the MANO model. The target hand orientation $R \in SO(3)$ is the minimum-angle rotation that aligns $\mathbf{e}_{mano}$ with $\mathbf{e}_h$:

$$R\,\mathbf{e}_{mano} = \mathbf{e}_h, \quad R = \arg\min_{\tilde{R}} \|\tilde{R} - I\|_F \ \text{ s.t. } \ \tilde{R}\,\mathbf{e}_{mano} = \mathbf{e}_h.$$

The predicted handover configuration specifies a fixed hand-to-end-effector transform $\mathbf{T}_{eh}$. Therefore, the target end-effector orientation is directly determined by R.

IV. USER STUDY

All participants provided written informed consent prior to participation.

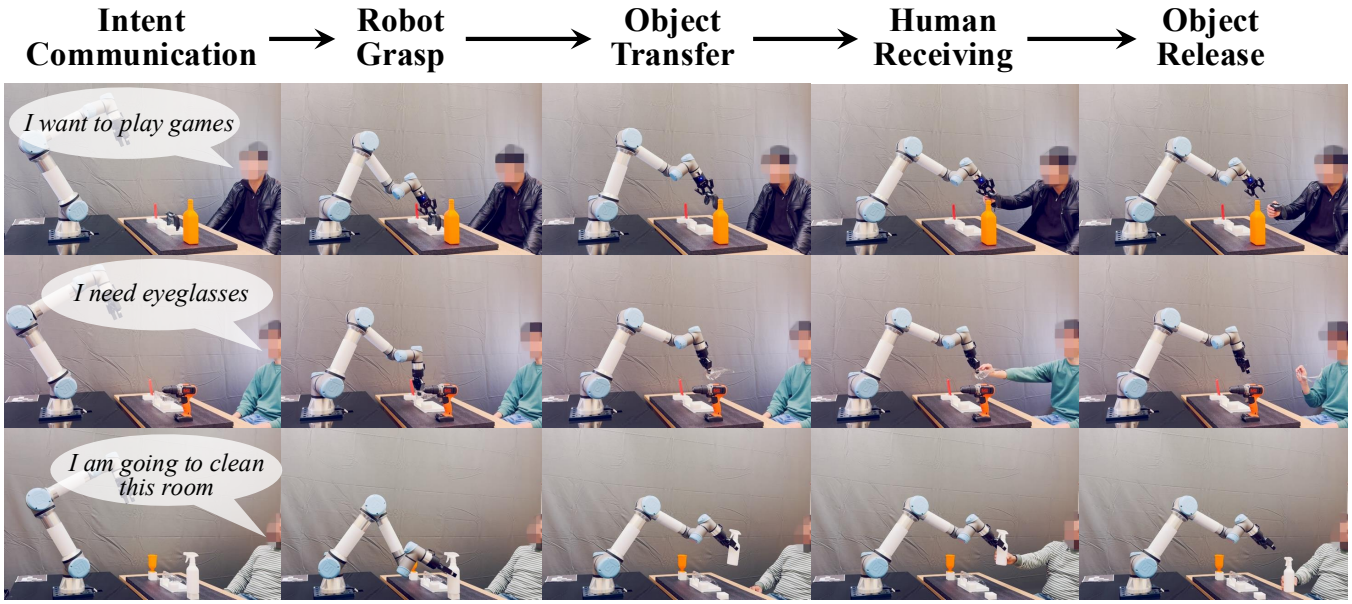

Fig. 4: **User Study Procedure.** Each trial follows five stages: intent communication, robot grasp, object transfer, human receiving, and object release. Each row shows a complete handover sequence for one participant. Three representative participants are shown.

A. Research Questions and Hypotheses

We isolate the effects of two grasp optimization strategies (human-usage region awareness and hand-gripper collision avoidance) through two research questions.

(**RQ1**) To what extent does the robot's grasp region affect user trust during handover? Correctly inferring the intended object alone is insufficient; grasping an inappropriate region may still undermine trust.

- **H1.** Users trust the robot more when it delivers the object in a grasp that is easy for them to use.

(**RQ2**) To what extent does failing to consider the user's hand affect perceived comfort and safety? Even if users can adjust their receiving pose, neglecting hand-gripper collision avoidance may reduce psychological safety.

- **H2.** Users feel less safe and comfortable when the robot does not optimize for hand-gripper collision avoidance.

B. Procedure

To evaluate the impact of our system on users' perceived trust and safety during physical handovers, we conducted a mixed-design user study. All participants experienced four system variants: the **Full Strategy** (**FS**) and three ablations:

- **Ablation 1 (A1): No Human-usage Region Awareness:** The robot does not consider whether a predicted grasp supports easy human use after handover. For example, it may grasp the blade of scissors rather than the handle, hindering user reception and use.
- **Ablation 2 (A2): No Hand-Gripper Collision Avoidance:** The robot disables constraints for hand-gripper collision avoidance, and the predicted grasp may result in contact with the user's hand.
- **Ablation 3 (A3): A1 + A2:** The robot selects a grasp solely based on whether it can grasp the object, ignoring both constraints.

The hardware platform consists of a Universal Robots UR5e arm equipped with a Robotiq 2F-85 parallel-jaw gripper. The object dataset comprises 16 everyday objects spanning tools, kitchenware, and household items. Object selection was randomized across trials such that every object appeared under every condition at least once across the full participant pool. In the experimental setup, three objects from this dataset are placed adjacently in front of the user with specific spatial intervals. Fig. 4 illustrates the five stages of a complete handover trial: intent communication, robot grasp, object transfer, human receiving, and object release. Some items, such as game controllers and knives, are elevated or supported, facilitating smooth grasp execution. Participants then express their intent in natural language by either naming the desired object or describing how they plan to use it (e.g., "*I want scissors*" or "*I want to cut paper*"). The system generates a corresponding grasp configuration, actuates the robotic arm to grasp the object from a fixed initial position, and delivers it to the human receiver. An embedded force-torque sensor within the end effector dictates the object release timing by triggering the gripper to release when the sensed force exceeds a preset threshold.

To mitigate order effects, the presentation order of the four conditions was independently randomized for each participant. Each participant completed 20 trials (five for each condition). After each block of five trials for a given condition, they completed a questionnaire.

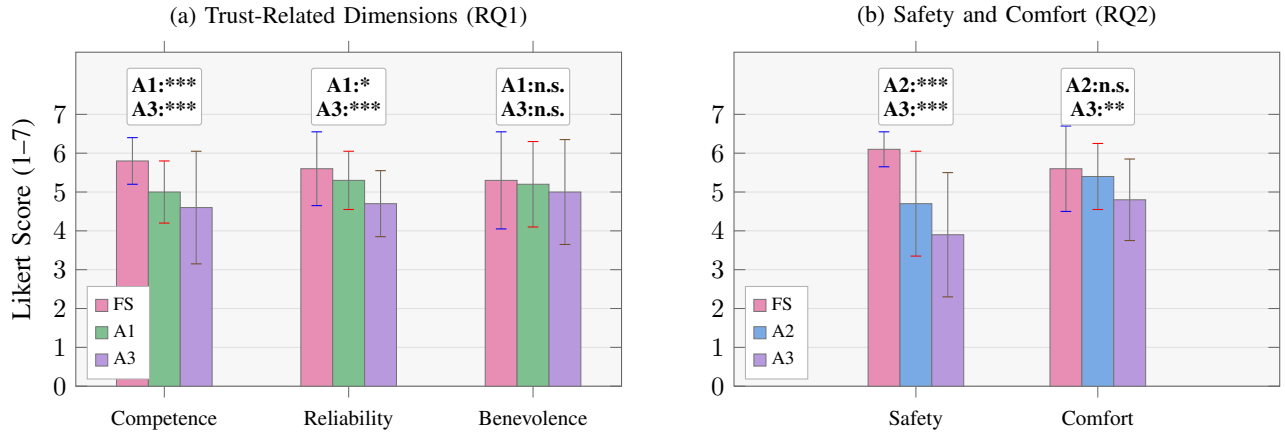

Fig. 5: **User-Study Results: Trust, Perceived Safety, and Interaction Comfort.** (a) Mean trust ratings (competence, reliability, benevolence) across FS, A1, and A3. (b) Mean perceived safety and interaction comfort ratings across FS, A2, and A3. Bars show mean Likert scores (1–7); error bars show ±SD. Annotation boxes show Bonferroni-corrected post-hoc comparisons against FS ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, n.s. not significant; all p-values are Bonferroni-adjusted). See TABLE I for full statistics.

TABLE I: Pairwise comparison results for the user-study conditions (7-point Likert means; $n = 30$, $\mathrm{df} = 29$; Bonferroni-corrected post-hoc comparisons following significant repeated-measures ANOVAs; n.s. not significant). Trust dimensions (competence, reliability, benevolence) are from MDMT v2; perceived safety and interaction comfort are from separate validated scales.

Measure	FS (M)	Abl. (M)	t	p	d_z	Effect
MDMT (Trust, FS vs A1)						
Competence	5.8	5.0	4.20	$< .001$	0.77	FS > A1
Reliability	5.6	5.3	2.45	.02	0.45	FS > A1
Benevolence	5.3	5.2	0.60	.55	0.11	n.s.
MDMT (Trust, FS vs A3)						
Competence	5.8	4.6	6.00	$< .001$	1.10	FS > A3
Reliability	5.6	4.7	5.20	$< .001$	0.95	FS > A3
Benevolence	5.3	5.0	2.10	.04	0.38	n.s.
Safety & Comfort (FS vs A2)						
Perceived Safety	6.1	4.7	6.58	$< .001$	1.20	FS > A2
Interaction Comfort	5.6	5.4	1.45	.16	0.26	n.s.
Safety & Comfort (FS vs A3)						
Perceived Safety	6.1	3.9	7.67	$< .001$	1.40	FS > A3
Interaction Comfort	5.6	4.8	3.65	$< .01$	0.67	FS > A3

C. Evaluation Measures

We assessed participants' trust, perceived safety, and interaction comfort using validated questionnaire instruments. After completing each system condition (FS, A1, A2, A3), participants responded to a questionnaire battery covering all measures, with each item rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). Shapiro-Wilk tests indicated no significant deviation from normality for any measure in any condition ($p > .05$). For each measure, we first conducted a one-way repeated-measures ANOVA across the four conditions; where a significant main effect was found, we performed Bonferroni-corrected pairwise comparisons.

To assess **RQ1** (trust), we employed the Multi-Dimensional Measure of Trust (MDMT) v2 [40], selecting three subscales: *competence* (perceived capability of grasping and delivering appropriately), *reliability* (perceived consistency across trials), and *benevolence* (perceived intent to act in the user's interest).

To assess **RQ2**, we measured *perceived safety* [8] (how safe participants felt regarding hand-gripper contact) and *interaction comfort* [19] (how natural and effortless it was to receive and immediately use the object).

V. RESULTS

A total of 30 participants were recruited to participate without compensation through local recruitment. Participants self-identified as men (n=17) and women (n=13). Ages of the participants ranged from 22 to 35 years (M = 26.3, SD = 2.5). For each measure, we first ran a one-way repeated-measures ANOVA across the four conditions; where a significant main effect was found ($p < .05$), we followed up with Bonferroni-corrected pairwise comparisons.

A. Effect of Grasp Region on Trust (RQ1)

To answer RQ1, we compared the full strategy (**FS**) against **A1** and **A3** using Bonferroni-adjusted pairwise tests (TABLE I). In the **FS** vs **A1** contrast, dropping human-usage region awareness lowers perceived *competence* ($p < .001$) and *reliability* ($p = .02$) but leaves *benevolence* unchanged ($p = .55$), meaning participants felt the robot was less capable and dependable, while the impression of considerateness stayed the same. In the **FS** vs **A3** contrast, the drops in *competence* ($p < .001$) and *reliability* ($p < .001$) are larger, while *benevolence* did not reach significance after Bonferroni correction ($p = .04 > .025$), meaning capability and dependability fall further when both strategies are absent, while

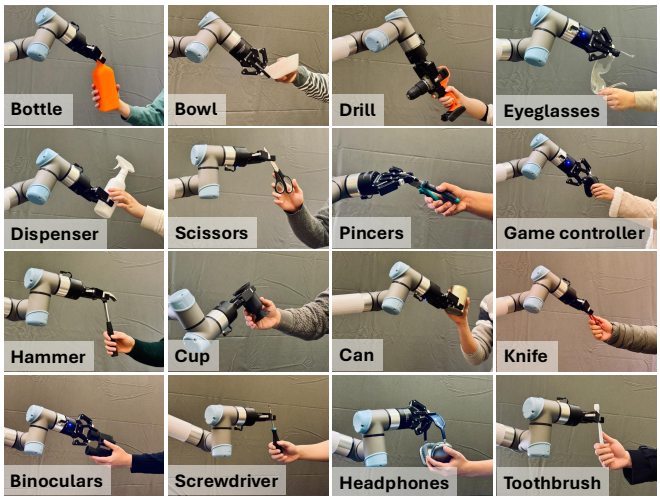

Fig. 6: **Representative Full-Strategy Handover Cases.** Each example shows an object selected from the user's spoken intent, grasped with human-usage region awareness, and delivered to the user's hand. Cases span multiple object categories and intent expressions, providing qualitative support for the objective results in TABLE II.

perceived care remains unaffected. Overall, *competence* is the most sensitive trust dimension; *benevolence* did not reach significance in either contrast after Bonferroni correction, though the FS vs A3 comparison ($p = .04$) suggests a possible trend that warrants further investigation. These results support **H1**. This is likely because participants treat the grasp region as a signal of "does the robot understand how I will use this," so *competence* and *reliability* respond to task-level failures, while the *benevolence* judgment remains stable regardless of the ablation. Results are visualized in Fig. 5.

B. Effect of Hand-Gripper Collision Avoidance on Safety and Comfort (RQ2)

To answer RQ2, we compared the full strategy (**FS**) against **A2** and **A3** using Bonferroni-adjusted pairwise tests (TABLE I). In the **FS** vs **A2** contrast, disabling hand-gripper collision avoidance sharply reduces *perceived safety* ($p < .001$) while *interaction comfort* does not change ($p = .16$), meaning people feel less safe but the handover can still feel similarly comfortable. In the **FS** vs **A3** contrast, **FS** is rated higher on both *safety* ($p < .001$) and *comfort* ($p < .01$), meaning comfort drops most when both hand-gripper collision avoidance and human-usage region awareness are absent. These findings partially support **H2**: disabling hand-gripper collision avoidance alone significantly reduces perceived safety, but comfort does not decrease until human-usage region awareness is also removed (A3). This is likely because people can adjust their receiving pose to compensate, preserving comfort; however, the perceived risk of collision still reduces safety. Comfort drops only when the grasp additionally makes the object awkward to receive.

TABLE II: Objective performance of **Intent-Handover (FS)** over 150 trials (30 participants, 5 trials each): pipeline success rates and mean processing times per stage.

Category	Metric	Value
Success Rates		
Intent identification	Correct object and grasp region inferred from natural language	88.00%
Handover execution	Robot delivered object to user's hand (given correct intent)	83.33%
Overall pipeline	Identification correct & handover successful	73.33%
Average Time		
Reasoning stage	Mean intent reasoning time	2.47 s
Execution stage	Mean handover execution time	8.83 s
Overall pipeline	Mean total time	11.30 s

C. Full-Strategy Pipeline Performance

TABLE II summarizes direct logs for **FS** over 150 trials (30 participants, 5 FS trials each), without questionnaire scoring. Intent identification succeeded in 88.00% of trials (132/150), handover execution succeeded in 83.33% of correctly identified trials (110/132), and the combined overall pipeline achieved a success rate of 73.33% (110/150). Intent identification failures mainly occurred when multiple objects in the scene were semantically compatible with the user's instruction (e.g., several objects could plausibly match the described use), causing the VLM to select an unintended target. Handover execution failures were primarily caused by the selected grasp pose leading to infeasible motion plans for the robot arm. Mean reasoning time was 2.47 s, mean execution time was 8.83 s, and mean total pipeline time was 11.30 s. Fig. 6 provides representative **FS** handover cases as qualitative support for these objective logs.

VI. CONCLUSIONS

We presented **Intent-Handover**, a robot-to-human handover system that performs handovers directly from natural language via vision–language intent inference and grasp optimization. Our user study indicates that human-usage region awareness mainly improves trust through perceived competence and reliability, hand-gripper collision avoidance mainly improves perceived safety, and comfort is best preserved when both are enabled. Future work will extend the system to richer language interaction, broader objects and users, and additional robot platforms.

REFERENCES

[1] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, "Object handovers: A review for robotics," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1855–1873, 2021.

[2] H. Duan, Y. Yang, D. Li, and P. Wang, "Human-robot object handover: Recent progress and future direction," *Biomimetic Intelligence and Robotics*, vol. 4, no. 1, p. 100145, 2024.

[3] P. Ardón, M. E. Cabrera, È. Pairet, R. P. A. Petrick, S. Ramamoorthy, K. S. Lohan, and M. Cakmak, "Affordance-aware handovers with human arm mobility constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3136–3143, 2021.

[4] A. Megyeri, N. Wiederhold, Y. Liu, S. Banerjee, and N. K. Banerjee, "Safety and naturalness perceptions of robot-to-human handovers performed by data-driven robotic mimicry of human givers," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1869–1875, IEEE, 2025.

[5] C. Meng, T. Zhang, D. Zhao, and T. L. Lam, "Fast and comfortable robot-to-human handover for mobile cooperation robot system," *Cyborg and Bionic Systems*, vol. 5, p. 0120, 2024.

[6] V. Ortenzi, F. Cini, T. Pardi, N. Marturi, R. Stolkin, P. Corke, and M. Controzzi, "The grasp strategy of a robot passer influences performance and quality of the robot-human object handover," *Frontiers in Robotics and AI*, vol. 7, p. 542406, 2020.

[7] F. Cini, V. Ortenzi, P. Corke, and M. Controzzi, "On the choice of grasp type and location when handing over an object," *Science Robotics*, vol. 4, no. 27, p. eaau9757, 2019.

[8] P. A. Lasota, T. Fong, and J. A. Shah, "A survey of methods for safe human-robot interaction," *Foundations and Trends in Robotics*, vol. 5, pp. 261–349, May 2017.

[9] E. Herrera, M. Lyons, J. Parron, R. Li, M. Zhu, and W. Wang, "Learning-finding-giving: A natural vision-speech-based approach for robots to assist humans in human-robot collaborative manufacturing contexts," in *2024 IEEE 4th International Conference on Human-machine Systems (ICHMS)*, pp. 1–6, 2024.

[10] D. Langer, F. Legler, P. Kotsch, A. Dettmann, and A. C. Bullinger, "I let go now! towards a voice-user interface for handovers between robots and users with full and impaired sight," *Robotics*, vol. 11, no. 5, p. 112, 2022.

[11] H. Fei, T. Xue, Y. He, S. Lin, G. Du, Y. Guo, and Z. Wang, "Large language model-driven natural language interaction control framework for single-operator bimanual teleoperation," *Frontiers in Robotics and AI*, vol. 12, p. 1621033, 2025.

[12] D. Lehotsky, A. Christensen, and D. Chrysostomou, "Optimizing robot-to-human object handovers using vision-based affordance information," in *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, IEEE, 2023.

[13] Z. Wang, Z. Liu, N. Ouporov, and S. Song, "ContactHandover: Contact-guided robot-to-human object handover," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9916–9923, IEEE, 2024.

[14] R. Zou, Y. Liu, Y. Li, G. Chu, J. Zhao, and H. Cai, "A novel human intention prediction approach based on fuzzy rules through wearable sensing in human–robot handover," *Biomimetics*, vol. 8, no. 4, p. 358, 2023.

[15] C. Liu, W. Wang, C. Li, R. Pang, Y. Shang, and Y. Gao, "Human-to-robot handovers based on multimodal perception," *Neurocomputing*, vol. 645, p. 130435, 2025.

[16] A. Tulbure, R. Zurbrügg, T. Grigat, and M. Hutter, "Llm-handover: Exploiting llms for task-oriented robot-human handovers," *IEEE Robotics and Automation Letters*, 2025.

[17] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschka, and D. Kragic, "Predicting human intention in visual observations of hand/object interactions," in *2013 IEEE International Conference on Robotics and Automation*, pp. 1608–1615, IEEE, 2013.

[18] S. Li, M. Bowman, H. Nobarani, and X. Zhang, "Inference of manipulation intent in teleoperation for robotic assistance," *Journal of Intelligent and Robotic Systems*, vol. 99, no. 1, pp. 29–43, 2020.

[19] K. W. Strabala, M. K. Lee, A. D. Dragan, J. L. Forlizzi, S. Srinivasa, M. Cakmak, and V. Micelli, "Towards seamless human-robot handovers," *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 112–132, 2013.

[20] K. Strabala, M. K. Lee, A. Dragan, J. Forlizzi, and S. S. Srinivasa, "Learning the communication of intent prior to physical collaboration," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pp. 968–973, IEEE, 2012.

[21] S. Roy and Y. Edan, "Investigating joint-action in short-cycle repetitive handover tasks: The role of giver versus receiver and its implications for human-robot collaborative system design," *International Journal of Social Robotics*, vol. 12, no. 5, pp. 973–988, 2020.

[22] Y. Lakhnati, M. Pascher, and J. Gerken, "Exploring a gpt-based large language model for variable autonomy in a vr-based human-robot teaming simulation," *Frontiers in Robotics and AI*, vol. 11, p. 1347538, 2024.

[23] C. Meng, T. Zhang, and T. lun Lam, "Fast and comfortable interactive robot-to-human object handover," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3701–3706, IEEE, 2022.

[24] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5031–5041, 2020.

[25] A. D. Christensen, D. Lehotský, M. W. Jørgensen, and D. Chrysostomou, "Learning to segment object affordances on synthetic data for task-oriented robotic handovers," in *The 33rd British Machine Vision Conference*, British Machine Vision Association, 2022.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.

[27] P. K. A. Vasu, F. Faghri, C.-L. Li, C. Koc, N. True, A. Antony, G. Santhanam, J. Gabriel, P. Grasch, O. Tuzel, and H. Pouransari, "Fastvlm: Efficient vision encoding for vision language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19769–19780, June 2025.

[28] M. Abdin, S. Agarwal, A. Awadallah, V. Balachandran, H. Behl, L. Chen, G. de Rosa, S. Gunasekar, M. Javaheripi, N. Joshi, P. Kauffmann, Y. Lara, C. C. T. Mendes, A. Mitra, B. Nushi, D. Papailiopoulos, O. Saarikivi, S. Shah, V. Shrivastava, V. Vineet, Y. Wu, S. Yousefi, and G. Zheng, "Phi-4-reasoning technical report," 2025.

[29] J. Cha, J. Kim, J. S. Yoon, and S. Baek, "Text2hoi: Text-guided 3d motion generation for hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

[30] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2o: Two hands manipulating objects for first person interaction recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10138–10148, October 2021.

[31] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision*, pp. 581–600, Springer, 2020.

[32] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] H. Li, W. Mao, W. Deng, C. Meng, H. Fan, T. Wang, P. Tan, H. Wang, and X. Deng, "Multi-GraspLLM: A multimodal LLM for multi-hand semantic guided grasp generation," Dec. 2024.

[34] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[35] M. Katayama and H. Hasuura, "Optimization principle determines human arm postures and "comfort"," in *SICE Annual Conference Program and Abstracts SICE Annual Conference 2003*, pp. 47–47, The Society of Instrument and Control Engineers, 2003.

[36] Q. Liu, J. Ren, Q. Zhang, and M. Hua, "Seated reach capabilities for ergonomic design and evaluation with consideration of reach difficulties," *Applied Ergonomics*, vol. 59, pp. 357–363, 2017.

[37] H. Hsiao and W. M. Keyserling, "Evaluating posture behavior during seated tasks," *International Journal of Industrial Ergonomics*, vol. 8, no. 4, pp. 313–334, 1991.

[38] N. A. Baker and K. Moehling, "The relationship between musculoskeletal symptoms, postures and the fit between workers' anthropometrics and their computer workstation configuration," *Work*, vol. 46, no. 1, pp. 3–10, 2013.

[39] W. L. Popp, L. Richner, O. Lambercy, C. Shirota, A. Barry, R. Gassert, and D. G. Kamper, "Effects of wrist posture and stabilization on precision grip force production and muscle activation patterns," *Journal of Neurophysiology*, vol. 130, no. 3, pp. 596–607, 2023.

[40] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in human-robot interaction*, pp. 3–25, Elsevier, 2021.