Point & Grasp: Flexible Selection of Out-of-Reach Objects Through Probabilistic Cue Integration

Xuejing Luo
Department of Information and Communication Engineering, School of Electrical Engineering
Aalto University
Helsinki, Finland
ELLIS Institute
Helsinki, Finland
xuejing.luo@aalto.fi

Hee-Seung Moon*
School of Computer Science and Engineering
Chung-Ang University
Seoul, Republic of Korea
hsmoon@cau.ac.kr

Christian Holz
ETH Zürich
Zurich, Switzerland
christian.holz@ethz.ch

Antti Oulasvirta
Department of Information and Communication Engineering, School of Electrical Engineering
Aalto University
Helsinki, Finland
ELLIS Institute
Helsinki, Finland
antti.oulasvirta@aalto.fi

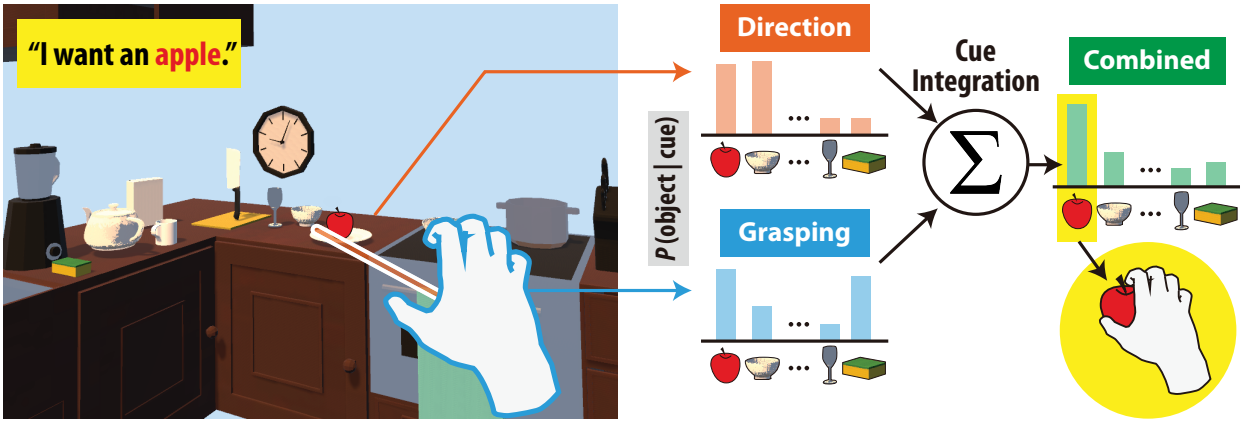

Figure 1: Point&Grasp is a technique for selecting out-of-reach objects in virtual reality. While pointing direction suffers in cluttered scenes and grasping gestures become ambiguous when objects share similar shapes, Point&Grasp combines both cues to enable more precise and usable target selection.

Abstract

Selecting out-of-reach objects is a fundamental task in mixed reality (MR). Existing methods rely on a single cue or deterministically fuse multiple cues, leading to performance degradation when the dominant cue becomes unreliable. In this work, we introduce a probabilistic cue integration framework that enables flexible combination of multiple user-generated cues for intent inference. Inspired by natural grasping behavior, we instantiate the framework with pointing direction and grasp gestures as a new interaction technique, Point&Grasp. To this end, we collect the Out-of-Reach Grasping (ORG) dataset to train a robust likelihood model of the gestural cue, which captures grasping patterns not present in existing in-reach datasets. User studies demonstrate that our selection method with cue integration not only improves accuracy and speed over single-cue baselines, but also remains practically effective compared to state-of-the-art methods across various sources of ambiguity. The dataset and code are available at https://github.com/drlxj/point-and-grasp.

*Corresponding author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790836

CCS Concepts

• **Human-centered computing → Interaction techniques**; **Mixed / augmented reality**; • **Mathematics of computing → Probabilistic inference problems**.

Keywords

Out-of-reach object selection, Bayesian cue integration, hand-object interaction dataset, mixed-reality.

ACM Reference Format:

Xuejing Luo, Hee-Seung Moon, Christian Holz, and Antti Oulasvirta. 2026. Point & Grasp: Flexible Selection of Out-of-Reach Objects Through Probabilistic Cue Integration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3790836

1 Introduction

In mixed reality (MR) environments, users often need to interact with objects that lie beyond their physical reach. For example, a designer may want to grab a virtual tool floating above a 3D workspace, or a player may need to select a distant item in a game scene. Such 3D out-of-reach object selection is a fundamental task in MR interactions [2, 10]. To support efficient and fluid selection, systems must accurately *infer* user intent. This is typically achieved by systems capturing behavioral signals produced as users act on their intent—in other words, user-generated *cues*.[1]

Among these, the most common are directional cues, usually implemented via raycasting [19, 53]. Targets can be indicated by pointing with a hand [10, 35] or controller [19, 53], or by gaze [55, 60, 73]. However, when targets are small, densely arranged, or occluded, directional cues often become ambiguous for intent inference [4, 32, 70]. Prior work addresses this through two main methods: immediate selection methods [3, 10, 32] that adjust the effective selection region, and progressive refinement methods [3, 11, 24] that iteratively narrow the candidate set. While these techniques mitigate spatial ambiguity, they remain direction-only and thus inherit the intrinsic noise of pointing; moreover, refinement-based approaches may add unnecessary overhead in simple scenes.

Therefore, a complementary line of research has explored gestural cues. Unlike direction, grasping gestures capture the geometry and functional affordances of objects [12, 22]. VirtualGrasp [67] demonstrates that people naturally produce consistent grasping gestures aligned with object semantics. However, existing approaches [39, 46, 63] often assume a one-to-one mapping between gesture and object, severely limiting scalability. In practice, grasping gestures support a many-to-many relationship, where a single gesture can apply to multiple semantically similar objects, increasing expressive power while introducing semantic ambiguity.

These limitations reveal a fundamental problem: *no single cue can reliably support out-of-reach selection across diverse MR scenarios.* While combining cues seems natural, most existing multi-cue approaches [13, 33, 52, 69] are rule-based—e.g., treating gaze as the primary cue and gestures as fixed command triggers [65]. Such deterministic designs become unreliable once the dominant cue turns ambiguous, since reliance cannot flexibly shift to alternatives.

In this work, we propose a probabilistic multi-cue integration framework for out-of-reach object selection in MR and introduce Point&Grasp as its implementation, integrating directional pointing for spatial disambiguation and grasping gestures for semantic ambiguity. Both cues are extracted from the same hand-tracking input, providing richer information without requiring additional sensors. Unlike existing deterministic multi-cue techniques, our approach treats both cues as probabilistic evidence: directional input is modeled as a likelihood over candidate objects, while a learned model estimates the gesture-object likelihood. These likelihoods are fused through Bayes' rule to produce a posterior distribution over candidate objects, enabling robust adaptation to noisy or ambiguous cues and generalization across interaction scenarios.

While directional cues can be readily modeled as probabilistic likelihoods, gestural cues lack such treatment. Prior methods often assume a fixed one-to-one mapping between a gesture and an object in a canonical pose, ignoring how pose and geometry affect grasp feasibility. To overcome this, we introduce a learning-based model that estimates pose-aware gesture–object likelihoods, providing context-sensitive evidence for Bayesian fusion and improving generalization across diverse interaction scenarios.

To support such modeling, we collected the Out-of-Reach Grasping (ORG) dataset that includes both in-reach and out-of-reach grasping gestures for the same objects. Prior work has shown that in-reach gestures capture object semantics, but it remains unclear whether out-of-reach gestures convey comparable semantic information and how they differ from in-reach gestures. Our dataset addresses this gap by systematically including both positive (matched) and negative (mismatched) gesture–object pairs. Training our probabilistic compatibility model on this dataset demonstrates robust generalization across subjects and partial transfer to previously unseen objects, providing a solid foundation for integrating grasping gestures into our probabilistic multimodal framework.

We conducted two user studies (Study 1 and Study 2) to investigate the benefits of Point&Grasp both against its constituent single cues and against state-of-the-art selection methods. Study 1 compares Point&Grasp with two single-cue baselines—direction-only and grasp-only—under systematically varied spatial and semantic ambiguities. Its goal is to identify the source of ambiguity where single-cue methods break down and how fusion overcomes these failures. Study 2 then benchmarks Point&Grasp against two leading selection techniques—BubbleRay [32] and Expand [11]. These represent two distinct state-of-the-art approaches for overcoming the limitations of directional selection—adaptive assistance for immediate selection and progressive refinements, respectively.

The results reveal four central findings. First, in Study 1, Point &Grasp consistently outperformed both single-cue baselines in selection time and completion rate, remaining robust even under the challenging spatial and semantic ambiguities. Second, Study 1 showed that Point&Grasp adapts cue reliance dynamically: gestural cues dominate under high spatial ambiguity, directional cues become more informative under high semantic ambiguity, and both cues converge in low-ambiguity scenarios, with fusion amplifying their combined strength. Third, subjective evaluations in Study 1 indicated that users found Point&Grasp easy to learn

[1] In the broader HCI literature, the term "cue" is bidirectional [1]—referring either to (1) system-to-user cues that guide or scaffold user behavior (e.g., [25]), or (2) user-to-system cues that reveal the user's intent (e.g., [49]). In this work, we focus exclusively on the latter: cues produced by users to convey their intent and interpreted by the system.

and judged gesture-based interaction natural and consistent with everyday grasping habits. Finally, Study 2 demonstrated that, when gestural cues provide reliable semantic information, Point&Grasp exhibits strong robustness to ambiguity, outperforming BubbleRay in high-spatial layouts and achieving faster selections than Expand in low-spatial layouts.

Together, these findings establish probabilistic cue integration as a principled and effective solution for out-of-reach object selection. By combining the complementary strengths of directional and gestural cues, Point&Grasp demonstrates how this framework improves selection robustness and flexibility, pointing toward more scalable multimodal interaction techniques for future MR systems.

In summary, the main contributions of this work are:

- We propose a probabilistic cue integration framework for out-of-reach object selection in MR and implement it as Point&Grasp, which fuses directional pointing and grasping gestures through Bayesian inference.
- We contribute Out-of-Reach Grasping (ORG) dataset, the first dataset for out-of-reach grasping in MR, with mid-air gestures aligned to in-reach virtual grasps as ground-truth contact references and explicit annotations of matched and mismatched gesture–object pairs.
- We develop a pose-aware gesture–object likelihood model based on the ORG dataset, which captures context-sensitive grasp feasibility, generalizes across subjects, and partially transfers to unseen objects.
- We validate Point&Grasp through two controlled user studies, demonstrating improved performance over single-cue baselines and superior robustness compared to state-of-the-art selection techniques across diverse ambiguity conditions.

2 Related Work

3D out-of-reach object selection is a fundamental component of MR interaction, widely used in 3D design, games, and everyday tasks [2, 50, 77]. A large body of research has focused on improving its accuracy and efficiency by enabling systems to identify users' intended targets through various interaction cues.

2.1 Out-of-Reach Object Selection via Directional Cue

Directional cues are widely used for out-of-reach object selection, where systems infer a target's spatial location from a user-generated ray, most commonly via raycasting [37]. The origin and direction of the ray can be defined using a handheld controller [19, 38, 53, 68], the posture of a bare hand [35, 37], head [37, 40, 62], or gaze [40, 62]. Some other approaches anchor the ray on specific body parts (e.g., eye, head, elbow) and extend it toward the fingertip [37, 40, 45]. Instead of a ray, a 3D cursor allows users to specify not only a direction but also disambiguate objects at different depths, such as the Go-Go technique [47, 71], which maps cursor direction and depth to the relative position between the user's chest and hand.

However, directional cues are prone to errors when objects are densely arranged, small, or partially occluded [4, 32, 70]. These errors stem from two types of pointing noise [36, 64]: constant noise, caused by geometric mismatches between the user's sightline and the emitted ray, limited field of view, or display distortion; and random noise, arising from signal-dependent motor variability and modulated by speed–accuracy trade-offs [20, 72].

To improve robustness, several immediate-selection techniques extend the basic pointing paradigm by modifying the selection endpoint or tolerance, such as cone-casting [29], or the BubbleRay [32], which guarantees a unique target within an adaptive region. While these methods help mitigate spatial ambiguity, they still rely on a single pointing action and therefore inherit user-dependent pointing noise and target-dependent biases. Other work adopts progressive refinement through multi-stage interaction to explicitly resolve ambiguity, including Expand [11], which enlarges the selectable region as the pointer approaches the target, and SQUAD [24], which separates coarse targeting from a confirmation step.

Since interaction-level approaches cannot fully eliminate pointing noise, a complementary line of research focuses on statistical modeling of ray distributions to compensate for target-dependent offsets. Prior work commonly models ray endpoints as Gaussian and predicts systematic deviations to improve selection accuracy [35, 36, 68, 76]. Building on this direction, we use distribution-based likelihoods of directional cues and integrate them with other modalities within a unified inference framework.

2.2 Out-of-Reach Object Selection via Gestural Cue

Gestures can serve as an alternative cue to directional information, helping disambiguate objects that spatially overlap [42]. Prior work has explored multiple strategies for using gestures to infer objects: Some methods explicitly assign a unique gesture to each object [39, 46, 63], while others use gestures to represent object features such as shape or size. For example, gestures that directly mimic the physical outline of objects have been investigated [43], emphasizing the importance of gestures being easy to discover and intuitive to learn in order to be effective [21].

Grasping is a natural way humans interact with objects, as it reflects both the stable hand configuration afforded by an object's shape and its functional use. This property enables grasping gestures to convey rich information about the user's intended object. Prior work has shown that observing grasping gestures alone can reveal object size and shape [58]. VirtualGrasp further demonstrated that grasping gestures can serve as a general selection cue in MR: gestures were easy to learn, and user studies showed strong cross-user agreement in how gestures mapped to objects [67]. However, when multiple objects share similar shapes, grasping gestures alone may not suffice for reliable disambiguation [67].

This paper addresses a research gap: how can grasping gestures, particularly performed for out-of-reach objects, be modeled probabilistically to infer user intention? Addressing this gap requires understanding differences between virtual and physical grasping, as well as between in-reach and out-of-reach contexts, which remain underexplored. In this work, we take the first step by constructing the ORG dataset for out-of-reach grasping gestures and developing a probabilistic inference model that enables grasping gestures to function as a quantitative cue for inferring user intention in out-of-reach selection tasks.

2.3 Target Inference from Multiple Cues

While each cue type offers unique advantages, no single cue performs reliably across all MR environments. This leads to our central research question: How can multiple cues be flexibly combined within a single interaction? Despite its importance, this question has received relatively little attention. Existing multi-cue approaches are typically limited, often assigning independent roles to each cue rather than integrating them in a principled way. For example, gaze-hand coordination techniques often use gaze as the primary cue for object selection (e.g., pre-select a group of objects), while gestures handle fine-grained selection within densely arranged targets [13, 33, 52, 59, 69]. Other studies explored speech-gesture combinations, where pointing identified the object while a user's verbal input specified its subsequent use with the object [65].

While these studies demonstrate the potential of multiple cues, our work directly addresses this gap by introducing a probabilistic cue integration technique for out-of-reach selection in MR environments. The closest approach to ours is by Wei et al. [62], who predicted a user's target of interest based on the distribution of eye endpoints and further showed that the head direction and the eye gaze distributions could be combined using Bayesian inference. However, because these two cues provided largely overlapping rather than complementary information, their multi-cue approach did not yield significant improvements over the single-cue method based solely on gaze. In contrast, we investigate whether integrating complementary cues with distinct discriminative power can yield measurable benefits. Specifically, we focus on probabilistically combining pointing and grasping gestures, which address different sources of ambiguity. Through their probabilistic integration, we aim to resolve both spatial and semantic ambiguities, thereby achieving robust out-of-reach object selection in MR.

2.4 Datasets for Hand-Object Interaction

Another dimension of this work concerns human-object interaction datasets, raising the question of whether existing datasets of physical, in-reach grasping can be utilized to support probabilistic modeling in out-of-reach interaction. A variety of datasets have been introduced for physical hand–object interaction [6, 16, 23, 31, 54, 61]. Representative examples include GRAB [54], which records full-body and hand motion during object manipulation; HOI4D [31], which captures egocentric videos with grasping hands and objects across daily activities; and DexGraspNet [61], which provides diverse dexterous grasp configurations for 3D object meshes.

These datasets have been highly valuable for studying physical in-reach grasping, supporting the training of models that predict and generate stable grasps for objects [28, 54, 66, 74, 75] and even infer which objects are suitable for grasping [44]. By contrast, out-of-reach grasping involves imaginary, mid-air gestures toward distant targets. Such gestures may differ from physical grasps due to perceptual uncertainty and noise in MR environments, yet little prior work has examined this distinction [7, 8]. Another challenge is the need for calibration in likelihood-based inference models, which require data covering both correct (positive) and incorrect (negative) object–gesture pairings [56]. Such balance is critical for disambiguating intent in cluttered scenes and avoiding false positives, but existing datasets overwhelmingly emphasize positive examples of feasible grasps.

To address these gaps, we contribute the ORG dataset of mid-air grasp gestures in out-of-reach selection scenarios, structured to support the calibrated training of probabilistic intent inference models. To our knowledge, this is the first dataset to systematically capture out-of-reach grasping behavior, providing a foundation for robust intention modeling beyond physical grasping contexts.

3 Probabilistic Cue Integration for Out-of-Reach 3D Object Selection

Our method formulates out-of-reach 3D object selection as a probabilistic intent inference problem from multiple cues. Unlike single-cue interaction, which often fails when cues become ambiguous in specific contexts (e.g., directional cues exhibiting spatial ambiguity when objects are clustered or occluded), our framework provides a principled way to integrate heterogeneous evidence sources.

Building on this formulation, we resolve such ambiguities by casting target inference as posterior computation under a Bayesian cue-integration framework, a formulation widely studied in human perception and decision-making [15, 26].

3.1 Probabilistic Inference via Multi-Cue Integration

We consider a set of candidate objects $O = \{o_1, \ldots, o_K\}$ in a given MR scene. From a user, we consider a set of observed cues $C = \{c_1, \ldots, c_M\}$, where each c_m corresponds to a different cue (e.g., pointing direction, hand gesture, eye gaze, speech). Our goal is to infer the posterior probability of each object being the intended target:

$$p(o \mid C) = p(o \mid c_1, c_2, \ldots, c_M).$$

Assuming conditional independence of cues given the target object (the naïve Bayes assumption),

$$c_i \perp c_j \mid o \quad \forall i \neq j,$$

the posterior distribution takes the form

$$p(o \mid C) = \frac{p(o) \prod_{m=1}^{M} p(c_m \mid o)}{\sum_{i=1}^{K} p(o_i) \prod_{m=1}^{M} p(c_m \mid o_i)}. \tag{1}$$

Here, $p(o)$ represents the prior over candidate objects, and $p(c_m \mid o)$ is the likelihood of observing cue c_m if the intended object were o. The denominator of Eq. 1 serves only as a normalizing constant, that is, the likelihoods play a decisive role in this inference.

Therefore, the inference procedure reduces to several steps: For each candidate object, the system computes the likelihood of the observed cues under the assumption that the object is the intended target. These likelihoods can be derived from previous computational models (e.g., Gaussian models of users' pointing offsets [36, 68]) that capture the statistical relationship between the intended target object and the user's resulting behavior. The likelihoods are then combined with the prior distribution. Finally, the object whose posterior probability is maximal is selected as the inferred target.

$$\hat{o} = \arg\max_{o \in O} \; p(o \mid C),$$

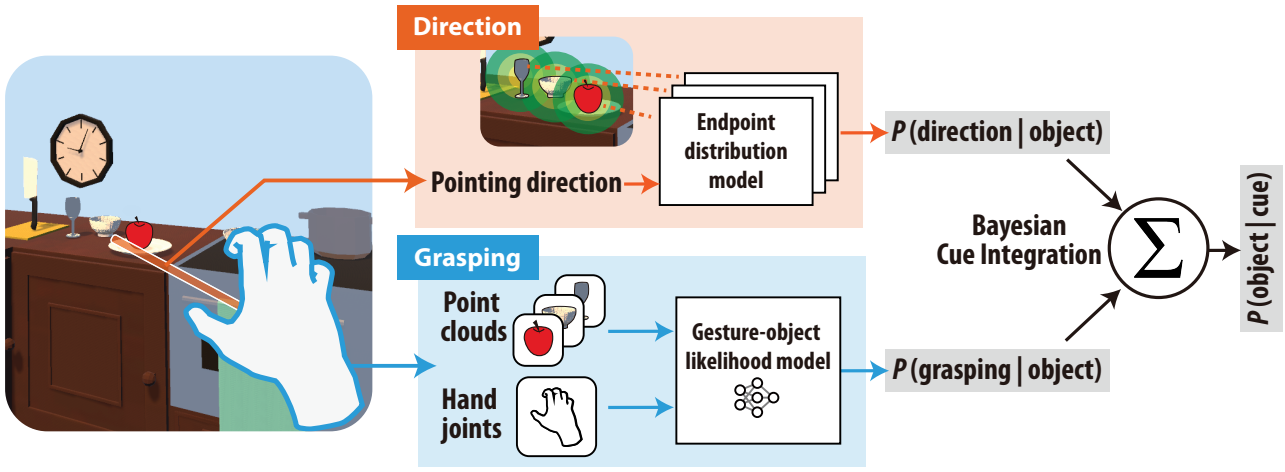

Figure 2: Overview of the proposed multi-cue inference framework for out-of-reach object selection. The framework combines an endpoint distribution model for directional cues and a gesture–object likelihood model for grasping gestural cues, which are integrated through Bayesian cue integration to infer the intended target object.

While the Bayesian framework is conceptually applicable to different cues, our work focuses on two cues that are both natural to MR interaction and complementary in their information content: *directional* and *gestural* cues.

- **Directional cue (c_D):** provides spatial evidence through pointing, but can be ambiguous in scenes where objects are small, cluttered, or occluded.
- **Gestural cue (c_G):** provides object-semantic evidence by modeling the compatibility between the user's hand configuration and object shape, but may be ambiguous among objects of similar shape or category.

To compute the posterior $p(o \mid c_D)$ and $p(o \mid c_G)$, we must therefore model $p(c_D \mid o)$ and $p(c_G \mid o)$.

3.2 Directional Cue Inference

The directional cue provides spatial evidence for inferring the intended target. As illustrated in Figure 2, we represent directional intent with a ray defined by an origin and a direction vector. The ray is cast from the user's hand toward the scene and intersects the reference plane of the objects at an endpoint p_e.

Following prior work [36, 68], we model the directional cue likelihood $p(c_D \mid o)$ as a Gaussian likelihood:

$$p(c_D \mid o) \propto \exp\left(-\frac{(p_e - p_o)^2}{2\sigma_d^2}\right),$$

Here p_o denotes the center of candidate object o, serving as the Gaussian mean, and σ_d controls the sensitivity to spatial deviation. This definition reflects the probability that the observed endpoint p_e would result if the user were actually intending to select object o. Consequently, objects whose centers are closer to the ray endpoint receive higher likelihood values, meaning the directional cue provides stronger support for them as the intended target.

3.3 Grasping Gestural Cue Inference

The grasping gestural cue captures semantic information about the object, reflecting its shape, size, and affordances. In other words, a grasping hand posture carries cues about which objects in the scene are physically compatible candidates, even without relying on precise directional pointing.

To make use of this information, we formulate a gesture–object likelihood model $p(c_G \mid o)$ that estimates the probability of a gesture c_G being intended for a candidate object o:

$$p(c_G \mid o) \approx f_\theta(c_G, o),$$

where f_θ is parameterized by a neural network that takes as input (1) the user's hand configuration (joint positions or gesture embedding) and (2) object features such as shape descriptors or category embeddings. The output is normalized to form a probability distribution over candidate objects. The choice of model architecture and the training objective for f_θ depends on available input data representations. For example, architectures may range from simple MLPs to graph- or attention-based models, depending on how hand and object features can be acquired in practice. Likewise, one may use contrastive objectives when the training dataset provides positive and negative gesture–object pairs, or a standard cross-entropy loss when only ground-truth target object labels are provided. Note that the specific architecture and training procedure used in this work are described in Section 3.4.2.

3.4 Implementation of Point&Grasp

Figure 2 provides an overview of how the Point&Grasp method is instantiated in our system. The general Bayesian formulation described earlier requires us to specify likelihood models for both directional and gestural cues. In our VR implementation, this involves defining how the pointing ray is constructed, how hand postures and objects are represented, and how these representations are processed to produce cue likelihoods that are ultimately combined in the final inference.

3.4.1 Directional Cue c_D. The directional cue fundamentally depends on how the ray is defined. Prior work [2, 27] has shown that head–hand direction is a common basis for defining pointing, where the ray originates at the user's head and passes through a hand-based endpoint. While some techniques [51] use the fingertip as this endpoint, our setting requires a reference point that remains consistent across different grasp poses. We therefore adopt the wrist position as the ray endpoint, which preserves the head–hand relationship while providing a stable spatial reference during grasping.

In our model, σ_d specifies the size of the region around the object's center where the directional cue retains meaningful probability. We use a constant σ_d to avoid unintentional differences in selection difficulty across objects, assuming graspable everyday objects whose visual sizes fall within a reasonably limited range. Based on prior observations in Yu et al. [68], we adopt $\sigma_d = 0.4$ as a representative value for objects with a visual angle of 1° to 2°. Note that size-dependent formulations (i.e., increasing σ_d according to the object size) can be considered in other settings.

3.4.2 Gestural Cue c_G. To effectively model the likelihood of gestural cues given an object, we require a representation that captures both the object and hand configuration. We represent the hand through its 3D joint positions and the objects through its point cloud. The point cloud is defined in the canonical coordinate system, where the object center is the object's origin and the canonical orientation defines the axes. The point cloud directly reflects the object's shape and pose; i.e., different orientations of the same object are expressed as different point coordinate sets. For the hand, we represent joint positions relative to the wrist as the origin, while using the same canonical axes as those of the object's point cloud.

In this way, the representation captures the relative geometry between hand and object: regardless of where the object is located, how it is oriented, or how far the hand is from it, the combination of point cloud and wrist-centered hand joints provides a consistent description of the object's shape and the hand–object relationship.

The gesture-object likelihood model comprises two encoders: one for the hand and one for the object.

- **Hand encoder**. The hand encoder takes the transformed joint positions (63-dimensional input) and processes them through a multi-layer perceptron (MLP). The encoder consists of two fully connected layers with 256 hidden units each. Each layer is followed by batch normalization and ReLU activation, with a dropout layer ($p = 0.2$) between the two blocks. The output is a 256-dimensional gesture embedding.
- **Object encoder**. The object encoder first converts the raw point cloud into a basis point set (BPS) representation [48], yielding a compact 1024-dimensional vector that captures the object's geometry independently of the number of points. This vector is then passed through an MLP with two fully connected layers (1024→512 and 512→256), each followed by batch normalization, ReLU, and dropout ($p = 0.2$). The output is a 256-dimensional object embedding, aligned in dimensionality with the hand embedding.

The gesture and object embeddings are then concatenated (512-dimensional vector) and fed into a fusion MLP. This consists of two hidden layers (512→128 and 128→64), each followed by batch normalization, ReLU, and dropout ($p = 0.2$). Finally, a linear layer maps to a single scalar output, which is squashed by a sigmoid activation to produce the probability $p(c_G \mid o)$ that the observed gesture corresponds to the candidate object.

The model is trained with a binary cross-entropy loss on each gesture–object pair, where $y \in \{0, 1\}$ indicates whether the gesture is intended for the object:

$$\mathcal{L} = -\big[y \log p(c_G \mid o) + (1 - y) \log(1 - p(c_G \mid o))\big],$$

3.4.3 Combining Cues. In practice, the Point&Grasp technique requires the user to simultaneously provide both types of evidence:

- By extending the arm and aligning the wrist with the target, the user produces a ray that provides the directional cue c_D.
- By shaping the hand into a natural grasp posture, the user conveys semantic information about the intended object, forming the gestural cue c_G.

During inference, the system evaluates for each candidate object o its directional likelihood $p(c_D \mid o)$ and its gestural likelihood $p(c_G \mid o)$. These two terms are then combined under the Bayesian formulation as Eq. 1. We assume $p(o)$ as uniform prior over candidate objects in the 3D scenes, meaning that each object is considered equally likely to be the target. This prior can also be flexibly adjusted; for example, to encode biases such as certain objects being more frequently used.

4 Out-of-Reach Grasping Dataset and Model Evaluation

Grasping gestures encode both the geometric properties of objects and their functional affordances, making them a powerful cue for inferring user intent. However, existing datasets such as GRAB [54] primarily focus on in-reach physical grasps and do not capture how grasping gestures are produced when objects are out of reach, a setting that is central to mid-air interaction in MR.

To address this gap, we introduce the Out-of-Reach Grasping (ORG) dataset, which captures both in-reach and out-of-reach grasping gestures. In-reach gestures provide ground-truth contact information, while their out-of-reach counterparts reveal users' intended grasps without physical interaction. We analyze the dataset to characterize the distribution, variability, and semantic structure of out-of-reach gestures, including their differences from in-reach gestures and their gesture-induced object confusability.

To ensure that our interaction technique receives stable and reliable probability $p(c_G \mid o)$ during runtime, we evaluate several network architectures. Because gesture and object features have distinct characteristics, we adopt a dual-branch encoder (Section 3.4.2) to extract their representations separately before fusing them. The rationale and effectiveness of this design are validated in Section 4.3.

4.1 Method: Data Collection

4.1.1 Participants. 19 participants (13 male, 6 female; aged 24–37) were recruited to participate in data collection. All participants were right-handed and had no motor or vision impairments based on self-report. Before the experiment, each participant received a detailed briefing and provided informed consent.

4.1.2 Experimental Design. Each *trial* was designed to sequentially collect three types of data for a given object from each participant

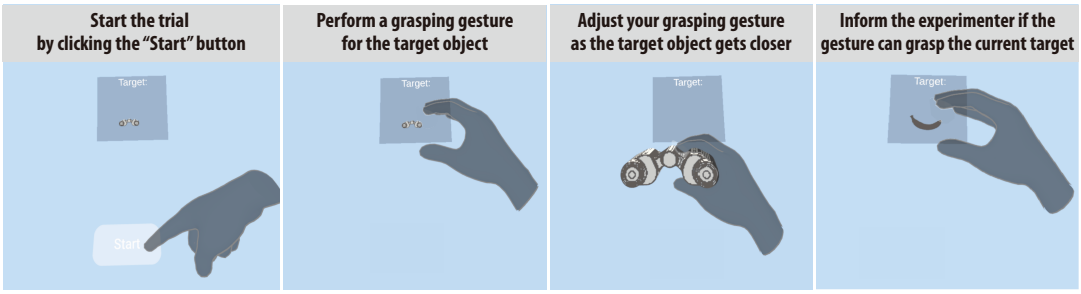

Figure 3: The trial procedure in the VR data collection task. (1) Participants initiated the trial by pressing a virtual "Start" button and performed the natural grasping gesture they would normally use to pick up the object. (2) The object was then placed within reach, and participants indicated the intended grasp location before refining their posture to align with the object. (3) The finalized grasping posture was tested for compatibility with five additional objects, for which participants verbally indicated whether their current gesture could plausibly grasp the object.

in the VR environment: (1) an out-of-reach grasping gesture, (2) an in-reach grasping gesture, which also served as a ground-truth contact reference, and (3) subjective evaluations of gesture–object compatibility with other objects of varying shapes. This design yields annotations of both positive and negative samples under out-of-reach and in-reach conditions.

We employed a *within-subjects design* with *target object* as the control factor, in which each participant completed trials with the same 30 target objects presented in a randomized order. To maximize the diversity of grasping behaviors, participants performed three trials per object, each with a distinct object pose. For each pose, they were asked to produce a grasping gesture they would normally use to grasp or interact with the object, from the perceived object shape and orientation.

Across the three trials, only object orientation varied. Each object was assigned a canonical orientation reflecting its typical usage (e.g., an upright mug, a horizontal knife). To introduce noticeable variation across trials while maintaining realism, we first sampled a base yaw angle from $\{0°, 120°, 240°\}$, and applied a random perturbation within $\pm 60°$, ensuring both systematic coverage of diverse grasping orientations and natural variability across repeated presentations.

The 30 target objects were common household items widely used in hand–object interaction research [14, 54], spanning diverse shapes and affordances to elicit varied grasping strategies. Each participant completed 90 trials (30 objects × 3 orientations), totaling over 1,700 trials across all users—providing broad coverage of gesture–object pairings while keeping each session under one hour.

4.1.3 Task. Each trial followed a sequence of three stages (see Figure 3). In the first stage, the target object appeared floating at eye level, 2 meters in front of the participant. The trial began when participants pressed a start button with their dominant hand, after which they were asked to perform the *"natural grasping gesture you would normally use if you were to pick up the object."* Because the object was intentionally placed out of reach, participants could not physically interact with it and therefore relied solely on prior experience to imagine the grasp. Upon completion, they verbally notified the experimenter. This out-of-reach stage enabled us to capture grasping intention without influence from physical contact.

In the second stage, the same target object was moved into the participant's reachable space (approximately at arm's length) while maintaining its orientation. Participants were asked to bring their hand to the spatial location they had previously imagined and then refine the posture as needed to naturally grasp the now reachable object. This stage recorded the in-reach gesture together with its precise spatial relationship to the object. Once the gesture was finalized, participants notified the experimenter.

In the third stage, participants maintained their finalized grasp while the target object was sequentially replaced with five objects randomly selected from the remaining pool. For each new object, they verbally indicated whether the current hand posture could plausibly grasp it ("yes," "no," or "unsure"), and the experimenter recorded the response. If the response was "yes," the object was moved into reachable space as in stage two, and participants reassessed whether the grasp was feasible. When the answer again was "yes," we recorded the grasp and its spatial configuration; otherwise, only the compatibility labels were stored.

4.1.4 Study Procedure. Before the experiment, participants provided written informed consent and were given time to familiarize themselves with the VR environment through practice trials. We also measured each participant's arm length to define the boundary of their reachable space and ensure correct object placement during in-reach stages. During data collection, participants progressed through the trials at their own pace and could rest between trials as needed. On average, each session (90 trials) lasted about one hour.

4.1.5 Apparatus and Materials. The experiment was conducted in a virtual environment developed with Unity3D and displayed on a Meta Quest 2 headset with its native hand tracking. The headset operated in a wired configuration connected to a desktop computer, ensuring stable rendering at 90 Hz and reliable data transfer throughout the experiment.

4.2 Out-of-Reach Grasping Dataset Analysis

4.2.1 Dataset Statistics. The ORG dataset comprises 10,260 gesture–object pairs, each annotated with binary compatibility labels for both in-reach and out-of-reach conditions. Most pairs were judged incompatible in both conditions (68.89%, 7,068 samples),

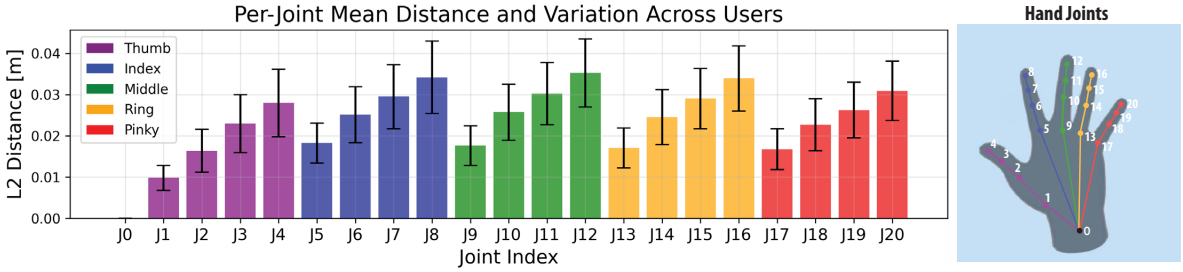

Figure 4: The results show that hand gesture for an out-of-reach object is significantly different from when it is within reach, and that variability across users is high. Bars indicate per-joint mean L2 distance grouped by finger, with error bars showing standard deviations. The illustration maps each joint index (J0–J20) to its anatomical position on the hand for reference.

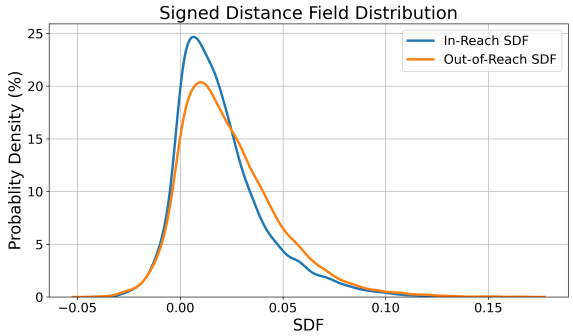

Figure 5: Signed Distance Field (SDF) distributions for in-reach and out-of-reach gestures. Out-of-reach gestures show a rightward shift, indicating more extended hand postures.

while clearly compatible cases account for 23.62% (2,423 samples). The remaining samples reflect asymmetric or uncertain judgments: 3.69% (379) were incompatible in-reach but compatible out-of-reach, 2.23% (229) incompatible in-reach but uncertain out-of-reach, and 1.57% (161) uncertain in-reach but compatible out-of-reach. Overall, the distribution contains a large proportion of clear positives and negatives, while also capturing natural ambiguity that is valuable for training probabilistic intent models.

4.2.2 Differences Between In-Reach and Out-of-Reach Gestures. To assess whether out-of-reach gestures are semantically comparable to in-reach gestures while revealing systematic differences, we compared their Signed Distance Field (SDF) distributions. SDF characterizes spatial occupancy by measuring the signed distance from a point to the nearest object surface, with positive values outside the object, negative values indicating penetration, and near-zero values reflecting surface contact [17, 41]. A two-sample Kolmogorov–Smirnov test [34] revealed a significant difference between in-reach and out-of-reach distributions ($p < 0.001$). As shown in Figure 5, out-of-reach gestures exhibit a slight rightward shift, indicating more extended hand postures at greater distances.

To further localize these differences, we computed per-joint Euclidean distances across conditions. As illustrated in Figure 4, most joints differ by 1.5–3 cm on average, with the largest deviations at distal joints (fingertips) and smaller differences at proximal joints. Participant variation remained limited, indicating a consistent deviation pattern across users.

4.2.3 Gesture-Induced Object Confusability. To examine the semantic discriminability of out-of-reach gestures, we constructed an object–object confusion matrix based on the compatibility judgments collected in Stage 3 of the experiment (Figure 6). In this stage, participants maintained a given hand posture while the target object was sequentially replaced with other objects, and they indicated whether the posture remained compatible. Aggregating these judgments across participants and trials yields an average similarity score for each object pair.

Overall, most objects are clearly separable in gesture space, suggesting that participants produced distinct hand postures reflecting object function and shape. However, certain object pairs exhibit high confusability, such as cup–mug–wineglass, which frequently share similar hand configurations. This pattern indicates a many-to-many relationship between gestures and objects, where a single gesture can be compatible with multiple objects and a single object can elicit diverse gestures. These findings delineate the semantic boundaries of out-of-reach gestures and motivate the inclusion of both positive and negative samples for probabilistic inference.

To further analyze the structure of this confusability, we applied the Silhouette method to the similarity matrix and clustered the 30 objects (Figure 7). The resulting clusters reveal systematic gesture-induced groupings: for example, fryingpan–hammer–knife form a coherent cluster associated with tool-like grasping postures, while cup–mug–wineglass–waterbottle cluster together due to shared cylindrical affordances. These results show that confusability is structured rather than random, reflecting functional and morphological regularities in gesture–object mappings.

4.3 Model Training and Evaluation

The goal of model evaluation is to assess the stability and suitability of the gesture–object likelihood model that supports our interaction technique Point&Grasp. During interaction, the model continuously estimates the probability $p(c_G \mid o)$, and therefore must produce consistent and reliable outputs across different users and different objects. To assess this, we evaluate the model under three complementary generalization settings:

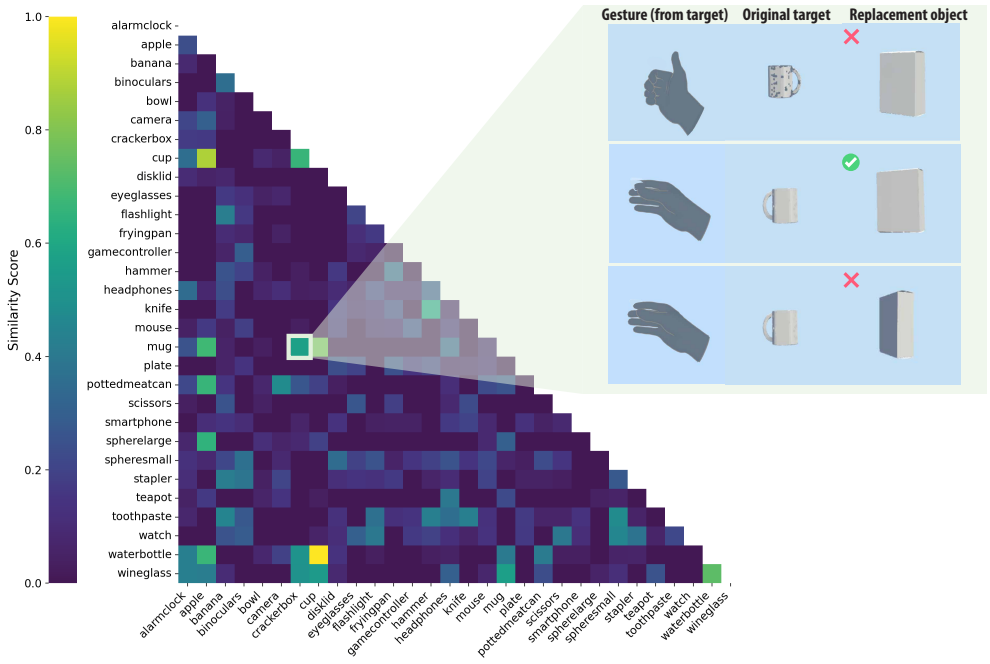

Figure 6: The object confusion matrix shows that when using grasping gesture cues alone, some objects are more easily and some less easily confused with each other. Each cell indicates how often a gesture elicited by the target object was also judged compatible with a replacement object; brighter values denote higher confusion likelihood. The examples on the right illustrate typical outcomes: (Row 1) the gesture for a mug was incompatible with a crackerbox; (Row 2) another gesture for a mug was judged compatible with a crackerbox; (Row 3) although the same gesture for a mug as in Row 2 was applied, the replacement object could not be grasped in the current hand pose due to its orientation. These cases highlight the many-to-many nature of gesture–object relationships, where one gesture can be judged compatible with multiple objects, and the same object can be grasped with different gestures depending on its orientation.

- Within-subject evaluation: training and testing on disjoint trials from the same participants.
- Cross-subject evaluation: training on a subset of participants and testing on unseen participants.
- Cross-object evaluation: training on a subset of objects and testing on unseen objects.

Given the substantial structural differences between gesture and object features, we adopt a dual-branch encoder to extract their representations separately before fusing them (see Section 3.4.2 for details). To validate the effectiveness of this design for our task, we compare it against a simplified single-branch alternative. These comparisons confirm the benefits of the dual-branch architecture and provide a reusable baseline for future research on gesture–object compatibility tasks.

4.3.1 Training Setup. We generated three train/test splits corresponding to the three generalization protocols described above.

- Within-subject split: each participant's trials were divided into an 8:2 train/test ratio.
- Cross-subject split: the ORG dataset is split by subjects, following a 7:3 train/test ratio.
- Cross-object split: objects were partitioned according to clustering analysis shown in Figure 7, with one representative object from each cluster held out for testing. Models were not exposed to these objects during training.

Implementation details of the single-branch baseline and the training configuration are provided in the supplementary material.

4.3.2 Evaluation Results. Table 1 summarizes the model performance across the three generalization settings. Overall, the dual-branch model outperforms the single-branch baseline on most metrics, indicating that separately encoding gesture and object features provides a clear advantage for this task. In the within-subject setting, both models achieve strong performance with only minor differences, suggesting that both architectures are able to capture individual users' compatibility judgment patterns when the training and testing data come from the same participant. In the cross-subject setting, the dual-branch model achieves higher accuracy, recall, and precision, while the single-branch baseline yields a slightly lower ECE. These results show that separating gesture and object encoding leads to more stable representations that generalize better to unseen users. The cross-object setting poses the greatest challenge, with performance decreasing for both models due to the difficulty of inferring compatibility for entirely unseen object shapes and affordances. Nevertheless, the dual-branch model consistently outperforms the single-branch baseline across all metrics,

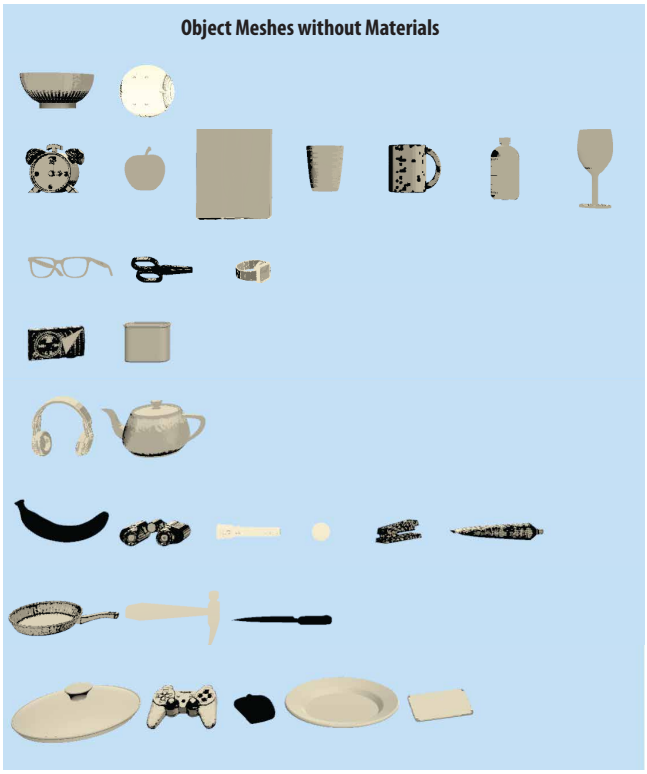

Figure 7: Clustering analysis shows that gesture-based confusability is systematic rather than random: objects that afford similar grasping postures, such as frying-pan–hammer–knife or cup–mug–wineglass–waterbottle, naturally group together. The clusters were derived by applying the Silhouette method to the similarity matrix of 30 objects: Cluster 1: bowl, large sphere. Cluster 2: alarm clock, cracker-box, apple, mug, water bottle, wineglass. Cluster 3: eyeglasses, scissors, watch. Cluster 4: camera, potted meat can. Cluster 5: headphones, teapot. Cluster 6: banana, binoculars, flashlight, small sphere, stapler, toothpaste. Cluster 7: frying pan, hammer, knife. Cluster 8: disklid, game controller, mouse, plate, smartphone.

demonstrating stronger shape-level generalization. Taken together, these results validate the effectiveness of the dual-branch architecture, particularly under cross-user and cross-object distribution shifts, while also revealing room for improvement when generalizing to completely novel objects.

5 Overview of User Studies

As discussed earlier, single-cue techniques for out-of-reach object selection have inherent limitations. Directional cues are intuitive but degrade under spatial ambiguity, such as when targets are densely clustered or partially occluded, leading a single pointing ray to intersect multiple candidates. In contrast, grasping gestural cues convey object semantics but become less discriminative under semantic ambiguity, when objects share similar shapes or functions and appear equally compatible with the same gesture.

Table 1: Evaluation results across three settings. Accuracy, Recall, and Precision (↑) are higher-is-better; ECE (↓) is lower-is-better.

Dataset	Accuracy ↑	Recall ↑	Precision ↑	ECE ↓
Within-subject				
Single-branch	0.813	**0.837**	0.791	**0.070**
Ours	**0.816**	0.807	**0.821**	0.085
Cross-subject				
Single-branch	0.811	0.835	0.763	**0.070**
Ours	**0.822**	**0.843**	**0.775**	0.090
Cross-object				
Single-branch	0.529	0.296	0.534	0.338
Ours	**0.566**	**0.368**	**0.590**	**0.299**

To investigate how these limitations can be mitigated, we evaluated Point&Grasp through two complementary user studies. The first examines its behavior and performance relative to single-cue baselines (Study 1), while the second compares our method against state-of-the-art 3D selection techniques (Study 2). Together, the studies address the following research questions:

- **RQ1:** Does the Interaction method Point&Grasp that integrates multiple cues outperform single-cue baselines (Point or Grasp) across various selection contexts?
- **RQ2:** How do users exploit the flexibility afforded by Point&Grasp across selection contexts? Which cues are used when resolving spatial or semantic ambiguity?
- **RQ3:** How do users subjectively perceive Point&Grasp?
- **RQ4:** How does Point&Grasp compare with state-of-the-art 3D selection techniques?

6 Study 1: Comparing with Single-Cue Baselines

We first examine how each single cue contributes to hand-based selection performance and what advantages arise from probabilistically integrating them. To this end, we compare Point&Grasp with its two single-cue variants—Point and Grasp—to assess accuracy, cue usage patterns, and subjective user preferences.

6.1 Method

6.1.1 Apparatus. The study was conducted using a Meta Quest 2 headset connected via Oculus Link, with participants remaining seated throughout the experiment. The right hand was used for directional input and grasping gesture, while the left hand confirmed selections via the space bar on a keyboard. A custom prototype was implemented in Unity and communicated in real time with a Python-based gesture–object likelihood model. Gesture data were transmitted from Unity to Python, where likelihoods were computed and returned to Unity. The system was executed on a desktop computer with an Intel Core i7-4790K CPU (4.00 GHz), 16 GB RAM.

6.1.2 Participants. Twelve participants (8 male, 4 female; aged 26–30) were recruited from the university community. All were

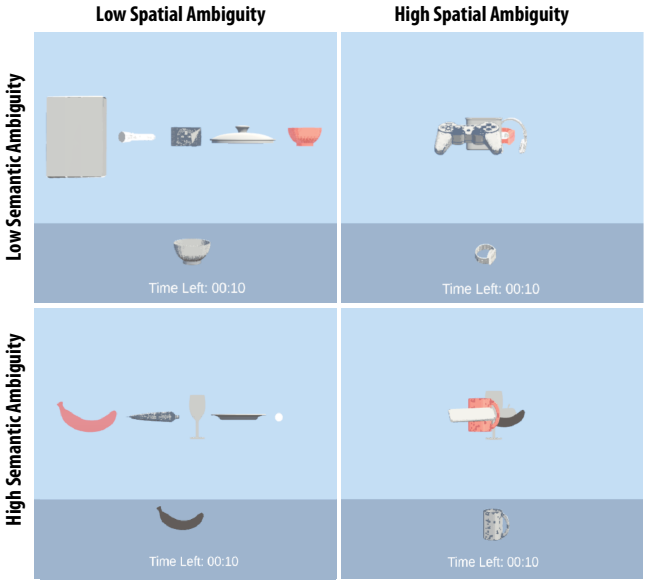

Figure 8: Example scenes illustrating the four ambiguity conditions that vary along semantic and spatial dimensions. In the low semantic × low spatial condition, objects (cracker box, flashlight, camera, disk lid, and bowl) are both semantically and spatially distinct, minimizing potential confusion. The low semantic × high spatial condition (game controller, meat can, flashlight, watch, and headphone) maintains semantic distinctiveness but increases spatial proximity, raising the likelihood of spatial ambiguity. Conversely, the high semantic × low spatial condition (banana, toothpaste, wineglass, plate, and small ball) introduces semantically similar items while keeping them spatially separated. Finally, the high semantic × high spatial condition (smartphone, mug, banana, apple, and wineglass) combines both semantic similarity and spatial proximity, creating the highest level of ambiguity.

right-handed with normal or corrected-to-normal vision, hearing, and motor abilities. Their self-reported VR experience was moderate ($M = 2.75$ on a 5-point scale, 1 = None, 5 = Expert). Participation was voluntary with informed consent.

6.1.3 Study Design. We used a within-subjects design with a $3 \times 2 \times 2$ factorial structure. Independent variables were:

- **Interaction Method:** Point (excluding gestural cues), Grasp (excluding directional cues), and Point&Grasp (our proposed Bayesian integration of both cues).
- **Spatial Ambiguity:** low (1° angular separation between adjacent objects, causing strong occlusion) vs. high (5° separation, ensuring no overlap).
- **Semantic Ambiguity:** low (objects from distinct clusters) vs. high (three objects from the same cluster and two from different clusters).

Crossing semantic and spatial ambiguity yielded four conditions: Low semantic × Low spatial, Low semantic × High spatial, High semantic × Low spatial, and High semantic × High spatial. Example

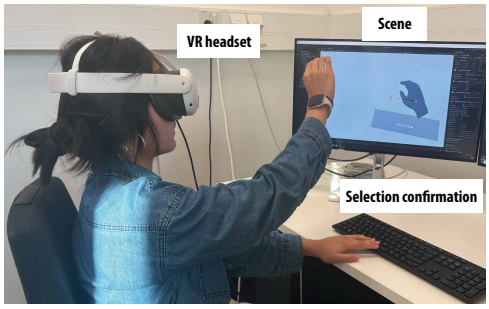

Figure 9: Study setup. Participants wore a VR headset with hand tracking. The scene displayed five candidate objects, with one highlighted in red as the target. Participants provided directional or grasping cues with their right hand and used the space bar with the left hand to confirm the selection.

scenes are shown in Figure 8. For semantic ambiguity, we prepared 16 object sets in total: 8 low-semantic sets and 8 high-semantic sets, each containing five objects (details in supplementary). For each set, two spatial layouts were instantiated (low vs. high). Each participant was assigned four low-semantic and four high-semantic sets, with each set appearing in both spatial layouts, resulting in 16 unique scenes per participant. Within each scene, interaction methods were tested, and the order of methods and semantic ambiguity conditions was counterbalanced across participants.

6.1.4 Task. Each trial presented five candidate objects in the VR scene, with one highlighted in red as the target (as illustrated in Figure 9). Participants selected the target using the assigned Interaction method and confirmed their choice by pressing the space bar with their left hand. Each scene comprised five trials, with each object serving once as the target. Objects were placed 2–3 meters away, and each trial had a 10-second time limit. After incorrect selections, participants could retry until selecting the correct target or reaching the time limit, at which point an auditory cue signaled the end of the trial and prompted them to proceed to the next trial.

6.1.5 Procedure. Upon arrival, participants provided informed consent and received a briefing about the study. They then practiced all three interaction methods with sample objects until comfortable. In the main study, each participant completed three sessions, one per interaction method, with session order counterbalanced. Within each session, they performed 16 scenes in randomized order, yielding five trials per scene. Participants could rest between scenes or sessions as needed. After finishing all sessions, participants filled out a brief questionnaire assessing the learning difficulty of the Grasp method, the naturalness of the gestures they produced compared to everyday grasping, and the ease of learning Point&Grasp.

6.1.6 Measures. We measured two objective performance metrics: target selection time (from trial onset to correct selection or the 10-second limit) and trial completion rate (the percentage of trials in which the target was correctly selected within the time limit). For each trial, we also logged the system's cue probabilities to analyze how directional and gestural evidence contributed to target inference, particularly in Point&Grasp.

Table 2: (Study 1) Statistical analysis of selection time and trial completion rate.

		Selection time		Completion rate	
	df	F-ratio	p-value	F-ratio	p-value
Method	(2, 22)	13.439	**<0.001**	4.754	**0.035**
Spatial Ambiguity	(1, 11)	165.407	**<0.001**	49.677	**<0.001**
Semantic Ambiguity	(1, 11)	58.537	**<0.001**	49.751	**<0.001**
Method × Spatial Ambiguity	(2, 22)	83.508	**<0.001**	38.436	**<0.001**
Method × Semantic Ambiguity	(2, 22)	19.282	**<0.001**	19.705	**<0.001**
Spatial Ambiguity × Semantic Ambiguity	(1, 11)	5.189	**0.044**	0.823	0.384
Method × Spatial Ambiguity × Semantic Ambiguity	(2, 22)	1.747	0.205	2.314	0.131

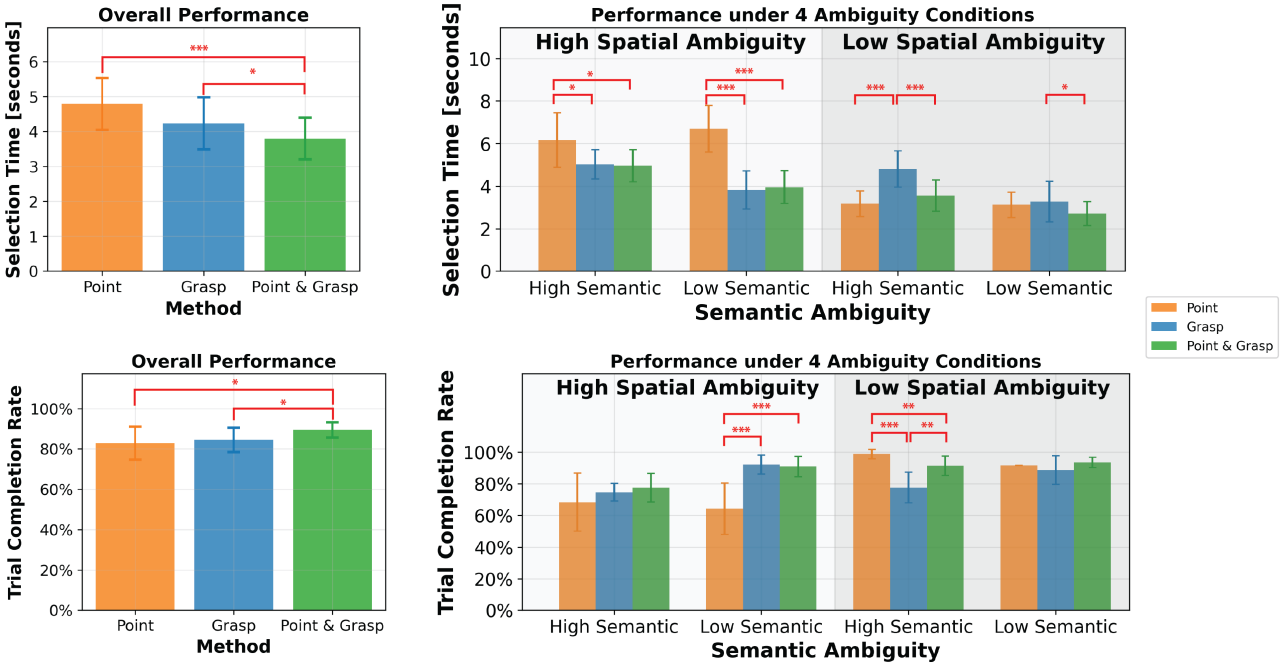

Figure 10: (Study 1) Point&Grasp achieves the fastest selection times and the highest trial completion rates, avoiding the slowdown of Point under spatial ambiguity and Grasp under semantic ambiguity. Across both measures, Point&Grasp remains robust under spatial and semantic ambiguity where single-cue methods degrade. Error bars show between-participant variation. Asterisks indicate statistically significant differences after Bonferroni correction (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

In addition, participants completed a post-study questionnaire providing subjective ratings (1) the ease of learning grasp-based selection, (2) the naturalness of the produced grasping gestures compared to everyday grasping, and (3) the ease of learning Point&Grasp, all on a 5-point Likert scale.

6.2 Results

We organize the results according to these questions. Section 6.2.1 compares the three Interaction methods, Section 6.2.4 analyzes cue agreement in the Point&Grasp method, and Section 6.2.5 reports participants' questionnaire responses.

6.2.1 Comparison of Methods. A three-way (Method × Spatial Ambiguity × Semantic Ambiguity) repeated-measures ANOVA with Greenhouse–Geisser correction revealed significant main effects of Method, Spatial Ambiguity, and Semantic Ambiguity on both selection time and completion rate (Table 2). We also found significant two-way interactions of Method × Spatial Ambiguity and Method × Semantic Ambiguity on both measures, and a smaller but reliable Spatial × Semantic Ambiguity interaction on selection time only. No significant three-way interaction was observed.

6.2.2 Selection Time. Figure 10 (top left) shows the overall selection time across methods. Point&Grasp was the fastest on average, while Point required the longest completion time. Post-hoc tests with Bonferroni correction confirmed that Point&Grasp was significantly faster than both Point ($p < 0.001$) and Grasp ($p = 0.016$). We break down the results by ambiguity condition, given that the ANOVA revealed significant interactions of Method × Spatial Ambiguity and Method × Semantic Ambiguity. The results confirmed

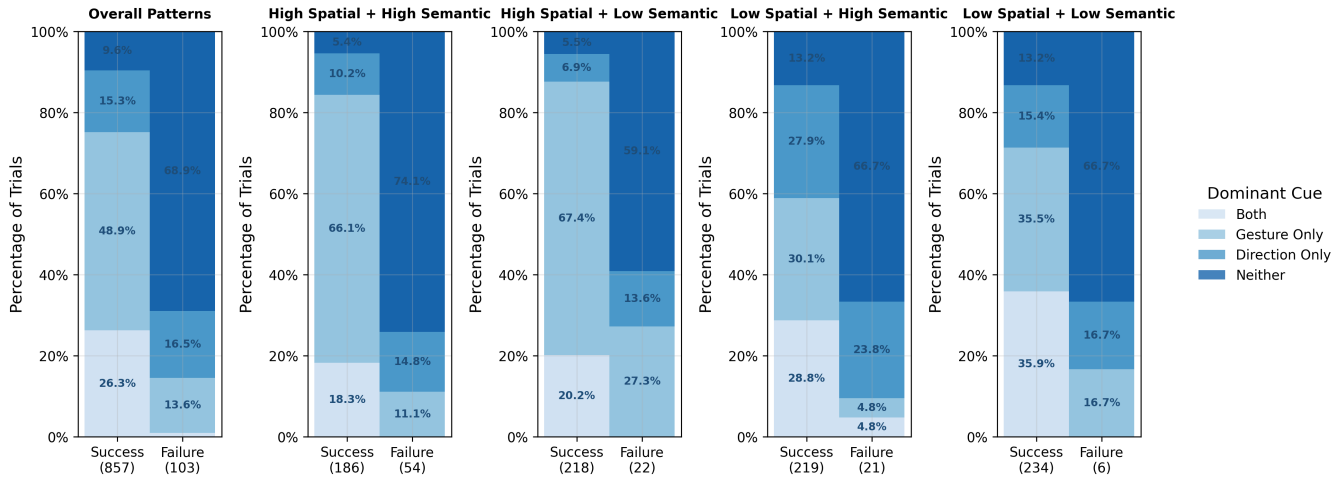

Figure 11: (Study 1) Users flexibly exploited gesture and direction cues in Point&Grasp, with successes often driven by a single dominant cue but still possible when neither cue was sufficient alone. Trials were labeled Both (both cues selected the target), Gesture Only (only gesture did), Direction Only (only direction did), or Neither (neither did). Left: overall patterns (success vs. failure). Right: breakdown by ambiguity conditions. Values denote percentages within each bar.

that each single-cue method was affected by a different source of ambiguity. Under high spatial ambiguity, Point showed markedly longer mean times than both Grasp ($p = 0.002$) and Point&Grasp ($p < 0.001$). In contrast, Grasp slowed under high semantic ambiguity, with longer durations than Point&Grasp ($p = 0.007$). Point&Grasp, with the use of the complementary cues, maintains robustly low selection times across conditions. Pairwise comparisons between methods under each ambiguity condition are presented in the top right panel of Figure 10.

6.2.3 Trial Completion Rate. Figure 10 (bottom left) summarizes the overall completion rates. Point&Grasp achieved the highest success rate, significantly outperforming both Point ($p = 0.024$) and Grasp ($p = 0.030$). When analyzed by ambiguity condition (Figure 10, bottom right), the results once again highlight the complementary strengths of the different modality cues. Under high spatial ambiguity, Point performance degraded substantially, resulting in significantly lower completion rates than both Grasp ($p = 0.002$) and Point&Grasp ($p < 0.001$). In contrast, under high semantic ambiguity, Grasp performance dropped, yielding significantly lower completion rates than Point&Grasp ($p = 0.007$), while Point remained relatively more robust. Across all conditions, Point&Grasp again consistently delivered the highest completion rates. These findings mirror the results with selection time, reinforcing that cue integration provides robustness against both spatial and semantic sources of uncertainty.

6.2.4 Cue Agreement Analysis of Point & Grasp Method. To unpack the mechanisms behind Point&Grasp's performance, we examined how agreement between the two cues related to task outcomes. For each trial, we checked whether the gestural and directional cues independently ranked the target the highest and grouped trials into four categories: Both, Gesture Only, Direction Only, and Neither. This analysis reveals whether robustness stems mainly from cue convergence or from the system's ability to rely on whichever cue is informative in context.

As illustrated in Figure 11, successes were dominated by cases where a single cue alone identified the target—Gesture Only accounted for 48.9% of successes, compared with 15.3% Direction Only and 26.3% Both. Neither was rare (9.6%), yet even in these cases the fused model sometimes recovered the correct target, demonstrating that Point&Grasp's robustness does not rely on cue agreement; rather, it hinges on the ability to capitalize on whichever cue provides strong evidence in context, and at times to accumulate weaker signals from both cues. Failures, by contrast, were largely driven by Neither (68.9%), indicating that errors most often occurred when both cues provided weak evidence.

Breaking down by ambiguity reveals when each cue is most informative. Under high spatial × low semantic ambiguity, successes were overwhelmingly Gesture Only (67.4%), with Direction Only contributing little (6.9%): spatial crowding undermines pointing, making grasp-compatibility the decisive signal. Under low spatial × high semantic ambiguity, contributions became balanced—Gesture Only 30.1%, Direction Only 27.9%, and Both 28.8%—highlighting that pointing regains discriminative power when objects are well separated but semantically similar. In the most difficult high spatial × high semantic condition, successes still leaned heavily on Gesture Only (66.1%), whereas failures were mostly Neither (74.1%), reflecting simultaneous cue degradation. Finally, under low–low ambiguity, Both peaked (35.9%), indicating frequent cue convergence and near-ceiling performance.

Overall, these patterns clarify Point&Grasp's advantage: Bayesian integration adapts by leaning on the cue that is informative—gesture under spatial ambiguity, direction under semantic ambiguity—and amplifies confidence when cues agree. This adaptive behavior directly explains the consistently faster selection times and higher completion rates observed in Section 6.2.1.

Figure 12: Evaluated techniques in Study 2. BubbleRay: A red ray is cast from the hand; the yellow disc marks the bubble tangent to the selected object; the green curve indicates the chosen target. Expand (Step 1 & Step 2): Coarse selection is performed by using a circular cursor to gather nearby objects; the collected candidates are then arranged into a screen-space grid for refined disambiguation. Point&Grasp: The green endpoint cursor provides directional feedback. The red cursor and semi-transparent target visualization indicate the selected object. A flat-hand posture, regardless of palm orientation, is used here as a "null" gesture, which the model interprets as yielding a uniform gesture–object likelihood.

6.2.5 Subjective Evaluation. Post-study questionnaire results indicate that participants found grasping gestures easy to learn (M = 1.92, SD = 0.51, 5-point scale, 1 = very easy, 5 = very hard) and fairly natural compared to everyday grasping habits (M = 3.67, SD = 0.65, 5-point scale, 1 = not at all, 5 = very matching). Participants also reported that Point&Grasp was very easy to learn (M = 1.08, SD = 0.28). Overall, these subjective findings align with the objective results, suggesting that grasping gestures are both intuitive and learnable, and can be smoothly integrated with pointing cues to support effective out-of-reach object selection.

7 Study 2: Comparing with State-of-the-Art 3D Selection Techniques

While Study 1 examined the benefits of cue integration by comparing Point&Grasp with its single-cue variants, Study 2 evaluates its performance against state-of-the-art 3D selection techniques. We selected two empirically strong methods from prior work—BubbleRay [32] and Expand [11]—which both rely on directional cues but employ distinct mechanisms to mitigate their limitations in dense target scenarios.

7.1 Method

7.1.1 Interaction Methods. We evaluated three interaction methods, illustrated in Figure 12.

- **BubbleRay** [32]. It performs immediate selection by dynamically shaping a bubble region around the pointing ray (inspired by the original Bubble Cursor [18]), and selects the object closest to the bubble boundary. Among the variants described by the authors, we implemented the version based on angular distance between the ray and candidate objects. This variant has been shown to outperform a wide range of established directional techniques, including 3D BubbleCursor [57], Go-Go [47], and SQUAD-style ray-based methods [24], making it a strong representative of state-of-the-art directional selection.
- **Expand** [11]. It adopts a two-stage selection strategy consisting of coarse filtering followed by grid-based refinement. During coarse selection, a circular cursor gathers nearby objects (in our setting, within a 10 cm Euclidean distance threshold) around the ray. The collected candidates are then arranged into a 2D screen-space grid to support disambiguation. Prior work has shown Expand to be significantly faster and more accurate than SQUAD [24] and Zoom [11]. More recently, in densely arranged 3D selection scenarios, a grid-based method inspired by Expand has also outperformed other contemporary baselines [70].
- **Point&Grasp (Ours)**. For Study 2, we introduced two refinements: (1) We added a circular cursor at the ray endpoint to ensure consistent directional feedback as with BubbleRay and Expand. Prior work suggests that explicit cursor representations help anchor user attention and reduce perceptual ambiguity [30]. (2) We retrained the gesture model to include a "null" flat-hand gesture that conveys no grasp semantics. This gesture produces a uniform likelihood on objects and and enables the system to correctly interpret cases where users do *not* intend to provide a semantic cue, allowing directional evidence to dominate posterior inference.

To ensure a fair comparison, all methods used the same definition of ray direction as described in Section 3.4.1.

7.1.2 Apparatus and Participants. The apparatus was identical to that used in Study 1, including the Meta Quest 2 headset connected via Oculus Link and the same desktop computer. We recruited a new set of twelve participants (5 male, 7 female; aged 25–35) from the university community. All participants had normal or corrected-to-normal vision and were right-handed.

7.1.3 Study Design. Study 2 adopted the same within-subjects 3 × 2×2 factorial design as Study 1, with identical independent variables: Interaction Method, Spatial Ambiguity, and Semantic Ambiguity. The only difference lies in the levels of the Interaction Method factor, which in Study 2 compared Point&Grasp against two state-of-the-art directional techniques, BubbleRay and Expand. Spatial and Semantic Ambiguity were carried over unchanged, resulting in the same four ambiguity configurations shown in Figure 8. Details

Table 3: (Study 2) Statistical analysis of selection time and trial completion rate.

		Selection time		Completion rate	
	df	F-ratio	p-value	F-ratio	p-value
Method	(2, 22)	.070	.885	27.132	**<0.001**
Spatial Ambiguity	(1, 11)	248.553	**<0.001**	138.078	**<0.001**
Semantic Ambiguity	(1, 11)	6.227	**0.030**	1.510	0.245
Method × Spatial Ambiguity	(2, 22)	44.936	**<0.001**	32.783	**<0.001**
Method × Semantic Ambiguity	(2, 22)	8.384	**0.008**	14.018	**<0.001**
Spatial Ambiguity × Semantic Ambiguity	(1, 11)	16.486	**<0.001**	22.172	**<0.001**
Method × Spatial Ambiguity × Semantic Ambiguity	(2, 22)	16.486	**<0.001**	22.172	**<0.001**

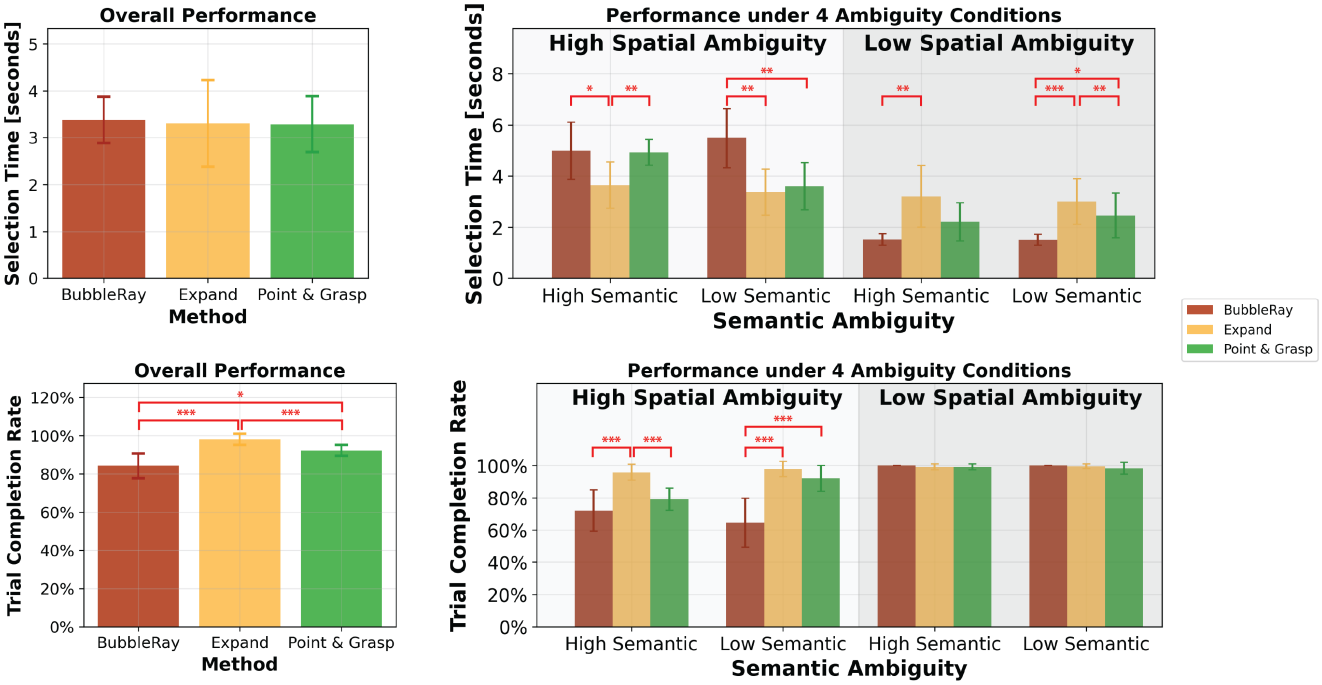

Figure 13: (Study 2) Point&Grasp achieves faster selection times than Expand under low spatial ambiguity and than BubbleRay under high spatial ambiguity. It also reaches near-Expand completion rates while significantly exceeding BubbleRay, particularly under low semantic ambiguity. Across both measures, Point&Grasp demonstrates competitive performance against state-of-the-art techniques under varying ambiguity conditions. Error bars show between-participant variation. Asterisks indicate statistically significant differences after Bonferroni correction (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

on scene construction, object sets, and ambiguity manipulation are provided in Section 6.1.3.

7.1.4 Task, Procedure, and Measures. The task and procedure were identical to those in Study 1, except that no subjective ratings were collected in Study 2. We evaluated the three Interaction methods with *target selection time* and *trial completion rate*. For full details of the task and procedure, please refer to Section 6.

7.2 Results

A three-way repeated-measures ANOVA (Method × Spatial Ambiguity × Semantic Ambiguity) revealed a significant main effect of Spatial Ambiguity on both selection time and completion rate, a significant main effect of Method on completion rate but not selection time, and a significant main effect of Semantic Ambiguity on selection time only (Table 3). Significant two-way interactions were observed for Method × Spatial Ambiguity, Method × Semantic Ambiguity, and Spatial × Semantic Ambiguity, as well as a significant three-way interaction for both outcome measures. Bonferroni-corrected post hoc tests were conducted for all significant effects, and the results reported below reflect these follow-up analyses.

7.2.1 Selection Time. Figure 13 (top left) shows no significant Method effect, with all techniques exhibiting similar overall speed. Condition-specific comparisons, however, revealed clear differences. Between the two state-of-the-art baselines, Expand was faster under

high spatial ambiguity ($p = 0.003$), whereas BubbleRay was faster under low spatial ambiguity ($p = 0.001$), reflecting the trade-off between refinement overhead and robustness in occluded layouts. Point&Grasp consistently fell between the two: it was faster than Expand under low spatial ambiguity ($p = 0.010$) and faster than BubbleRay under high spatial ambiguity ($p = 0.035$). Importantly, semantic ambiguity further highlighted Point&Grasp's advantage. Under low semantic ambiguity—when gesture cues become reliable and informative—it showed significantly faster performance ($p = 0.013$) and maintained stable speed across spatial conditions. For example, under the low-semantic × high-spatial condition, Point&Grasp achieved selection speed comparable to Expand, but surpassed it under the low-semantic × low-spatial condition ($p = 0.004$).

7.2.2 Trial Completion Rate. Figure 13 (bottom left) shows the overall completion rates across the three techniques. Expand achieved the highest accuracy, followed by Point&Grasp, with BubbleRay performing least accurately. Breaking the results down by ambiguity condition clarifies the source of Expand's advantage. Under low spatial ambiguity, all three techniques performed near ceiling (i.e., 100%), showing little separation. Under high spatial ambiguity, however, Expand's refinement step provided a clear benefit, yielding substantially higher completion rates than the other two methods ($p < 0.001$ for both). For Point&Grasp, performance again varied with semantic ambiguity ($p = 0.006$): under the high semantic × high spatial condition,Point&Grasp performed significantly worse than Expand ($p < 0.001$), whereas under low semantic ambiguity it achieved completion rates comparable to Expand (no significant difference) and significantly outperformed BubbleRay ($p < 0.001$).

7.2.3 Summary. The results of Study 2 suggest that **multi-cue integration reduces Point&Grasp's dependence on the spatial-ambiguity dimension by incorporating semantic information.** BubbleRay and Expand showed a clear trade-off along this spatial dimension—BubbleRay performing better in low-spatial scenes and Expand excelling in high-spatial scenes. However, when semantic cues were reliable (i.e., under low semantic ambiguity), Point&Grasp demonstrated strong robustness to spatial interference. It matched Expand's speed and accuracy under high spatial ambiguity, while avoiding Expand's refinement overhead in low spatial ambiguity, resulting in faster performance in those layouts. This highlights how incorporating a complementary cue dimension enables Point&Grasp to maintain stable and reliable performance where state-of-the-art directional-cue methods exhibit trade-offs.

8 Discussion

Both studies demonstrate clear benefits of probabilistic cue integration for out-of-reach object selection. Based on these findings, the key features of Point&Grasp can be summarized as follows:

- **Overall performance:** In Study 1, Point&Grasp surpassed both single-cue baselines in speed and accuracy. In Study 2, it achieved performance comparable to state-of-the-art directional techniques, consistently landing between BubbleRay and Expand, which are in trade-off along spatial ambiguity.
- **Robustness under different ambiguity:** The advantage of cue integration became evident under different sources of ambiguity. Study 1 showed that Point and Grasp degrade under different ambiguity sources, while Point&Grasp remained stable across both. Study 2 further demonstrated that, compared to state-of-the-art techniques, multi-cue integration reduces Point&Grasp's dependence on the spatial-ambiguity dimension by incorporating semantic information. Under low semantic ambiguity, Point&Grasp maintained robust performance even in highly cluttered scenes, matching Expand in high-spatial layouts, while avoiding its refinement overhead in low-spatial layouts.
- **Context-dependent cue dominance:** As directional and grasping gestural cues are complementary, the relative contribution of each changes with the type of ambiguity. Bayesian cue integration naturally lets the more discriminative cue dominate. The cue agreement analysis showed that the dominance of grasping cues increased as scenes became more spatially ambiguous and less semantically ambiguous.
- **Probabilistic rescue:** The Bayesian framework enabled successful selection even when neither cue alone maximized likelihood. Approximately 10% of Study 1 trials fell into this category, showing that probabilistic integration extends performance beyond the limits of single-cue methods.
- **Real-time operation:** All probabilistic computations, including the deployed gesture-object likelihood model, were lightweight enough to support real-time object selection, ensuring practical applicability in MR environments.

Below, we discuss broader implications and open questions.

Are in-reach grasping datasets applicable to out-of-reach interaction? The ORG dataset reveals statistically significant differences between in-reach and out-of-reach grasping gestures performed by the same users on the same objects (see Section 4.2.2). In out-of-reach conditions, fingertip positions shifted by around 3 cm and grasp apertures became noticeably narrower—likely reflecting perceptual and motor uncertainties when imagining grasps on distant objects. These deviations indicate that directly applying physical in-reach grasp datasets such as GRAB to out-of-reach interaction can introduce systematic bias in gesture–object likelihood modeling.

How does multi-cue integration shape user strategies under ambiguity? Our cue agreement analysis in Study 1 (see Section. 6.2.4) suggests that users may adaptively shift their selection strategy depending on scene ambiguity. Under high spatial ambiguity, around 67% of successful trials were driven by gesture cues, but this proportion dropped to 30% in low-ambiguity settings. These shifts reflect not only changes in cue informativeness with scene context but also users strategically privileging the cue that best reduces uncertainty. When spatial crowding made directional cues unreliable, they leaned on grasp cues; when objects were semantically similar but spatially distinct, users adopted a more balanced strategy that leveraged both directional and grasp information.

How can users recover from mis-selections? Although probabilistic cue integration allows Point&Grasp to handle cases where neither cue alone is decisive, Study 2 reveals substantial room for user-driven recovery. The two refinements introduced in Study 2—an endpoint cursor for clearer directional feedback and a null gesture to intentionally disable the gestural cue—significantly reduced

mis-selections, raising completion rates in low-ambiguity scenes to near 100%. More broadly, because the Bayesian model maintains posterior probabilities over all candidates, our probabilistic cue integration offers an opportunity for additional corrective mechanisms: the system can detect when posterior scores are ambiguous between candidates and trigger lightweight refinement steps only when needed (e.g., a brief confirmatory step within the candidates). Such mechanisms offer a practical way to resolve semantic and spatial ambiguities among nearby objects in real-world use.

What additional modalities could be considered? Beyond pointing direction and grasping gestures, our probabilistic framework can be extended with other modalities that complement user intent. Gaze pointing is a natural candidate to combine with hand pointing, often used to mitigate eye–hand misalignment in VR; similar to directional cues from the hand, gaze endpoints can be modeled as Gaussian likelihoods over candidate objects [62]. Speech provides an orthogonal cue, where verbal references to objects can be modeled as categorical likelihoods over candidates whose labels or attributes match the utterance. This idea dates back to Bolt's "Put-that-there" system [9], which demonstrated the power of combining speech with deictic gestures for object specification. Finally, action history may offer contextual priors: while not specific to MR, routine modeling work [5] suggests that past behavior can inform predictions of future actions, and in our setting, such history could slightly bias the prior toward previously selected objects.

How can Point&Grasp transition from controlled experiments to real MR use? Our study employed controlled design choices to investigate the benefits of cue integration in isolation, but each can extend naturally to real MR scenarios.

First, we provide which item to grasp via object-indicating visualization (see Figure 8). Real MR environments include comparable situations—both when users can locate a target through visual search and when they cannot see all details yet still know the intended object (e.g., reaching for a screwdriver to tighten a bolt in a cluttered toolbox). In such cases, users can recall the object's shape and produce a meaningful grasp without explicit visualization. Second, for directional cue, we fixed σ_d for evaluation. Real systems can scale it by object size, simply adjusting directional dispersion without altering the fusion mechanism. Third, target objects were placed at a fixed depth to focus on semantic grasp forms. In real MR settings, users can perceive distance and infer the geometry of familiar objects; nevertheless, viewing distance may shape grasp articulation—through perceived scale or required hand aperture. Investigating these effects remains valuable future work. Finally, we used a space-bar press for controlled confirmation, but in real MR settings this can be replaced by native gestures such as a left-hand controller click or a bare-hand pinch gesture, allowing the interaction to remain fully *controller-free*.

Future work. Several directions remain open for future research. First, the gesture–object likelihood model is limited by the size and diversity of our dataset. Collecting richer gesture data across more objects and users would improve generalization to unseen objects and grasping styles. Second, personalization remains an open challenge, as users differ in how they balance directional and gestural cues. Adaptive cue weighting through brief calibration or online adjustment could better align the system with individual strategies. Finally, although this work focuses on selection, the same probabilistic framework could extend to downstream MR tasks such as object manipulation, cooperative handover, or multi-user collaboration, where ambiguity is also pervasive.

9 Conclusion

In this paper, we introduced a probabilistic framework for out-of-reach object selection in MR that integrates directional and grasping gestural cues via Bayesian inference. To support this approach, we built a dataset of out-of-reach grasping gestures and developed a gesture–object likelihood model. Our evaluation showed that a single user motion can simultaneously express complementary yet separate cues—pointing direction and grasping gesture—that, when fused probabilistically, resolve both spatial and semantic ambiguity. By leveraging such complementary cues, our proposed Point&Grasp method achieved higher accuracy and speed than single-cue baselines consistently under the different ambiguities. We further showed that, by incorporating semantic information, Point&Grasp achieves more robust performance across diverse spatial layouts compared with empirically strong directional techniques. Our result contributes to a line of work in HCI showing the benefits of addressing multimodal input as a probabilistic inference problem. While prior work has shown strong results in areas like text entry where there is a single end-effector, in this work we have found out a way to exploit the same principle in out-of-reach selection, where one needs to deal with the full kinematic chain of the hand. We envision our insights to pave the way to more expressive selection techniques that better exploit the information humans can express with their body.

Acknowledgments

This work was supported by the Research Council of Finland (FCAI, 328400, 345604, 341763; Subjective Functions, 357578), the ERC Advanced Grant (101141916), and the National Research Foundation of Korea (RS-2025-00521470). We thank Yutong Du for assistance with early dataset collection and Jiayi He for support in developing the initial VR object selection prototype.

References

[1] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2571–2582. doi:10.1145/3025453.3026033

[2] Ferran Argelaguet and Carlos Andujar. 2013. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics* 37, 3 (2013), 121–136. doi:10.1016/j.cag.2012.12.003

[3] Felipe Bacim, Regis Kopper, and Doug A Bowman. 2013. Design and evaluation of 3D selection techniques based on progressive refinement. *International Journal of Human-Computer Studies* 71, 7-8 (2013), 785–802. doi:10.1016/j.ijhcs.2013.03.003

[4] Marc Baloup, Thomas Pietrzak, and Géry Casiez. 2019. Raycursor: A 3d pointing facilitation technique based on raycasting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3290605.3300331

[5] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K Dey. 2016. Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 248–260. doi:10.1145/2858036.2858557

[6] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2022. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15935–15946. doi:10.1109/cvpr52688.2022.01547

[7] Andreea Dalia Blaga, Maite Frutos-Pascual, Chris Creed, and Ian Williams. 2021. Freehand grasping: An analysis of grasping for docking tasks in virtual reality. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 749–758. doi:10.1109/vr50410.2021.00102

[8] Andreea Dalia Blaga, Maite Frutos-Pascual, Chris Creed, and Ian Williams. 2025. VR-Grasp: a human grasp taxonomy for virtual reality. *International Journal of Human–Computer Interaction* 41, 7 (2025), 4406–4422. doi:10.1080/10447318.2024.2351719

[9] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270. doi:10.1145/965105.807503

[10] Doug A Bowman and Larry F Hodges. 1997. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Proceedings of the 1997 symposium on Interactive 3D graphics*. 35–ff. doi:10.1145/253284.253301

[11] Jeffrey Cashion, Chadwick Wingrave, and Joseph J LaViola Jr. 2012. Dense and dynamic 3d selection for game-based virtual environments. *IEEE transactions on visualization and computer graphics* 18, 4 (2012), 634–642. doi:10.1109/tvcg.2012.40

[12] Umberto Castiello. 2005. The neuroscience of grasping. *Nature Reviews Neuroscience* 6, 9 (2005), 726–736. doi:10.1038/nrn1744

[13] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+ gesture: Expressive, precise and targeted free-space interactions. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 131–138. doi:10.1145/2818346.2820752

[14] Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, et al. 2024. Dense Hand-Object (HO) GraspNet with Full Grasping Taxonomy and Dynamics. In *European Conference on Computer Vision*. Springer, 284–303. doi:10.1007/978-3-031-73007-8_17

[15] Marc O Ernst and Martin S Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 6870 (2002), 429–433. doi:10.1038/415429a

[16] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12943–12954. doi:10.1109/cvpr52729.2023.01244

[17] Sarah F Frisken, Ronald N Perry, Alyn P Rockwood, and Thouis R Jones. 2000. Adaptively sampled distance fields: A general representation of shape for computer graphics. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 249–254. doi:10.1145/344779.344899

[18] Tovi Grossman and Ravin Balakrishnan. 2005. The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 281–290. doi:10.1145/1054972.1055012

[19] Tovi Grossman and Ravin Balakrishnan. 2006. The design and evaluation of selection techniques for 3D volumetric displays. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. 3–12. doi:10.1145/1166253.1166257

[20] Christopher M Harris and Daniel M Wolpert. 1998. Signal-dependent noise determines motor planning. *Nature* 394, 6695 (1998), 780–784. doi:10.1038/29528

[21] Chris Harrison, Robert Xiao, Julia Schwarz, and Scott E Hudson. 2014. TouchTools: leveraging familiarity and skill with physical tools to augment touch interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2913–2916. doi:10.1145/2556288.2557012

[22] Marc Jeannerod. 1984. The timing of natural prehension movements. *Journal of motor behavior* 16, 3 (1984), 235–254. doi:10.1080/00222895.1984.10735319

[23] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. 2023. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14713–14724. doi:10.1109/iccv51070.2023.01352

[24] Regis Kopper, Felipe Bacim, and Doug A Bowman. 2011. Rapid and accurate 3D selection by progressive refinement. In *2011 IEEE symposium on 3D user interfaces (3DUI)*. IEEE, 67–74. doi:10.1109/3dui.2011.5759219

[25] Byungjoo Lee. 2022. Cue integration in input performance. *Bayesian methods for interaction and design* (2022), 287–307. doi:10.1017/9781108874830.015

[26] Byungjoo Lee, Sunjun Kim, Antti Oulasvirta, Jong-In Lee, and Eunji Park. 2018. Moving target selection: A cue integration model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3173574.3173804

[27] Sangyoon Lee, Jinseok Seo, Gerard Jounghyun Kim, and Chan-Mo Park. 2003. Evaluation of pointing techniques for ray casting selection in virtual environments. In *Third international conference on virtual reality and its application in industry*, Vol. 4756. SPIE, 38–44. doi:10.1117/12.497665

[28] Jiaman Li, Jiajun Wu, and C Karen Liu. 2023. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–11. doi:10.1145/3618333

[29] Jiandong Liang and Mark Green. 1994. JDCAD: A highly interactive 3D modeling system. *Computers & graphics* 18, 4 (1994), 499–506. doi:10.1016/0097-8493(94)90062-0

[30] Chiuhsiang Joe Lin, Benedikta Anna Haulian Siboro, and Wen-Ting Tsai. 2026. Interaction performance of mid-air touch with and without cursor in augmented reality environment. *International Journal of Industrial Ergonomics* 112 (2026), 103894.

[31] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21013–21022. doi:10.1109/cvpr52688.2022.02034

[32] Yiqin Lu, Chun Yu, and Yuanchun Shi. 2020. Investigating bubble mechanism for ray-casting to improve 3d target acquisition in virtual reality. In *2020 IEEE Conference on virtual reality and 3D user interfaces (VR)*. IEEE, 35–43. doi:10.1109/vr46266.2020.00021

[33] Mathias N Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. 2022. Gaze-hand alignment: Combining eye gaze and mid-air pointing for interacting with menus in augmented reality. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–18. doi:10.1145/3530886

[34] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78. doi:10.2307/2280095

[35] Sven Mayer, Valentin Schwind, Robin Schweigert, and Niels Henze. 2018. The effect of offset correction and cursor on mid-air pointing in real and virtual environments. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13. doi:10.1145/3173574.3174227

[36] Sven Mayer, Katrin Wolf, Stefan Schneegass, and Niels Henze. 2015. Modeling distant pointing for compensating systematic displacements. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 4165–4168. doi:10.1145/2702123.2702332

[37] Mark R Mine. 1995. Virtual environment interaction techniques. *UNC Chapel Hill CS Dept* (1995).

[38] Hee-Seung Moon, Yi-Chi Liao, Chenyu Li, Byungjoo Lee, and Antti Oulasvirta. 2024. Real-time 3d target inference via biomechanical simulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3613904.3642131

[39] Marco Moran-Ledesma, Oliver Schneider, and Mark Hancock. 2021. User-defined gestures with physical props in virtual reality. *Proceedings of the ACM on Human-Computer Interaction* 5, ISS (2021), 1–23. doi:10.1145/3486954

[40] Kai Nickel and Rainer Stiefelhagen. 2003. Pointing gesture recognition based on 3d-tracking of face, hands and head orientation. In *Proceedings of the 5th international conference on Multimodal interfaces*. 140–146. doi:10.1145/958432.958460

[41] Stanley Osher, Ronald Fedkiw, and Krzysztof Piechor. 2004. Level set methods and dynamic implicit surfaces. *Appl. Mech. Rev.* 57, 3 (2004), B15–B15. doi:10.1115/1.1760520

[42] Brandon Paulson, Danielle Cummings, and Tracy Hammond. 2011. Object interaction detection using hand posture cues in an office setting. *International journal of human-computer studies* 69, 1-2 (2011), 19–29. doi:10.1016/j.ijhcs.2010.09.003

[43] Siyou Pei, Alexander Chen, Jaewook Lee, and Yang Zhang. 2022. Hand interfaces: Using hands to imitate objects in ar/vr for expressive interactions. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16. doi:10.1145/3491102.3501898

[44] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. 2023. Object pop-up: Can we infer 3d objects and their poses from human interactions alone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4726–4736. doi:10.1109/cvpr52729.2023.00458

[45] Jeffrey S Pierce, Andrew S Forsberg, Matthew J Conway, Seung Hong, Robert C Zeleznik, and Mark R Mine. 1997. Image plane interaction techniques in 3D immersive environments. In *Proceedings of the 1997 symposium on Interactive 3D graphics*. 39–44. doi:10.1145/253284.253303

[46] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-defined gestures for augmented reality. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 955–960. doi:10.1145/2468356.2468527

[47] Ivan Poupyrev, Mark Billinghurst, Suzanne Weghorst, and Tadao Ichikawa. 1996. The go-go interaction technique: non-linear mapping for direct manipulation in VR. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*. 79–80. doi:10.1145/237091.237102

[48] Sergey Prokudin, Christoph Lassner, and Javier Romero. 2019. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4332–4341. doi:10.1109/iccv.2019.00443

[49] Tye Rattenbury and John Canny. 2007. CAAD: an automatic task support system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 687–696. doi:10.1145/1240624.1240731

[50] Gang Ren and Eamonn O'Neill. 2013. 3D selection with freehand gesture. *Computers & Graphics* 37, 3 (2013), 101–120. doi:10.1016/j.cag.2012.12.006

[51] Valentin Schwind, Sven Mayer, Alexandre Comeau-Vermeersch, Robin Schweigert, and Niels Henze. 2018. Up to the finger tip: The effect of avatars

on mid-air pointing accuracy in virtual reality. In *Proceedings of the 2018 annual symposium on computer-human interaction in play*. 477–488. doi:10.1145/3242671.3242675

[52] Rongkai Shi, Yushi Wei, Xueying Qin, Pan Hui, and Hai-Ning Liang. 2023. Exploring gaze-assisted and hand-based region selection in augmented reality. *Proceedings of the ACM on Human-Computer Interaction* 7, ETRA (2023), 1–19. doi:10.1145/3591129

[53] Frank Steinicke, Timo Ropinski, and Klaus Hinrichs. 2006. Object selection in virtual environments using an improved virtual pointer metaphor. In *Computer Vision and Graphics: International Conference, ICCVG 2004, Warsaw, Poland, September 2004, Proceedings*. Springer, 320–326. doi:10.1007/1-4020-4179-9_46

[54] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 581–600. doi:10.1007/978-3-030-58548-8_34

[55] Vildan Tanriverdi and Robert JK Jacob. 2000. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 265–272. doi:10.1145/332040.332443

[56] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*. PMLR, 3459–3467. doi:10.48550/arXiv.1902.06977

[57] Lode Vanacken, Tovi Grossman, and Karin Coninx. 2007. Exploring the effects of environment density and target visibility on object selection in 3D virtual environments. In *2007 IEEE symposium on 3D user interfaces*. IEEE. doi:10.1109/3dui.2007.340783

[58] Radu-Daniel Vatavu and Ionuţ Alexandru Zaiţi. 2013. Automatic recognition of object size and shape via user-dependent measurements of the grasping hand. *International Journal of Human-Computer Studies* 71, 5 (2013), 590–607. doi:10.1016/j.ijhcs.2013.01.002

[59] Uta Wagner, Matthias Albrecht, Andreas Asferg Jacobsen, Haopeng Wang, Hans Gellersen, and Ken Pfeuffer. 2024. Gaze, wall, and racket: Combining gaze and hand-controlled plane for 3D selection in virtual reality. *Proceedings of the ACM on Human-Computer Interaction* 8, ISS (2024), 189–213. doi:10.1145/3698134

[60] Uta Wagner, Mathias N Lystbæk, Pavel Manakhov, Jens Emil Sloth Grønbæk, Ken Pfeuffer, and Hans Gellersen. 2023. A fitts' law study of gaze-hand alignment for selection in 3d user interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15. doi:10.1145/3544548.3581423

[61] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. 2023. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11359–11366. doi:10.1109/icra48891.2023.10160982

[62] Yushi Wei, Rongkai Shi, Difeng Yu, Yihong Wang, Yue Li, Lingyun Yu, and Hai-Ning Liang. 2023. Predicting gaze-based target selection in augmented reality headsets based on eye and head endpoint distributions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–14. doi:10.1145/3544548.3581042

[63] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-defined gestures for surface computing. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1083–1092. doi:10.1145/1518701.1518866

[64] Jacob O Wobbrock, Kristen Shinohara, and Alex Jansen. 2011. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1639–1648. doi:10.1145/1978942.1979181

[65] Erik Wolf, Sara Klüber, Chris Zimmerer, Jean-Luc Lugrin, and Marc Erich Latoschik. 2019. " Paint that object yellow": Multimodal interaction to enhance creativity during design tasks in VR. In *2019 International conference on multimodal interaction*. 195–204. doi:10.1145/3340555.3353724

[66] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. 2023. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14928–14940. doi:10.1109/iccv51070.2023.01371

[67] Yukang Yan, Chun Yu, Xiaojuan Ma, Xin Yi, Ke Sun, and Yuanchun Shi. 2018. Virtualgrasp: Leveraging experience of interacting with physical objects to facilitate digital object retrieval. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–13. doi:10.1145/3173574.3173652

[68] Difeng Yu, Hai-Ning Liang, Xueshi Lu, Kaixuan Fan, and Barrett Ens. 2019. Modeling endpoint distribution of pointing selection tasks in virtual reality environments. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13. doi:10.1145/3355089.3356544

[69] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2021. Gaze-supported 3d object manipulation in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3411764.3445343

[70] Difeng Yu, Qiushi Zhou, Joshua Newn, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2020. Fully-occluded target selection in virtual reality. *IEEE transactions on visualization and computer graphics* 26, 12 (2020), 3402–3413. doi:10.1109/tvcg.2020.3023606

[71] Shumin Zhai, William Buxton, and Paul Milgram. 1994. The "Silk Cursor" investigating transparency for 3D target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 459–464. doi:10.1145/191666.191822

[72] Shumin Zhai, Jing Kong, and Xiangshi Ren. 2004. Speed–accuracy tradeoff in Fitts' law tasks—on the equivalency of actual and nominal pointing precision. *International journal of human-computer studies* 61, 6 (2004), 823–856. doi:10.1016/j.ijhcs.2004.09.007

[73] Chenyang Zhang, Tiansu Chen, Eric Shaffer, and Elahe Soltanaghai. 2024. FocusFlow: 3D gaze-depth interaction in virtual reality leveraging active visual depth manipulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3613904.3642589

[74] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. 2024. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*. Springer, 386–403. doi:10.1007/978-3-031-73347-5_22

[75] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. 2024. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 235–246. doi:10.1109/3dv62453.2024.00016

[76] Yawen Zheng, Jin Huang, Hao Zhang, Yulong Bian, Juan Liu, Chenglei Yang, Feng Tian, and Xiangxu Meng. 2025. 3D Ternary-Gaussian model: Modeling pointing uncertainty of 3D moving target selection in virtual reality. *International Journal of Human-Computer Studies* 198 (2025), 103454. doi:10.1016/j.ijhcs.2025.103454

[77] Qian Zhou, George Fitzmaurice, and Fraser Anderson. 2022. In-depth mouse: Integrating desktop mouse into virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17. doi:10.1145/3491102.3501884