Supplementary Materials for FAM-HRI: Foundation-Model Assisted Multi-Modal Human-Robot Interaction Combining Gaze and Speech

Yuzhi Lai, Shenghai Yuan, Boya Zhang, Benjamin Kiefer, Peizheng Li and Andreas Zell

I. List of Symbols and Their Meanings

To provide a comprehensive reference for the mathematical notations used throughout our paper, we include Tab. I in the supplementary materials. This table details all symbols, spaces, sets, and corresponding meanings, covering essential aspects of frame transformations, gaze reconstruction, scene observations, intention fusion, control system variables, and policy generation.

The notation system is structured to facilitate clarity in understanding multi-view geometric transformations, gaze-language fusion, and LLM-based policy generation. Specifically, it includes:

1) Frame representations for robot base, cameras, and gaze.
2) Scene Observations for both human and robot view, ensuring a unified representation for multi-view alignment.
3) Control system parameters, including action primitives, parameters, and planning policies.

This structured notation serves as a foundation for understanding FAM-HRI's multi-modal interaction system, enabling reproducibility and further development within the research community.

II. Explanation of the 3D Gaze Reconstruction

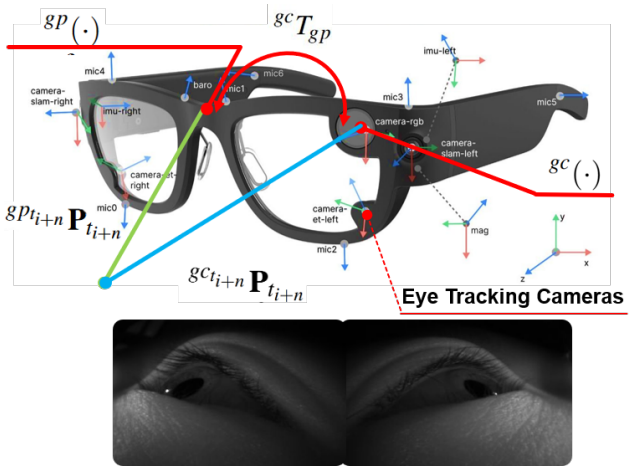


Fig. 1. Gaze Estimation System on META ARIA Glasses.

$$^{gc_{t_{i+n}}}\mathbf{P}_{t_{i+n}} = {}^{gc}T_{gp}\,{}^{gp_{t_{i+n}}}\mathbf{P}_{t_{i+n}} \tag{1}$$

In our system, gaze estimation and reconstruction are performed to ensure accurate and efficient intention fusion. As described in Eq. 1, first the raw gaze vector in glasses pupil frame $^{gp_{t_{i+n}}}\mathbf{P}_{t_{i+n}}$, estimated from the eye-tracking cameras (as shown in Fig. 1), is transformed into the glasses camera frame using the pre-calibrated transformation $^{gc}T_{gp}$.

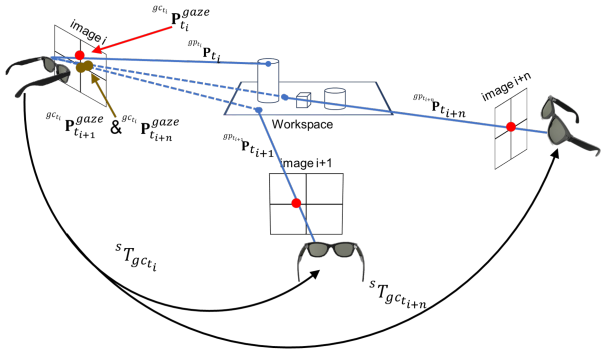


Fig. 2. Schematic of 3D gaze estimation and reconstruction.

$$\begin{aligned} ^{gc_{t_i}}\mathbf{P}_{t_{i+n}} &= ({}^{s}T_{gc_{t_i}})^{-1}({}^{s}T_{gc_{t_{i+n}}})\,{}^{gc_{t_{i+n}}}\mathbf{P}_{t_{i+n}} \\ ^{gc_{t_i}}\mathbf{p}^{gaze}_{t_{i+n}} &= \mathcal{K}\,{}^{gc_{t_i}}\mathbf{P}_{t_{i+n}} \end{aligned} \tag{2}$$

Then, all gaze vector across different timestamps in gaze fixation time period Δt are transformed onto the image plane of the user-centric camera at t_i, the first timestamps of Δt. As shown in Fig. 2 and Eq. 2, this is achieved by leveraging glasses pose estimation via SLAM, applying the transformations $^{s}T_{gc_{t_i}}$ and $^{s}T_{gc_{t_{i+n}}}$ to convert each gaze vector from the glasses camera frame at t_{i+n} to the glasses camera frame at t_i, donated as $^{gc_{t_i}}\mathbf{P}_{t_{i+n}}$. Finally, the reconstructed 3D gaze vectors are projected onto the 2D image plane of the user-centric camera using the camera intrinsic matrix $\mathcal{K}$. This projection allows gaze-language fusion to operate on a single reference image, minimizing the impact of gaze drift and reducing computational overhead by avoiding per-frame scene observation generation.

TABLE I

LIST OF SYMBOLS AND THEIR MEANINGS

Symbol	Space	Meaning
${}^{r}()$	-	Frame: Robot Base
${}^{c}()$	-	Frame: Robot Camera
${}^{gc}()$	-	Frame: Glasses Camera
${}^{gp}()$	-	Frame: Glasses Pupil
$\mathcal{K}$	$\mathbb{R}^{3\times3}$	Basic Geometry: Camera Intrinsic Matrix
$\mathbf{p}$	$\mathbb{R}^2$	Basic Geometry: 2D Position of a 3D coordinate
$\mathbf{P}$	$\mathbb{R}^3$	Basic Geometry: 3D coordinate
${}^{gc}T_{gp}$	$\mathbb{SE}(3)$	Transformation from Glasses Camera Frame to Glasses Pupil Frame
${}^{s}T_{t_i}^{head}$	$\mathbb{SE}(3)$	Transformation Representing Head Pose at Time t_i
β	$\mathbb{R}^4$	Environmental Bounding Box Representation ($[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$)
M	$\mathbb{R}^{H\times W}$	Environmental Segmentation Mask (binary or probability map)
$\mathcal{S}$	Σ^*	Human Input Speech Sequence in Text
$\mathcal{G}$	$\mathbb{R}^{H\times W}$	Human input of Eye tracking camera
$\mathcal{U}$	$\mathbb{R}^{H\times W*(t_n-t_1)}$	Human view: Image from User-Centric Camera on glasses
$\mathcal{Z}_H$	$\Sigma^* \times \mathbb{R}^{H_1\times W_1\times(t_n-t_1)} \times \mathbb{R}^{H_2\times W_2\times(t_n-t_1)}$	Human Input State Vector / Descriptor
${}^{gc_{t_i}}\mathbf{P}_{t_i}$	$\mathbb{R}^3$	Human view: 3D Gaze Vector at Time t_i in Glasses Carmea Frame at Time t_i
${}^{gc_{t_i}}\mathbf{P}_{t_{i+n}}$	$\mathbb{R}^3$	Human view: 3D Gaze Vector at Time t_{i+n} in Glasses Carmea Frame at Time t_i
$\mathcal{Z}_g$		Human view: Scene Observations Set $\{c_{target},\ \{\mathbf{p}_{g_i},\ \beta_{g_i},\ M_{g_i}\}_{i=1}^{N}\}$
$\mathbf{p}^{gaze}$	$\mathbb{R}^2$	Human View: 2D Projected Gaze Point on the Image Plane of $\mathcal{U}$
${}^{gc_{t_i}}\mathbf{p}_{t_{i+n}}^{gaze}$	$\mathbb{R}^2$	Human View: Projected Gaze Point at Time t_{i+n} on Image at Time t_i
$\overline{\mathcal{Z}_g}$	$(\overline{\mathbf{p}_g},\ \overline{\beta_g},\ \overline{M_g})$	Human view: Referred Object
$\mathcal{C}$	$\mathbb{R}^{H\times W}$	Robot view: Image from Camera of Robot
$\mathcal{Z}_r$		Robot view: Scene Observations Set $\{c_{target},\ \{\mathbf{p}_{r_i},\ \beta_{r_i},\ M_{r_i}\}_{i=1}^{N}\}$
$\overline{\mathcal{Z}_r}$	$(\overline{\mathbf{p}_r},\ \overline{\beta_r},\ \overline{M_r})$	Robot view: Referred Object
$\mathbf{O}_1$	$\{\mathcal{L},\ c_{target},\ \Delta t\}$	Control system: LLM Output for Gaze Time Period Prediction
$\mathcal{L}$	Σ^*	Control System: Target Property Descriptor (e.g., Object, Position)
c_{target}	Σ^*	Control System: Target Object Category from User Command
Δt	$\mathbb{R}$	Control System: Time Period of User's Gaze Fixation on Target Object
X	$\{\overline{\mathcal{Z}_r},\ \mathbf{O}_1,\ \mathcal{S}\}$	Control System: State Space
$\mathbf{O}_2$	π	Control System: LLM Output for Policy Generation
θ	$\mathbb{R}^n$	Control System: Action Parameter Space (Rotation, Translation, position, etc.)
a	Σ^*	Control System: Action Space Primitives
$\mathcal{A}$	$\{a(\theta)\}$	Control System: Set of Action Primitives
α	$\mathbb{R}$	Weighted Decay Factor for Temporal Gaze Analysis
π	$\pi: X \mapsto\ \mathcal{A}\times\Theta$	Control System: Parameterized Planning Policy for Executing Actions

III. EFFECT OF WEIGHT FACTOR

$$\overline{\mathcal{Z}_g} = \underset{\mathbf{p}_{g_i}\in\mathcal{Z}_g}{\arg\min} \sum_{n=0}^{N} e^{\alpha(n-N)} \left\| {}^{gc_{t_i}}\mathbf{p}_{t_{i+n}}^{gaze} - \mathbf{p}_{g_i} \right\| \tag{3}$$

$$\alpha = \begin{cases} 0, & \text{if } N = 2 \\ \min(0.65,\ 0.1N), & \text{otherwise} \end{cases}$$

In our proposed FAM-HRI, the human view intention alignment relies on computing the closest distance to the target objects. A weighted distance function, as shown in Eq. 3, ensures that recent gaze points contribute more to the decision, while older points still retain influence but gradually decay in importance.

The reasoning behind this design of the weight factor α can be observed in the Fig. 3:

1) **Ensuring Recent Gaze Points Have Higher Contribution:** The exponential function $e^{\alpha(n-N)}$ weights recent gaze points higher while decaying older ones.
2) **Avoiding Over-Reliance on a Single Gaze Point:** The

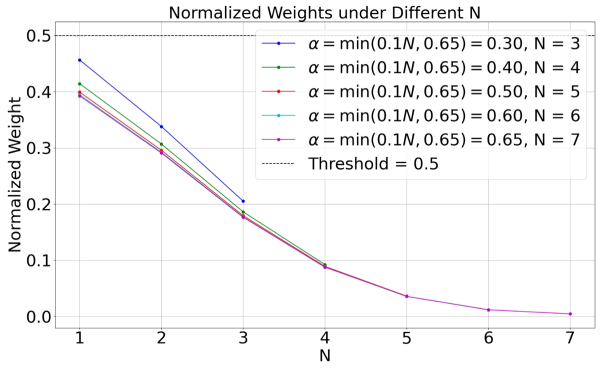
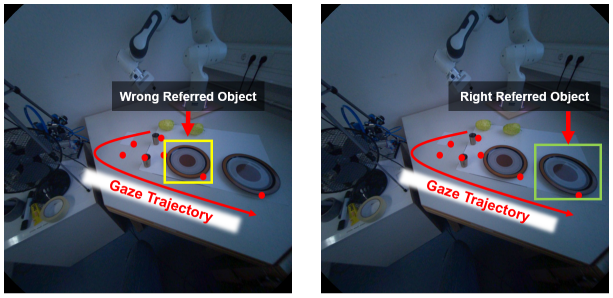


Fig. 3. Normalized Weights under Different N.

upper threshold of 0.65 prevents a single gaze point from dominating the selection.



(a) Human View Intention Alignment without Weight Factor.

(b) Human View Intention Alignment with Weight Factor.

Fig. 4. Comparison of Intention Alignment with and without the Weight Factor.

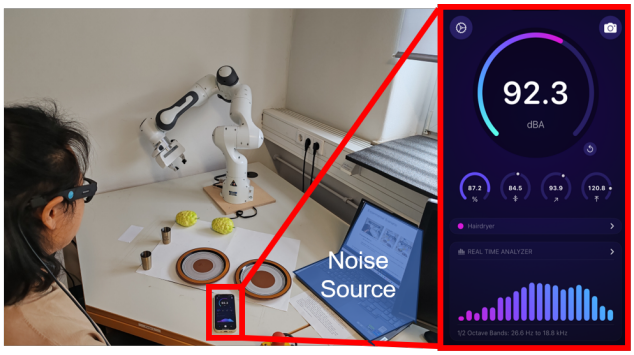
The Fig. 4 illustrates an example of intention alignment with and without weight factor. The user issued the command "put this on the plate", and at the fixation time period "plate", the user's gaze points was directed towards the rightmost plate. However, without a proper weight function, older gaze points retained too much influence, leading the system to mistakenly select the left plate (Fig. 4 (a)). By applying the exponential weight factor, recent gaze points were given higher contribution, effectively reducing the influence of old fixations. As a result, our system successfully identified the correct referred object (Fig. 4 (b)), demonstrating the effectiveness of our approach in handling gaze drift and improving multi-modal intent fusion.

IV. Latency of LLMs Inference

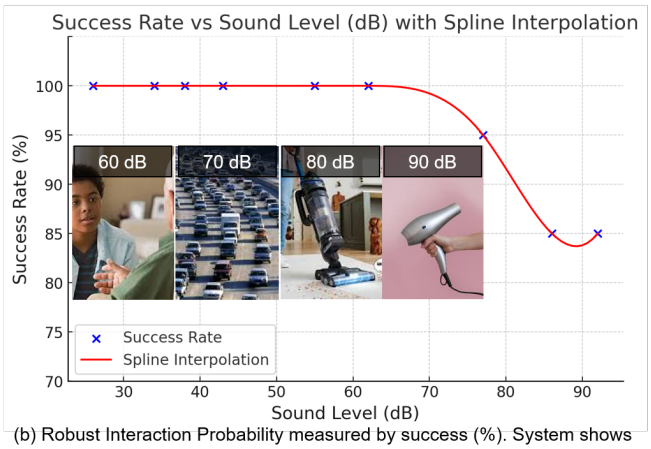
Tab. II presents the inference latency of the two LLM Agent under different experimental scenarios. The reported values include the average response time for human view command processing and planning policy generation, measured across all user trials.

The results show that command processing latency varies between 1353ms and 2202ms, with increased latency in more complex scenarios due to additional language reasoning. Similarly, the policy generation latency ranges from 1281ms to 1997ms, reflecting the increasing complexity of task planning as the number of required action primitives grows. Despite these variations, our proposed FAM-HRI maintains a reasonable inference latency, demonstrating its feasibility for human-robot collaboration, with further optimizations possible through model distillation or on-device inference acceleration.

V. Experiments in Noisy Environment



(a) Noise Environment Experiment for rooms with 92.3 dB.



(b) Robust Interaction Probability measured by success (%). System shows high level of robustness even with extreme higher sound level of noise

Fig. 5. System robustness evaluation under varying background noise.

To evaluate the robustness of FAM-HRI in real-world conditions, we conducted experiments under varying ambient noise levels. As shown in Fig. 5(a), the background noise was introduced using a laptop placed near the user, playing a pre-recorded dialogue video. The content of the video simulated a real-indoor conversation scenario, ensuring that the noise characteristics are similar to those encountered in household or public settings. A sound level meter was positioned next to the user to measure the real-time noise intensity. The noise intensity ranged from 20 dB to 90 dB, cover common noise conditions, such as conversation, traffic, vacuum cleaner, and hairdryer. According to the Level Comparison Chart [1], hearing loss may result from sustained exposure to 90dB or more.

The results, presented in Fig. 5(b), illustrate that our proposed FAM-HRI maintains a high success rate even under

[1] http://ehs.yale.edu/sites/default/files/files/decibel-level-chart.pdf

TABLE II

INFERENCE LATENCY UNDER DIFFERENT CONDITIONS

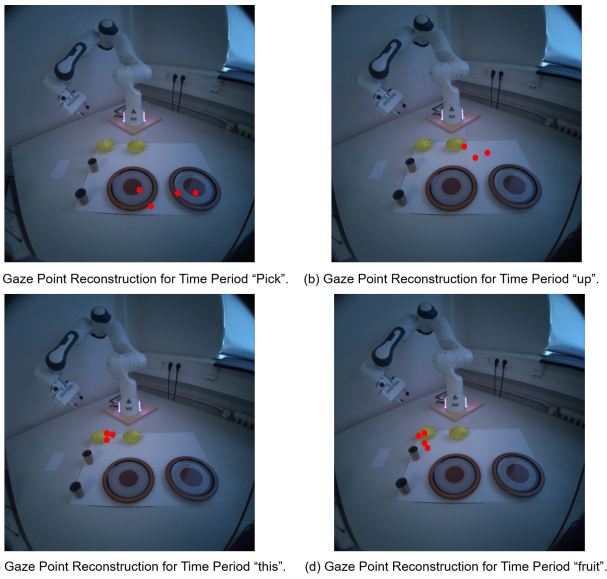
Scenario	Average Latency of Human View Command Processing	Average Latency of Policy Generation
S_1	1353 ± 753ms	1281 ± 692ms
S_2	1624 ± 660ms	1535 ± 946ms
S_3	1855 ± 725ms	1803 ± 891ms
S_4	2202 ± 936ms	1997 ± 861ms

increasing noise levels. Here the user needs to complete a pick-placing task. The system consistently achieved 100% success rate up to 80 dB, indicating strong resilience to moderate noise conditions. Even at 90+ dB, where typical speech recognition systems often fail, FAM-HRI still maintained an 85% success rate, demonstrating its robustness in handling speech input under challenging conditions.

The strong noise resistance is attributed to two aspects:

1) **Hardware:** Unlike traditional microphones that are directly exposed to ambient noise, the ARIA glasses' microphones are embedded within the frame. This placement significantly reduces the impact of background noise, allowing the user's voice to be clearly captured, even in high-noise environments.
2) **Software:** Our two LLM-Agent framework further enhances robustness against background noise. The first LLM agent incrementally analyzes task-relevant verbs, nouns, and pronouns in the user's command, filtering out irrelevant words and recognized background noise before passing a cleaned and structured instruction to the agent for policy generation.

VI. Effects of Gaze Fixation



(a) Gaze Point Reconstruction for Time Period "Pick". (b) Gaze Point Reconstruction for Time Period "up".

(c) Gaze Point Reconstruction for Time Period "this". (d) Gaze Point Reconstruction for Time Period "fruit".

Fig. 6. Effects of Gaze Fixation Time Periods.

The key challenge in gaze-language intention alignment is determining the time period for gaze fixation. In this section, we analyzed the relationship between gaze fixation time period and verbal references, as visualized in Fig. 6.

In the experiments we found that the users do not immediately fixate on the referred object at the begining of their command. Instead, they first scan the environment, briefly shifting their gaze across multiple objects, introducing dynamic noise. As the user's command progresses, users tend to fixate on the referred object when verbalizing pronouns (e.g., "this") or specific object categories. However, after the object reference, gaze may shift away as users transition to the next phase of their command. These observations confirm that gaze-language alignment follows a structured temporal pattern rather than being uniformly distributed across the command. Our LLM-Agent for human view command processing leverages this insight to dynamically estimate the most probable fixation time period based semantic cues in speech input, filtering out irrelevant words and improving success rate for task execution.

VII. Analysis of User Study

We performed a user study to investigate two main hypotheses:

1) **H1:** FAM-HRI is preferred by participants compared to all other baseline methods.
2) **H2:** FAM-HRI is quantitatively more effective and accurate compared to all other baseline methods.

H1: As shown in the Fig. 6 in the submitted paper, participants rated FAM-HRI higher than all baseline methods as "light and fiexible" ($p < 0.001$), "modern" ($p < 0.001$), and "simple to use" ($p < 0.001$). Additionally all participants indicated that our method was the most effective for users with limited mobility or motor impairments.

H2: In the submitted paper, the Tab. I shows the success rate and the average interaction time of our proposed FAM-HRI and all baseline methods. In all experimental scenarios, our method demonstrates a high success rate and the shortest interaction time. The participants further rated the execution of their personal preferences, FAM-HRI, as "accurate" ($p < 0.001$) and "efficient" ($p < 0.001$).

VIII. Object Selection Among Similar Items

In Scenario S_1, the system evaluates its ability to distinguish between visually similar objects through multi-view alignment. As shown in Fig. 7, our approach integrates gaze-language intention alignment with a feature-based multi-view matching strategy to ensure precise object selection.

This method significantly improves selection accuracy and efficiency, particularly when objects share similar appearances and are densely arranged. Unlike methods that rely solely on language-based interaction, our approach ensures

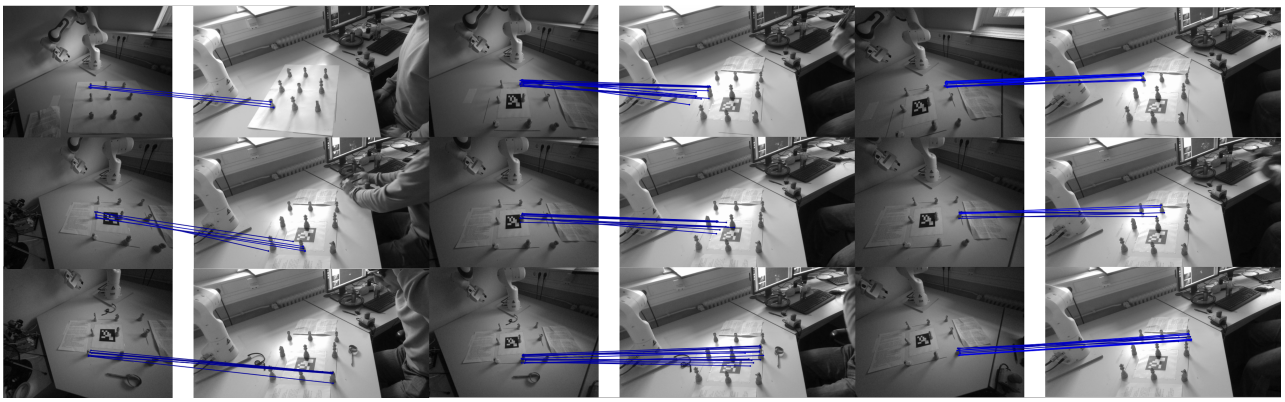


Fig. 7. Multi-View Alignment for Object Selection Among Similar Items (S_1).

robust selection without requiring explicit row-column specifications or predefined object labels.

IX. Task Execution in Complex Environment

Table III presents the success rates and average interaction times for various tabletop manipulation tasks across different scenarios. A key observation is that tasks executed without a referred object generally exhibit faster interaction times due to shorter command lengths. However, in Scenario S_1, the success rate for commands without explicit object references is lower. This is primarily because small, detailed parts of the chess pieces were sometimes misdetected by VLM. Consequently, during multi-view alignment, insufficient keypoints in these small detected parts were found, reducing accuracy of object selection.

In contrast, for other scenarios, the success rate without explicit object references remains higher than with explicit object references. This is attributed to the fact that shorter commands avoid potential speech recognition errors, thereby preventing incorrect misclassification of object categories.

X. Failure Modes

The failure cases of our system can be categorized into four key aspects: human view input, scene observation, multi-view alignment, and LLM reasoning.

1) **Human View Input:** For human view input, there are two sources of failure, speech recognition errors and inaccurate gaze estimation. The speech recognition error in our system primarily due to misinterpretation of similar-sounding words. Gaze estimation errors are largely influenced by improper wearing of the glasses. Specifically, when the nose bridge of the ARIA glasses does not correctly placed on the user's nose, the accuracy of gaze estimation decrease significantly. However, to mitigate this, users were instructed to adjust their glasses properly before proceeding with the experiment, ensuring that such errors do not propagate into the system.
2) **Scene Observation:** For scene observation, failures mainly come from more or less detected objects. As shown in Fig. 8, in Scenario S_1, where users need to select a specific chess piece without explicitly stating its category, smaller parts of the chess pieces were sometimes incorrectly detected as separate objects. Additionally, speech recognition errors could result in the misclassification of the target object, causing the system to detect an incorrect category and subsequently select the wrong object.
3) **Multi-View Alignment:** For multi-view alignment, the main reason for failure is insufficient or unclear feature correspondence between human and robot views. Feature matching using superglue becomes unreliable in the presence of weak object textures, repetitive patterns, or partial occlusions, leading to incorrect correspondences.
4) **LLM reasoning:** For LLM Reasoning, the primary failure modes come from formatting inconsistencies and hallucinations in the generated output. Since FAM-

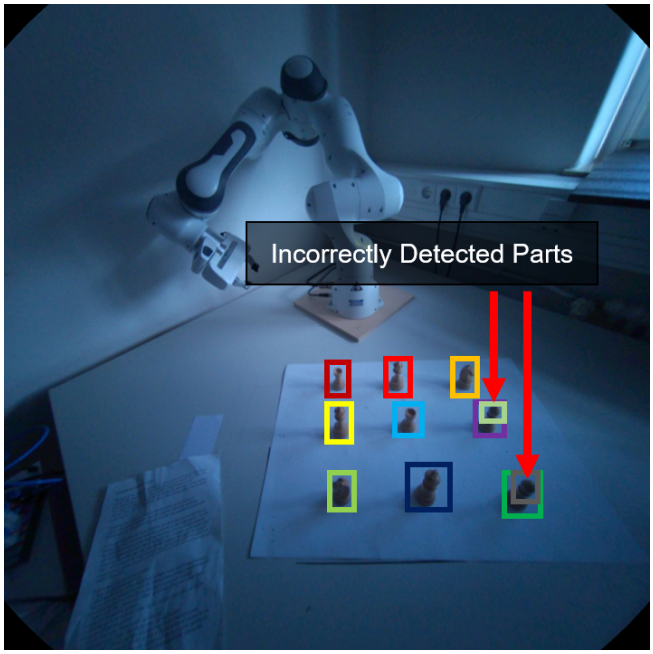


Fig. 8. More/ Less Detected Objects.

TABLE III

DETAILED SIMULATION TABLETOP MANIPULATION SUCCESS RATE (%) AND INTERACTION TIME (S) ACROSS DIFFERENT TASK SCENARIOS

	Success Rate (%)	Average Interaction Time (s)
With referred object		
pick up the <object>	100	1.3
grab the pieces (S_1)	96	1.8
put the <object> on the <recepticle-plate>	100	2.2
put the <object> on the <recepticle-plate> then pour some thing from the <recepticle-cup> on it	94	5.9
put this <object> there	89	2.5
put the <object1> and <object2> on the <recepticle-plate>	92	3.8
put the <object1> on the <recepticle-plate1> then put the <object2> on the <recepticle-plate2>	90	6.1
Without referred object		
pick up this	100	1.1
grab this (S_1)	87	0.8
put this on that	98	1.7
put this on that then pour some thing from this on it	96	4.2
put this there	93	1.9
put this and this on that	96	3.1
put this on this then put this on that	92	4.2

HRI requires the LLM's response to strictly follow a predefined prompt format, any deviation renders the output unusable by subsequent system modules. In rare cases, the LLM misinterprets the required action parameters, leading to incorrect parameter values or an incorrect ordering of action arguments. Additionally, logic errors happen in policy generation due to LLM hallucinations. For instance, in placement tasks, the generated policy may not open the gripper after reaching the target position, leading to task failure. However, such errors were observed infrequently in our experiments, as structured prompt design and output constraints significantly reduced the occurrence of these issues.

XI. PARAMETER TUNING AND REPRODUCIBILITY

TABLE IV

LIST OF PARAMETERS

Grounding Dino SAM2	
Box Threshold	0.3
Text Threshold	0.3
Grounding-Model	Tiny
SAM2 Model	Large
SuperGlue	
Max Keypoints	10000
Keypoints Threshold	1e-5
Match Threshold	1e-5
Resize	No
Weights	indoor

Our approach involves several parameters that are manually set or empirically tuned based on experimental observations.The Weight Factor in intention alignment was determined through extensive testing to balance recent gaze points' influence while preventing a single point from dominating the selection process. While these parameters have been carefully adjusted to optimize system performance, we acknowledge that they are not mathematically proven to be optimal.

To ensure reproducibility, we provide explicit parameter settings for key components of our system in Tab. IV, including Grounding Dino for Scene Observation and SuperGlue for Multi-View Alignment.

XII. LIMITATION AND FUTURE WORK

Although our proposed FAM-HRI has strong performance, there are still some challenges. Future work will focus on improving inference efficiency, gaze noise handling, and deployment on edge devices to enhance system adaptability and real-time performance.

One of the primary limitations of our system is the high inference latency for human view command processing and policy generation. To address this, future improvements could involve Model Distillation and Retrieval-Augmented Generation (RAG).

A key challenge in gaze-based interaction is the variability in gaze behavior across users. Some users prefer to fixate steadily on the target object, while others tend to shift their gaze from one object to another as a way of indicating selection. In our approach, we mitigate this issue by using a distance function with weight factor that dynamically adjusts the contribution of gaze points over time. Future work will explore an end-to-end Vision-Language Model (VLM) approach, where gaze and speech signals are fused directly within a deep learning framework trained on large-scale multi-modal data.

Currently, FAM-HRI relies on high-performance GPUs, which limits deployment on low-power embedded systems. To make the system more practical for real-world assistive applications, future research will focus on creating lightweight models, and enabling real-time, autonomous operation on embedded systems.