Identity-Conditioned Preference-Aware Table Tidying with LLM-in-the-Loop

Bojun Long
School of Robotics, Xi'an Jiaotong-Liverpool University
Suzhou, China
Bojun.Long21@student.xjtlu.edu.cn

Zhenhao Guo
School of Robotics, Xi'an Jiaotong-Liverpool University
Suzhou, China
Zhenhao.Guo21@student.xjtlu.edu.cn

Fan Zhu*
School of Robotics, Xi'an Jiaotong-Liverpool University
Suzhou, China
Fan.Zhu@xjtlu.edu.cn

Abstract: This paper presents the Hierarchical Identity-conditioned Table-Tidying Framework (HITF), a lightweight embodied AI system for personalized organization in shared-desk environments via identity-conditioned (per-user history–conditioned) reasoning. HITF integrates YOLOv5 and CLIP for multi-attribute object perception, SFace-based facial recognition for user identification, a relational database for persistent user–object–action linkage, and an LLM-in-the-loop reasoning module for habit inference and action generation. Experiments across multiple users and everyday objects demonstrate that HITF achieves up to 91.3% user satisfaction under low computational constraints. These results highlight the framework's effectiveness and extensibility for scalable, data-efficient personalized tidying in shared-desk scenarios with per-user history–conditioning.

Keywords: Personalized Robotics, User Preference Learning, Multimodal Perception

1 Introduction

In daily life and work scenarios, shared spaces such as desks, dining tables, meeting tables, and bedrooms are often alternately used by different users for different purposes. For example, a parent may use the desk as an office area during the day, and in the evening rearrange it for their child's study or play. This multi-user, multi-purpose, and multi-scene usage pattern poses significant challenges for robots, as they are not only required to organize and tidy objects but also to recognize who the current user is, understand their habits, and meet their personalized organizational preferences. This makes it essential for robots to interpret both the user and the environment in order to execute tasks effectively.

Traditional service robots often rely on predefined rules or fixed task scripts [1], which allow them to perform basic organization but lack adaptability. Some studies have explored learning from human organizational habits to predict reasonable placements of objects in household contexts. Research on personalization has further introduced methods such as collaborative filtering [2], learn-

*Corresponding author.

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

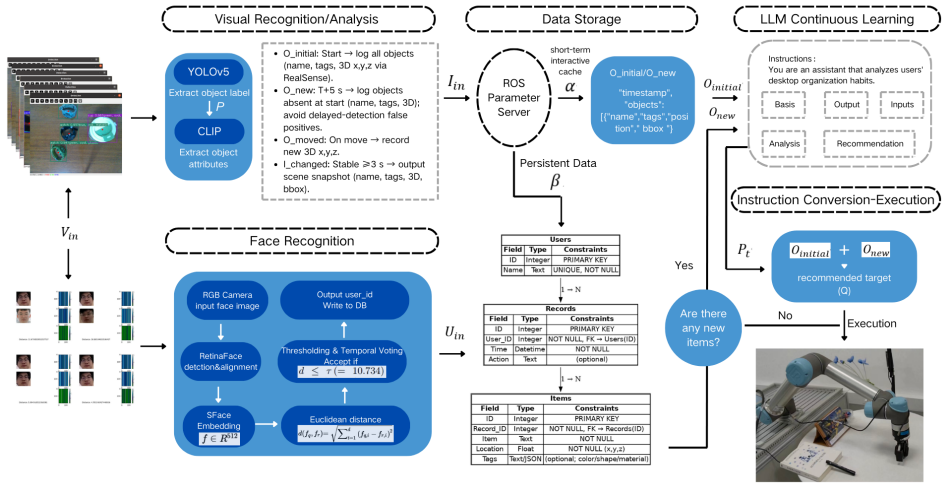


Figure 1: **Overview of HITF.** Top-left: a lightweight vision cascade (YOLOv5→CLIP) operates on RGB-D frames to produce object classes and multi-attribute cues (color/shape/material). Bottom-left: user identity is obtained via a RetinaFace-based detection/alignment pipeline and SFace embeddings with robust temporal voting. Both streams, together with scene-change events (e.g., O_{initial}, O_{new}), are written to the ROS Parameter Server for short-term buffering and dispatch, and are persisted in a structured store (Users/Records/Items). Right: an LLM-in-the-loop reasons over the identity-conditioned history to infer user preferences, produces a tidying target, and drives the instruction conversion module to command a robotic arm, enabling personalized organization in shared-desk environments.

ing latent preference vectors through graph-based or representation learning approaches [3], and scene–context association learning [4]. While these studies demonstrate the feasibility of extracting user preferences from example data, they require extensive data collection and training, leading to high human and computational resource costs.

Subsequent works have incorporated large language models (LLMs) to help infer user preferences. For example, [5] proposed TidyBot, which combines LLMs to generalize user organization rules from a few examples. Although it demonstrates certain generalization capabilities, its application remains limited and focuses only on single-user preferences, lacking the ability to handle multiple users with different needs simultaneously. These existing research directions still suffer from several key limitations: the inability to distinguish between multiple users, reliance on manually annotated data, difficulty in autonomously learning preferences from real behaviors, and poor handling of vague or abstract instructions. As a result, current systems often still require human intervention, limiting their ability to achieve truly intelligent and personalized organization.

To address these challenges, this paper proposes a unified framework named *Hierarchical Identity-conditioned Table-Tidying Framework* (HITF). The framework integrates a YOLOv5 vision model trained on a small dataset with the lightweight foundation model CLIP to extract and classify key environmental information from video frames, which is then stored in structured form. This is combined with large language model reasoning to enable fast and continuous learning of users' organizational habits without relying on external prompts or interventions. Specifically, for each individual user, HITF first generates a generic organizational strategy based on environmental information and then, through facial recognition, creates a dedicated user database. By continuously observing the user's organizational behavior, HITF updates the database and rapidly adapts organizational strategies to better align with the user's real preferences. Our contributions can be summarized as follows: 1) We propose a desktop tidying framework that is lightweight and can be deployed under extremely low computational constraints, yet adapts to diverse user organization needs in real-world scenarios. 2) Our approach enlarges salient features of target objects to facilitate subsequent learning under

limited computation. 3) We establish a user database linked to face recognition for individualized habit modeling. 4) We leverage LLM reasoning to infer organizational preferences. Furthermore, by adjusting the training set of the vision model and modifying the control instructions of the execution module, our architecture can be extended to a wider range of organizational tasks. Finally, we outline future research directions and potential improvements to this framework. A 2-minute anonymized demo of HITF is provided in Appendix A.

2 Related Work

In recent years, embodied AI has witnessed rapid progress in task research within indoor environments. Researchers have proposed a variety of household benchmarks to evaluate agents' perception and planning capabilities in complex spaces [6]. Among these tasks, *object rearrangement* has been recognized as one of the key challenges, where the central goal is to achieve a target environmental configuration through a sequence of pick-and-place operations [7]. Early approaches typically relied on explicit human specification, such as assigning target layouts for each object [8] or using interactive gestures to determine placement locations [1]. Although effective in achieving task completion, these methods required heavy human annotation, making them difficult to scale to larger real-world applications. To reduce manual intervention, subsequent works attempted to learn from human organizational habits to predict reasonable placements of objects in household settings [9, 10]. However, these methods often captured only generalized organizational patterns and lacked the ability to model individual differences and personalized requirements.

To address inter-user variability, some studies introduced personalized preference learning. [2] proposed a collaborative filtering framework that modeled "object–user" relationships as a matrix factorization problem, augmented with active probing and semantic hierarchy expert models to alleviate the cold-start issue. This work demonstrated the feasibility of personalized organization, but it relied heavily on large-scale crowdsourced data and focused primarily on static single-user scenarios. [3] further proposed the NeatNet model, which represents organizational scenes as graphs and leverages graph neural networks with variational autoencoders to learn low-dimensional user preference vectors. This enabled robots to infer user habits from few examples and generalize to unseen objects. [5] proposed TidyBot, leveraging LLMs to induce organizational rules from limited "object–location" data with strong real-world performance, but restricted to single-user settings. To scale personalization, [4] introduced the PARSEC benchmark (110K examples, 72 users) and ContextSortLM, which integrates demonstrations into structured preference representations with contextual reasoning. Yet, it too remains limited to single-user scenarios, without addressing conflicts or collaboration in shared environments.

Personalized organization further requires robust user distinction. Recent models such as ArcFace [11], SFace [12], and speaker embeddings (x-vector, ECAPA-TDNN) achieve strong identity recognition. To enhance robustness, [13] proposed a multimodal incremental Bayesian network integrating facial and soft biometric traits, while [14] introduced a Transformer-based model leveraging gaze and speech for multi-party response inference. However, these methods focus on identity verification or response targeting, without addressing preference modeling or tidying behaviors.

Meanwhile, large language models (LLMs) have shown significant potential in robotic semantic reasoning. [15] decomposes high-level instructions into executable actions filtered by value functions, enabling robots to perform long-horizon multi-step tasks. SayPlan [16] further extends planning to complex multi-room environments by leveraging 3D scene graphs. RT-2 [17] and PaLM-E [18] unify vision, language, and action through integrated modeling, thereby supporting cross-modal zero-shot reasoning and control. These works highlight the semantic reasoning abilities of LLMs, but their generated plans remain largely generic and lack adaptation to personalized preferences or multi-user contexts.

In summary, prior research has made substantial progress in personalized organization, multimodal user recognition, and LLM-based reasoning. However, these directions remain largely fragmented: personalization methods mostly assume single-user scenarios, multimodal recognition methods are

not integrated with preference modeling, and LLM reasoning approaches have not been adapted to multi-user differentiated needs. To address these gaps, this work proposes a unified framework that integrates identity-conditioned (per-user history–conditioned) preference modeling, multimodal recognition with behavioral attribution, and LLM-based semantic reasoning, enabling robots to interpret ambiguous instructions more naturally and execute personalized organization tasks in shared spaces.

3 Method

To rapidly collect item-level information in everyday desktop environments and to differentiate tidying preferences across users via identity-conditioned retrieval over a per-user store, we design a hierarchical architecture named the **Hierarchical Identity-conditioned Table-Tidying Framework (HITF)**. The architecture consists of four primary modules: (i) a vision recognition/analysis module, (ii) a data storage module, (iii) a large-language-model (LLM) learning module, and (iv) a command translation-and-execution module.

3.1 Vision Recognition/Analysis

Vision recognition/analysis. The module consumes frame-wise observations V_{in} and *branches* them into desktop-environment signals I_{in} from an RGB–D camera and user-facial signals U_{in} from an RGB camera. I_{in} is processed by a "YOLOv5 $\rightarrow$ CLIP (ViT-B/32)" cascade: YOLOv5 with non-maximum suppression (NMS) produces bounding boxes (bbox); for each *YOLOv5 bbox crop*, CLIP serves as a zero-shot attribute head to assign Top-1 labels for color/shape/material (per vocabulary); RealSense deprojection with camera intrinsics recovers the 3D coordinates (x, y, z). This yields *multi-attribute, 3D-anchored* object descriptors that provide lightweight, informative signals to the LLM. Conditioned on the current scene state, four object-centric signals are distilled from I_{in}:

- $O_{\textbf{initial}}$: At task start, record each item's name, tags, and position (x, y, z).
- $O_{\textbf{new}}$: Beginning $5\,\text{s}$ after task start, when a previously absent item appears, record its name, tags, and (x, y, z) to avoid treating late detections as genuinely new.
- $O_{\textbf{moved}}$: When an item is moved by the user, record its new (x, y, z).
- $I_{\textbf{changed}}$: Whenever any item is judged to have moved and then stabilizes for $3\,\text{s}$, commit a full scene snapshot including each item's name, tags, (x, y, z), and bbox.

Face recognition and identity binding. U_{in} is fed to a face-recognition pipeline: RetinaFace detects faces and performs five-point landmark alignment to obtain a standardized crop; SFace then extracts a 512-dimensional embedding $\mathbf{f} \in \mathbb{R}^{512}$. We use the Euclidean distance $d(\mathbf{f}_q, \mathbf{f}_r) = \|\mathbf{f}_q - \mathbf{f}_r\|_2$ to compare the query against enrolled templates; if $d < \tau = 10.734$ the two are deemed the same identity and the unique identifier `user_id` is returned. The threshold τ is selected via cross-validation on pilot experiments and a held-out validation set to balance the false accept rate (FAR) and false reject rate (FRR). To enhance robustness under identity switching and transient perturbations, we apply a temporal sliding-window majority vote with window length W and vote threshold T (e.g., $W{=}10$, $T{=}7$); an identity is confirmed as the *current active user* only if predictions are consistent across consecutive frames and the mean confidence exceeds a preset threshold. The recognized `user_id` is then temporally aligned and bound to co-period environmental perception (object names, attribute tags, and 3D coordinates) and immediately written to persistent storage (β), forming a traceable user–action–object linkage that provides a conditional prior for subsequent preference modeling and LLM-based semantic reasoning.

3.2 Data Storage

All textual outputs from the vision recognition/analysis module are routed to the data-storage module for categorized storage and subsequent dispatch. The data-storage module comprises two components: (i) a short-term exchange cache (α) implemented on the ROS Parameter Server, and (ii) a

persistent per-user database (β). Hosted by the ROS Master, the ROS Parameter Server provides a global, hierarchically namespaced key–value store; in this project, we use it as a lightweight exchange hub for control flags and structured scene snapshots, with values encoded as YAML/JSON. Parameters are written from Python via `rosparam`, and downstream nodes read them on demand.

We write the YOLOv5–CLIP cascade outputs—the initial environment information O_{initial} and the new-object information O_{new}—into the ROS Parameter Server as the short-term exchange cache (α). For events that occur only once within each task episode (after a new object is logged it is bound in the in-memory dictionary `object_memory`; when the same label is seen again, `object_memory` appends a record and returns `is_new=false`, so the object is no longer treated as new), we commit a scene snapshot to the corresponding variables in the form `{"timestamp","objects":[{"name","tags","position","bbox"}]}`. These parameters retain only the most recent snapshot/event and are overwritten when the current task ends and a new task begins (the current ROS node terminates and a new node is launched); they are not intended for long-term storage.

For persistent data (β)—namely the changed-environment information I_{changed}—the user database continuously monitors this variable after the task starts and, upon confirming its materialization, extracts its contents and commits them to persistent storage. Within the data-storage module, we adopt a relational three-table schema (**Users**, **Records**, **Items**) and establish hierarchical, chain-like associations via primary/foreign-key constraints (see Fig. 2). This design guarantees unified management and traceability across user identities, interaction events, and scene objects.

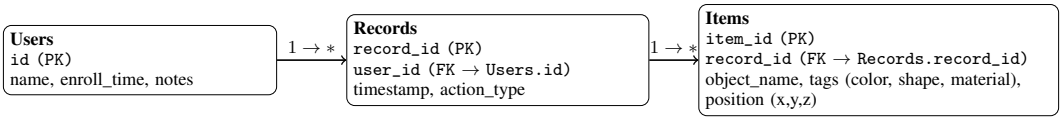


Figure 2: Entity–relationship diagram linking *Users* → *Records* → *Items*. PK = Primary Key, FK = Foreign Key.

- **Users**: Stores minimal identity attributes with `id` as the primary key, providing the root entity for all interaction events and ensuring data traceability to each user.
- **Records**: Logs user–system interactions, with `record_id` as primary key and `user_id` as foreign key to `Users(id)`. The $1:N$ constraint allows one user to have multiple records, each containing `timestamp` and `action_type`.
- **Items**: Stores object-level details, linked to `Records(record_id)` via a foreign key. A single record may involve multiple items ($1:N$), each described by `item_id`, `object_name`, `tags` (e.g., color, shape, material), and 3D $\text{position}(x, y, z)$.

This schema ensures temporal and entity consistency, forming a complete *user–action–object* mapping that supports both LLM-based reasoning and preference analysis.

History-driven reorganization (no-new-object). When the scene remains stable and no new-object event O_{new} is detected, the system bypasses language reasoning and retrieves the user's historical scene records from the storage layer to determine target placements. For each item i, we compute the empirical distribution over previously observed regions $r \in \mathcal{R}$ and select the most frequent region as the target:

$$r_i^\star \;=\; \arg\max_{r \in \mathcal{R}} \text{freq}(i, r).$$

The resulting per-item map $\{\, i \mapsto r_i^\star \,\}$ is cached and passed to the executor. This history-driven policy resides precisely between storage and execution: the storage layer supplies user-specific placement statistics, while the executor consumes them to maintain the user's habitual layout without additional inference latency.

3.3 LLM Learning

Scope & trigger. This module is activated only when O_{new} is present in the current round; if O_{new} is absent, the LLM is not invoked and control is handed off directly to the *Command Translation & Execution* module.

Trigger & pipeline. Upon detection of O_{new}, the node assembles a fixed-template prompt from the current scene summary (detected objects, 3D positions, and multi-attribute tags) and, using the current active `user_id` supplied by the recognition pipeline, queries the storage layer (*Users* $\rightarrow$ *Records* $\rightarrow$ *Items*) for that user's historical scene records, which cover every post-change snapshot (I_{changed}) during the user's active periods. These per-user histories are aggregated into a user-specific summary and injected—together with the current scene summary—into the prompt as prior evidence for preference inference. The LLM (GPT-4o) is then called, and the resulting recommendation is written to `/placement_prompt` (P_t). After each inference, the relevant triggers and keys are reset to ensure idempotence and a well-defined system state.See Appendix B for the full prompt template and placeholder definitions used when O_{new} is present.

3.4 Command Translation & Execution

Inputs and triggers. The executor is gated by lightweight ROS parameters (optionally set by an ASR interface): `recommend=1` starts the opening routine; `/recommended_position_flag=1` triggers a single pick-and-place. *Decision rule.* The target source is selected by the presence of O_{new} in the current episode: (i) if O_{new} exists, consume an **LLM target** produced from `/placement_prompt` (P_t) and written to `/recommended_position`; (ii) otherwise, consume a **history-derived** per-item map for the active `user_id`, obtained by filtering the persistent database by `user_id` and joining **Users**$\rightarrow$**Records**$\rightarrow$**Items**, then resolving the final target via a fixed `letter_to_coord` map. When `user_id` changes, the cache is flushed; flags are reset after each cycle.

Alignment, placement, and planning. Depth-backed centroids and calibration correspondences yield an SVD rigid transform $(\mathbf{R}, \mathbf{t})$ that maps camera to robot frame, $\mathbf{x}_r = \mathbf{R}\mathbf{x}_c + \mathbf{t}$. We synthesize pre-grasp, grasp, and place poses using a table-normal approach with small vertical clearance. For side-by-side placements, the translator chooses a side, estimates a safe lateral offset from object widths and depth, forms the target $\mathbf{Q}$, and publishes it to `/recommended_position`. MoveIt computes collision-checked joint trajectories (limits and velocities enforced) and executes *home* $\rightarrow$ *pre-grasp* $\rightarrow$ *grasp* $\rightarrow$ *lift* $\rightarrow$ *place* $\rightarrow$ *retreat*. Transient occlusion or planning failure triggers bounded retries; outcomes (task, target, status, latency, active `user_id`) are logged to **Records**. Soft workspace bounds, approach-height margins, and gripper-force limits are enforced. On completion or timeout, `/recommended_position` is cleared and gating flags are reset.

4 Experimental Results

4.1 Recognition Accuracy

We first evaluate the recognition module of our framework in real-world settings. Four everyday categories (*cup*, *pen*, *perfume*, *watch*) are selected, and each category contains three instances that differ by at least one attribute, allowing us to assess CLIP's attribute-detection accuracy. We also study the effect of lighting by comparing bright vs. dim conditions. Under dim lighting, YOLOv5 achieves an overall accuracy of 86.56%, while CLIP attains an average accuracy of 70.9% across the three attributes.

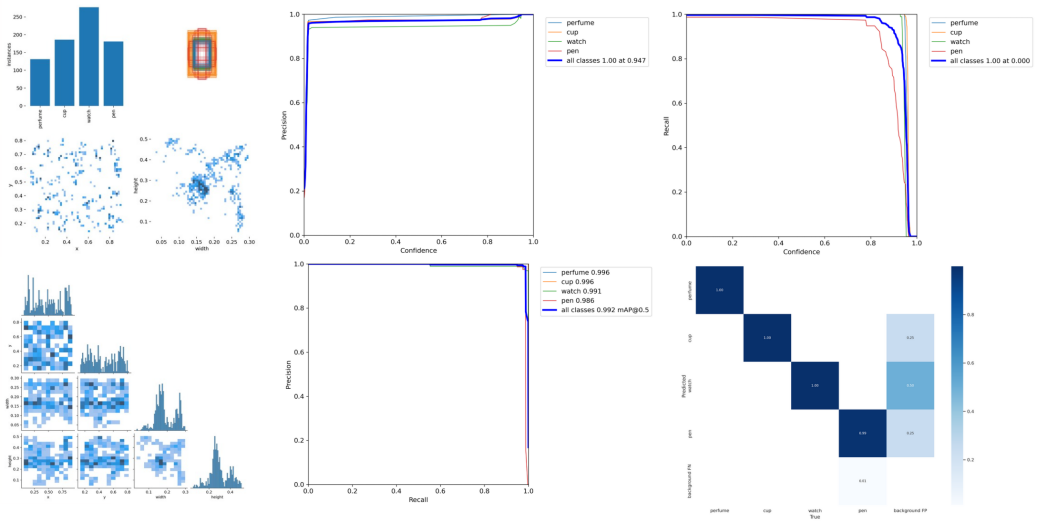


Figure 3: Vision-detection diagnostics for the recognition module. (a) Per-class sample counts under *bright* vs. *dim* lighting; (b) precision and recall as functions of the YOLOv5 confidence threshold (with NMS applied); (c) class-aggregated precision–recall curve summarizing performance across thresholds; (d) confusion matrix over {*cup*, *pen*, *perfume*, *watch*} on the combined test split.

class Average	cup	pen	perfume	watch
Precision (Bright) 86.56%	79.17%	91.89%	90.00%	85.19%
Recall (Bright) 72.80%	76.00%	91.89%	62.79%	60.53%
Precision (Dim) 88.84%	92.31%	85.71%	88.46%	88.89%
Recall (Dim) 81.41%	97.30%	80.00%	85.19%	63.16%

Table 1: Detection performance (%) under bright and dim lighting. Overall recognition in dim lighting is slightly higher than in bright lighting, likely due to specular highlights on glossy objects (e.g., cups and perfume bottles) under strong illumination, which can reduce detector confidence and localization quality.

	color	shape	material	overall
Accuracy (Bright)	73.63%	87.27%	51.81%	70.90%
Accuracy (Dim)	75.67%	78.37%	34.23%	62.76%

Table 2: Attribute accuracy under different lighting conditions. *Material* is consistently the hardest attribute: finishes with anisotropic/specular reflectance (e.g., brushed metal) and optically ambiguous surfaces (e.g., frosted glass, glossy plastic) confound CLIP labels and degrade RGB–D returns, yielding elevated errors compared to color/shape.

4.2 User Satisfaction and Decision Latency

We evaluate user satisfaction and decision latency across three configurations: **S1** (YOLOv5 + LLM), **S2** (YOLOv5+CLIP + LLM), and **S3** (S2 augmented with per-user history from the persistent database (β); when a novel item O_{new} is detected, the LLM conditions on retrieved per-item placements for the active user). To enable controlled, cross-configuration comparisons, we use the same standardized desk scenes and hold environmental factors fixed. User satisfaction is measured

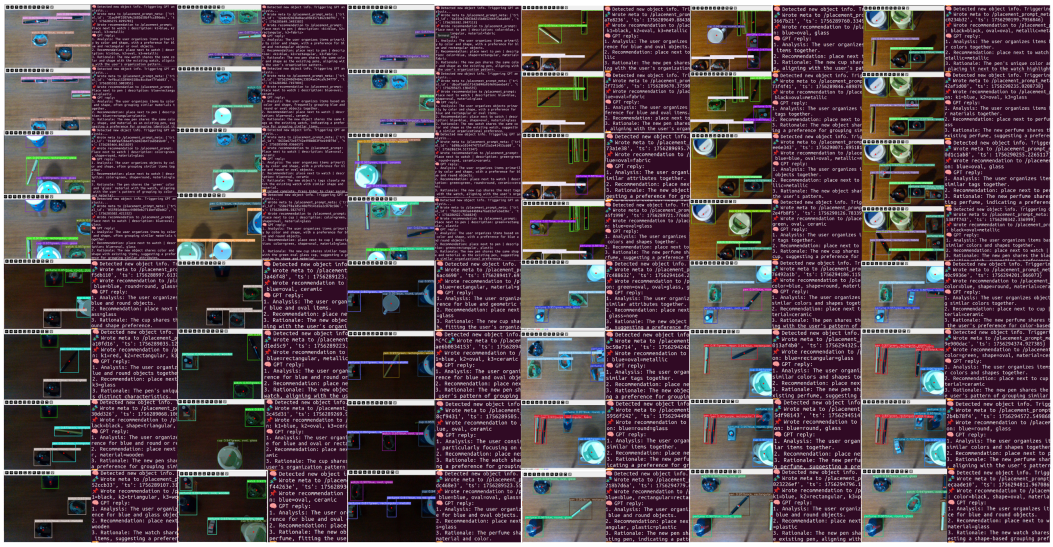


Figure 4: Selected vision-recognition and user-satisfaction results. We evaluate detection accuracy under different lighting conditions and user satisfaction with the recommended placements.See Appendix C for partial detailed results.

on a 5-point Likert scale. We report *decision latency* from trigger to the time a placement recommendation is produced; mechanical execution and dynamic multi-user disambiguation (face/speaker identification) are disabled in this study due to resource constraints prior to the submission timeline.

	User satisfaction (%)	Decision Latency (s)
S1	53.79	2.40
S2	88.27	3.20
S3	91.35	5.90

Table 3: User satisfaction and decision latency. The configuration that includes both attribute cues and the database (*S3*) attains the highest satisfaction (91.35%) but incurs higher latency, primarily because the LLM must process richer historical context.

Adding CLIP-derived attribute cues (S2) improves preference inference without external instructions, yielding 88.27% satisfaction and larger gains under *dim* lighting. The LLM (GPT-4o) leverages co-location patterns among items sharing similar attributes to recommend placements for novel objects. Further improvement from S3 over S2 indicates that incorporating per-user history into the inference context benefits preference learning with only a modest lookup overhead. Residual failures are dominated by misclassification of the *material* attribute, which can mislead the LLM; other attributes partially mitigate this effect.

5 Conclusion

We presented HITF, a hierarchical table-tidying framework that couples a lightweight vision cascade (YOLOv5 $\rightarrow$ CLIP with multi-attribute, 3D-anchored descriptors) with a Users$\rightarrow$Records$\rightarrow$Items memory and LLM-based semantic reasoning, executed by a calibrated manipulator. The central insight is that explicitly injecting object attributes and user history into the inference loop makes preference reasoning robust enough for real, cluttered desktops, while keeping computation and engineering overhead modest. End-to-end evaluations across S1/S2/S3 and varying lighting conditions show that attribute cues improve preference inference and user satisfaction over a perception-only baseline, and that a history-driven path—activated when no new object

O_{new} is observed—recovers habitual placements with bounded latency costs. A voice (ASR) gate that writes lightweight ROS parameters aligns execution timing with human intent, mitigating idle polling and unnecessary motion.

From a deployment perspective, our design emphasizes evaluability and reproducibility: a short-term exchange cache (α) on the ROS Parameter Server, a persistent per-user database (β) with a transparent schema, and deterministic interfaces (e.g., `/placement_prompt` (P_t), `/recommended_position`, and a fixed `letter_to_coord` resolver) make the data flow inspectable and easy to ablate. Safety is enforced through conservative approach heights, soft workspace bounds, gripper-force limits, and MoveIt-based, collision-checked planning; outcomes are logged to Records for audit and analysis. Altogether, HITF provides a practical template for evaluating and deploying preference-aware organization in shared spaces, illustrating how LLMs can be grounded by vision and user history to produce reliable, human-aligned behaviors.

Limitations

- *Attribute recognition (*material*).* In our multi-attribute pipeline (YOLOv5 $\rightarrow$ CLIP with 3D anchoring), the *material* attribute remains the weakest cue. Subtle, illumination-dependent reflectance/texture patterns and failures on transparent/reflective surfaces degrade accuracy, which in turn causes a small fraction of recommendations to diverge from user preferences.
- *Command translation & execution (grasp generality).* The current target acquisition and pose synthesis adopt a table-normal, top-down approach with small vertical clearance and a fixed wrist orientation. This design enables reliable side-by-side placements but effectively restricts the arm to vertical grasps; the policy does not adapt the grasp pose to object geometry or contact affordances. Consequently, tall, elongated, or irregular objects can lead to grasp failures despite a correct target location.
- *Identity and privacy.* Disambiguating the active `user_id` relies on RGB facial capture. Although facial data are processed locally (on-device) and not transmitted beyond the system boundary, certain application and compliance contexts may still judge this insufficient for handling identifiable biometrics (e.g., data minimization, purpose limitation, revocable consent, retention controls). Opting out of face-based identification degrades the history-derived path that queries the persistent database (β) for per-user placements.

Future work. We will strengthen *material* recognition with illumination- and sensor-robust cues (e.g., learned reflectance priors, polarization/NIR augmentation, and tactile/force feedback during contact), coupled with self-training on in-the-wild scenes to reduce vocabulary brittleness. For manipulation, we plan to replace the fixed top-down policy with shape-aware grasp synthesis (6D pose or category-level pose estimation, diffusion-based grasp proposals with collision filtering, and regrasp strategies) so that wrist orientation and approach vectors adapt to object geometry and local clutter. For identity and privacy, we will support privacy-preserving identification (on-device face embeddings with template protection and retention controls), face-free identity options (e.g., voluntary tokens or an ASR-gated passphrase), and a consent-aware fallback that maintains functionality without storing biometric data while still enabling limited, local history.

Appendix

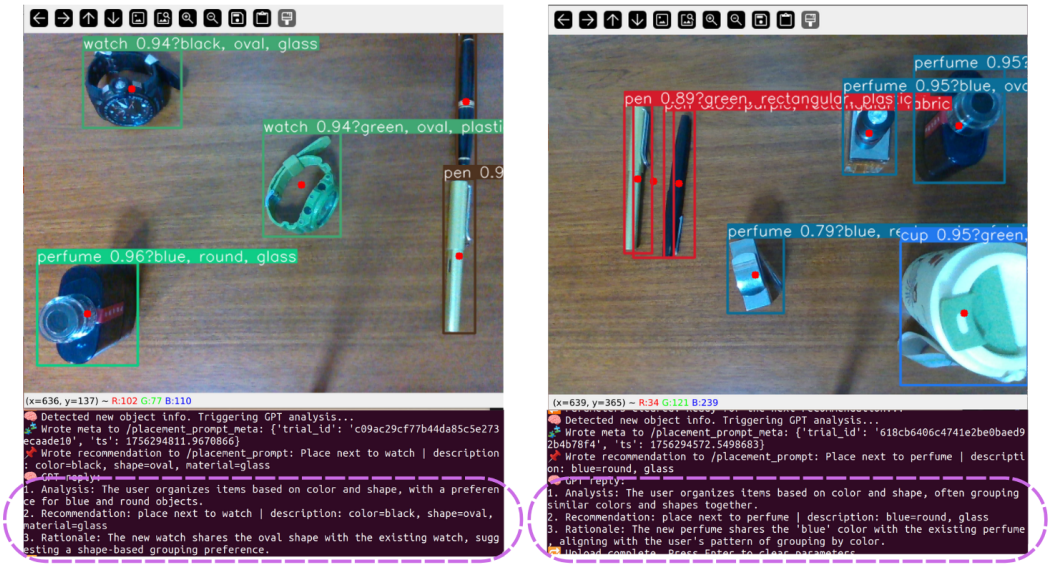
A Demo Video

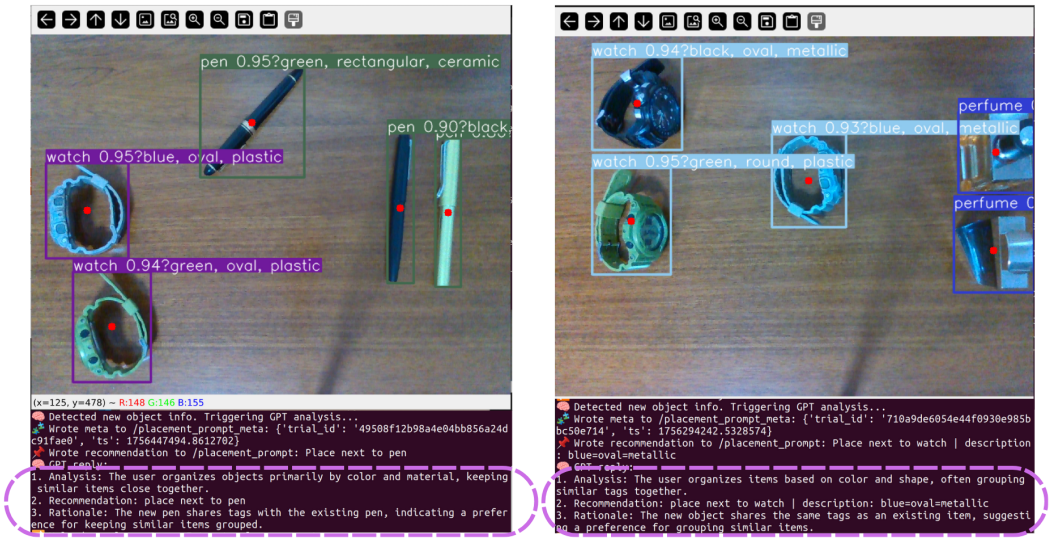
An anonymized 2-minute demo is available at https://youtu.be/k3Rpqisbwy8.

B LLM Prompt

```
You are an assistant that analyzes users' desktop organization habits.

Step 1: Based on the following history, summarize the user's organization patterns:
{history_summary}

Step 2: These are the items currently on the desk:
{current_objects_summary}

A newly detected object:
- name: {new_object['name']}
- tags: {new_object['tags']}

Please answer with EXACTLY the following three lines:

1. Analysis: <summarize the user's organization logic>
2. Recommendation: place next to <OBJECT_NAME>
3. Rationale: <one sentence explaining the reason>
```

C User Satisfaction (Partial Detailed Results)




pen 0.95?green, rectangular, ceramic
watch 0.95?blue, oval, plastic
watch 0.94?green, oval, plastic
(x=125, y=478) ~ R:148 G:146 B:155
Detected new object info. Triggering GPT analysis...
Wrote meta to /placement_prompt_meta: {'trial_id': '49508f12b98a4e04bb856a24dc91fae0', 'ts': 1756447494.8612702}
Wrote recommendation to /placement_prompt: Place next to pen
GPT reply:
1. Analysis: The user organizes objects primarily by color and material, keeping similar items close together.
2. Recommendation: place next to pen
3. Rationale: The new pen shares tags with the existing pen, indicating a preference for keeping similar items grouped.
watch 0.94?black, oval, metallic
watch 0.93?blue, oval,
watch 0.95?green, round, plastic
perfume
metallic
perfume
Detected new object info. Triggering GPT analysis...
Wrote meta to /placement_prompt_meta: {'trial_id': '710a9de6054e44f0930e985bbc50e714', 'ts': 1756294242.5328574}
Wrote recommendation to /placement_prompt: Place next to watch | description: blue=oval=metallic
GPT reply:
1. Analysis: The user organizes items based on color and shape, often grouping similar tags together.
2. Recommendation: place next to watch | description: blue=oval=metallic
3. Rationale: The new object shares the same tags as an existing item, suggesting a preference for grouping similar items.
pen 0.91?yellow, rectangular, fabric
watch 0.89?black, oval, metallic
perfume 0.93?blue, oval,
(x=638, y=366) ~ R:62 G:28 B:18
Detected new object info. Triggering GPT analysis...
Wrote meta to /placement_prompt_meta: {'trial_id': 'ed35200363f54bca8c6205fac0234b32', 'ts': 1756290199.7956846}
Wrote recommendation to /placement_prompt: Place next to watch | description: black=black, oval=oval, metallic=metallic
GPT reply:
1. Analysis: The user organizes items by color and shape, often grouping similar colors together.
2. Recommendation: place next to watch | description: black=black, oval=oval, metallic=metallic
3. Rationale: The pen's unique color and shape contrast with existing items, so placing it next to the watch highlights its distinctiveness.
pen 0.60?green
perfume 0.78
perfume 0.85?blue,
(x=571, y=444) ~ R:62 G:47 B:21
Detected new object info. Triggering GPT analysis...
Wrote meta to /placement_prompt_meta: {'trial_id': '29154779f25a40fba2cd7cea573f4f41', 'ts': 1756289846.6898704}
Wrote recommendation to /placement_prompt: Place next to watch | description: black=oval=metallic
GPT reply:
1. Analysis: The user organizes items by color and shape, often grouping similar tags together.
2. Recommendation: place next to watch | description: black=oval=metallic
3. Rationale: The new object shares the same tags as the existing watch, suggesting a preference for grouping similar items.
watch 0.90?blue, oval, glass
cup 0.94?b
perfume 0.91?blue, round, glass
perfume 0.81?green, rectangular, metallic
(x=638, y=396) ~ R:14 G:7 B:1
Detected new object info. Triggering GPT analysis...
Wrote meta to /placement_prompt_meta: {'trial_id': '6cb15b00830b4b08972fed0a2dc468e3', 'ts': 1756289523.5548444}
Wrote recommendation to /placement_prompt: Place next to watch | description: blue=blue, oval=oval, glass=glass
GPT reply:
1. Analysis: The user organizes items based on color and shape, with a preference for blue and oval objects.
2. Recommendation: place next to watch | description: blue=blue, oval=oval, glass=glass
3. Rationale: The perfume shares the most tags with the watch, particularly the material and color.
watch 0.92?b
perfume 0.86?green, oval, wooden
perfume 0.93?blue, rectangular, fabric
cup 0.91?blue, ov
(x=219, y=318) ~ R:62 G:223 B:1
Detected new object info. Triggering GPT analysis...
Wrote meta to /placement_prompt_meta: {'trial_id': '9aae7ebef5004d14978f9d660f44263e', 'ts': 1756289305.6795146}
Wrote recommendation to /placement_prompt: Place next to perfume | description: blue=oval, ceramic
GPT reply:
1. Analysis: The user organizes items primarily by color and shape, with a preference for blue and oval or round objects.
2. Recommendation: place next to perfume | description: blue=oval, ceramic
3. Rationale: The new object shares the same color and material as the existing perfume, fitting the user's pattern.

References

[1] R. Rasch, D. Sprute, A. Pörtner, S. Battermann, and M. König. Tidy up my room: Multi-agent cooperation for service tasks in smart environments. *Journal of Ambient Intelligence and Smart Environments*, 11(3):261–275, 2019. doi:10.3233/AIS-190524.

[2] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard. Robot, organize my shelves! tidying up objects by predicting user preferences. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[3] I. Kapelyukh and E. Johns. My house, my rules: Learning tidying preferences with graph neural networks. In *Conference on Robot Learning (CoRL)*, 2022.

[4] K. Ramachandruni and S. Chernova. Personalized robotic object rearrangement from scene context. *arXiv preprint arXiv:2505.11108*, 2025. doi:10.48550/arxiv.2505.11108.

[5] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023. doi:10.1007/s10514-023-10139-z.

[6] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning (CoRL)*, 2022.

[7] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020. doi:10.48550/arXiv.2011.01975.

[8] Z. Yan, N. Crombez, J. Buisson, Y. Ruichck, T. Krajnik, and L. Sun. A quantifiable stratification strategy for tidy-up in service robotics. In *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, 2021.

[9] Y. Kant, A. Ramachandran, S. Yenamandra, I. Gilitschenski, D. Batra, A. Szot, and H. Agrawal. Housekeep: Tidying virtual households using commonsense reasoning. *arXiv preprint arXiv:2205.10712*, 2022. doi:10.48550/arXiv.2205.10712.

[10] A. Taniguchi, S. Isobe, L. El Hafi, Y. Hagiwara, and T. Taniguchi. Autonomous planning based on spatial concepts to tidy up home environments with service robots. *arXiv preprint arXiv:2002.03671*, 2021. doi:10.48550/arxiv.2002.03671.

[11] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. doi:10.1109/TPAMI.2021.3087709.

[12] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE Transactions on Image Processing*, 30:2587–2598, 2021. doi:10.1109/TIP.2020.3048632.

[13] B. Irfan, M. G. Ortiz, N. Lyubova, and T. Belpaeme. Multi-modal open world user identification. *ACM Transactions on Human-Robot Interaction*, 11(1):1–23, 2021. doi:10.1145/3477963.

[14] H. Zhu, R. Miyoshi, and Y. Okafuji. Whom to respond to? a transformer-based model for multi-party social robot interaction. *arXiv preprint arXiv:2507.10960*, 2025. doi:10.48550/arxiv.2507.10960.

[15] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Levine, I. Mordatch, A. Rai, P. Sermanet, J. Singh, J. Tan, A. Tung, T. Xiao, P. Xu, M. Yan, and A. Zeng. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. doi:10.48550/arXiv.2204.01691.

[16] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *arXiv preprint arXiv:2307.06135*, 2023. doi:10.48550/arxiv.2307.06135.

[17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. A. Gomez, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, E. Jang, N. Joshi, K. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, I. Mordatch, C. Parada, P. Pastor, A. Rai, D. Sadigh, M. S. Ryoo, P. Sermanet, J. Singh, J. Tan, J. Tenenbaum, A. Tung, V. Vanhoucke, V. Vasudevan, Q. Vuong, S. Welker, T. Xiao, P. Xu, Z. Xu, M. Yan, S. Yenamandra, A. Zeng, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. doi:10.48550/arXiv.2307.15818.

[18] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, J. Dean, P. Abbeel, C. Finn, D. Pathak, and P. Florence. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. doi:10.48550/arXiv.2303.03378.