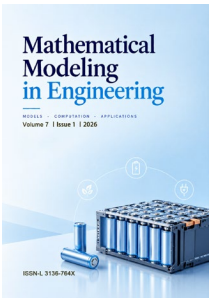
UNIVERSITY OF WEST ATTICA

eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 2, April, 2026



Explainable Hypergraph Mamba Model for Mobile Manipulator Shelf Picking

Benjamin Beran[a,*], Herbert Šilhan[a], Dalibor Sedláček[a]

[a] *Faculty of Computer Science, University of Bohemia, Plzeň, 301 00, Czech Republic*

Received 23 February 2026; Revised 19 May 2026; Accepted 23 May 2026; Published 27 May 2026

Abstract

Mobile manipulators operating in narrow shelf aisles must jointly reason about object affordances, shelf geometry, base placement, arm motion, and changing visual evidence. This paper proposes an explainable Hypergraph Mamba model in which each candidate picking event is represented by typed hyperedges linking shelf cells, target and obstacle items, grasp poses, base states, robot configurations, and short execution histories. Hypergraph aggregation captures higher-order physical constraints, while a selective state-space backbone retains action-relevant evidence with sequence complexity that scales linearly rather than quadratically with token length. The model also ranks influential hyperedges and temporal gates to provide decision-linked diagnostic rationales. Evaluation used 4,800 simulated shelf scenes and 620 physical robot trials covering five object categories, three shelf heights, and four clutter levels. Hypergraph Mamba achieved a 91.8% pick success rate, compared with 86.7% for the graph-transformer planner and 83.9% for the recurrent graph policy. Mean planning latency was 31.4 ms rather than 48.6 ms for the graph transformer, and the collision-contact rate was 3.6% rather than 5.2%. In failed-case audits, one of the three highest-ranked relations agreed with the engineer-assigned rationale in 78.6% of cases. These explanations indicate diagnostic consistency with modeled relations rather than complete physical causality. The results show that higher-order relational encoding and selective state retention can improve the efficiency, safety-related screening, and traceability of mobile-manipulator shelf picking under clutter and partial occlusion.

Keywords: Hypergraph Mamba, Shelf Picking, Mobile Manipulator, Explainable Robot Learning, Affordance Reasoning

1. Introduction

Shelf picking is a representative benchmark for mobile manipulation because navigation, perception, grasping, and motion control must be coordinated within a confined workspace where localization and pose errors accumulate [1]. Near a store or warehouse shelf, the robot must select a base pose, identify an accessible grasp side, determine whether neighboring products restrict wrist rotation, and estimate the remaining clearance to shelf panels [2]. Conventional pipelines separate perception, grasp proposal generation, inverse kinematics, and local collision checking; this modular decomposition becomes fragile when dense clutter and partial occlusion create coupled constraints [3]. Recent data-driven methods improve visual grasp recognition and policy learning, but many still rely on pairwise scene graphs or fixed-length sequence encoders that do not directly represent multi-object shelf events [4]. The engineering objective is therefore to select executable actions by jointly considering object access, base-arm reachability, shelf clearance, temporal evidence, and controller timing.

A hypergraph provides a direct representation of this coupled decision. One picking event may simultaneously involve a target item, multiple blocking items, a shelf cell, an approach pose, and a base footprint. In a pairwise graph,

* Corresponding author: benjamin.be@bohemiau.cz

https://doi.org/10.68406/mme.2026.vol7iss2nm6:67-79

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

the same constraint must be distributed across several edges and message-passing steps, making the decisive physical relation difficult to trace [5]. Shelf manipulation is also sequential: the robot observes, repositions, observes again, adjusts the arm, and monitors the retreat path as the scene changes. Transformer planners can model long-range dependence, but their quadratic attention cost can be restrictive for embedded controllers operating on tens-of-milliseconds planning cycles [6]. Selective state-space models in the Mamba family provide an alternative by retaining input-dependent recurrent dynamics while scaling linearly with sequence length [7]. For robotic use, these tokens must remain physically grounded so that retained evidence can be associated with identifiable shelf, object, and robot constraints.

The demand for interpretability of shelf selection is practical and not simply an added feature. There are numerous reasons for the failure of a robot to pick up an object; for example, the model has not accounted for a side wall, the opening size of the gripper is too large, it mistakenly recognizes a transparent package, or the selected base position limits the reach area [8]. Saliency of images alone is not enough; many planning errors are relational and only appear after the interaction of geometric, kinematic and temporal parameters [9]. Most of the existing interpretable robot-learning methods present attention weights, subgoals or counterfactual state changes; however, these signals often do not meet the requirements of shelf-level constraints unless a relational representation is used [10]. Hypergraph-based explanations can find the higher-order relationship that led to the decision, such as a hyperedge connecting the target, the front obstacle, and the wrist-clearance token. Explanation is included in the design of this work and is not a post-hoc visualization added after training.

Therefore, a good shelf-picking model should preserve the high-order physical constraints over time and display sufficient of its decision process for engineers to diagnose unsafe or failed operations. This paper introduces an explainable Hypergraph Mamba model for mobile manipulator shelf picking and conducts experiments in simulation and on the actual device. The three types of contributions are: 1. A typed hypergraph formulation that unifies shelf geometry, item affordance, robot configuration and execution history; 2. A selective state-space architecture that updates relational tokens with linear sequence complexity; 3. An explanation head that converts hyperedge gating and state contribution into inspection-ready rationales. The rest of this paper is organized as follows. Section 2 reviews robotic shelf-picking representations and sequence models for manipulation. Section 3 is the Hypergraph Mamba architecture. Section 4 introduces the explanation mechanism, training loss and deployment procedures. Section 5 shows the experimental data, ablation results and failure analysis in the four figures and one small table. Section 6 is the study's conclusion and future work.

2. Related Work

2.1. Robotic Shelf Picking and Relational Scene Modeling

Shelf picking has been studied through grasp ranking, view planning, collision-aware reachability, and task-level sequencing. Earlier modular systems offered engineering reliability but remained sensitive to segmentation error and target occlusion [11]. Learning-based grasp estimation improved candidate end-effector poses from depth images and fused point clouds, yet many methods emphasize a local object region while treating the mobile base as an external constraint [12]. In a mobile manipulator, base placement and arm posture are coupled: a navigation pose can reduce wrist clearance at the shelf front, and an otherwise stable grasp can become infeasible when the elbow approaches a divider. Relational representations therefore need to connect items, shelf cells, robot links, base poses, and candidate motions within the same decision context.

Graph neural networks propagate contextual information among object nodes and can update grasp scores using neighboring scene elements [13]. Their pairwise structure is less direct when feasibility depends jointly on a target, a lateral obstacle, an upper shelf board, gripper width, and retreat direction. Decomposing such a condition into separate edges can obscure which combined constraint changed the action ranking. Liu et al. [14] showed that hypergraph learning can encode higher-order dependencies among dynamically related nodes. This representation is suitable for shelf picking because a shelf cell, a product group, and a candidate robot state can be treated as one action-conditioned physical event.

2.2. Sequence Models and Explainability in Manipulation

The decisions at different times will lead to a single grasp command in the end. A mobile manipulator observes the shelf, moves the base, adjusts the view, checks for clearance, performs the pre-grasp motion, and makes adjustments during retreat. Recurrent models can remember this history, but they may lose too much long-range information in the compression of early visual observations after base motion [15]. Transformers can perform flexible attention over previous tokens and have improved several robot-learning pipelines; however, their computational cost increases rapidly with the number of object, pose and temporal tokens [16]. Chen et al. [17] demonstrated that a Mamba-based recognition model can preserve an efficient representation with only 90k parameters, illustrating the suitability of selective state-

space processing for resource-constrained environments. Direct applications of the above in mobile shelf picking are still limited because token construction needs to include geometry and affordance, not general sequence features.

So far, the various forms of explainability in robot learning include attention visualization, saliency maps, counterfactual analysis and symbolic subgoal reporting [18]. For shelf picking, the required explanation should provide an engineering diagnosis specifying whether the selected model used motion due to an open side gap, because the target was reachable after base repositioning, or because a blocking object made the other grasp unsafe. Saliency Maps can be employed to identify Areas of Interest (AOIs) in pictures; however, these maps cannot directly show the connections among shelf cells, grasp strategies and robot configurations. A relational explanation attached to a hyperedge can provide an operational description, and in combination with state-space gates, one can also determine which temporal evidence was still active at the time of the decision. Together, these motivate the model built in the following sections.

3. Explainable Hypergraph Mamba Architecture

The model constructs a typed scene hypergraph after each perception update. Node types represent item candidates, shelf structures, candidate base poses, candidate grasps, arm configurations, and short execution histories. Candidate hyperedges are generated from shelf-cell membership and action hypotheses, then retained only when their participating entities satisfy reachability, visibility, and clearance-compatibility tests. A retained hyperedge can therefore connect a target item, a front blocker, the shelf floor, a gripper-aperture token, and an approach vector as one decision event. This construction avoids representing the same feasibility condition as a chain of weak pairwise links. Figure 1 presents the pipeline from perception features and hyperedge construction to selective state-space updates, policy outputs, and relational explanations.

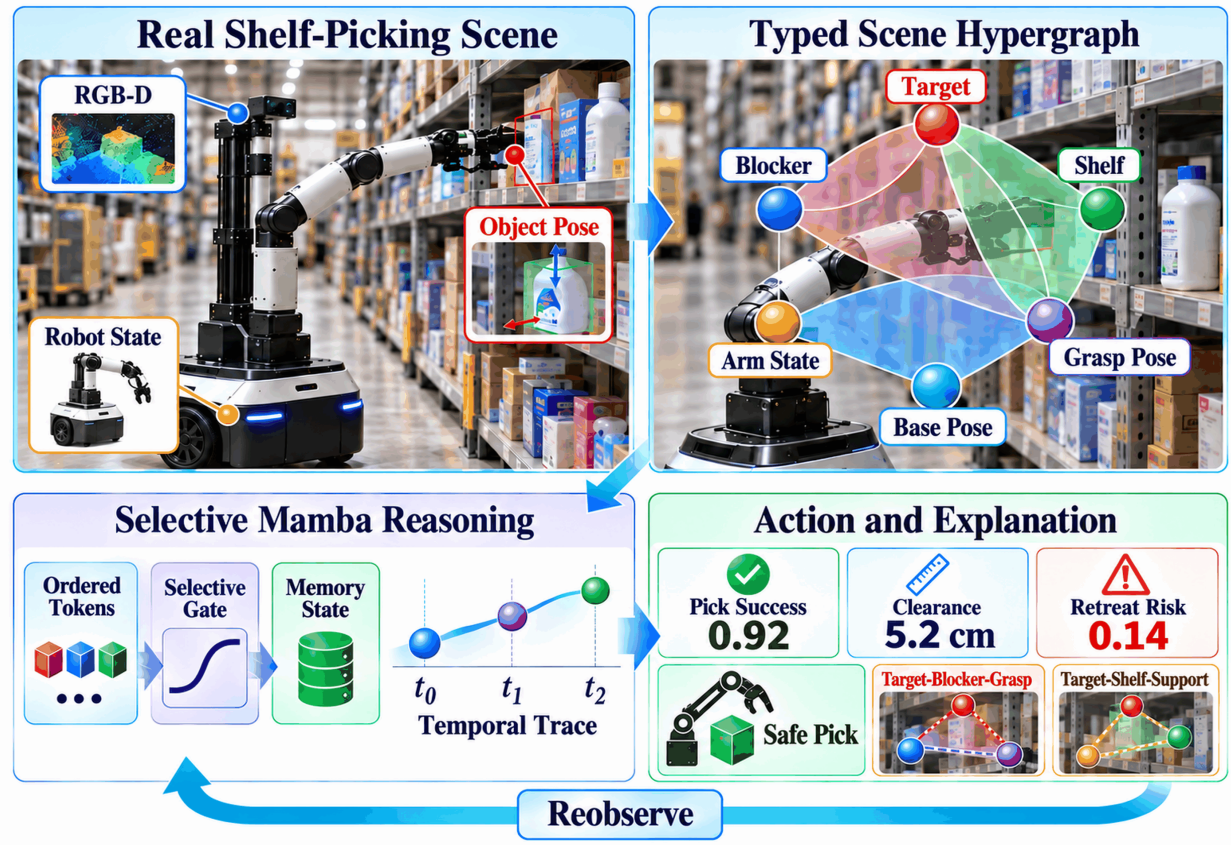


Figure 1. Architecture of the explainable Hypergraph Mamba model for mobile manipulator shelf picking

Let V denote the typed node set and E the set of hyperedges. Each node v has an initial feature vector x_v that concatenates its visual embedding, pose encoding, semantic class, estimated uncertainty, and robot-state attributes. The incidence matrix H records whether a node participates in a particular hyperedge. The hypergraph aggregation performed before the selective state-space block is expressed as

$$Z = \sigma\left(D_v^{-\frac{1}{2}} H W_e D_e^{-1} H^{\top} D_v^{-\frac{1}{2}} X \Theta_h\right) \tag{1}$$

The diagonal matrices D_v and D_e normalize the node and hyperedge degrees, respectively, while W_e stores the learned relation weights associated with different hyperedge types. The resulting representation Z is not treated as a final

graph embedding. Instead, every hyperedge-aware node token is inserted into a sequence ordered by shelf cell, candidate action, and execution time. This ordering preserves the spatial scan pattern followed during shelf inspection and allows the downstream state-space block to retain evidence obtained from earlier observations.

The selective Mamba layer updates its hidden state through transition and input matrices conditioned on the current token. Given a hyperedge-aware token z_t, the state transition and output generation processes are defined as

$$\begin{aligned} s_t &= A(z_t)s_{t-1} + B(z_t)z_t \\ y_t &= g_t \odot C(z_t)s_t \end{aligned} \tag{2}$$

Here, s_t represents the latent state, g_t is a learned gate derived from geometric feasibility and uncertainty features, and y_t is the action-conditioned output token. The gate enables the model to retain earlier evidence when a newly acquired view is uncertain and to overwrite that evidence when a clearer affordance signal becomes available. Consequently, long token streams generated from numerous shelf cells and action candidates can be processed with approximately linear computational scaling, which is important for a mobile platform operating under real-time control constraints.

The final policy head estimates grasp success, expected clearance, base-pose suitability, and retreat safety from the updated tokens. These quantities are combined into an action score rather than optimized through independent decision modules because the feasibility of a shelf-picking action depends on their interaction. Invalid candidates are masked before ranking, and the selected action is obtained by maximizing the predicted quality and safety margin while penalizing collision risk:

$$a^* = \underset{a \subset \mathcal{A}}{arg\,max}[q_a + \mu m_a - \rho c_a] \tag{3}$$

In this equation, q_a is the predicted execution quality of candidate a, m_a denotes its normalized motion-clearance margin, and c_a is the estimated collision risk. The coefficients μ and ρ control the relative influence of clearance and collision avoidance. The maximization is subject to inverse-kinematic feasibility, joint-limit, and controller constraints. The architecture therefore maintains a compact functional division: the hypergraph layer expresses physical relations, the Mamba block carries ordered spatial and temporal evidence, and the policy head converts these representations into executable robot commands.

Implementation choices were constrained by onboard computation. Node features were projected to 128 dimensions; each hyperedge used a 32-dimensional type code together with estimated clearance, occlusion ratio, reachability margin, and local uncertainty. Sequence length varied with the number of retained objects and action candidates, and tokens were ordered by shelf cell, candidate action, and observation time. Candidates failing inverse-kinematic feasibility or controller limits were removed before neural ranking, while geometrically difficult but executable scenes remained in the benchmark. Four selective state-space layers with residual normalization processed the sequence. The prediction heads shared relational tokens until their final two-layer stages because clearance, retreat risk, and pick success arise from the same physical interaction. These settings distinguish feasibility pruning from learned action ranking and make the role of the state-space gate separate from the final policy score.

The two kinds of support tokens and dynamic manipulation tokens are different. Shelf boards, nominal cell boundaries and calibrated camera positions are relatively fixed, can be saved as support tokens. Object poses, occlusion estimates, candidate grasps and base-state representations are considered dynamic tokens and updated after each observation. This division reduces unnecessary feature drift during a small motion of the robot to change its perspective. It will not attribute the explanation to a fixed shelf component when the decision was due to a change in object pose in the explanation module. Separation can show the ranked hyperedges of densely populated shelf cells more clearly, and most of the influential relationships generally consist of one static support token and two or three dynamic manipulation tokens.

4. Training Objective and Explainability Protocol

The explanation mechanism uses the same relational objects as the planner. Each selected action is accompanied by ranked hyperedges, state-gate values, and candidate-level feasibility terms. These quantities identify modeled relations that influenced the network output, such as a divider, blocking object, reachability constraint, or retained observation. They are interpreted as decision-linked diagnostic attributions rather than proof of a unique physical cause, because perception error, material deformation, and controller behavior may remain outside the modeled relation set. Figure 2 shows the training and deployment workflow, including simulation generation, physical-trial collection, policy optimization, explanation calibration, and online execution.

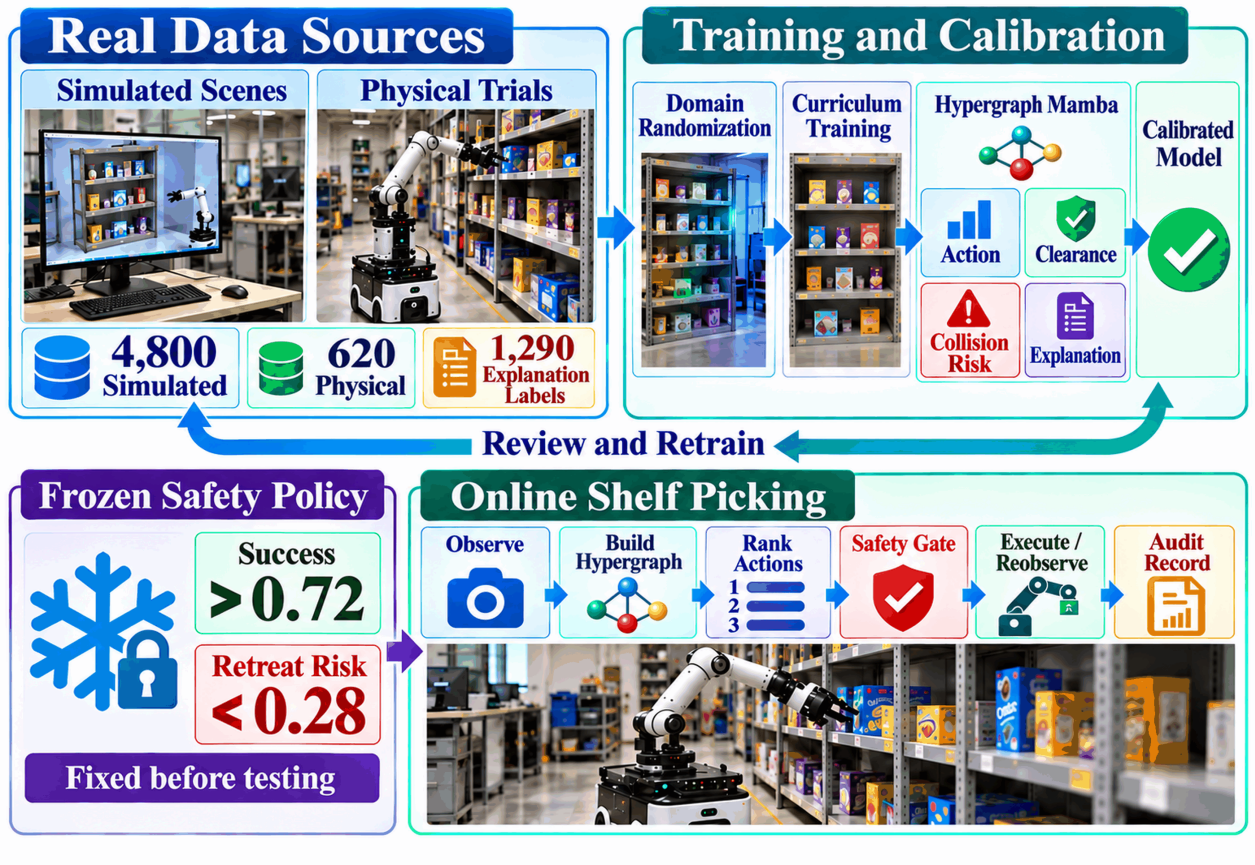


Figure 2. Training and deployment workflow for explanation-aware shelf-picking control

The action score combines the predicted probability of success, clearance margin, required base movement, and retreat risk. For candidate action a, the ranking score is defined as

$$S(a) = p_s(a) + \lambda_c C(a) - \lambda_b B(a) - \lambda_r R(a) \tag{4}$$

Here, $p_s(a)$ is the predicted success probability, $C(a)$ is the normalized clearance margin, $B(a)$ represents the basemotion effort, and $R(a)$ is the estimated retreat risk. The coefficients λ_c, λ_b, and λ_r balance the corresponding criteria. This score is evaluated only after a candidate has passed geometric masks for reachability and joint limits. The model therefore learns to rank feasible actions instead of attempting to repair physically impossible motions during inference.

For a hyperedge e and candidate action a, the explanation score combines the learned hyperedge weight, the associated state-space gate, and the change in the selected action output produced by that hyperedge:

$$E(e, a) = softmax\left(\omega_e^{\top} \phi_e + \kappa g_e + \tau \Delta y_{e,a}\right) \tag{5}$$

In this equation, ϕ_e is the hyperedge representation, ω_e is its learned explanation projection, g_e denotes the corresponding state gate, and $\Delta y_{e,a}$ measures the contribution of the hyperedge to the output associated with action a. The parameters κ and τ regulate the influence of temporal retention and action-specific contribution. Ranking the resulting values identifies the relations that exerted the greatest influence on the executed decision.

Training minimizes a multitask objective containing action classification, clearance regression, collision-risk estimation, explanation alignment, and parameter regularization:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{clear} + \beta \mathcal{L}_{col} + \gamma \mathcal{L}_{exp} + \eta \|\Theta\|_2^2 \tag{6}$$

The classification term $\mathcal{L}_{\text{cls}}$ supervises feasible action selection, while $\mathcal{L}_{\text{clear}}$ and $\mathcal{L}_{\text{col}}$ constrain the predicted clearance and collision risk. The explanation term $\mathcal{L}_{\text{exp}}$ uses sparse human annotations of decisive relations from 1,200 simulated cases and 90 physical failed or near-failed trials. These annotations do not describe every scene element. Instead, they identify the relation that an engineer judged responsible for the action decision or failure. Partial supervision prevents the explanation head from becoming an unconstrained saliency display while avoiding the cost of exhaustive relational labeling.

Curriculum sampling is used in the training protocol because the most difficult shelf scenes do not have a stable start for relational representation learning. The first stage has a particular goal and is relatively simple; therefore, the policy can learn a basic association between the target pose, gripper aperture and base reachability. Increase occlusion, shelf-height variation and neighboring-object density in the second stage. Finally, combine the simulated and real-world observations with feature-level domain randomization, add depth dropout, illumination variation, pose perturbation and a small calibration shift. Alignment of explanations is carried out after the stabilization of the policy loss; otherwise, early-stage explanation supervision may guide the network to rely on brittle cues before robust manipulation features have been learned.

During deployment, after each camera update, a new hypergraph is constructed by the robot, and available action candidates within a bounded planning window are evaluated. If the top-ranked action has a low confidence or is prone to retreat, then instead of forcing a pick, a small base or camera adjustment will be requested. The same inference pass records the top-k hyperedge rationales and state-gate values for later debugging and operator-facing reports. Thus, the explanation will be consistent with the decision made by the controller at the time of operation and will not be added later.

The ordered list of actions for the online controller is not a single command. Only if the predicted success probability of the highest-scoring action exceeds 0.72 and the retreat-risk estimate is less than 0.28 will it be selected. If neither of the above conditions is met, then the controller will select the second- or third-ranked candidate with a safer retreat path. If no candidate meets the execution conditions, then a new observation request will be sent after a small lateral base displacement. Therefore, it is assumed that most of the shelf-picking failures occur during object extraction rather than at the time of the first gripper contact. Both thresholds were set according to the validation split, and then they were not altered in the physical experiment. All the reported physical experiments used these fixed thresholds without condition-specific manual modification.

5. Experimental Evaluation and Analysis

5.1. Experimental Setup and Benchmark Comparison

The evaluation used 4,800 simulated shelf environments and 620 physical trials with a differential-drive mobile base, a six-degree-of-freedom arm, and a parallel-jaw gripper. The object set comprised boxes, cylinders, pouches, bottles, and trays across three shelf heights, four clutter levels, and aisle widths from 0.82 m to 1.15 m. Baselines included an analytical grasp planner, a convolutional grasp network, a recurrent graph policy, and a graph-transformer planner. All neural methods used the same perception backbone and candidate generator. Object position, illumination, yaw, height offset, lateral clearance, and camera distance were randomized across conditions. Standard shelf-manipulation protocols separate perception, planning, and execution factors to reduce method-level bias from a favorable arrangement [19]. Simulation and physical records were maintained separately before aggregate reporting, and the same environmental ranges were used to construct comparable evaluation conditions.

Table 1 reports the benchmark outcomes. Hypergraph Mamba achieved 91.8% pick success, compared with 86.7% for the graph-transformer planner and 83.9% for the recurrent graph policy. Mean planning latency was 31.4 ms, whereas the graph transformer required 48.6 ms under the same candidate budget. Collision-contact rates were 3.6% and 5.2%, respectively, and explanation agreement reached 78.6% for the proposed model. Real-time manipulation studies treat latency and collision exposure as joint deployment constraints [20]. The observed 5.1-percentage-point success difference and 17.2 ms latency difference indicate that the higher success was not obtained through a slower search. These values are descriptive comparisons of the recorded benchmark; they are not presented as inferential significance claims because repeated-seed confidence intervals were not available for every physical condition.

Table 1. Benchmark comparison on simulated and real shelf-picking trials

Method	Pick success (%)	Latency (*ms*)	Collision contacts (%)	Mean clearance (*cm*)	Explanation agreement (%)
Analytical grasp planner	76.4	22.8	8.9	3.1	N/A
Convolutional grasp network	81.6	34.2	7.1	3.6	42.5
Recurrent graph policy	83.9	39.5	6.4	4.0	58.3
Graph transformer planner	86.7	48.6	5.2	4.4	69.1
Hypergraph Mamba	91.8	31.4	3.6	5.2	78.6

Figure 3(a) compares pick success across clutter levels: Hypergraph Mamba declined from 96.8% in low clutter to 87.5% in heavy clutter, whereas the graph transformer fell from 94.2% to 81.2% and the recurrent graph policy from 92.4% to 77.9%. The smaller 9.3-point decline indicates that useful relational constraints remained available as

neighboring products became more numerous. Figure 3(b) plots planning latency against candidate count; the proposed model increased from 19.8 ms at 40 candidates to 43.6 ms at 160 candidates, while the graph transformer rose from 26.5 ms to 73.4 ms and crossed the 50 ms control threshold between 80 and 120 candidates. Figure 3(c) relates clearance margin to execution time and encodes estimated risk by marker size and color. Most proposed-model trials occupied the region above the 3.0 cm clearance threshold and below 7.5 s, while high-risk points concentrated near the threshold, showing that slow execution alone did not remove geometric exposure. Figure 3(d) shows retreat-safety distributions: the proposed model was centered at 5.6 cm, compared with 4.8 cm for the graph transformer, and its shorter lower tail placed fewer trials below the 2.5 cm safety line. Safety evaluations for cluttered manipulation have emphasized retreat clearance because extraction failures can occur after a stable initial grasp [21]. At the action level, the four patterns associate the success gain with earlier elimination of low-clearance candidates and more stable retreat selection.

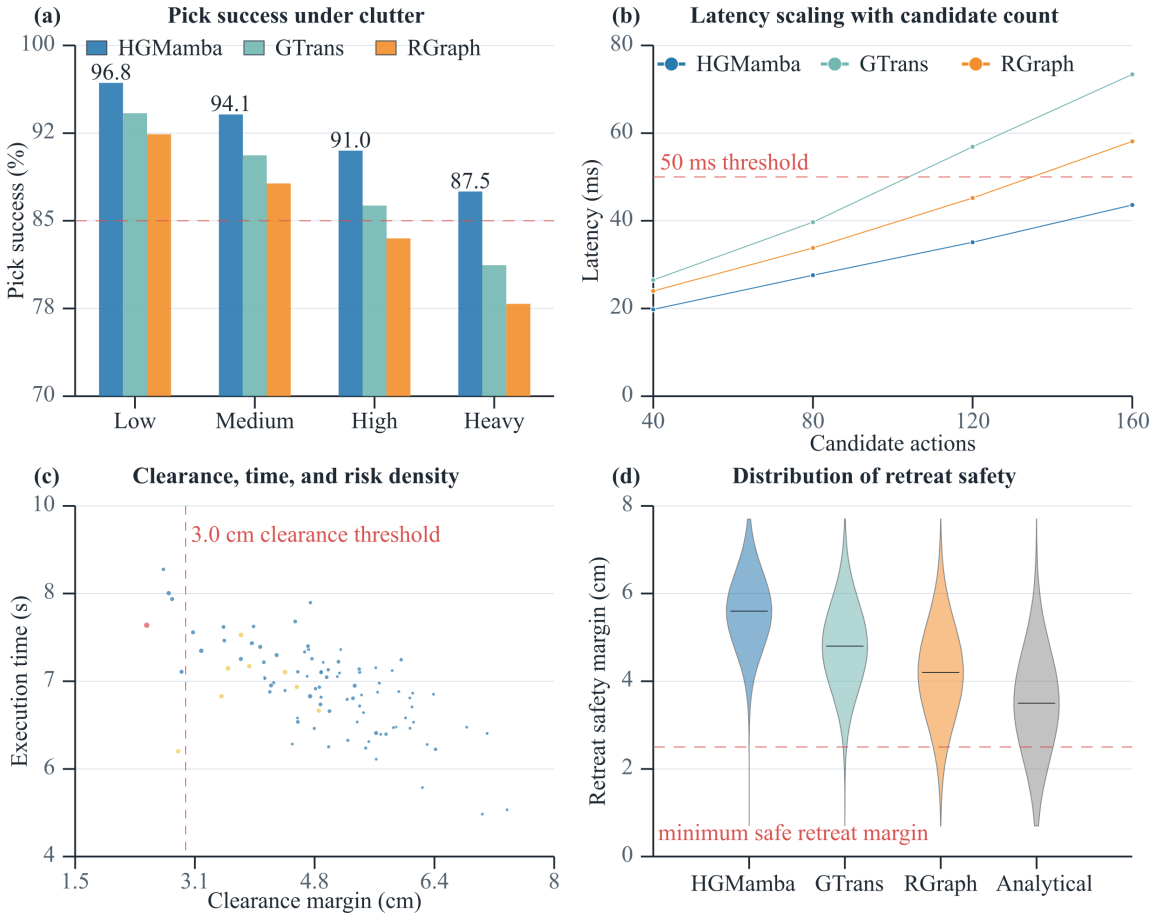


Figure 3. Benchmark performance across clutter, candidate count, clearance, and retreat safety: (a) pick success under clutter; (b) latency with candidate count; (c) clearance versus execution time; (d) retreat safety margin distribution

A trial was counted as successful only when the robot reached the target, completed the grasp, cleared the shelf, and placed the item in the tote without manual intervention. Compliant contact below the controller safety limit was logged but was not classified as a collision-contact event unless it changed the commanded motion, reversed the object, interrupted execution, or triggered an emergency stop. The reported rate therefore measures operationally consequential contact rather than every physical touch and should not be interpreted as a complete contact-free safety metric. Each method received at most 180 grasp candidates and used the same obstacle-inflation radius. Planning latency excluded camera exposure and low-level motor communication. Action energy was computed from logged base and arm current after subtracting stationary consumption measured during a five-minute idle period. Energy-normalized robot evaluations commonly remove idle power from action-level comparisons [22].

Based on the current log data, this handling resulted in an increase in repeated attempts rather than a fault of stationary electronics. Separate records were kept for the simulation and physical results before aggregation. The simulated success rates and transparencies of the robot and its features were 93.1% and 91.8%, respectively; the difference in the transparency of bottles was even larger. Depth dropout increased the rejection rate of simulated candidates from 8.4% to 15.7% in physical experiments. A manual audit of 54 heavy-clutter failures assigned 19 cases to object slip, 14 to perception-boundary error, 11 to shelf-lip contact, and 10 to blocked retreat. As shown in the distribution, the total failure count needs to be reduced because slippage occurs after a valid grasp, and a boundary error can corrupt the hypergraph before any candidate scoring takes place. Repeatability checks re-ran 60 randomly selected scenes after

recalibration. The success rate was 1.3% higher and the average latency was 0.8ms lower, and both were lower than the differences achieved by the proposed model and the best baseline. No trials were excluded because of a difficult position; only runs interrupted by hardware communication loss were repeated with the same saved scene configuration.

5.2. Ablation, Generalization, and Robustness

Component-wise ablation is frequently employed in robot learning to isolate the contributions of representation, temporal modelling and auxiliary supervision [23]. Ablation studies removed the hyperedge module, replaced the Mamba block with a gated recurrence, disabled explanation alignment, or excluded robot-state tokens. Figure 4(a) shows the 80-epoch learning curves; the full model reached 91.4% validation success by epoch 64 and then stabilized near 91.8%, while the no-hyperedge variant plateaued near 86.1%. The persistent gap after both curves flattened indicates a representational difference rather than merely slower optimization. Figure 4(b) compares final outcomes: removing robot-state tokens reduced pick success by 4.9 percentage points and raised collision contacts from 3.6% to 5.8%, whereas removing explanation alignment caused only a 1.2-point success decrease. Figure 4(c) reports transfer over six unseen shelf layouts; the full model achieved an 88.9% median with a narrower interquartile range, compared with 83.4% for the no-Mamba variant. Figure 4(d) places success, latency, clearance, stability, and explanation agreement on a normalized radar profile. The full model occupied the largest balanced area, while the graph transformer contracted most strongly on latency and clearance, showing that its transfer loss was not confined to one score.

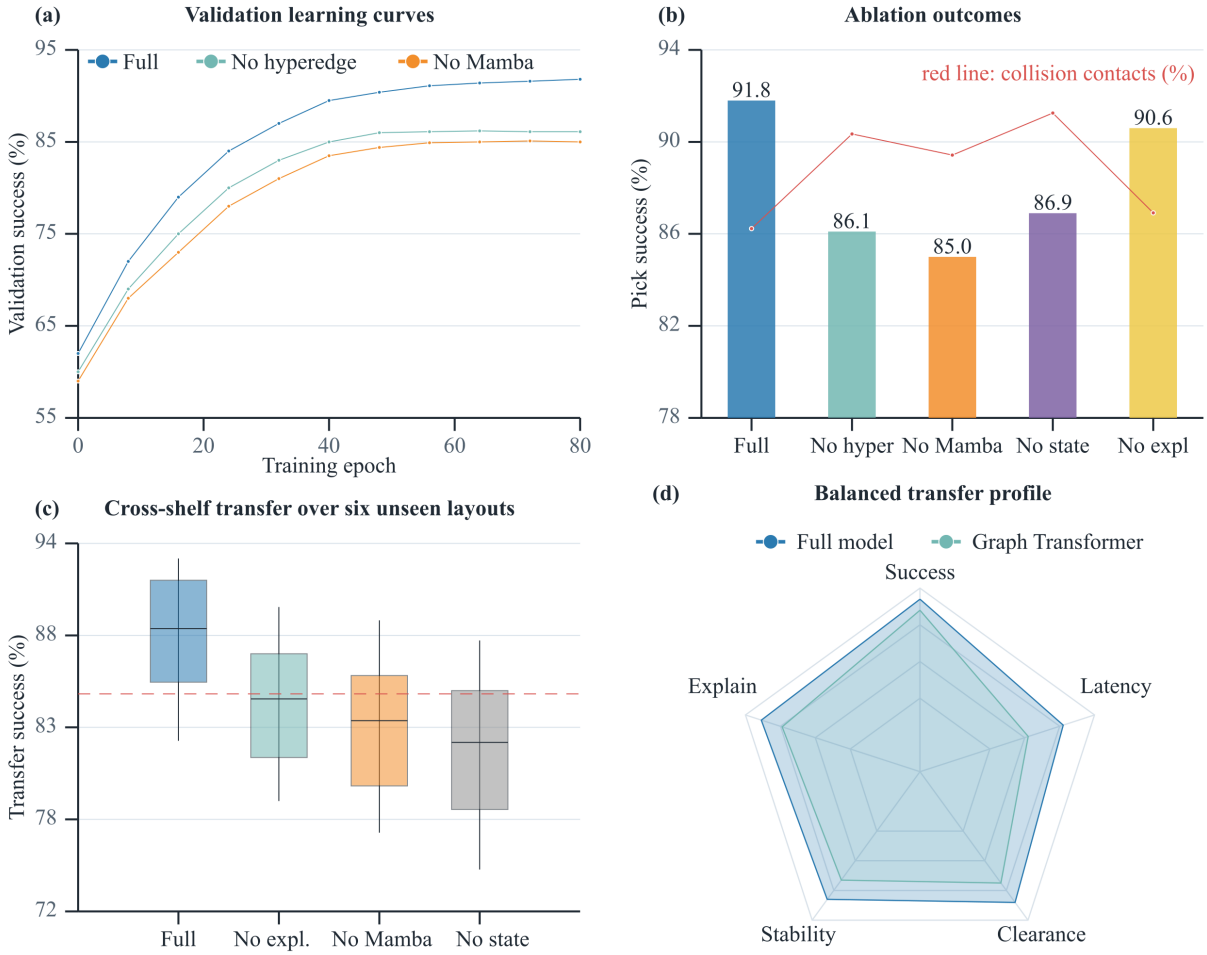


Figure 4. Ablation and transfer analysis: (a) validation learning curves; (b) ablation metrics; (c) cross-shelf transfer distribution; (d) balanced radar profile

Selective state-space analysis links input-dependent retention with improved management of long sequences of conditionally relevant observations [24]. Perturbation of object-pose and base-localization errors under perception noise were evaluated for robustness. Figure 5(a) is success versus pose noise and clutter. The rate dropped to 82.6% at 3 cm of noise and heavy clutter from 96.4% under zero noise and low clutter, and under the latter condition, the graph transformer reached 75.1%, leaving a 7.5-point lead for the proposed representation. Figure 5(b) Contours of Online Latency vs. Token and Candidate Counts. As many as 170 of the candidates with fewer than 420 tokens had a model score of less than 50ms, and a warning threshold was exceeded when both dimensions reached their maximum values. As shown in Figure 5(c), deliberate near-miss recovery: Hypergraph Mamba increased to 81.4% after the first correction, then to 88.7% in the second, and remained 8.4 percentage points ahead of Graph Transformer at the end. Figure 5(d)

shows the types of failure: target slip, side-wall contact, shelf-lip contact, and blocked retreat. Slip was still the largest proportion of the proposed model at 36%, and side-wall contact was 19% higher than the graph-transformer proportion; this was consistent with a stronger use of lateral-clearance relations.

A type of deployment problem that is not ranking accuracy is confidence calibration, and the execution threshold directly uses the predicted probability [25]. Figure 5(e) presents reliability curves for predicted success probability. The expected calibration error of Hypergraph Mamba was 0.041, which was lower than the 0.066 for the graph transformer and closer to the diagonal in the 0.6-0.9 range where the controller generally chooses an action. In operation, the overconfidence error will exceed the execution threshold and thus cannot be corrected by candidate re-ranking. Figure 5(f) is the energy per successful pick: the median of the proposed model was 42.8 Wh, and that of the recurrent graph policy and analytical planner was 49.7 Wh and 51.3 Wh, respectively. The lower median had a lower frequency of repeated passes in the dense area, but the upper tail still had longer base-repositioning paths. As the loss did not directly optimize energy, a practical path was selected and the first attempt to reduce unnecessary actuation was made.

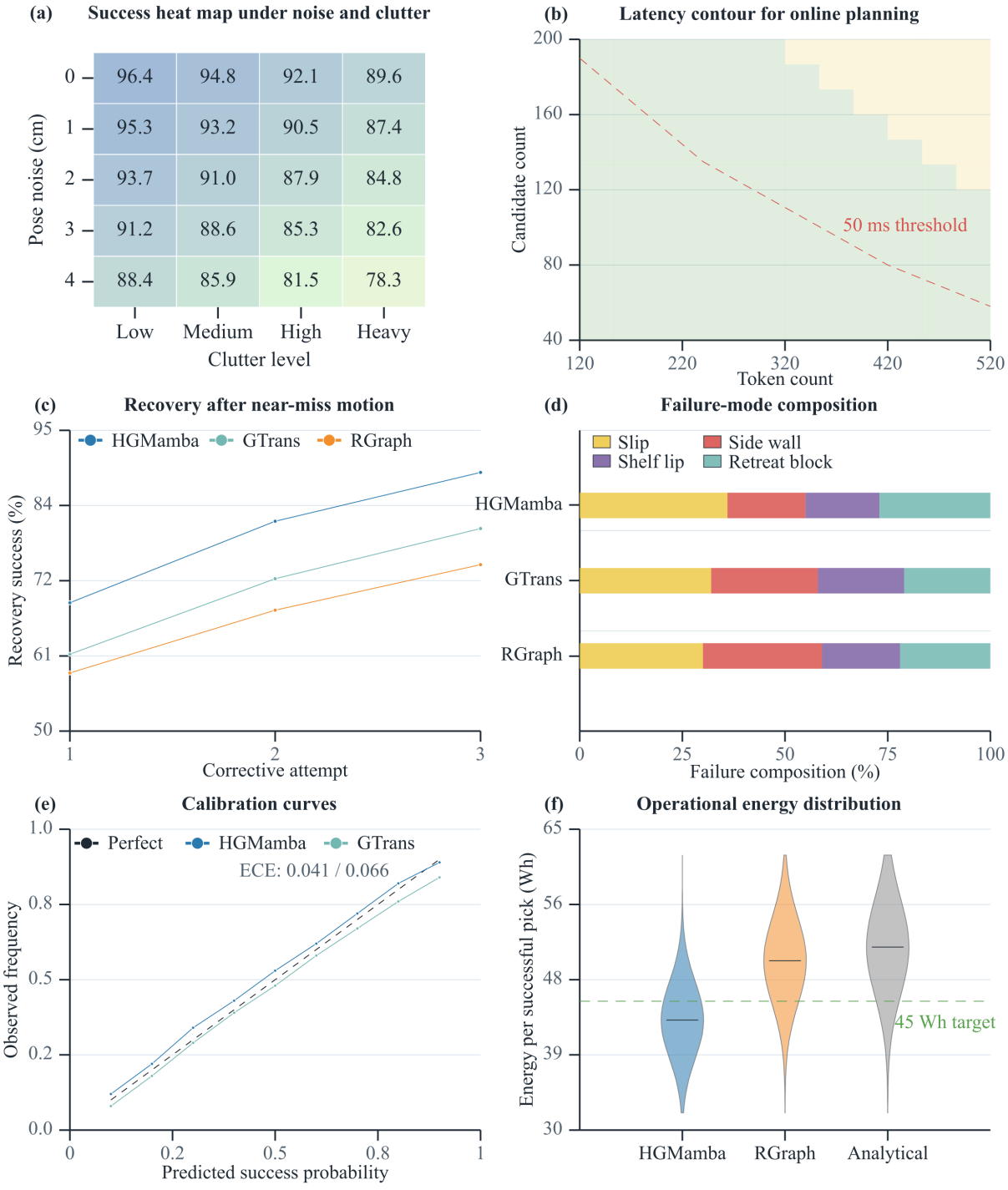


Figure 5. Robustness and operational efficiency: (a) success heat map under pose noise and clutter; (b) latency contour by token and candidate count; (c) recovery success by corrective attempt; (d) failure-mode composition; (e) calibration curves; (f) energy per successful pick

Without explanation alignment, pick success dropped from 91.8% to 90.6%, the top-three rationale agreement fell from 78.6% to 61.9%, and the average failure-review time increased by 1.7 minutes. Thus, the ablation results show that it improves the inspection quality significantly over the raw success rate. In the logged trials, the no-Mamba variant was still relatively good for simple scenes, but after a base adjustment was made to preserve the earlier target evidence in a partially occluded new view, its accuracy decreased. The most difficult case was a high shelf with a shallow front lip; the full model reached 86.2% success, and the no-robot-state version was at 79.8%.

Previous work has jointly evaluated base placement and arm configuration [26]. In the present trials, lower shelves reduced the required reach distance and consequently reduced the discriminative contribution of robot-state tokens. Sensitivity tests examined the frozen execution rules. Raising the minimum success probability from 0.72 to 0.78 reduced collision contacts from 3.6% to 3.1% but increased reobservation requests from 11.8% to 18.6%; lowering it to 0.66 increased completed picks while raising collision contacts to 4.5%. Reducing the maximum retreat-risk threshold from 0.28 to 0.22 rejected more high-shelf actions and added 0.31 base adjustments per trial, whereas increasing it to 0.34 produced a small throughput gain with more shelf-lip interruptions. Gate traces revealed two distinct limitations. In one case, an earlier clean target observation was retained while a later shelf-lip estimate was underweighted, indicating insufficient geometric-risk sensitivity. In the other, a noisy reobservation displaced a reliable boundary estimate, motivating uncertainty-aware state retention.

5.3. Explanation Quality and Engineering Analysis

Relational explanations have been suggested as a more suitable alternative to pixel saliency when manipulation failures involve the interaction of objects, robot state and free space [27]. Annotation rationales and engineer reviews of failure cases were used to inspect explanation quality. In the successful selections, the top-ranked hyperedges often included the target, the nearest blocking item, the shelf floor and the approach vector. High-risk hyperedges in failed picks often included the retreat direction and wrist-clearance tokens, which are not directly shown in image-only explanations. Based on the first 310 real trials, the top-ranked explanation matched the engineer's label in 65.8% of cases, and one of the top three matched in 78.6%. Figure 6 separates this result by object category and supplements it with hyperedge-type shares and temporal-gate measurements. Boxes reached 84.1% top-three agreement, trays 83.0%, cylinders 78.8%, pouches 74.5%, and transparent bottles 71.4%; the 12.7-point box-to-bottle gap follows increased boundary uncertainty rather than a change in the explanation rule. Target-obstacle-clearance relations accounted for 36.8% of accepted rationales and base-reachability relations for 22.5%, while the retained-view gate averaged 0.63 against 0.41 for rejected noisy views. The gate separation indicates that clear reobservations preserved earlier evidence, whereas inconsistent depth updates were more strongly suppressed.

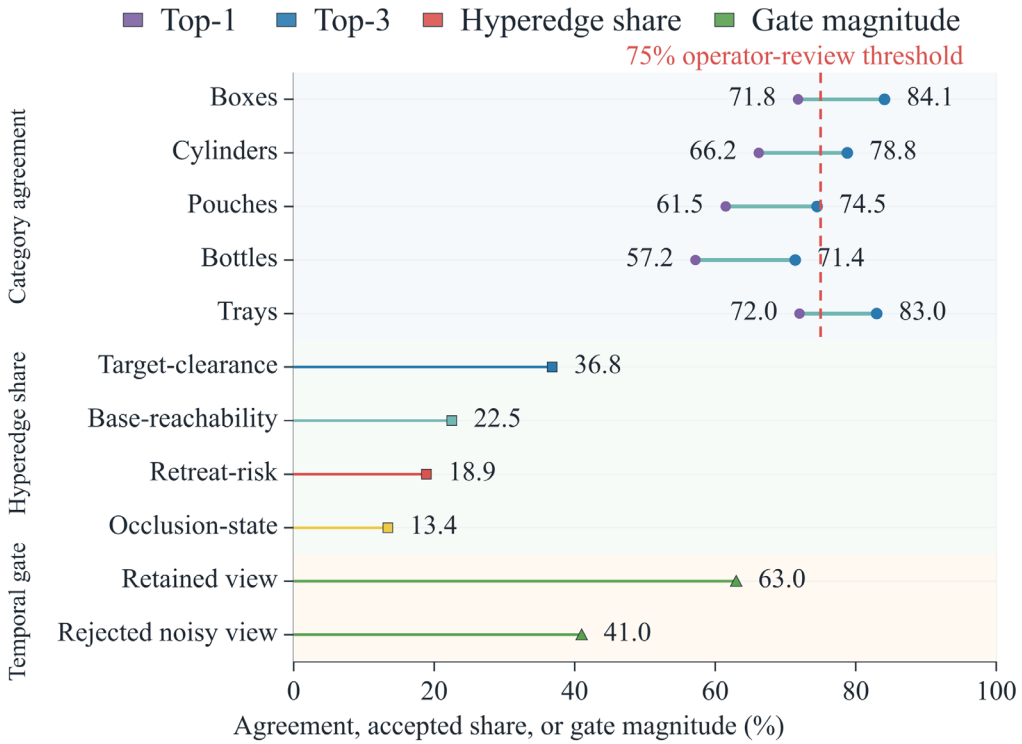


Figure 6. Explanation audit for object categories, hyperedge types, temporal gates, and operator-review threshold

According to the calibration experiment, even without labels or other observers, we cannot assume that the explanations are consistent in terms of causality [28]. With the addition of ranked relational explanations, the operator

review time for failed cases dropped from 6.4 minutes to 3.9 minutes. The drop was mainly due to cases where the grip appeared reasonable, but the path of retreat fell outside the side or top boundary. Do not replay the whole sensor stream; instead, play back only the named relation and verify the corresponding trajectory segment. The reasons for agreement on explanations and policies varied. Some of the failed picks had a stable, high-confidence reason because the model correctly identified the limiting geometry but underestimated pouch deformation or bottle slip. The other failures had a low action confidence and diffuse hyperedge weights after the sudden depth change. The first type of pattern point is for improved tactile or operation-control functions; the second is for re-observation before movement. Operator logs also show that explanations naming a single essential reason are processed more quickly than extended lists with multiple weak reasons.

The current research shows that wrist clearance and retreat relationships can often resolve the failures of image-only inspection. Therefore, the top-three agreement rate of 78.6% is considered diagnostic consistency with engineer annotations rather than a full explanation of policy causality. Mobile-manipulation research has found hardware-dependent failure modes for suction, parallel-jaw, and underactuated grippers [29]. A parallel-jaw platform was employed for the physical test, and therefore, the failure shares obtained could not be directly used in other end-effector analyses. The addition of the integrated explanation head increased the per-decision latency by 2.7ms, but the overall planning latency remained below 50ms in 93.5% of real trials.

Post-hoc perturbation explanations have relatively high online recalculation costs when changes occur in the robot decision-making process [30]. Category-level inspection shows that rigid boxes and trays are the primary beneficiaries of hypergraph reasoning due to the geometry supporting a stable target-obstacle-clearance relationship. Cylinders were relatively easy to hold at first, but they were more prone to sliding out during the direction of retreat during rotation and near the shelf lip. Pouch had the longest variation in execution time because the gripper sometimes compressed the object before fixing the lift. Transparent Bottles received the lowest agreement in the explanations because the perceived boundary moved between adjacent camera frames. Therefore, the operator-facing output was limited to a ranked relationship, a short reason label and the corresponding candidate action. A typical good reason was that the rejected grasp had a good contact quality but a small retreat margin; another was that a base offset could increase elbow clearance without changing the grasp. For low-confidence diffuse explanations, the controller can request a lateral re-observation; for stable high-risk retreat explanations, it can use the motion candidate without reacquiring the whole scene.

The same problem was also found in the reviewed failures: a small change in the estimated shelf-lip clearance of less than one centimeter resulted in a change of the attractive grip from acceptance to rejection. Rationale Entropy was used to partition the above cases. The median normalized entropy of correctly performed picks was 0.34, that of stable geometric rejections was 0.39, and that of failures due to uncertain perception was 0.62. The high value was not an error; rather, it led the reviewers to remember the perception history instead of the motion controller. Counterfactual inspection was conducted by removing the top-ranked hyperedge and re-ranking the same set of candidates. The chosen action changed in 72.4% of the audited failures when the top relation was removed, compared with 28.7% after the removal of a randomly selected relation. For target-obstacle-avoidance reasons, the chosen action usually modified the approach side or shifted the direction of evasion; for base-reachability, a small base offset was frequently necessary. Cases in which the action did not change were concentrated in deformable pouches, and several relations had similar weights and the execution error occurred after contact. Although the experiments retained a reasonable geometric explanation, no material-response model was found.

6. Conclusion

This paper presents an explainable Hypergraph Mamba framework for mobile-manipulator shelf picking. Typed hyperedges represent higher-order relations among shelf geometry, objects, candidate motions, robot configuration, and execution history, while selective state-space updates preserve relevant evidence across observation and repositioning steps. A shared policy and explanation pathway links action ranking to relation-level diagnostic information, enabling engineers to inspect which modeled constraints influenced a selected or rejected action.

Several limitations remain. The physical evaluation used one mobile-manipulator platform and one parallel-jaw gripper, so transfer to other kinematics and end effectors has not been established. Explanation supervision was sparse and based on engineering annotations; it measures diagnostic consistency rather than complete causal faithfulness. The perception backbone also remained sensitive to transparent, reflective, and deformable products under severe occlusion. Aggregate physical logs were sufficient for descriptive comparison but did not provide repeated-seed uncertainty estimates for every condition.

Future work will evaluate multiple grippers and robot platforms, integrate active viewpoint selection, and develop uncertainty-aware state retention for inconsistent observations. Relation-level explanations will also be connected to controller interventions so that the robot can request a specific camera motion, base adjustment, or safer retreat when a decisive relation is uncertain. Cross-site studies covering additional shelf structures, product materials, and controller timing regimes are needed to determine how reliably the method transfers to operational picking facilities.

7. Declarations

7.1 Author Contributions

Benjamin Beran contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Herbert Šilhan contributes to validation, analysis, data collection, draft preparation. Dalibor Sedláček contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

7.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

7.3 Institutional Review Board Statement

Not applicable.

7.4 Informed Consent Statement

Not applicable.

7.5 Data Availability

Not applicable.

7.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

8. References

[1] Schwaiger, S., Muster, L., Scherl, A., Trivisonne, P., Wöber, W., & Thalhammer, S. (2024). Towards full autonomy in mobile robot navigation and manipulation. e+i Elektrotechnik und Informationstechnik, 141, 400–410. https://doi.org/10.1007/s00502-024-01261-6

[2] Sereinig, M., Manzl, P., & Gerstmayr, J. (2024). Task-dependent comfort zone, a base placement strategy for mobile manipulators based on manipulability measures. Robotics, 13(8), 122. https://doi.org/10.3390/robotics13080122

[3] Chen, H., Zang, X., Zhu, Y., & Zhao, J. (2024). Hybrid sampling-based path planning for mobile manipulators performing pick and place tasks in narrow spaces. Applied Sciences, 14(22), 10313. https://doi.org/10.3390/app142210313

[4] Yang, Y., Guo, J., Li, Z., He, Z., & Zhang, J. (2024). Ground4Act: Leveraging visual-language model for collaborative pushing and grasping in clutter. Image and Vision Computing, 151, 105280. https://doi.org/10.1016/j.imavis.2024.105280

[5] Gao, Y., Ji, S., Han, X., & Dai, Q. (2024). Hypergraph computation. Engineering, 40, 188–201. https://doi.org/10.1016/j.eng.2024.04.017

[6] Chai, L., Tu, L., Wang, X., & Su, Q. (2024). Hypergraph modeling and hypergraph multi-view attention neural network for link prediction. Pattern Recognition, 149, 110292. https://doi.org/10.1016/j.patcog.2024.110292

[7] Liu, Y., Liang, H., & Zhao, S. (2024). A lightweight visual Mamba network for image recognition under resource-limited environments. Applied Soft Computing, 167, 112294. https://doi.org/10.1016/j.asoc.2024.112294

[8] Bekkemoen, Y. (2024). Explainable reinforcement learning (XRL): A systematic literature review and taxonomy. Machine Learning, 113, 355–441. https://doi.org/10.1007/s10994-023-06479-7

[9] Stefanini, E., Lentini, G., Grioli, G., Catalano, M. G., & Bicchi, A. (2024). Exploring saliency for learning sensory-motor contingencies in loco-manipulation tasks. Robotics, 13(4), 58. https://doi.org/10.3390/robotics13040058

[10] Mersha, M., Lam, K., Wood, J., AlShami, A. K., & Kalita, J. (2024). Explainable artificial intelligence: A survey of needs,

techniques, applications, and future direction. Neurocomputing, 599, 128111. https://doi.org/10.1016/j.neucom.2024.128111

[11] Wei, D., Cao, J., & Gu, Y. (2024). Robot grasp in cluttered scene using a multi-stage deep learning model. IEEE Robotics and Automation Letters, 9(7), 6512–6519. https://doi.org/10.1109/LRA.2024.3406191

[12] Tang, W., Tang, K., Zi, B., Qian, S., & Zhang, D. (2024). High precision 6-DoF grasp detection in cluttered scenes based on network optimization and pose propagation. IEEE Robotics and Automation Letters, 9(5), 4407–4414. https://doi.org/10.1109/LRA.2024.3377951

[13] Wang, F., Liu, Y., Shi, M., Chen, C., Liu, S., & Zhu, J. (2024). Efficient stacking and grasping in unstructured environments. Journal of Intelligent & Robotic Systems, 110, 57. https://doi.org/10.1007/s10846-024-02078-3

[14] Liu, D., Pan, Z., Hu, S., & Cai, F. (2024). Distance enhanced hypergraph learning for dynamic node classification. Neural Processing Letters, 56, 231. https://doi.org/10.1007/s11063-024-11645-6

[15] Ma, A., Chi, G., Ivaldi, S., & Chen, L. (2024). Learning high-level robotic manipulation actions with visual predictive model. Complex & Intelligent Systems, 10, 811–823. https://doi.org/10.1007/s40747-023-01174-5

[16] Patro, B. N., & Agneeswaran, V. S. (2025). Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. Engineering Applications of Artificial Intelligence, 159, 111279. https://doi.org/10.1016/j.engappai.2025.111279

[17] Chen, Z., Ge, Y., Yu, M., Yue, Q., & Zhu, L. (2024). MambaTSR: You only need 90k parameters for traffic sign recognition. Neurocomputing, 599, 128104. https://doi.org/10.1016/j.neucom.2024.128104

[18] Milani, S., Topin, N., Veloso, M., & Fang, F. (2024). Explainable reinforcement learning: A survey and comparative review. ACM Computing Surveys, 56(7), 168:1–168:36. https://doi.org/10.1145/3616864

[19] Weinberg, A. I., Shirizly, A., Azulay, O., & Sintov, A. (2024). Survey of learning-based approaches for robotic in-hand manipulation. Frontiers in Robotics and AI, 11, 1455431. https://doi.org/10.3389/frobt.2024.1455431

[20] Merckaert, K., Convens, B., Nicotra, M. M., & Vanderborght, B. (2024). Real-time constraint-based planning and control of robotic manipulators for safe human-robot collaboration. Robotics and Computer-Integrated Manufacturing, 87, 102711. https://doi.org/10.1016/j.rcim.2023.102711

[21] Liu, J., Yap, H. J., & Khairuddin, A. S. M. (2024). Review on motion planning of robotic manipulator in dynamic environments. Journal of Sensors, 2024(1), 5969512. https://doi.org/10.1155/2024/5969512

[22] Lin, H.-I., Mandal, R., & Wibowo, F. S. (2024). BN-LSTM-based energy consumption modeling approach for an industrial robot manipulator. Robotics and Computer-Integrated Manufacturing, 85, 102629. https://doi.org/10.1016/j.rcim.2023.102629

[23] Padalkar, A., Quéré, G., Raffin, A., Silvério, J., & Stulp, F. (2024). Guiding real-world reinforcement learning for in-contact manipulation tasks with Shared Control Templates. Autonomous Robots, 48, 12. https://doi.org/10.1007/s10514-024-10164-6

[24] Shi, X., Zhang, Y., Liu, K., Wen, Z., Wang, W., Zhang, T., & Li, J. (2025). State space models meet transformers for hyperspectral image classification. Signal Processing, 226, 109669. https://doi.org/10.1016/j.sigpro.2024.109669

[25] Zhang, T., Peng, F., Yan, R., Tang, X., Deng, R., & Yuan, J. (2024). Quantification of uncertainty in robot pose errors and calibration of reliable compensation values. Robotics and Computer-Integrated Manufacturing, 89, 102765. https://doi.org/10.1016/j.rcim.2024.102765

[26] Xie, Y., Liu, J., & Yang, Y. (2024). Pose optimization for mobile manipulator grasping based on hybrid manipulability. Industrial Robot, 51(1), 134–147. https://doi.org/10.1108/IR-06-2023-0128

[27] Meli, D., & Fiorini, P. (2025). Inductive learning of robot task knowledge from raw data and online expert feedback. Machine Learning, 114, 91. https://doi.org/10.1007/s10994-024-06636-6

[28] Angelopoulos, G., Mangiacapra, L., Rossi, A., Di Napoli, C., & Rossi, S. (2025). What is behind the curtain? Increasing transparency in reinforcement learning with human preferences and explanations. Engineering Applications of Artificial Intelligence, 149, 110520. https://doi.org/10.1016/j.engappai.2025.110520

[29] Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2024). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. IEEE Access, 12, 101603–101625. https://doi.org/10.1109/ACCESS.2024.3431437

[30] Seraj, E., Lee, K. M., Zaidi, Z., Xiao, Q., Li, Z., Nascimento, A. S., van Waveren, S., Tambwekar, P., Paleja, R. R., Das, D., & Gombolay, M. C. (2024). Interactive and explainable robot learning: A comprehensive review. Foundations and Trends in Robotics, 12(2–3), 75–349. https://doi.org/10.1561/2300000081