SEE, FORESEE, ACT: INTENT-AWARE WORLD MODELING FOR PROACTIVE RETAIL ASSISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

World models allow agents to anticipate action consequences, but proactive assistance also requires interpreting incomplete observations and deciding whether to intervene before a user asks for help. We propose See–Foresee–Act and instantiate it as the Proactive Intent World Model **(PIWM)** for retail assistance. See preserves role-relevant observable evidence alongside AIDA–BDI customer states; Foresee predicts one-step consequences for candidate actions; Act explicitly compares those consequences, including non-intervention. GuidanceSalesBench supplies predominantly synthetic supervision linking observations, states, candidate outcomes, and decisions. On 509 Full test examples, consequence-based selection achieves 63.65% decision accuracy, compared with 45.38% for direct decision output, an improvement of 18.27 percentage points. These results support candidate-consequence comparison for decision reasoning in this retail task. Our code is available at [1].

1 INTRODUCTION

World models enable agents to predict action-conditioned state changes and use these predictions to guide decisions (Hafner et al., 2020; 2023). Proactive assistance additionally requires inferring service needs from incomplete observations and deciding whether to intervene before a user makes an explicit request (Yang et al., 2025; Horvitz, 1999). A retail customer may compare products, inspect labels, or repeatedly examine the same item. These behaviors may indicate a need for assistance or independent exploration. We study proactive retail assistance as a decision problem: given sparse pre-interaction observations, an agent must infer the customer's service needs and decide whether and how to intervene by considering the consequences of both assistance and silence(Fu et al., 2026).

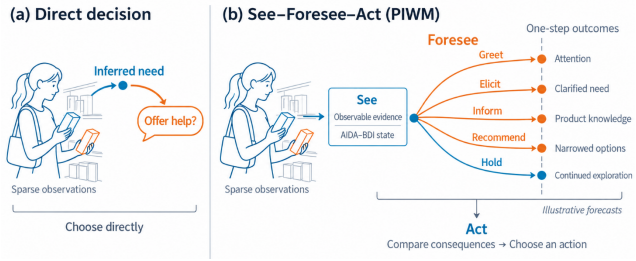


Figure 1: Direct decision versus PIWM. See preserves observable evidence and inferred states; Foresee predicts one-step outcomes; Act compares alternatives, including non-intervention (Hold).

This problem presents two challenges: grounding customer-state inference in observable evidence and selecting interventions under uncertain service needs. First, observations must be translated into task-relevant states rather than generic descriptions (Yang et al., 2025). Intent alone can compress away the evidence needed to distinguish a request for information from independent comparison: similar inferred intentions can arise from different gaze, engagement, and product-handling patterns. Second, action selection must weigh assistance benefits against uncertain needs and interruption costs (Horvitz, 1999; Fu et al., 2026). This requires considering what happens if the agent remains silent as well as what happens if it acts. We formulate the task at a single decision point, linking sparse-video observations to structured customer states, action-conditioned normative outcomes, and intervention choices that include remaining silent.

[1] https://anonymous.4open.science/r/piwm-review-materials-2026-63BD/README.md

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

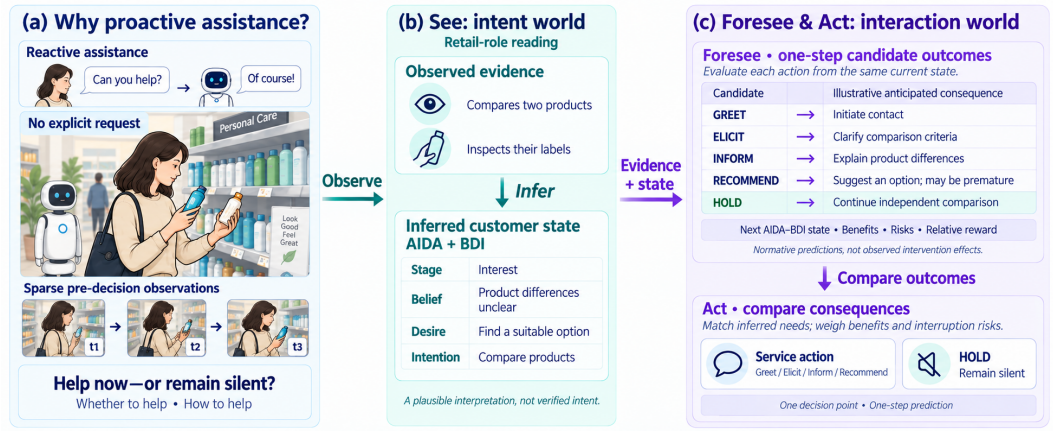


Figure 2: Motivation and task scope of PIWM. Proactive retail assistance requires deciding whether and how to help before an explicit customer request. The intent world represents observed human–object evidence and inferred AIDA–BDI states, while the interaction world associates candidate service actions with normative next-state targets. We evaluate sparse-video *See*, action-conditioned *Foresee*, and intervention selection including HOLD at a single decision point.

We propose the **See–Foresee–Act framework** and instantiate it as **PIWM** (Proactive Intent World Model) for retail assistance. See extracts role-relevant human–object evidence and predicts a structured customer state from sparse video. Foresee predicts one-step normative outcomes for each candidate service action. Act receives the current evidence, inferred state, and candidate forecasts, then explicitly compares alternatives including non-intervention (Hold). The common state and action interfaces support both end-to-end assistance decisions and controlled comparisons of evidence use and consequence prediction.

PIWM combines **AIDA(Attention-Interest-Desire-Action)** purchasing stages (Barry, 1987) with **BDI** fields for **beliefs, desires, and intentions** (Rao & Georgeff, 1995), while retaining explicit evidence about engagement, gaze, posture, and object interaction. The inferred fields express a task interpretation; the evidence fields preserve the observations supporting it (Rabinowitz et al., 2018; Bai et al., 2025). Both enter consequence prediction and action selection. This design lets downstream modules compare service alternatives against the same task interpretation and its visible basis.

Hold is modeled through the same consequence interface as Greet, Elicit, Inform, and Recommend. An available silence option establishes what the agent may choose; a forecast describes why waiting could preserve independent exploration or avoid premature interruption. We therefore test the contribution of the Hold forecast while retaining the action itself and the non-intervention instruction. This connects the value of non-intervention to the comparison used for other service actions (Horvitz, 1999; Lu et al., 2024).

To support systematic evaluation of this task, we construct **GuidanceSalesBench**, linking sparse pre-interaction videos, structured customer states, candidate-action outcomes, and normative action labels. Rule-guided, expert-reviewed synthetic examples supervise observation-to-state, state–action-to-consequence, and state-to-action mappings. The learned components are composed at inference by supplying generated candidate consequences to Act. The benchmark supports component-level evaluation of state inference, consequence prediction, and action selection, together with full-pipeline evaluation (Full) of video-to-action decisions.

On 509 Full test examples, consequence comparison achieves 63.65% accuracy (324/509), versus 45.38% (231/509) for direct decision output. This improves accuracy by 18.27 percentage points, a net increase of 93 correct decisions. Removing explicit evidence fields yields 27.70% (141/509); withholding the Hold forecast yields 47.94% (244/509). PIWM achieves 62.20%, 74.43%, and 89.80% on See, Foresee, and Act, respectively.

Our main contributions are as follows:

1. **A task formulation for proactive retail assistance.** We study whether and how to intervene from sparse pre-interaction observations, connecting observable evidence and inferred customer states with action-conditioned normative outcomes at a single decision point.
2. **GuidanceSalesBench for structured evaluation.** We construct a benchmark linking sparse observations, customer states, normative candidate outcomes, and preferred actions, supporting both component-level and full-pipeline evaluation.
3. **A See–Foresee–Act framework instantiated as PIWM.** The framework connects evidence-grounded state inference, one-step candidate-consequence prediction, and comparative action selection through common state and action interfaces.
4. **Empirical findings on evidence, consequence comparison, and non-intervention.** Controlled comparisons show how candidate forecasts, explicit evidence fields, and the Hold forecast contribute to decision accuracy on GuidanceSalesBench.

2 Related Work

2.1 Proactive assistance and intervention timing.

Mixed-initiative interaction balances assistance benefits against uncertainty and interruption costs (Horvitz, 1999). Proactive Agent studies human acceptance and rejection, LlamaPIE separates response timing from generation, and ProAssist evaluates streaming visual assistance (Lu et al., 2024; Chen et al., 2025; Zhang et al., 2025b). ContextAgent connects sensory and personal context to service initiation and tool use (Yang et al., 2025); PRISM studies risk-sensitive, uncertainty-aware proactive deliberation (Fu et al., 2026). These works motivate deciding whether assistance is warranted. PIWM focuses on a complementary question: how explicit retail evidence and candidate-specific forecasts inform a shared choice among five interventions, including Hold.

ReAct and Reflexion combine reasoning, action, and feedback, while trajectory and process refinement improve behavior through exploration and supervision (Yao et al., 2023; Shinn et al., 2023; Song et al., 2024; Xiong et al., 2024; Wang et al., 2025a). PIWM evaluates a single decision point rather than an executed interaction trajectory. Its non-intervention control holds the candidate set and instruction fixed while removing only the Hold forecast, directly examining the value of anticipating silence in action comparison.

2.2 Intent modeling and structured state.

Partial-observation control distinguishes observable inputs from underlying state (Åström, 1965). Machine theory of mind models other agents from observed behavior (Rabinowitz et al., 2018). AIDA describes purchasing stages and BDI organizes beliefs, desires, and intentions (Barry, 1987; Rao & Georgeff, 1995). VAGEN separates visual state estimation and transition modeling within multi-turn world-model reasoning (Wang et al., 2025b). PIWM distinguishes two parts of the current representation: an inferred task state and explicit visible evidence. Its evidence-removal experiment tests whether the latter contributes beyond retaining stage, intent, and BDI, rather than treating state structure alone as sufficient for decisions.

Interactive environments support evaluating actions through execution in household, scientific, and shopping tasks (Puig et al., 2018; Shridhar et al., 2021; Wang et al., 2022; Yao et al., 2022). PIWM instead links state interpretations to candidate interventions at a single observation point. Its evaluation primarily measures agreement with a constructed schema, rather than behavioral validation of inferred mental states or successful interactive task completion.

2.3 World models and decision utility.

World models support prediction and control through learned transitions (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2023). RAP couples a language-model world model with search, while other language-agent methods learn or construct planning models from experience and world knowledge (Hao et al., 2023; Xiang et al., 2023; Guan et al., 2023; Qiao et al., 2024; Wang et al., 2025c; Fang et al., 2025). PIWM learns normative one-step retail outcomes and uses them for explicit candidate

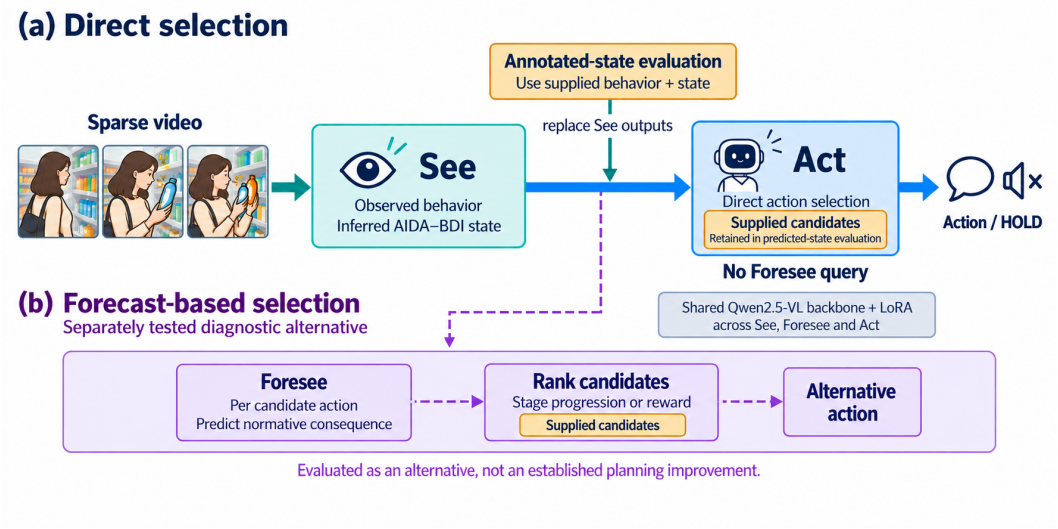


Figure 3: See–Foresee–Act inference. See constructs a customer state from video; Foresee predicts one-step consequences for each candidate, including Hold; Act compares the candidate–consequence pairs to select an intervention. The architecture image is retained for collaborator revision.

comparison. Its question concerns the information needed for a proactive service decision, including the consequence of not intervening.

Closed-loop evaluation links world-model predictions to downstream task performance (Zhang et al., 2025a). We compare approaches along four dimensions: the information represented from observation, the supervision for candidate outcomes, the treatment of non-intervention, and the decision setting. PIWM combines explicit retail evidence with AIDA–BDI, learns candidate outcomes from synthetic normative supervision, predicts Hold through the same interface as active actions, and evaluates single-step decision accuracy. Its direct-decision and forecast-based conditions test the utility of this combination on the same Full set.

3 Method

3.1 Task definition and framework

PIWM addresses service decisions before an explicit customer request. Given sparse pre-decision video observations o and a retail role description, the agent selects an action from candidates C: Greet, Elicit, Inform, Recommend, or Hold. The inference pipeline first constructs a customer state, predicts a one-step consequence for each candidate, and then explicitly compares those consequences to select an action.

The three interfaces are:

$$(\hat{z}, \hat{s}) = \mathrm{See}_\theta(o), \quad \hat{y}^{(a)} = \mathrm{Foresee}_\theta(\hat{z}, \hat{s}, a), \quad \hat{a} = \mathrm{Act}_\theta\Big(\hat{z}, \hat{s}, \{(a, \hat{y}^{(a)})\}_{a \in C}\Big).$$

Here $\hat{z}$ summarizes observable evidence, $\hat{s}$ represents the customer state, and $\hat{y}^{(a)}$ describes the anticipated consequence of candidate a.

3.2 See: role-oriented evidence and customer state

See represents the current observation through explicit evidence and inferred customer state. Four evidence fields, `visible_evidence`, `engagement_pattern`, `gaze_and_attention`, and `body_and_hands`, describe visible scene and behavior, engagement and interaction, attention, and posture and hand movements. The retained state fields are `stage`, `intent_label`,

`belief`, `desire`, and `intention`. AIDA describes task progress, while BDI organizes beliefs, desires, and intentions. The inferred fields can also encode information derived from visual observations.

Consider a customer repeatedly comparing two products and inspecting their labels. A generic description records the movements; role-oriented reading organizes the comparison targets and information-seeking cues. The state representation expresses a possible need to understand product differences and provides a common starting point for evaluating interventions. This running example illustrates the interfaces rather than reporting an experimental prediction.

3.3 FORESEE: ONE-STEP CONSEQUENCES OF CANDIDATE ACTIONS

Foresee receives the current evidence, customer state, and one candidate action. It predicts the next AIDA–BDI state, anticipated benefits and risks, and a relative reward. Each candidate is evaluated from the same current state. Hold is evaluated alongside active interventions, allowing the model to compare the consequences of acting and remaining silent.

In the running example, Elicit may clarify the comparison criteria, Inform may address a specific information gap, and an early Recommend may precede a clear understanding of preferences. Hold represents continued independent comparison. Foresee expresses these alternatives as candidate-specific next-state and benefit–risk descriptions for Act to compare. Field schemas and inference interfaces appear in Appendices B and C.

3.4 ACT: EXPLICIT CONSEQUENCE COMPARISON AND NON-INTERVENTION

Act receives current evidence and state together with the candidate–consequence pairs produced by Foresee. It compares how well each action addresses the inferred need, how it may advance the interaction, and what interruption risks it introduces, then generates a candidate label and a service realization. In the example, the choice between Elicit and Inform depends on whether the comparison need has become sufficiently specific; Hold remains available when independent exploration is appropriate.

Non-intervention is an explicit decision alternative whose predicted consequence participates in the same comparison as service actions. This makes the decision concern both how to help and whether help is currently warranted. The implementation uses a single decision point and one-step candidate predictions.

3.5 TRAINING AND INFERENCE

PIWM parameterizes the three prediction tasks with a Qwen2.5-VL-7B-Instruct backbone and LoRA adaptation. Rather than training a monolithic video-to-action predictor, PIWM decomposes supervision according to the model-based decision process: current-state estimation, action-conditioned transition prediction, and comparative action selection.

The central world-modeling objective is carried by *Foresee*. Given the current structured representation and a candidate action, Foresee is supervised to predict the corresponding one-step customer state together with its anticipated benefit and risk. Training on these state–action–outcome tuples therefore teaches PIWM an action-conditioned transition model over the structured customer state, rather than only a direct policy from observations to actions. *See* provides the grounded current-state estimate on which these transitions are conditioned, while *Act* learns to select among candidate actions using their predicted outcomes.

At inference time, the learned components are composed without access to annotated states or consequences. See first constructs the current representation from the sparse video. Foresee then evaluates each candidate action from this shared current state, including HOLD, and generates its predicted one-step outcome. Act finally compares the resulting candidate–outcome pairs and selects the intervention. This composition turns the learned transition model into an explicit model-based decision procedure: actions are selected according to their predicted consequences rather than from the current state alone.

Data resource	Records	Purpose
See	1,043	Multimodal observation representation
Foresee	2,001	Future-state prediction
Act	680	Decisions from textual states
Full	3,054	Multimodal-input-to-decision training data
Real	418	Sales-expert-annotated real-store scenes; evaluation only

Table 1: Task-specific data resources. Full is split into 3,054 training, 509 validation, and 509 test examples. Real is excluded from training.

4 GUIDANCESALESBENCH

4.1 CONSTRUCTION AND SUPERVISION

Construction proceeds through four linked stages. **Knowledge distillation** extracts 27 compact principles and 78 source-linked conditional rules from retail and consumer-behavior material; these counts describe an internal rule ledger, not independently established laws. **Scenario specification** instantiates product context, viewpoint, visible cues, and AIDA–BDI fields in a scene manifest. **Candidate annotation** pairs a parent state with alternative service realizations and LLM-assisted, rule-guided consequence targets, followed by the expert review described in the construction records. **Task export** turns the same parent material into video-to-state See records, state–action-to-consequence Foresee records, best-action Act records, and predicted-state evaluation inputs.

4.2 DATASET COMPOSITION AND SPLITS

GUIDANCESALESBENCH contains task-specific resources for state estimation, future-state prediction, and action selection, together with an end-to-end multimodal collection. Table 1 summarizes the corresponding data resources.

The component resources use different supervision units. See contains 1,043 multimodal observation records, Foresee contains 2,001 state–action records, and Act contains 680 textual-state decision records. Because these resources are task-specific exports, their record counts should not be interpreted as counts of independent scenes or summed with the Full collection.

A separate Real set contains 418 sales-expert-annotated real-store records. These examples are excluded from training and reserved for evaluation.

4.3 A COMPLETE ANNOTATION EXAMPLE

Figure 4 illustrates a complete GuidanceSalesBench example across the four supervision layers. The sparse video frames show a customer remaining near the terminal, shifting gaze across display regions, and leaning slightly forward, without making an explicit request.

See records these observable cues and infers an AIDA *Interest* state, together with the belief that relevant options may exist, the desire to compare before choosing, and the intention to explore available options. These inferred fields are task interpretations rather than directly observed facts.

Given this state, Foresee associates each candidate action with a normative consequence target. Elicit is annotated as clarifying the customer's comparison need with mild interruption risk; Inform may reduce basic uncertainty but can remain generic; Recommend may narrow the search but risks premature preference pressure; and Hold preserves independent exploration while leaving uncertainty unresolved. The preferred action is therefore Elicit.

The example illustrates how a single parent observation is transformed into observable evidence, an inferred customer state, candidate-specific consequence targets, and a gold best-action label. The consequence targets are normative annotations rather than outcomes observed after real interventions.

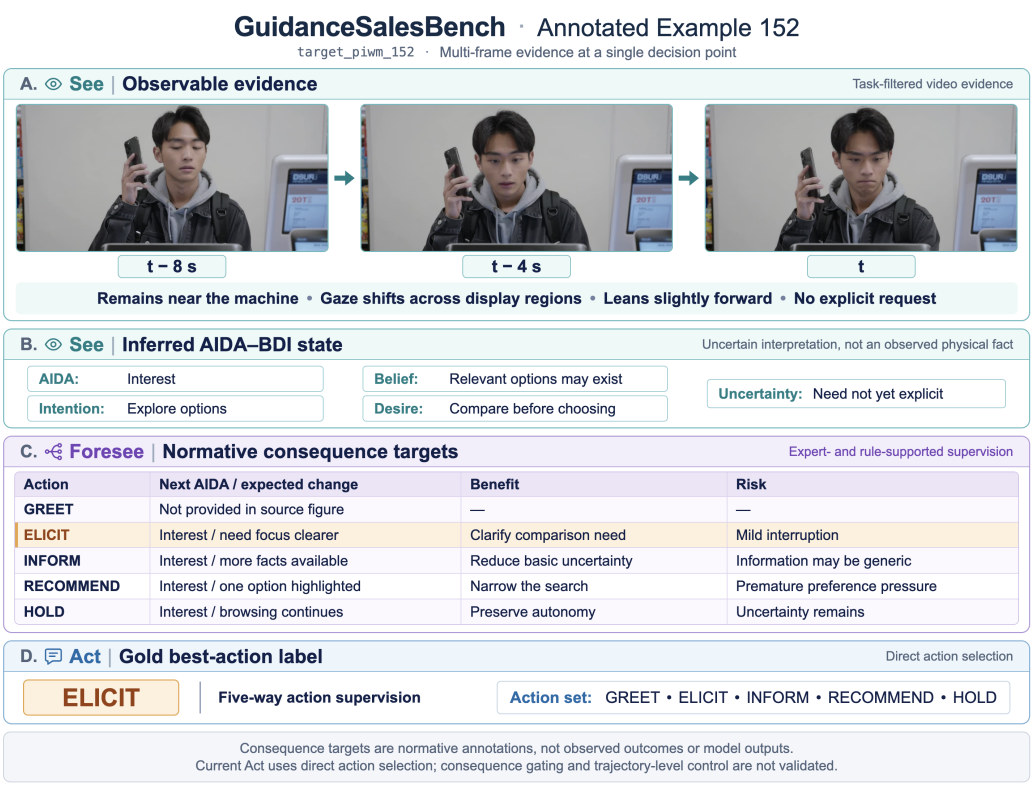


Figure 4: A complete GUIDANCESALESBENCH sample. Observed video evidence supports *See*; the AIDA–BDI state is an inferred interpretation; each candidate action is paired with an expert- and rule-supported normative consequence target for *Foresee*; and the gold best-action label supplies *Act* supervision.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate PIWM at both the component and end-to-end levels. The component evaluations isolate the three structured prediction tasks defined in Section 3. *See* is evaluated on 209 held-out examples and is counted as correct only when both the predicted AIDA stage and intent label match the reference annotation. *Foresee* evaluates action-conditioned next-state prediction on 309 state–action records, using next-AIDA-stage accuracy over four classes. *Act* evaluates intervention selection on 245 examples where the reference customer state is provided.

The *Full* evaluation measures the complete video-to-action pipeline on the 509-example held-out test split described in Section 4.2. Unlike the component evaluations, Full begins from sparse visual observations: See first predicts the current customer representation, Foresee generates candidate-specific outcomes when enabled, and Act produces the final intervention. Consequently, Full captures the effect of composing the learned components rather than evaluating any one prediction task in isolation. Invalid or missing predictions are counted as errors.

For all tasks, we report accuracy:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i^\star].$$

The theoretical random baselines follow the corresponding label spaces: 3.13% for the joint four-stage/eight-intent See target, 25.00% for the four-class Foresee target, and 20.00% for the five-action Act and Full decisions.

Model	See (209)	Foresee (309)	Act (245)	Full (509)
Random Baseline	3.13	25.00	20.00	20.00
Qwen2.5-VL-7B	43.54	9.39	13.47	24.75
Gemini 2.5 Flash	—	—	—	**69.74**
GPT-4o	—	—	—	68.96
PIWM (direct decision)	**62.20**	**74.43**	**89.80**	45.38
PIWM (counterfactual)	**62.20**	**74.43**	**89.80**	63.65

Table 2: Component and full-pipeline accuracy (%). The two PIWM settings share the same See, Foresee, and Act evaluations and differ only in Full inference: direct decision selects from the predicted current representation without using Foresee outputs, whereas the counterfactual setting supplies predicted candidate consequences to Act. Bold indicates the highest reported score in each column.

Inference setting	Correct / total	Decision accuracy	Difference from full pipeline
Predicted state + Foresee + Act	324 / 509	63.65%	—
Predicted state → Act directly	231 / 509	45.38%	−18.27 pp
Without explicit evidence fields	141 / 509	27.70%	−35.95 pp
Hold forecast unavailable	244 / 509	47.94%	−15.72 pp

Table 3: Ablations on the 509-example Full test set. Differences are measured against the complete predicted-state + Foresee + Act pipeline.

5.2 Overall performance

PIWM improves substantially over the untuned Qwen2.5-VL-7B backbone on all three component tasks. See increases from 43.54% (91/209) to 62.20% (130/209), indicating that the adapted model more reliably recovers the joint AIDA-stage and intent representation from sparse visual observations. Foresee increases from 9.39% (29/309) to 74.43% (230/309). This task directly evaluates the state–action-conditioned next-state prediction used by the world-modeling component of PIWM. Act reaches 89.80% (220/245), compared with 13.47% (33/245) for the base model, when the structured current state is provided.

The end-to-end Full evaluation tests whether these learned components remain useful when composed from video input. Qwen2.5-VL-7B obtains 24.75% accuracy (126/509). PIWM reaches 45.38% (231/509) when Act selects directly from the predicted current representation without receiving explicit candidate forecasts. When Foresee predictions for the candidate actions are supplied to Act, accuracy rises to 63.65% (324/509), corresponding to 93 additional correct decisions and an improvement of 18.27 percentage points. GPT-4o and Gemini 2.5 Flash achieve 68.96% (351/509) and 69.74% (355/509), respectively, on Full.

5.3 Ablation analysis

We further examine three design choices in the composed pipeline: explicit comparison of candidate consequences, retention of observable evidence alongside the inferred customer state, and explicit prediction of the consequence of non-intervention. All ablations use the same 509-example Full test set and preserve the five-action decision space. Table 3 summarizes the results.

Consequence comparison. The complete pipeline predicts the current customer representation, generates one-step consequences for the five candidate actions, and passes the resulting candidate–consequence pairs to Act. The direct-decision condition removes these forecasts and instead asks Act to select an intervention directly from the predicted current state. Accuracy decreases from 63.65% to 45.38%, a drop of 18.27 percentage points and 93 correct decisions. Because the underlying PIWM components and final action space are unchanged, this comparison isolates the decision-time contribution of explicitly using action-conditioned forecasts. The result supports the central

premise of See–Foresee–Act: predicting future states is useful when those predictions participate in the final action comparison.

Explicit evidence. The evidence ablation removes `visible_evidence`, `engagement_pattern`, `gaze_and_attention`, and `body_and_hands` from the predicted current representation, while retaining `stage`, `intent_label`, `belief`, `desire`, and `intention`. Foresee then regenerates all candidate consequences from this reduced representation before Act performs the final comparison. Accuracy falls to 27.70% (141/509), 35.95 percentage points below the full pipeline and 183 fewer correct decisions.

This ablation therefore measures the end-to-end contribution of preserving explicit observable evidence rather than only whether Act uses several additional text fields. The retained intent and BDI fields may still contain information inferred from the visual input, so the comparison does not remove all visually derived information. Instead, it isolates the additional value of retaining explicit descriptions of gaze, engagement, posture, and object interaction alongside the structured customer state.

Non-intervention forecast. The Hold ablation retains all five candidate actions—Greet, Elicit, Inform, Recommend, and Hold—together with the full predicted current state and the explicit instruction that non-intervention is a valid choice. The Hold forecast is withheld from Act, while forecasts for the four active interventions remain available.

With the Hold forecast withheld, accuracy decreases to 47.94% (244/509), a drop of 15.72 percentage points and 80 correct decisions. Because Hold remains selectable under both conditions, the comparison distinguishes access to an anticipated outcome for waiting from merely including Hold in the candidate set. Within the evaluated pipeline, explicitly modeling the expected consequence of non-intervention provides useful information for candidate comparison.

6 Discussion: extending single-step decisions over time

PIWM connects task-oriented observation, one-step consequence prediction, and explicit action comparison. Extending this formulation to sustained interaction calls for tracking how new observations revise customer state and how an earlier intervention changes later opportunities. Multi-step trajectories and the value of waiting are natural directions for extending the current single-decision framework.

7 Limitations and Next Validation Steps

The experiments study single-decision prediction and selection on a 509-example Full test set with synthetic, normative supervision. The evidence ablation changes both Foresee and Act inputs and regenerates forecasts, measuring their combined response to explicit evidence removal. The Hold control measures the effect of a missing Hold forecast while retaining the action and non-intervention instruction. These results characterize decision accuracy within the evaluated framework; longer interactions and customer outcomes require other evaluation settings.

Future work can examine generalization across source scenes and real-store interactions, and trace how state estimates and candidate forecasts affect individual decisions. These extensions connect the current single-step evaluation to sustained proactive assistance.

8 Conclusion

PIWM organizes proactive retail assistance through role-oriented state construction, one-step consequence prediction, and explicit action comparison. On the 509-example Full test set, consequence-based selection achieved 63.65% accuracy, compared with 45.38% for direct decision output. Removing explicit evidence fields reduced accuracy to 27.70%; withholding only the Hold forecast reduced it to 47.94%. Together, these results support preserving observable evidence and explicitly comparing the anticipated consequences of intervention and non-intervention.

9 AI USE STATEMENT

Generative models were used in constructing synthetic data and normative annotations. AI assistance was also used for literature discovery, manuscript restructuring and drafting, interpretation of archived experimental reports, and suggestions for additional experiments. The current draft does not assert that newly suggested experiments have been performed. Exact data-generation model versions and prompts require reconciliation with the construction records before submission; the authors remain responsible for verifying the final text, citations, data provenance, and claims.

10 ETHICS STATEMENT

Retail observations can expose personal behavior, and inferred intent may be wrong. The real-store material in this study is staged; this work does not evaluate covert deployment. Assistance should preserve the option to decline and should not infer sensitive characteristics from appearance. Data release requires verified consent and media rights. Neither conformity to a sales-stage label nor an increase in inferred purchase intention is a substitute for user welfare. The present experiments do not establish customer benefit or deployment safety.

11 REPRODUCIBILITY STATEMENT

Appendix A describes the current resources and Full partition. Appendices B and C specify the structured interfaces and inference controls. Appendix D gives task-specific scoring definitions and correct counts. Appendix E describes the Real evaluation coverage and anonymous-material version.

REFERENCES

Karl Johan Åström. Optimal control of markov processes with incomplete state information I. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. doi: 10.1016/0022-247X(65)90154-X. URL https://doi.org/10.1016/0022-247X(65)90154-X.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Thomas E Barry. The development of the hierarchy of effects: An historical perspective. *Current issues and Research in Advertising*, 10(1-2):251–295, 1987.

Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A. Smith, and Shyamnath Gollakota. LlamaPIE: Proactive in-ear conversation assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13801–13824, 2025. doi: 10.18653/v1/2025.findings-acl.710. URL https://aclanthology.org/2025.findings-acl.710/.

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. WebEvolver: Enhancing web agent self-improvement with co-evolving world model. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8959–8975, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.454. URL https://aclanthology.org/2025.emnlp-main.454/.

Yuxuan Fu, Xiaoyu Tan, Teqi Hao, Chen Zhan, and Xihe Qiu. PRISM: Festina lente proactivity—risk-sensitive, uncertainty-aware deliberation for proactive agents. *arXiv preprint arXiv:2602.01532*, 2026. URL https://arxiv.org/abs/2602.01532.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79081–79094, 2023.

David Ha and Jürgen Schmidhuber. World Models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507/.

Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Proactive agent: Shifting LLM agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*, 2024. URL https://arxiv.org/abs/2410.12361.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8494–8502, 2018.

Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Agent planning with world knowledge model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 114843–114871, 2024.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Anand S Rao and Michael P Georgeff. BDI agents: From theory to practice. In *Icmas*, volume 95, pp. 312–319, 1995.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.409. URL https://aclanthology.org/2024.acl-long.409/.

Hanlin Wang, Jian Wang, Chak Tou Leong, and Wenjie Li. STeCa: Step-level trajectory calibration for LLM agent learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11597–11614, Vienna, Austria, July 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.604. URL https://aclanthology.org/2025.findings-acl.604/.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Fei-Fei Li, Yejin Choi, and Manling Li. VAGEN: Reinforcing world model reasoning for multi-turn VLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 172871–172933. Curran Associates, Inc., 2025b. doi: 10.52202/085713-5754. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fc6688d75adde86b9df910769c1d02e3-Abstract-Conference.html.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Siyin Wang, Zhaoye Fei, Qinyuan Cheng, Shiduo Zhang, Panpan Cai, Jinlan Fu, and Xipeng Qiu. World modeling makes a better planner: Dual preference optimization for embodied task planning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21518–21537, Vienna, Austria, July 2025c. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1044. URL https://aclanthology.org/2025.acl-long.1044/.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75392–75412. Curran Associates, Inc., 2023. doi: 10.52202/075280-3295. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ee6630dcbcff857026e474fc857aa9f0-Abstract-Conference.html.

Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. Watch every step! LLM agent learning via iterative step-level process refinement. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1556–1572, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.93. URL https://aclanthology.org/2024.emnlp-main.93/.

Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. ContextAgent: Context-aware proactive LLM agents with open-world sensory perceptions. *Advances in Neural Information Processing Systems*, 38: 167509–167543, 2025.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, et al. World-in-World: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025a.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12044–12068, 2025b. doi: 10.18653/v1/2025.emnlp-main.605. URL https://aclanthology.org/2025.emnlp-main.605/.

A Data resources, review, and partitioning

A.1 Resource units and evaluation sets

GuidanceSalesBench organizes supervision around observations, customer states, candidate consequences, and service decisions. See evaluates multimodal observation representation; Foresee evaluates next-state prediction; Act evaluates decisions conditioned on textual reference states; Full evaluates the complete multimodal-input-to-action pipeline. Real supplies sales-expert-annotated records from staged real-store scenes.

Resource	Resource records	Evaluation records	Evaluation target
See	1,043	209	Jointly correct AIDA stage and intent
Foresee	2,001	309	Next AIDA stage
Act	680	245	Selected action category
Full	3,054 training	509 test	Final video-to-action decision
Real	418	418	Real-store action choice

Table 4: Current resource and evaluation counts. Full additionally has 509 validation examples. Real is reserved exclusively for evaluation.

The reviewed Full pool contains 4,072 examples retained from an initial pool of 6,000 synthetic examples after review by large language models and sales experts. Its partition is 3,054 training, 509 validation, and 509 test examples, corresponding to a 6:1:1 ratio. The Full results and inference-time ablations in this paper use the 509-example test set. The validation split is distinct from the reported test split. Real records do not participate in training.

The component counts refer to task-specific resources and their corresponding evaluation sets. A Foresee record is a state–action pair, whereas a Full record covers a complete observation-to-decision example. The Full split ratio does not define the component splits.

A.2 Annotation structure

Each example connects four levels: a pre-interaction observation, a structured current state, candidate-conditioned consequences, and a preferred service action. The current state contains explicit observable evidence together with an inferred AIDA stage, intent label, and BDI description. Consequence annotations specify the next state and the anticipated benefit, risk, and reward associated with a candidate action.

A consequence target describes the expected one-step outcome under the benchmark's retail-service interpretation. The preferred action is selected within that same interpretation. This connects state estimation and action-conditioned prediction to the decision label used for evaluation.

A.3 Real-store evaluation

Real contains 418 sales-expert-annotated records from staged real-store scenes. It is used to examine service decisions under real-store visual observations. The evaluation target is the selected service action. Its records are kept separate from synthetic training data and from the Full test set.

B Structured interfaces and a running example

B.1 Current-state representation

The state interface separates observable evidence from inferred task state.

See predicts four AIDA stages and eight intent categories. Its evaluation uses joint correctness of the two categorical fields. The evidence and BDI fields remain part of the downstream representation even though the See score is defined over stage and intent.

The interface can be serialized as follows:

Field group	Fields	Function
Observable evidence	visible_evidence, engagement_pattern, gaze_and_attention, body_and_hands	Describe the scene, engagement, attention, posture, and hand movements
Task labels	stage, intent_label	Represent purchasing stage and intent category
BDI interpretation	belief, desire, intention	Express the inferred customer state

Table 5: Current-state fields supplied to downstream prediction and decision modules.

```
<stage>...</stage>
<intent_label>...</intent_label>
<visible_evidence>...</visible_evidence>
<engagement_pattern>...</engagement_pattern>
<gaze_and_attention>...</gaze_and_attention>
<body_and_hands>...</body_and_hands>
<belief>...</belief>
<desire>...</desire>
<intention>...</intention>
```

B.2 Candidate consequences and action output

Foresee takes the current state and one candidate action. Its output specifies the next AIDA stage, next BDI state, risk, benefit, and relative reward. Risk and benefit use low, medium, and high categories; reward is represented on the interval [-1, 1]. Each of the five service actions, including Hold, has its own consequence record.

```
<next_stage>...</next_stage>
<next_belief>...</next_belief>
<next_desire>...</next_desire>
<next_intention>...</next_intention>
<risk>low|medium|high</risk>
<benefit>low|medium|high</benefit>
<reward>...</reward>
```

Act receives the current state and candidate–consequence pairs. Its output includes a rationale, selected candidate, action realization, and intervention utterance.

```
<rationale>...</rationale>
<chosen>...</chosen>
<intervention_action>...</intervention_action>
<intervention_utterance>...</intervention_utterance>
```

These blocks specify output interfaces. See, Foresee, and Act have different input requirements: See receives multimodal observations, Foresee receives a state–action pair, and comparative Act receives the current state and candidate forecasts.

B.3 A running example

Consider a customer who pauses at a terminal, looks across the display, and compares available options. See records these visible cues and infers an Interest stage with an exploratory intent. A possible BDI interpretation is that relevant products may be available, the customer wants to compare their suitability, and the immediate intention is to continue exploring.

Foresee evaluates candidate interventions from this same current state. Elicit may clarify comparison criteria; Inform may address a specific information gap; Recommend may be premature before preferences become clear; Hold allows independent exploration to continue. Act compares these alternatives using the current evidence, inferred need, and anticipated consequences. This example illustrates the interfaces and does not assign a model success score.

A useful consequence need not change the AIDA label. An intervention can clarify preferences while leaving the customer in the Interest stage. Accordingly, Foresee's next-stage score and Full decision accuracy measure different parts of the pipeline.

C Model configuration and single-step inference

C.1 Model and task interfaces

PIWM uses a Qwen2.5-VL-7B-Instruct backbone with LoRA adaptation. The three tasks share the structured state and action interfaces. See estimates the current state from multimodal observations. Foresee learns action-conditioned one-step consequences. Act performs action selection from the information provided by its decision configuration.

The two PIWM rows in the main table share the See, Foresee, and Act component evaluations. Their Full configurations differ in whether predicted consequences are supplied to the decision step. The direct-decision configuration does not invoke Foresee during Full inference.

C.2 Candidate comparison

The candidate action categories are Greet, Elicit, Inform, Recommend, and Hold. The full pipeline first computes the current state, then predicts the consequence of each candidate, and finally supplies the state and candidate–consequence pairs to Act. Each forecast starts from the same current state. Predictions are not recursively rolled forward into a multi-step imagined trajectory.

The following pseudocode specifies the input transformations used by the four configurations:

```
state = See(observations)
if condition == "direct":
    decision = Act(state, candidates, consequences=None)
else:
    if condition == "without_explicit_evidence":
        state = remove_four_evidence_fields(state)
    forecasts = {a: Foresee(state, a) for a in candidates}
    if condition == "hold_forecast_unavailable":
        forecasts["Hold"] = {"prediction_unavailable": True}
    decision = Act(state, candidates, forecasts)
```

C.3 Controlled input changes

Configuration	State supplied downstream	Consequences supplied to Act	Hold selectable
Full pipeline	Complete predicted state	Five predicted outcomes	Yes
Direct decision	Complete predicted state	None	Yes
Without explicit evidence	Four evidence fields removed	Regenerated from the reduced state	Yes
Hold forecast unavailable	Complete predicted state	Four outcomes and a missing marker for Hold	Yes

Table 6: Inputs of the four inference configurations.

The evidence ablation removes visible_evidence, engagement_pattern, gaze_and_attention, and body_and_hands. It retains stage, intent_label, belief, desire, and intention. Both Foresee and Act receive the reduced state, and consequences are generated again from that state. It therefore measures the combined downstream effect of retaining explicit evidence.

The Hold ablation replaces only the Hold forecast with {"prediction_unavailable": true}. The complete current state, other four forecasts, five-action candidate set, and explicit non-intervention instruction are retained. Hold remains a legal decision.

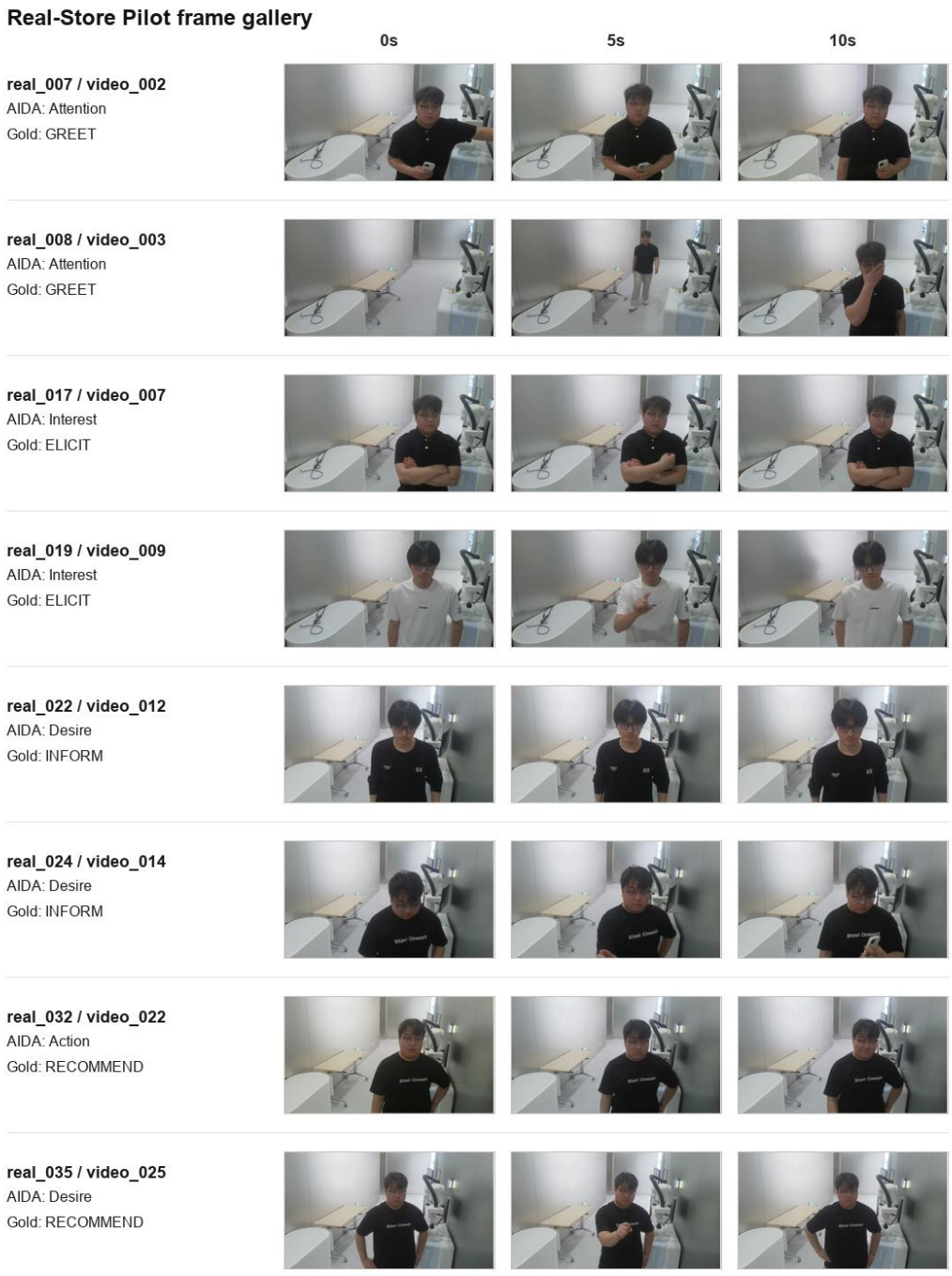


Figure 5: Additional frame examples from the real-store pilot. Each row shows a distinct paid-participant session, with rows spanning different gold-standard AIDA stages and best-action labels. All participants signed release forms authorizing public release of the recorded materials (see Ethics Statement).

D EVALUATION DEFINITIONS AND DETAILED RESULTS

D.1 TASK-SPECIFIC ACCURACY

Let c_i denote whether example i is correctly predicted under its task-specific target. Accuracy is

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} c_i.$$

For See, stage and intent must both be correct:

$$c_i^{\text{See}} = \mathbf{1}[\hat{s}_i = s_i]\,\mathbf{1}[\hat{\ell}_i = \ell_i], \qquad N_{\text{See}} = 209.$$

A correct stage with an incorrect intent, or the reverse, receives zero credit. Missing or invalid target fields also receive zero credit.

Foresee compares the predicted next AIDA stage with its reference over 309 state–action records. Act compares the chosen action category with the reference over 245 textual reference-state examples. Full compares the final action produced from multimodal observations with the reference over 509 test examples. Real uses expert action labels over its 418 evaluation records.

All records remain in the task denominator, including invalid or missing predictions. Risk, benefit, reward, and generated BDI text are part of the consequence representation; they are not additional terms in the reported Foresee accuracy.

D.2 COMPONENT RESULTS

Model	See	Foresee	Act
Qwen2.5-VL-7B	91/209 (43.54%)	29/309 (9.39%)	33/245 (13.47%)
PIWM	130/209 (62.20%)	230/309 (74.43%)	220/245 (89.80%)

Table 7: Correct counts and accuracies for the current component evaluations.

The random baseline assumes uniform sampling of the corresponding categorical targets. Its theoretical accuracies are 3.13% for joint See prediction, 25.00% for Foresee, and 20.00% for five-action decisions.

D.3 FULL RESULTS

Model or decision configuration	Correct / total	Accuracy
Qwen2.5-VL-7B	126 / 509	24.75%
Gemini 2.5 Flash	355 / 509	69.74%
GPT-4o	351 / 509	68.96%
PIWM, direct decision	231 / 509	45.38%
PIWM, counterfactual reasoning	324 / 509	63.65%

Table 8: Correct counts for the current Full test set.

D.4 INFERENCE ABLATIONS

Consequence comparison adds 93 correct decisions relative to direct output. Removing explicit evidence reduces the total by 183 correct decisions; withholding the Hold forecast reduces it by 80. These are aggregate differences over the same test-set size.

Configuration	Correct / total	Accuracy	Difference from full pipeline
Direct decision	231 / 509	45.38%	−18.27 pp
Without explicit evidence	141 / 509	27.70%	−35.95 pp
Hold forecast unavailable	244 / 509	47.94%	−15.72 pp
Full pipeline	324 / 509	63.65%	—

Table 9: Inference ablations on the 509-example Full test set.

For paired correctness indicators, let b count examples correct only under condition A and c examples correct only under condition B. The accuracy difference is $(b-c)/N$. This distinction separates the net difference in correct counts from the number of predictions that change in each direction.

E Real-store material and reproducibility

E.1 Real-store observations

The Real collection uses staged retail interactions and sales-expert action annotations. Its 418 records are reserved for evaluation. The task remains a single-decision service choice from the observed customer context, using the same five action categories as the synthetic setting.

Real-store material can contain identifiable people and background information. The scope of research access, paper illustration, and public media redistribution is handled separately. The main Ethics Statement describes the conditions governing data release.

E.2 Available Real results

Model	Correct / matched records	Accuracy
GPT-4o	54 / 237	22.78%
Gemini 2.5 Flash	62 / 237	26.16%
Qwen2.5-VL-7B	5 / 237	2.11%
PIWM	264 / 418	63.16%

Table 10: Available results matched to Real records.

These are mapped, record-weighted results from the existing real-store evaluation. The table reports each model's available coverage separately.

E.3 Reproduction record structure

The experimental organization links task records, reference labels, predictions, and scores by example identifier. A Full trace contains the input observation identifier, predicted current state, each candidate forecast, the final decision input, the raw decision output, and the scored action. Ablation identifiers distinguish the four inference configurations.

The following record map specifies the information needed to associate each result with its evaluation unit.

E.4 Anonymous materials

Anonymous materials are hosted at:

`https://anonymous.4open.science/r/piwm-review-materials-2026-63BD/`

The hosted snapshot contains previously released data and tools. Its historical results are separate from the current evaluation tables in this paper. Snapshot updates require the current split manifests, configurations, and prediction files.

Evaluation	Records	Linked information
See	209	Observation, stage and intent references, predicted state
Foresee	309	Current state, candidate action, next-stage reference and prediction
Act	245	Textual reference state, candidate actions, chosen action
Full	509 per configuration	Observation, state, forecasts, decision, reference action
Real	418	Real-store observation, expert reference action, decision

Table 11: Record-level links for reproducing the current evaluation.

Anonymous identifiers should remain consistent across inputs, labels, and predictions. Media paths are relative to the release directory, while private reverse mappings and credentials remain outside the anonymous package.