See, Infer, Intervene: Proactive World Modeling for Goal-Oriented Social Intelligence

Anonymous ACL submission

Abstract

Multimodal retail agents should not only recognize what a customer is doing, but also decide whether and how to assist before an explicit request is made. We study this setting through the See–Infer–Intervene (SII) framework, where a device must see pre-interaction behavior, infer latent customer intent, and act by selecting an appropriate service intervention or choosing to wait. We instantiate SII with the Proactive Intent World Model (PIWM), which represents customer state with AIDA (Attention, Interest, Desire, Action) purchasing phases and BDI (belief, desire, intention) psychological fields, predicts action-conditioned intent transitions, and selects from five response classes: GREET, ELICIT, INFORM, RECOMMEND, and HOLD. We further construct GuidanceSalesBench, a smart-retail benchmark containing state manifests, pre-interaction videos, candidate responses, action-conditioned outcomes, and best-action labels. When conditioned on ground-truth customer state to isolate action selection, PIWM achieves 0.641 macro F1 on 30 held-out target videos, outperforming a zero-shot Qwen2.5-VL-7B baseline and training variants without balanced action supervision; end-to-end video-only selection drops to 0.295, below the 5-class balanced random baseline of 0.414, identifying video-to-state grounding as the dominant deployment-time bottleneck. A preliminary staged real-store pilot (recorded with paid participants performing scripted customer behaviors) reaches 0.579 action macro F1 on 20 fully annotated videos, with 10 additional accessible videos released with index-level labels.

1 Introduction

Intelligent retail systems are increasingly expected to provide assistance before customers issue explicit requests (Deng et al., 2023; Yang et al., 2025). This setting differs from conventional command-driven interaction: a customer may pause in front of a device, compare products, shift gaze across shelves, approach the screen, or hesitate before purchase, but none of these behaviors is a click, scan, or verbal query. A system that waits passively can miss the service window, whereas a system that intervenes too early can interrupt the customer or create sales pressure (Bailey et al., 2000, 2001). The central problem is therefore not only recognition, but proactive decision making: from a short pre-interaction observation, the agent must infer the customer's state and decide whether an intervention would be helpful (Novick and Sutton, 1997; Horvitz, 1999).

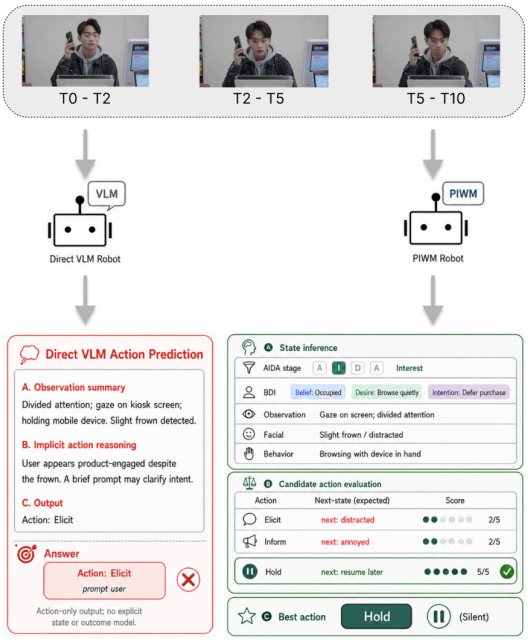

Figure 1: Motivating example for proactive retail intervention.

Current multimodal models(Lin et al., 2024; Maaz et al., 2024) are strong observers but are not designed as proactive retail agents(Mathur et al., 2024). Given an image or video, they can describe visible behavior or answer questions, yet this descriptive ability does not directly specify what the device should do next. Proactive assistance re-

quires three additional capabilities: modeling latent customer intent beyond visible motion, comparing multiple candidate responses, and predicting how each response may change the customer's future state. Without this action-conditioned reasoning, a model may produce plausible descriptions while still choosing an intrusive or mistimed service action(Yao et al., 2022b).

We organize this problem as an SII framework: a device should *see* customer behavior, *infer* latent intent, and *intervene* only after evaluating candidate responses. The key technical view is to treat proactive retail assistance as intent-level world modeling. World models predict future states under candidate actions (Ha and Schmidhuber, 2018; Hafner et al., 2019), and recent World Action Models connect state prediction with action generation (Wang et al., 2026; Ye et al., 2026). However, existing work mainly targets physical control, where actions are motor commands and outcomes are physical state changes. In retail interaction, actions are communicative service moves, and the relevant outcomes are changes in intent, uncertainty, receptiveness, and purchase readiness.

We instantiate SII with the **Proactive Intent World Model (PIWM)**. Given a pre-interaction observation, PIWM estimates a structured customer state, predicts the outcome of each candidate device response, and selects the best response from a constrained intervention set. The state combines AIDA purchasing phases with BDI psychological fields (Barry, 1987; Rao et al., 1995). The action space contains five device responses: GREET, ELICIT, INFORM, RECOMMEND, and HOLD. Treating HOLD as an action is important because non-intervention can be the correct decision when the customer is browsing smoothly.

To train and evaluate this formulation, we construct **GuidanceSalesBench**, a smart-retail benchmark that records the structured decision process rather than only the observed scene (Rouhi et al., 2026; Singh et al., 2016). Each instance links a customer state manifest, a pre-interaction video prompt and video, an AIDA-conditioned candidate response set, action-conditioned outcomes, and a best-action label. Experiments on Target-Test (n=30, 30 held-out target-domain videos) show that PIWM reaches 0.641 macro F1 for best-action selection when given the ground-truth customer-state representation, compared with 0.313 for Qwen2.5-VL-7B (zero-shot) and 0.390 for a Small-Scale Action Training variant; without oracle state, the Target-Test E2E setting reaches 0.295. We additionally evaluate transfer on a real-store pilot: 30 accessible paid-participant videos are released as sparse frames, 20 have full session-level annotations for the primary sim-to-real metric, and 10 have index-level labels only.

Our contributions are threefold. First, we formulate multimodal retail marketing as an SII problem that requires seeing customer behavior, inferring intent, and acting through low-intrusion response selection. Second, we propose PIWM, an intent-level world model that predicts action-conditioned customer-state transitions over an AIDA–BDI schema and a five-class device action space. Third, we introduce GuidanceSalesBench and an evaluation protocol for state prediction, outcome simulation, and proactive action selection.

2 Related Work

2.1 World Models and World Action Models

World models learn environment representations and transition dynamics for prediction, planning, and policy learning. Ha and Schmidhuber (Ha and Schmidhuber, 2018) learn a latent model that compresses visual observations and predicts future latent states with recurrent dynamics, while Dreamer (Hafner et al., 2019) learns behaviors from imagined latent trajectories. Recent World Action Models further combine dynamics modeling with action generation (Wang et al., 2026; Ye et al., 2026), targeting embodied AI, robotic manipulation, autonomous driving, and other physical control tasks.

PIWM also uses action-conditioned future simulation, but differs in task grounding. Vision-language-action systems such as RT-2 (Brohan et al., 2023), OpenVLA (Kim et al., 2024), and π_0 (Black et al., 2024) map observations and instructions to motor actions. PIWM instead targets communicative service interventions in a single-customer retail setting, where success is best response selection under customer-state constraints and non-intervention is explicitly represented as an action.

2.2 Proactive Interaction and Mixed-Initiative Systems

Proactive interaction studies systems that initiate assistance before explicit user requests(Bi et al., 2024). Related work includes proactive dialogue systems(Yan et al., 2023), context-aware agents,

and mixed-initiative interaction, where control is shared between user and system (Novick and Sutton, 1997). A key challenge is timing: assistance can reduce uncertainty or effort, but poorly timed interventions may interrupt ongoing tasks or increase cognitive load (Bailey et al., 2000).

PIWM formalizes this timing problem as structured selection over simulated intervention outcomes. Rather than directly triggering an action from the current context, PIWM predicts the next customer state under each candidate response and assigns a relative action score. Compared with broad proactive assistants such as ContextAgent (Yang et al., 2025), PIWM targets a narrower but more decision-critical setting: selecting which low-intrusion service move should be used for a third-party customer observed from retail video.

2.3 Intent Modeling in Human-Centered Interaction

Intent recognition typically infers user goals from text, speech, behavior sequences, or visual cues(Rabinowitz et al., 2018). In retail environments, intent is rarely explicit and must be inferred from dwelling, gaze movement, comparison, approaching, hesitation, and device operation. We represent customer intent with AIDA for macro-level purchasing phase (Barry, 1987) and BDI for micro-level psychological state (Rao et al., 1995).

Unlike static intent recognition, PIWM uses intent as a transition state for action-conditioned simulation. The model predicts how candidate interventions change the customer's purchasing phase and psychological state, making intent representation part of the decision process rather than only an output label.

2.4 Retail Behavior Understanding and Smart Retail Datasets

Retail behavior understanding has been studied through shopping action recognition, item-interaction detection, trajectory analysis, and multi-view video understanding. The MERL Shopping Dataset focuses on behaviors such as browsing, picking up items, and returning items (Singh et al., 2016), while PRISM provides richer multi-view retail environments for embodied vision-language understanding (Rouhi et al., 2026).

These datasets primarily describe what customers do or what occurs in the scene. Shopping-agent benchmarks such as WebShop (Yao et al., 2022a) and Shopping MMLU (Jin et al., 2024) focus on product navigation or knowledge reasoning. PIWM targets a different decision layer: how a smart retail device should intervene given an observed customer state, so GuidanceSalesBench includes candidate interventions, predicted next states, relative action scores, and best-action labels.

3 Problem Definition

We study proactive intervention selection for smart retail devices. Given a short pre-interaction video V_t, the system predicts a structured customer intent state, simulates the outcomes of candidate device responses, and selects an action before an explicit request is made.

3.1 Intent State

At timestamp t, the intent state is defined as

$$s_t = (g_t, b_t, x_t), \tag{1}$$

where $g_t \in \{\text{attention}, \text{interest}, \text{desire}, \text{action}\}$ is the AIDA purchasing phase, $b_t = (\text{belief}_t, \text{desire}_t, \text{intention}_t)$ is the BDI state, and x_t describes observable evidence such as behavior, expression, posture, and short-term temporal events.

3.2 Intervention Outcome

Given s_t, the device considers a state-dependent candidate set $\mathcal{A}(s_t) = \{a_1, \ldots, a_K\}$. Each a_i is a high-level service response from $\{\textsc{Greet}, \textsc{Elicit}, \textsc{Inform}, \textsc{Recommend}, \textsc{Hold}\}$. For each candidate, PIWM predicts

$$o_t(a_i) = (g_{t+1}^{(i)}, b_{t+1}^{(i)}, y_i), \tag{2}$$

where $g_{t+1}^{(i)}$ and $b_{t+1}^{(i)}$ are the predicted next intent state, and $y_i \in \{1, 2, 3, 4, 5\}$ is a relative action score. A score of 5 denotes the best candidate for the current state.

3.3 Low-Intrusion Decision

The selected response is the highest-scoring candidate:

$$a_t^* = \arg\max_{a_i \in \mathcal{A}(s_t)} y_i. \tag{3}$$

The score is not a free-form preference statement: it is assigned under a fixed rubric in which only one candidate receives 5, and all candidates must be valid for the current AIDA phase. This formulation allows HOLD to be selected when intervention is predicted to be unnecessary or disruptive.

4 Proactive Intent World Model

We use AIDA (Barry, 1987) and BDI (Rao et al., 1995) because they provide a compact state space for pre-interaction retail assistance. AIDA separates pre-commitment stages from purchase action, while BDI captures inferred uncertainty, preference, and action readiness. Both schemas are textual and schema-constrained, making them suitable intermediate targets for VLM training and diagnostic evaluation.

PIWM is the modeling component of SII. It turns the see–infer–intervene view into a schema-constrained prediction and decision pipeline:

$$\hat{s}_t = f_\theta(V_t), \tag{4}$$

$$\hat{o}_t(a_i) = h_\phi(\hat{s}_t, a_i), \quad a_i \in \mathcal{A}(\hat{s}_t), \tag{5}$$

$$a_t^* = \pi(\{\hat{o}_t(a_i)\}_{a_i \in \mathcal{A}(\hat{s}_t)}). \tag{6}$$

Figure 2 gives an overview of the GuidanceSalesBench construction and PIWM training pipeline, from scenario seeds to AIDA–BDI manifests, candidate actions, counterfactual outcomes, verification, and the final SFT training set.

Here, f_θ estimates customer state, h_ϕ predicts action-conditioned outcomes, and π selects the candidate with the highest predicted score. We implement both worlds with a single Qwen2.5-VL-7B backbone adapted by LoRA(Bai et al., 2025). The same adapter is trained to model a user-intent world, which maps sparse video frames to AIDA, intent labels, and BDI state, and a sales-action world, which predicts the next customer state under a candidate action.

Table 1: PIWM structured outputs and AIDA-conditioned candidate actions.

Type	Item	Description or values
State	$\hat{g}_t$	AIDA phase
State	$\hat{b}_t$	BDI state
State	$\hat{x}_t$	Observable evidence
Action	a_i	Candidate action class
Outcome	$\hat{g}_{t+1}^{(i)}$	Next AIDA phase
Outcome	$\hat{b}_{t+1}^{(i)}$	Next BDI state
Outcome	$\hat{y}_i$	Relative action score in 1–5
Decision	a_t^*	Selected response
Candidates	attention	GREET, ELICIT, INFORM, HOLD
Candidates	interest	ELICIT, INFORM, RECOMMEND, HOLD
Candidates	desire	INFORM, RECOMMEND, HOLD
Candidates	action	GREET, RECOMMEND, HOLD

4.1 State Estimation and Candidate Set

The state estimator separates observable evidence from inferred mental state. Visual cues such as gaze shifts and repeated comparison are represented in $\hat{x}_t$, while uncertainty, preference, or purchase intention is represented in $\hat{b}_t$. Given $\hat{s}_t$, PIWM constructs candidates from a finite inventory $\mathcal{R} = \{$GREET, ELICIT, INFORM, RECOMMEND, HOLD$\}$. The candidate set is constrained by AIDA phase, and HOLD is always included as a non-intervention baseline.

4.2 Outcome Modeling and Selection

For each candidate response, h_ϕ predicts the next AIDA phase, the next BDI state, and a score from 1 to 5. This differs from direct action classification, which estimates $p(a \mid V_t)$ or $p(a \mid s_t)$ without modeling intervention-specific consequences(Ahn et al., 2022). PIWM instead exposes the intermediate future state under each candidate and then selects the candidate whose outcome receives the highest score.

The selector π is deterministic: it returns the candidate with score 5. Low-intrusion behavior is encoded through the candidate inventory and scoring rubric rather than an additional hand-designed cost function. For example, HOLD can receive the highest score when the predicted intervention would disturb smooth browsing, while ELICIT or INFORM can be preferred when the customer shows uncertainty or an information gap.

PIWM is best understood as a world-model-augmented proactive action policy. The intent-level world-model objective h_ϕ shapes the policy during training by supervising state estimation and action-conditioned intent transitions, while the learned policy remains the strongest decision-time component in the current system. Our empirical analysis (Section 6.5) shows that direct action selection consistently outperforms inference-time counterfactual planning over predicted outcomes; we therefore retain the structured formulation in Equations (4)–(6) primarily because it (1) supports the dataset construction protocol (Section 5), (2) enables intermediate-state diagnostics (Appendix B), and (3) provides a foundation for future work where stronger next-state prediction could make decision-time planning competitive.

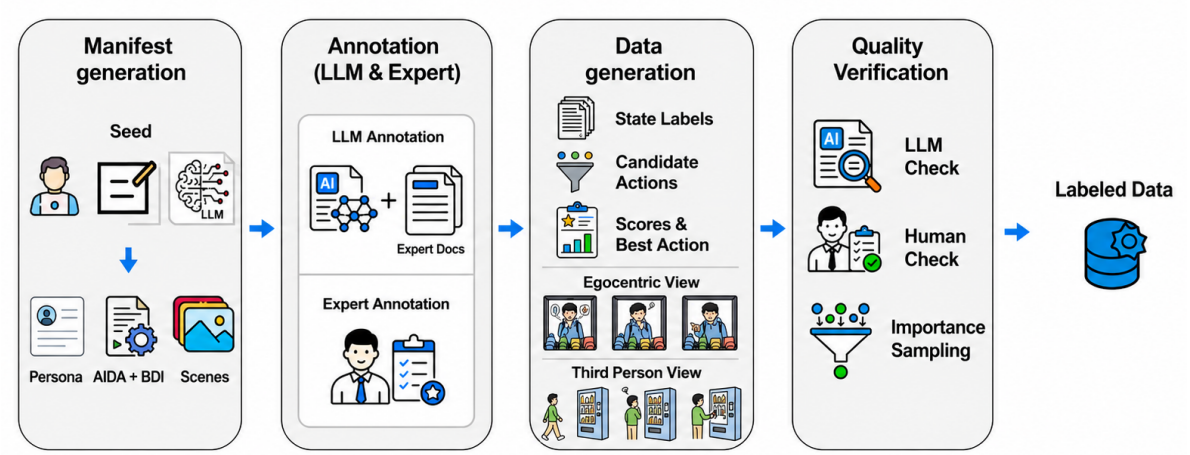

Figure 2: Overview of GuidanceSalesBench construction and PIWM training. The pipeline generates AIDA–BDI manifests, candidate actions, action-conditioned outcomes, and best-action labels before producing the final SFT training set.

5 GuidanceSalesBench Construction

GuidanceSalesBench provides supervision for state estimation, outcome prediction, and response selection. It is constructed through two linked chains:

$$\text{seed} \rightarrow \text{manifest} \rightarrow \text{prompt} \rightarrow \text{video}, \quad (7)$$

$$\text{manifest} \rightarrow \text{candidate actions} \rightarrow \text{outcomes} \rightarrow \text{scores} \rightarrow \text{best action}. \quad (8)$$

Manifests define customer states, prompts and videos provide pre-interaction visual inputs, and labeled outcomes provide action-conditioned supervision.

5.1 State and Video Generation

Each instance starts from a natural-language seed describing a retail scenario. We use schema-constrained structured generation to produce a manifest containing the customer persona, visual appearance, AIDA phase, BDI state, observable behavior, facial expression, posture, and a 10-second timeline. In dataset construction, this generator instantiates f_θ; in experiments, the same schema becomes the target for video or text-based state prediction.

Each manifest is rendered into a 10-second video prompt. The prompt describes only the pre-interaction state and excludes candidate actions, scores, and best-action labels, preventing the generated video from leaking the target response.

5.2 Candidate Actions and Outcome Labels

Actions are represented by five class labels rather than free-form text: GREET, ELICIT, INFORM, RECOMMEND, and HOLD. For each manifest, the annotation system first selects the candidate set allowed by the current AIDA phase, always including HOLD. A prompt-based structured annotator then predicts each candidate's next AIDA phase, next BDI state, and relative score.

The score is a coarse expert label assigned under a fixed rubric. It ranges from 1 to 5: 5 denotes the best action for the current state, 4 denotes a useful but less fitting action, 3 is neutral, 2 is weak, and 1 is inappropriate or disruptive. The annotation validator enforces that every candidate has exactly one outcome, exactly one candidate receives score 5, and the best-action label matches that candidate.

Table 2: Dataset and supervision sources.

Source	Role	Count
Sales rules	action design	78
General synthetic scenes	state/outcome data	543
General multitask samples	multitask SFT	2,544
Target train videos	action SFT	71
Target test videos	evaluation	30
Greet augmentation	balancing	15
Balanced action data	final action SFT	190
Real-store pilot videos	transfer set	30
fully annotated	sim-to-real metric	20
index-labeled	auxiliary release	10
Missing-video candidates	documented only	20

5.3 Statistics and Validation

Table 2 summarizes the data sources used for training and evaluation. The general synthetic retail data provides supervision for customer understanding and action consequence prediction, while the target vending-machine data supports action selection and evaluation in the deployment scenario.

The action-world supervision is counterfactual by construction. Across 543 retail scenes, we generate 2,001 scene-action samples, with an average of 3.685 candidate actions per scene and full candidate-set coverage for each scene. We apply validation at three stages: manifests must contain complete AIDA, BDI, observable behavior, and timeline fields; video prompts must exclude action labels; and outcome labels must satisfy candidate-count, HOLD-baseline, stage-allowed action, score-range, and best-action consistency constraints. Additional dataset distributions and real-store frame examples are provided in Appendix A and Appendix G.

6 Experiments

We evaluate whether the SII formulation improves proactive retail action selection. The experiments answer three questions: whether target-domain action training improves response selection, whether AIDA-conditioned candidate filtering is useful, and where the remaining errors arise.

6.1 Experimental Setup

The main GuidanceSalesBench evaluation contains Target-Test (n=30) in the smart vending-machine setting, with 6 examples for each action class. The 30 test samples are stratified to achieve balanced best-action class distribution while preserving an AIDA stage distribution representative of the target deployment (Appendix A). Each example is manually QA-checked and represented by three sparsely sampled frames. We report macro F1 over the five action classes, which avoids allowing frequent actions to dominate the score. The real-store pilot is reserved for transfer evaluation and is not used for training; its 20 fully annotated videos form the primary sim-to-real metric, and 10 additional index-labeled videos are released as auxiliary real data.

Training has two stages. The first stage trains the shared LoRA adapter on 2,544 multitask samples derived from 543 general retail scenes, covering customer understanding and action consequence prediction. The second stage continues training on 190 balanced target-domain action-selection samples. These 190 samples combine 71 original target examples, 15 Greet augmentations, and 104 balanced samples from the general pool to reduce action-class imbalance. Each main training run takes approximately 9 A100-hours on A100 80GB hardware; full configuration details (LoRA hyperparameters, optimizer, decoding) are provided in Appendix D.

6.2 Tasks and Metrics

We evaluate four linked tasks in the PIWM pipeline. AIDA prediction measures whether the model identifies the customer's sales-funnel stage. Intent prediction measures whether it captures core customer intent. Next-state prediction measures whether it predicts how a candidate action changes the customer's state. Best-action prediction measures whether it selects the correct response among GREET, ELICIT, INFORM, RECOMMEND, and HOLD.

Our primary metric is macro F1 for best-action selection on Target-Test (n=30). We additionally report per-action F1 and collapse HOLD versus the other four actions to evaluate the binary intervention decision.

6.3 Baselines

We compare PIWM against Random Baseline and Constant-Action Baselines, Qwen2.5-VL-7B (zero-shot) without project-specific training, the State-Outcome Model trained only on customer state and action consequences, Small-Scale Action Training variants, and three closed-source baselines evaluated through the same role-separated prompts. Table 3 shows that balanced target-domain action training yields the strongest Cross-Domain oracle performance.

PIWM achieves 0.641 macro F1 with oracle customer state on Target-Test and 0.734 on Cross-Domain, outperforming Qwen2.5-VL-7B (zero-shot) and the State-Outcome Model. In the video-only E2E setting, performance drops to 0.295, showing that state grounding is the main bottleneck. On the fully annotated Real-Store Pilot, PIWM reaches 0.579 action macro F1, providing preliminary evidence that the learned action selection capability is not confined to synthetic data; however, this subset contains no HOLD cases and therefore does not evaluate real-world non-intervention decisions.

The closed-source comparison is mixed across settings. Gemini 2.5 Flash slightly exceeds PIWM on Target-Test oracle (0.659 vs. 0.641), while PIWM exceeds all three 302ai closed-source baselines on Cross-Domain oracle (0.734 vs. 0.638). GPT-4o is stronger on Target-Test E2E (0.354 vs. 0.295), and Claude Sonnet 4.6 is stronger on Real-

Table 3: Best-action macro F1 across evaluation scenarios. Oracle uses ground-truth customer state; E2E predicts state from video. Closed-source baselines are evaluated via 302ai with the same prompts; GPT-5.5 is a diagnostic Codex CLI reference.

Model	Target-Test (n=30) Oracle	Cross-Domain (n=60) Oracle	Target-Test E2E (n=30)	Real-Store Pilot (n=20)
Random Baseline	0.414	0.466	0.414	-
Constant-Action Baselines	0.067	-	-	-
Qwen2.5-VL-7B (zero-shot)	0.313	0.142	-	0.111
State-Outcome Model	0.240	0.259	-	0.217
Small-Scale Action Training	0.390	-	-	-
Gemini 2.5 Flash	0.659	0.638	0.216	0.655
GPT-4o	0.445	0.503	0.354	0.535
Claude Sonnet 4.6	0.579	0.569	0.294	0.833
GPT-5.5 (Codex CLI, diagnostic)	0.622	0.648	-	-
PIWM (full)	0.641	**0.734**	0.295	0.579

Note: Real-Store Pilot uses 20 fully annotated videos; 10 index-labeled and 20 inaccessible planned videos are excluded.

Store Pilot (0.833 vs. 0.579), indicating that scale and model family still matter for video grounding and real-store transfer.

The State-Outcome Model drops from 0.641 to 0.240 on Target-Test, while the Small-Scale Action Training variant reaches 0.390. This comparison suggests that world-model supervision is most useful when paired with balanced target-domain action supervision. GPT-5.5 through the Codex CLI agent surface is included only as a diagnostic reference because its agent interface is not directly comparable to API-level baselines.

6.4 Action-level Analysis

The best model performs well on active intervention actions, with parsed-output F1 scores of 0.800 for GREET, 0.750 for ELICIT, 0.769 for INFORM, and 0.600 for RECOMMEND. HOLD remains the weakest class (parsed-output F1 0.286; strict F1 0.250), and most HOLD errors are over-interventions, especially predictions of ELICIT. When HOLD is collapsed into non-intervention and the other actions into intervention, binary macro F1 is 0.592.

6.5 Ablation Studies

We perform ablations on the main design choices: action-selection supervision, AIDA-conditioned candidate filtering, inference-time counterfactual planning, and oracle versus end-to-end state conditioning. Table 4 summarizes the results.

Table 4: Compact ablation results on Target-Test (n=30).

Variant	Macro F1
No action-selection training	0.240
71 target + 15 GREET augmentation	0.390
Balanced 190-sample supervision (PIWM full)	**0.641**
Choose freely from all five actions	0.504
AIDA-conditioned candidates (PIWM full)	**0.641**
AIDA candidates + HOLD bias correction	0.633
Direct action selection (PIWM full)	**0.641**
Counterfactual Planning (stage-reward)	0.171
Counterfactual Planning (model-reward)	0.265
State-Outcome Model + CF Planning (model-reward)	0.418
Oracle state	**0.641**
End-to-end	0.295

Balanced action supervision improves macro F1 from 0.390 to 0.641, and AIDA-conditioned candidates improve over unconstrained five-action selection (0.504 to 0.641). Inference-time counterfactual planning underperforms direct action selection, suggesting that the current world-model objective is more effective as training-time supervision than as a standalone decision-time planner. The E2E drop from 0.641 to 0.295 confirms that video-to-state grounding remains the main bottleneck. Additional schema ablations in Appendix B show that removing BDI weakens action selection and next-state prediction, while removing observable evidence improves oracle action selection but hurts E2E and transition metrics.

6.6 Sim-to-Real Pilot Evaluation

To assess transfer from synthetic training data to real footage, we evaluate PIWM on a small staged real-store pilot recorded by paid employee participants enacting scripted customer-behavior scenarios. The primary real-store metric uses 20 fully annotated videos, while 10 additional accessible videos have index-level labels only. As shown in

Table 3, PIWM reaches 0.579 action macro F1 on Real-Store Pilot, compared with 0.217 for the State-Outcome Model and 0.111 for Qwen2.5-VL-7B (zero-shot). Stage prediction remains challenging, and the fully annotated subset contains no HOLD cases; we therefore treat this result as preliminary evidence of sim-to-real transfer rather than a complete real-store benchmark.

7 Conclusion

This paper studies multimodal retail marketing as an SII problem: a device must see customer behavior, infer latent intent, and intervene before the customer makes an explicit request. We instantiate this view with PIWM, an intent-level world-model-augmented proactive action policy that learns AIDA–BDI customer state estimation and action-conditioned intent transition prediction during training, and selects from an AIDA-conditioned response set at decision time. Experiments on held-out target videos show that balanced action supervision and candidate constraints substantially improve best-action macro F1. Specifically, PIWM reaches 0.641 macro F1 on Target-Test ($n = 30$) with ground-truth state, 0.295 in the end-to-end video-only setting, and 0.579 on a 20-video staged real-store pilot, with the world-model training framework providing a substantial gap over the State-Outcome baseline (0.240).

We also introduce GuidanceSalesBench, a structured benchmark that links pre-interaction observations with state manifests, candidate interventions, predicted next states, relative action scores, and best-action labels. Comparisons with closed-source models show a mixed but informative picture: PIWM is strongest on Cross-Domain oracle selection, while larger models remain stronger in some E2E and real-store settings. The current formulation focuses on short single-customer clips and a finite set of high-level device responses. Extending PIWM to longer interactions, richer multimodal device behaviors, and online feedback from real deployments is an important direction for future work.

Limitations

PIWM has several limitations.

1. **Oracle-state dependence.** The strongest best-action result conditions on ground-truth customer state to isolate response selection. End-to-end video-only performance drops from 0.641 to 0.295, showing that video-to-state grounding remains the main bottleneck.

2. **Closed-source models remain stronger in some settings.** PIWM is competitive on oracle-state action selection, but it does not uniformly exceed larger closed-source systems: Gemini 2.5 Flash is slightly higher on Target-Test oracle, GPT-4o is stronger in E2E selection, and Claude Sonnet 4.6 is stronger on the small real-store pilot.

3. **Non-intervention remains difficult.** HOLD is the weakest class on Target-Test, indicating that deciding when not to intervene is harder than choosing among active service actions. Larger HOLD-rich evaluation sets are needed to measure low-intrusion behavior more reliably.

4. **The real-store pilot is preliminary.** The Real-Store Pilot uses staged recordings from paid employee participants and only 20 fully annotated videos. The subset contains no HOLD cases, so real-world non-intervention behavior and naturalistic customer variation remain open evaluation targets.

Ethics Statement

This work involves video recordings of human participants performing retail customer behaviors and computational modeling of inferred customer states. We address ethical considerations as follows.

Data sources and participant recruitment. The synthetic GuidanceSalesBench training data is generated entirely by text-to-video models from scenario seeds and contains no real individuals. The real-video evaluation set (Real-Store Pilot, $n = 20$ fully annotated plus 10 group-labeled videos) was collected through a collaborating retailer. Recordings were performed by paid employees of the retail partner, working from scripted customer-behavior scenarios that the research team prepared in advance. Participants were not actual store customers, and the videos are staged enactments rather than naturalistic observation of real customer behavior. Approximately 5–10 individual employees participated, with each individual performing multiple scripted scenarios.

Informed consent and compensation. All participants were informed in advance that recordings would be used for research on proactive retail assistance, including the development of vision-language models for customer-state inference and action selection. Participants signed release forms covering both the use of their image in research and the public release of the recorded materials. Participants were compensated for their time. We do not release any biographical information, and individual identification beyond what is visible in the released video frames is not retained or distributed.

Ethics review. Formal IRB-style ethics review was not conducted; the project followed the compliance practices of the collaborating retail partner, including the release-form procedure described above. We acknowledge this as a limitation: future iterations of this work, particularly any deployment study involving real customers, will require formal ethics-board approval.

Data release. The synthetic GuidanceSalesBench dataset, including all customer manifests, candidate actions, and action-conditioned outcomes, will be released publicly under an open license upon acceptance. The 30 real-store videos and the frame-and-annotation subset used for sim-to-real evaluation will also be released, in accordance with the participant release forms. Code is made available anonymously during review (see Data and Code Availability).

Potential risks of proactive sales intervention. PIWM is designed to select from a constrained set of service responses, including HOLD (non-intervention), explicitly to discourage over-intervention. Nevertheless, deploying proactive sales systems may exert commercial pressure on customers, particularly when intervention is triggered before explicit interest. We caution that any deployment of PIWM-like systems should: (1) preserve the customer's ability to ignore or dismiss interventions, (2) avoid leveraging psychological vulnerabilities such as decision fatigue or indecision, and (3) provide transparency that the customer is interacting with an AI-driven assistant rather than human staff.

Intent inference and psychological profiling. PIWM infers AIDA stage and BDI (belief/desire/intention) state from short pre-interaction observations. These inferences are intentionally coarse-grained—four AIDA classes and short schema-constrained BDI fields—and serve as intermediate signals for action selection rather than as persistent customer profiles. We do not retain inferred state across sessions or link inferred state to identity in our evaluation setup; any downstream deployment must commit to the same constraint.

Demographic representativeness. Our real-video set covers a limited demographic range, drawing from the employees available at one retail partner. We have not evaluated PIWM for fairness across gender, age, ethnicity, or disability. Broader fairness evaluation is required before any deployment and is left as future work.

Compute footprint. Each main PIWM training run consumes approximately 9 A100-hours on A100 80GB hardware. The complete project (main run plus four ablations) totals approximately 45 A100-hours, substantially below the pretraining cost of foundation vision-language models.

Data and Code Availability

The synthetic GuidanceSalesBench dataset, including all customer manifests, prompts, candidate actions, and action-conditioned outcomes, will be released publicly under an open license upon acceptance. The real-store video benchmark was recorded with paid participants who consented to research use and release. Anonymous code, configuration, prompt templates, per-video frames for 30 accessible real-store videos, full session-level annotations for 20 videos, group-level index labels for 10 videos, and a manifest documenting 20 planned videos with missing files are available at `https://anonymous.4open.science/r/piwm-emnlp2026-anonymous-D692/`.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Brian P Bailey, Joseph A Konstan, and John V Carlis. 2000. Measuring the effects of interruptions on task performance in the user interface. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, volume 2, pages 757–762. IEEE.

Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Interact*, volume 1, pages 593–601.

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

Shuxian Bi, Wenjie Wang, Hang Pan, Fuli Feng, and Xiangnan He. 2024. Proactive recommendation with iterative preference guidance. In *Companion Proceedings of the ACM Web Conference 2024*, pages 871–874.

Kevin Black, Noah Brown, Danny Driess, Chelsea Finn, Karol Hausman, Brian Ichter, Sergey Levine, Karl Pertsch, Lucy Xiaoyang Shi, Quan Vuong, Ury Zhilinsky, and 1 others. 2024. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750*.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

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

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, and 1 others. 2024. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. *arXiv preprint arXiv:2410.20745*.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in ai agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560.

David G Novick and Stephen Sutton. 1997. What is mixed-initiative interaction. In *Proceedings of the AAAI spring symposium on computational models for mixed initiative interaction*, volume 2, page 12.

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

Anand S Rao, Michael P Georgeff, and 1 others. 1995. Bdi agents: from theory to practice. In *Icmas*, volume 95, pages 312–319.

Amirreza Rouhi, Parikshit Sakurikar, Satya Sai Reddy, Narsimha Menga, Anirudh Govil, Sri Harsha Chittajallu, Rajat Aggarwal, Anoop Namboodiri, and Sashi Reddi. 2026. Prism: A multi-view multi-capability retail video dataset for embodied vision-language models. *arXiv preprint arXiv:2603.29281*.

Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, and 1 others. 2026. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*.

Cen Yan, Jun Bai, Yanmeng Wang, Wenge Rong, Yuanxin Ouyang, and Zhang Xiong. 2023. Goal-oriented conditional variational autoencoders for proactive and knowledge-aware conversational recommender system. *Computer Speech & Language*, 79:101468.

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

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, and 1 others. 2026. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*.

A Additional Dataset Statistics

Tables 5 and 6 report the detailed distributions used to contextualize GuidanceSalesBench. The balanced action-selection set addresses the raw target split's absence of HOLD examples by supplementing the 71 target examples and 15 GREET augmentations with balanced general-domain samples.

Table 5: AIDA distribution by data source.

Source	Att.	Interest	Desire	Action	Total
General Synthetic	49	289	128	77	543
Target Train	18	31	17	5	71
Target-Test	3	10	11	6	30
Balanced Action	63	60	44	23	190

Table 6: Best-action distribution by data source.

Source	Greet	Elicit	Inform	Recom.	Hold	Total
Target Train	11	14	41	5	0	71
Balanced Action	26	41	41	41	41	190
Target-Test	6	6	6	6	6	30
Real-Store Pilot	5	5	5	5	0	20

The remaining coverage statistics are as follows. General Synthetic contains 177 `confirm_choice`, 94 `explore_options`, 16 `compare_value`, and 256 other intent labels. Target-Test contains 6 `confirm_choice`, 15 `explore_options`, 6 `compare_value`, and 3 `no_clear_intent` examples. Candidate-set sizes over the 543 general scenes are distributed as 87 two-action, 59 three-action, 335 four-action, and 62 five-action sets, with mean size 3.685. The user-intent and next-state-prediction tasks overlap on all 543 scenes, while 110 scenes also appear in the five-action action-selection task.

B Additional Experimental Diagnostics

Tables 7–10 provide diagnostic numbers omitted from the main text for space. These diagnostics support the same conclusion as the main ablation: action selection benefits from balanced supervision and candidate constraints, while video-to-state grounding remains the main end-to-end bottleneck.

Table 7: PIWM per-action F1.

Split	Greet	Elicit	Inform	Recom.	Hold
Target-Test	0.800	0.750	0.769	0.600	0.286
Cross-Domain	0.800	0.905	0.571	0.667	0.727
Real-Store	0.500	0.545	0.778	0.778	N/A

Table 8: State and transition diagnostics.

Model	AIDA	Intent	Next-State
Random Baseline	0.235	0.138	0.233
Qwen2.5-VL-7B	0.190	0.053	0.108
State-Outcome	0.349	0.111	**0.587**
PIWM	**0.350**	**0.114**	0.565

Table 9: Schema ablations.

Variant	Target	Cross	E2E	AIDA	Intent	Next
Full PIWM	0.641	0.734	0.295	0.350	0.114	0.565
No BDI	0.526	0.693	0.219	0.428	0.120	0.163
No evidence	0.778	0.796	0.229	0.370	0.116	0.205

Table 10: Frame-count ablations.

Frames	Target	Cross	E2E	AIDA	Intent	Next
1 frame	0.589	0.723	0.189	0.370	0.074	0.165
3 frames	0.641	0.734	0.295	0.350	0.114	0.565
5 frames	0.627	0.770	0.234	0.412	0.115	0.222

Removing BDI weakens action selection and next-state prediction. Removing observable evidence improves oracle-state action F1 but hurts E2E and transition metrics, so we treat this result as a prompt/supervision artifact rather than a recommended design change.

C Prompt Templates

All prompts use the following system message.

```
You are a multimodal sales-guidance agent trained on retail
    pedagogy.
You observe customers in physical retail stores via a
    streaming camera
```

```
and decide whether and how to intervene. Always output your reasoning
in the structured tag format requested by the user prompt. Do not output
free-form prose outside the requested tags.
```

C.1 State Estimation Prompt

Current code source: `piwm_train/prompts.py::build_user_intent_prompt`.

```
You are observing a short customer-behavior window.

Scene: {scene}

Below are {n_frames} frames sampled from a streaming camera, in chronological order.

<image>
<image>
<image>

Infer the customer's current user state from the visible behavior only. Do not choose a sales action. Do not output candidate actions, rewards, or recommendations.

Output the following fields, in this exact order, each on its own line:
<stage>...</stage>
<intent_label>...</intent_label>
<visual_summary>...</visual_summary>
<engagement_pattern>...</engagement_pattern>
<gaze_and_attention>...</gaze_and_attention>
<body_and_hands>...</body_and_hands>
<belief>...</belief>
<desire>...</desire>
<intention>...</intention>

- stage must be one of: attention, interest, desire, action.
- intent_label must be one existing PIWM intention category, such as confirm_choice, explore_options, request_demonstration, seek_reassurance, negotiate_price, leave_without_purchase, compare_value_for_money, or no_clear_intent.
- visual_summary, engagement_pattern, gaze_and_attention, and body_and_hands must cite visible evidence.
- belief, desire, and intention must be short customer-state clauses, not salesperson advice.
```

C.2 Outcome Prediction Prompt

Current code source: `piwm_train/prompts.py::build_deliberation_prompt`.

```
You are observing a customer in a retail store. Below are {n_frames} frames sampled from a streaming camera, in chronological order.

<image>
<image>
<image>

The customer's current state is:
- stage: {aida_stage}
- visible evidence: {visual_summary}
- engagement pattern: {engagement_pattern}
- gaze and attention: {gaze_and_attention}
- body and hands: {body_and_hands}
- belief: {belief}
- desire: {desire}
- intention: {intention}

Consider one candidate intervention: {candidate_action}

Dialogue-act layer for this candidate:
- act: {act}
- params: {params}

Terminal realization for this candidate:
- surface_text: {surface_text}
- screen: {screen}
- voice_style: {voice_style}
- light: {light}

Concrete execution plan for this candidate:
- physical action: {physical_action}
- utterance: {utterance}

Predict how this candidate intervention will change the customer's state in the next decision step. Output the following fields, in this exact order, each on its own line:
<next_stage>...</next_stage>
<next_belief>...</next_belief>
<next_desire>...</next_desire>
<next_intention>...</next_intention>
<risk>...</risk>
<benefit>...</benefit>
<reward>...</reward>

- next_stage must be one of: attention, interest, desire, action.
- risk and benefit must each be one of: low, medium, high.
- reward must be a number in [-1.00, 1.00] with two decimal places.
- all textual spans must be a single short clause in English.
```

C.3 Best-Action Selection Prompt

Current code source: `piwm_train/prompts.py::build_action_prompt_no_leak(...,five_act_only=True)`. The current repository does not define a separate function named `build_best_action_prompt`; this is the best-action prompt used for the 5-act action-selection rows and closed-model evaluation set.

```
You are observing a customer in a retail store. Below are {n_frames} frames sampled from a streaming camera, in chronological order.

<image>
<image>
<image>

The Stage-1 customer-state representation is:
- stage: {aida_stage}
- intent_label: {intent_label}
- visible evidence: {visual_summary}
- engagement pattern: {engagement_pattern}
- gaze and attention: {gaze_and_attention}
- body and hands: {body_and_hands}
- belief: {belief}
- desire: {desire}
- intention: {intention}

Candidate interventions are listed below. They include only the action identity, parameters, and concrete realization. They do not include gold rewards, predicted next states, risks, or benefits.

{candidate_block_no_leak}

Choose the best intervention and explain your reasoning briefly.

Output the following fields, in this exact order:
<rationale>...</rationale>
<chosen>...</chosen>
<intervention_action>...</intervention_action>
<intervention_utterance>...</intervention_utterance>

- chosen must be one of the candidate labels listed above, exact string match.
- rationale should use the customer state and action fit, not hidden reward values.
- intervention_action must describe the concrete salesperson or terminal behavior.
- intervention_utterance must be a short customer-facing sentence, or (silent) for silent Hold.
- In the 5-act setting, Hold is a real no-intervention decision, not an uncertainty fallback. Choose Hold only when the visible customer state clearly supports waiting or silence.
```

D Reproducibility

Training configuration. All trained PIWM variants use Qwen2.5-VL-7B-Instruct as the base vision-language model and ms-swift as the fine-tuning framework. The main checkpoint is trained with LoRA adaptation (`tuner_type=lora`) using rank 16, alpha 32, dropout 0.05, and no LoRA bias. The ms-swift configuration specifies `target_modules=all-linear`; the saved PEFT adapter materializes this over the Qwen projection and MLP modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` under `model.*`). The optimizer is AdamW (`adamw_torch`) with learning rate 2×10^{-5}, weight decay 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$, cosine learning-rate schedule, max gradient norm 1.0, no warmup (`warmup_steps=0`, `warmup_ratio=0.0`), bf16 precision, gradient checkpointing, per-device batch size 4, gradient accumulation 2, and 3 training epochs. The random seed and data seed are both 42.

Data and schedule. The main PIWM training file is the joint Stage-1 plus Stage-2 ms-swift file with 2,734 rows: 543 user-intent rows, 2,001 next-state-prediction rows, and 190 balanced action-selection rows. The archived run arguments record `max_length=8192` and `max_pixels=1003520`. Stage-1-only variants use the same user-intent and next-state objectives without the balanced action-selection rows. Zero-shot Qwen2.5-VL-7B is evaluated without loading a LoRA adapter.

Hardware and compute. The project used an A100 80GB server pool. Each main training run is budgeted at approximately 9 A100-hours (A100 80GB) using ms-swift LoRA fine-tuning. Counting the main PIWM run plus four ablations gives an estimated total of about 45 A100-hours.

Inference configuration. All local checkpoint evaluations use greedy decoding. In `scripts/eval_ms_swift_checkpoint.py`, generation calls `model.generate(..., do_sample=False)` with default `max_new_tokens=256`. In `scripts/eval_e2e_decision_loop_checkpoint.py`, perception, deliberation, and action calls also use `do_sample=False` and default to 128 new tokens for each step. The ms-swift checkpoint metadata stores `temperature=0.0` and `num_beams=1`. Parse failures are counted as errors in strict metrics unless a table explicitly states a parsed-output convention.

Frame sampling. The standard input contains three sparse frames per video. The frame extractor uses a cue-timeline sampling plan with default timestamps at 2.0s (`cue_onset`), 5.0s (`cue_peak`), and 8.0s (`cue_resolution`) when a prompt does not provide a custom `frame_sampling_plan`. Extracted frames are stored as `000.jpg`, `001.jpg`, and `002.jpg`, and the prompt presents them in chronological order.

E Dataset Sample

The following example is drawn from `data/GuidanceSalesBench/general/main_schema.jsonl` and keeps the task-relevant manifest, frame references, candidate actions, action-conditioned outcomes, and best-action label.


```
{
  "state_id": "piwm_09df40aae2",
  "product_category": "jewelry",
  "viewpoint": "salesperson_observable",
  "manifest": {
    "aida_stage": "interest",
    "latent_state": "active_evaluation",
    "intent": "request_demonstration",
    "persona": {
      "description": "A first-time buyer considering a higher-stakes
       purchase and needing reassurance.",
      "intent_tier": "exploring",
      "type": "first_time_high_consideration"
    },
    "bdi": {
      "belief": "A companion's approval may influence whether the choice
       feels safe.",
      "desire": "see how the product works",
      "intention": "ask for a demonstration"
    },
    "observable_cues": ["asking_companion_opinion"],
    "visual_state": {
      "summary": "The customer asks a companion for feedback, indicating
       socially influenced evaluation at a jewelry counter.",
      "engagement_pattern": "The customer remains engaged with the item
       while using companion feedback to support the decision.",
      "gaze_and_attention": "Gaze alternates between the product and the
       companion while waiting for a reaction.",
      "body_and_hands": "Hands and body posture orient the product toward
        the companion for inspection."
    },
    "proactive_score": 3
  },
  "frames": [
    {"index": 0, "relative_path": "Archive_generated_priority24/
       piwm_09df40aae2/frames/000.jpg", "timestamp_sec": null},
    {"index": 1, "relative_path": "Archive_generated_priority24/
       piwm_09df40aae2/frames/001.jpg", "timestamp_sec": null},
    {"index": 2, "relative_path": "Archive_generated_priority24/
       piwm_09df40aae2/frames/002.jpg", "timestamp_sec": null}
  ],
  "candidate_actions": [
    "Elicit_b1166d372e5e",
    "Inform_5ff00ba15ca5",
    "Hold_eda24b4bb712",
    "Recommend_8d7f8993e333"
  ],
  "outcomes": {
    "Elicit_b1166d372e5e": {
      "dialogue_act": "Elicit",
      "act_params": {"openness": "open", "slot": "need_focus"},
      "next_aida_stage": "desire",
      "next_bdi": {
        "belief": "The salesperson may help resolve the decision.",
        "desire": "explore available options",
        "intention": "continue browsing and comparing"
      },
      "risk": "low",
      "benefit": "high",
      "reward": 0.83
    },
    "Inform_5ff00ba15ca5": {
```

```
      "dialogue_act": "Inform",
      "act_params": {"content_type": "demo", "depth": "brief"},
      "next_aida_stage": "desire",
      "next_bdi": {
        "belief": "The salesperson may help resolve the decision.",
        "desire": "explore available options",
        "intention": "continue browsing and comparing"
      },
      "risk": "low",
      "benefit": "high",
      "reward": 0.8
    },
    "Hold_eda24b4bb712": {
      "dialogue_act": "Hold",
      "act_params": {"mode": "silent"},
      "next_aida_stage": "interest",
      "next_bdi": {
        "belief": "Several options remain worth comparing.",
        "desire": "see how the product works",
        "intention": "ask for a demonstration"
      },
      "risk": "low",
      "benefit": "low",
      "reward": 0.2
    },
    "Recommend_8d7f8993e333": {
      "dialogue_act": "Recommend",
      "act_params": {"pressure": "soft", "target": "item"},
      "next_aida_stage": "attention",
      "next_bdi": {
        "belief": "The salesperson may be applying too much pressure.",
        "desire": "avoid further engagement",
        "intention": "leave without buying"
      },
      "risk": "medium",
      "benefit": "low",
      "reward": -0.3
    }
  },
  "best_action": "Elicit_b1166d372e5e",
  "best_action_spec": {"act": "Elicit", "params": {"openness": "open", "
      slot": "need_focus"}},
  "best_action_realization": {
    "physical_action": "The smart retail terminal presents this candidate
        response through screen, voice, and light cues.",
    "utterance": "Would you like to look at price, functionality, or
      usage scenario first?"
  }
}
```


F Failure Case Analysis

We summarize three representative PIWM errors on Target-Test (n=30). We report them as short qualitative cases rather than a wide table because the useful information is the error pattern, not the sample identifier.

HOLD misread as ELICIT. In sample `target_piwm_810`, the gold action is HOLD, while PIWM predicts ELICIT. The customer is in the interest stage with intent `explore_options`; the visible evidence shows steady relaxed scanning without confusion or help-seeking gestures. PIWM treats sustained engagement as an information-need signal, whereas the gold annotation treats it as autonomous exploration that should not be interrupted. This is the main over-intervention pattern.

RECOMMEND misread as INFORM. In sample `target_piwm_760`, the gold action is RECOMMEND, while PIWM predicts INFORM. The customer repeatedly looks at the same item and shows minor anxious behavior, such as lightly rubbing the phone edge. PIWM defaults to information provision under uncertainty cues, but the gold annotation interprets the behavior as readiness to commit and expects a soft confirming recommendation.

GREET misread as HOLD. In sample `target_piwm_721`, the gold action is GREET, while PIWM predicts HOLD. The customer faces the terminal and raises a hand to operate it. PIWM applies a conservative non-intervention heuristic because the customer is already acting, but the gold annotation expects a lightweight availability cue. This boundary case highlights the difficulty of distinguishing "do not interrupt" from "briefly acknowledge presence."

G Real-Store Pilot Frame Examples

We include additional frames from the 20 fully annotated real-store videos in Figure 3. Each row shows three sparse timestamps (0s, 5s, 10s) from one paid-participant session, covering different AIDA stages and action labels.

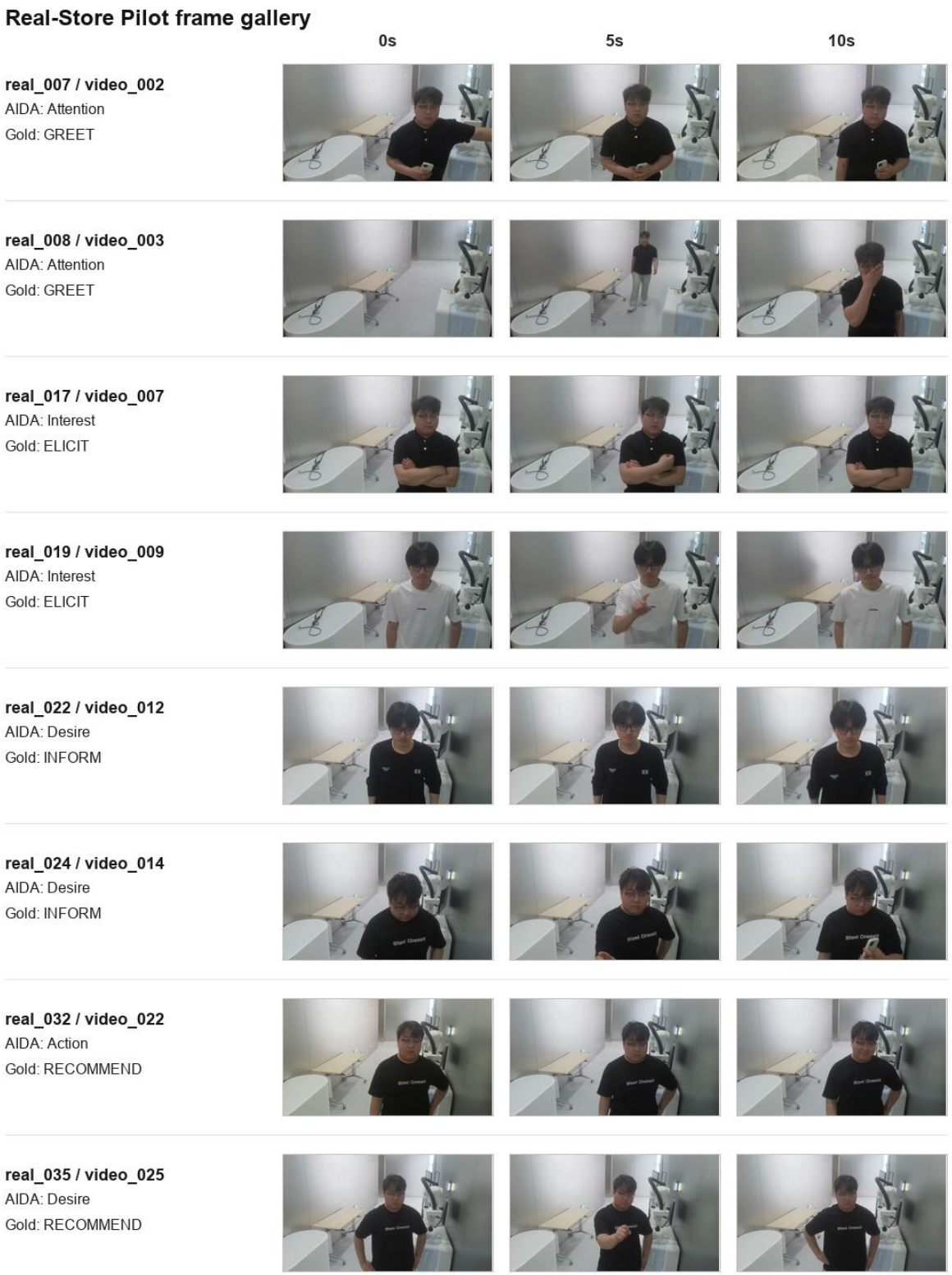

Figure 3: Additional real-store pilot frame examples. Rows correspond to distinct paid-participant sessions covering different gold AIDA stages and best-action labels. All participants signed release forms covering public release of recorded materials (see Ethics Statement).