MULTIMODAL BEHAVIOR MODELING FOR LLM-BASED VIRTUAL STANDARDIZED PATIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective physician–patient communication is central to clinical decision-making and patient-centered care. While Standardized Patients (SPs) are widely regarded as the gold standard for communication training and assessment, they are costly and difficult to scale. Recent LLM-based SP simulators offer a scalable alternative, but largely focus on verbal dialogue, overlooking nonverbal behaviors such as discomfort and hesitation that are essential to realistic patient interaction and clinical judgment. To bridge this gap, we introduce **Patient Records with Integrated Signals across Modalities** (**PRISM**), a multimodal SP behavior dataset covering 14 medical specialties and more than 8.6k dialogue turns, annotated along 18 behavioral dimensions. Building on PRISM, we propose **Behavior-Aware Patient Simulator** (**BAPS**), an LLM-based simulator that generates realistic multimodal patient behaviors. BAPS is optimized by our proposed self-improving **context-evolution** framework that iteratively develops a structured skill library covering verbal behavior, information disclosure, and nonverbal expression. Specifically, the context evolution proceeds through iterative simulation-diagnosis-refinement cycles. Firstly, at the simulation stage, BAPS predicts multimodal patient responses using the current skill library. Secondly, the diagnosis stage compares this prediction against real-SP response annotations to identify behavioral discrepancies and distill actionable insights. Finally, at the refinement stage, these insights are consolidated into the skill library to improve the subsequent simulations. After context evolution, the established skill library is used by BAPS at inference time to guide more behaviorally faithful patient simulation. Extensive experiments demonstrate that BAPS substantially outperforms existing methods across verbal quality, disclosure control, and multimodal behavioral fidelity, improving the overall multimodal behavioral fidelity score by 0.208 (a 48.6% relative gain) over the strongest baseline. Blinded human evaluation further confirms BAPS's superiority using MaSP, a validated instrument for assessing the quality of standardized patients. Together, PRISM and BAPS provide a scalable foundation for building virtual SPs that more faithfully reproduce the multimodal behaviors of real patients.

1 INTRODUCTION

Effective physician–patient communication is central to clinical decision-making and patient-centered care. Standardized Patients (SPs), typically trained actors who portray patients according to predefined clinical scenarios, are widely used to train and assess such communication skills (Flanagan & Cummings, 2023; Malidze et al., 2026). However, SP-based training is costly, labor-intensive, and difficult to scale, motivating increasing interest in virtual SPs as a more accessible alternative (Pollak et al., 2026; Nawathe et al., 2024).

Recent advances in large language models (LLMs) have enabled virtual SPs to sustain increasingly natural and context-aware clinical conversations (Schlager et al., 2026; Jiang et al., 2026; Du et al., 2025b; Wu et al., 2026). Existing systems can generate plausible patient responses following predefined clinical profiles. Yet most approaches primarily model *what* a patient says, while paying substantially less attention to *how* the patient behaves during the interaction. Real patient encounters are inherently multimodal (Henry et al., 2012; Zimmermann et al., 2007). Patients may hesitate before disclosing sensitive information, exhibit discomfort when discussing particular symptoms,

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.

or express emotional states that are only partially reflected in their words. These behavioral signals provide important context for interpreting patient states and adapting communication for reliable clinical decision-making and patient-centered care (Zimmermann et al., 2007). This gap matters because realistic patient simulation requires more than generating clinically plausible dialogue. Verbal responses, information disclosure, and nonverbal behavior should evolve coherently with the interaction. For instance, a patient may initially avoid discussing a sensitive issue, show signs of hesitation, and disclose further information only after sufficient rapport has been established.

Developing such simulators poses two key challenges. First, existing SP resources provide limited supervision for jointly modeling verbal behavior, information disclosure, and nonverbal expression. Without aligned behavioral annotations, it is difficult to characterize how these signals co-occur and evolve throughout a clinical encounter. Second, directly prompting an LLM with generic behavioral instructions provides no systematic mechanism for identifying recurring simulation errors and converting them into reusable behavioral knowledge.

To address these challenges, we introduce **Patient Records with Integrated Signals across Modalities** (**PRISM**), a multimodal SP behavior dataset covering 14 medical specialties and more than 8.6K dialogue turns, with each turn annotated along 18 behavioral dimensions. These annotations jointly characterize verbal behavior, information disclosure, and nonverbal expression, providing a unified basis for modeling and evaluating patient behavior beyond dialogue content alone.

Building on PRISM, we propose the **Behavior-Aware Patient Simulator** (**BAPS**), an LLM-based simulator for generating multimodal patient behaviors. BAPS is optimized via a proposed self-improving **context-evolution** framework that develops a structured skill library through iterative *simulation–diagnosis–refinement* cycles. At the simulation stage, BAPS first generates patient responses using its current skill library. The diagnosis stage then compares the generated behaviors against real-SP annotations to identify discrepancies and distill actionable insights. At the refinement stage, these insights are subsequently consolidated into the skill library, allowing behavioral knowledge to accumulate across iterations. The evolved library is then used at inference time to guide verbal responses, disclosure decisions, and nonverbal expressions toward more behaviorally faithful patient simulation.

We evaluate BAPS across verbal quality, disclosure control, and multimodal behavioral fidelity. BAPS consistently outperforms existing methods across these dimensions (Schlager et al., 2026; Zhang et al., 2026a; Kyung et al., 2025). Blinded human evaluation using MaSP (Wind et al., 2004), a validated instrument for assessing standardized patient quality, further confirms the improvements achieved by BAPS.

Our main contributions are:

- We introduce **PRISM**, a multimodal SP behavior dataset spanning 14 medical specialties, more than 8.6K dialogue turns, and 18 behavioral dimensions covering verbal behavior, information disclosure, and nonverbal expression.
- We propose **BAPS**, an LLM-based patient simulator that jointly models verbal, disclosure, and nonverbal behaviors to improve multimodal behavioral fidelity.
- We develop a self-improving **context-evolution** framework that distills discrepancies between simulated and real-SP behaviors into a structured skill library through simulation–diagnosis–refinement cycles to enable progressively more behaviorally faithful patient simulation.
- Extensive automatic and blinded human evaluations demonstrate consistent improvements in verbal quality, disclosure control, and multimodal behavioral fidelity over existing methods.

2 Related Work

2.1 LLM-based Patient Simulation

Recent work has increasingly used LLMs to simulate standardized patients (SPs) in interactive doctor–patient conversations (Liao et al., 2024; Du et al., 2025b; Zhang et al., 2026a; Xu et al., 2026). Existing studies primarily improve realism and controllability through persona modeling and information disclosure. Patient diversity is introduced through predefined psychometric or structured profiles, as in PWP and MSPRP (Schlager et al., 2026; Jiang et al., 2026), or interaction-driven

Table 1: Comparison of data resources relevant to patient simulation. ✓: available; △: partially available; ×: not available; n/r: not reported. [a]Hospital service types recorded in the `services` table (e.g., cardiac medical, orthopaedic, psychiatric), rather than consultation specialties. [b]Only one subset is role-played by a volunteer lay person; the others are expert-written scripts. [c]Patient roles are played by medical students rather than trained SPs. [d]Audio for all dialogues; video for 85 of 240, available upon request.

		Data Source		Turn-Level Patient Annotations				
Dataset	**#Specialties**	Human SP	Audio/Video	Verbal	Facial	Bodily	Vocal	Affective
MIMIC-IV Family (Johnson et al., 2023c;a;b)	21[a]	×	×	×	×	×	×	×
ACI-Bench (Yim et al., 2023)	n/r	△[b]	×	✓	×	×	×	×
French OSCE (Bonzi et al., 2026)	12	△[c]	△[d]	✓	×	×	×	×
PRISM (Ours)	**14**	✓	✓	✓	✓	✓	✓	✓

adaptation, as in EvoPatient (Du et al., 2025b). Disclosure is typically controlled through intention- or query-conditioned mechanisms (Zhang et al., 2026a; Schlager et al., 2026), explicit gating rules (Wu et al., 2026), or state-aware control (Liao et al., 2024). In contrast, behavioral realism remains substantially less explored. Hauck et al. (2026) show that isolated facial-expression adjustments are insufficient to improve perceived realism, with expert evaluation suggesting that realistic patient simulation requires coordinated verbal and non-verbal cues. Existing work therefore largely models who the patient is and what they disclose, while how patients express themselves through coordinated multimodal behavior remains underexplored.

2.2 Context Engineering and Iterative Refinement

Context engineering improves LLM behavior by optimizing inference-time information and strategies rather than model parameters, enabling reuse across different LLM backbones without retraining (Suzgun et al., 2026; Zhang et al., 2026b). Such context-based optimization can rival reinforcement-learning-based methods on selected tasks (Agrawal et al., 2026). Recent work further extends static prompt optimization toward persistent context evolution: Agentic Context Engineering (ACE) maintains an evolving playbook through incremental updates (Zhang et al., 2026b), while Meta Context Engineering (MCE) co-evolves context-engineering skills and context artifacts (Ye et al., 2026).

2.3 Datasets for Patient Simulation

Patient-simulation methods primarily draw on clinical records for patient grounding and conversational data for interaction modeling. The MIMIC family provides rich clinical information, including demographics, medical histories, medications, diagnoses, and clinical notes (Johnson et al., 2023c;a;b), and has supported case-grounded patient profiles in systems such as PatientSim (Kyung et al., 2025) and EvoPatient (Du et al., 2025a). Conversational resources include ACI-Bench (Yim et al., 2023), which contains doctor–patient visit dialogues for clinical note generation, and the French OSCE dataset (Bonzi et al., 2026), which provides audio-recorded student–patient OSCE interactions with structured station materials. However, few resources jointly provide structured patient profiles, complete turn-by-turn interactions, and fine-grained multimodal behavioral annotations such as facial expression, gaze, gesture, posture, and vocal behavior.

3 PRISM Multimodal SP Dataset

In this work, we introduce **Patient Records with Integrated Signals across Modalities** (**PRISM**), a multimodal dataset of complete clinical consultations involving human standardized patients (SPs). Each consultation combines (1) a structured patient case profile and (2) turn-level dialogue aligned with audiovisual signals and fine-grained behavioral annotations. A comparison with existing resources relevant to patient simulation is provided in Table 1.

PRISM is constructed through five stages, as illustrated in Figure 1.

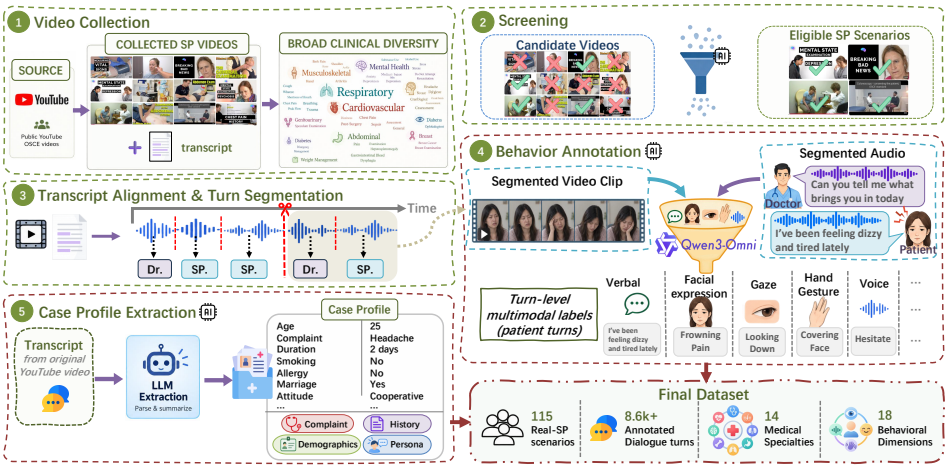


Figure 1: **Overview of the PRISM dataset construction pipeline.** Candidate SP consultation videos are screened, temporally aligned, and segmented into doctor–patient turns. Multimodal patient behavior is annotated at the turn level, while full transcripts are parsed into structured case profiles. The resulting dataset contains 115 consultations, 8,623 dialogue turns, and 17.4 hours of interaction across 14 medical specialties.

Video collection. We conduct an extensive survey of publicly available SP consultation videos, including OSCE training recordings, and identify more than 1,200 candidate videos spanning over 20 medical specialties.

Screening. Candidate videos undergo LLM-assisted pre-screening to identify complete consultations with sustained doctor–SP interaction. Human reviewers subsequently verify dialogue completeness and audiovisual quality, resulting in 115 retained consultations across 14 specialties as specified in Appendix F.

Temporal alignment and turn segmentation. For each retained consultation, the video, audio, and transcript are temporally aligned and segmented into speaker-attributed doctor and SP turns. This process produces synchronized audiovisual and textual segments for each dialogue turn.

Multimodal behavioral annotation. The synchronized video, audio, and transcript segments of each patient turn are jointly processed by Qwen3-Omni to produce structured annotations across 18 behavioral dimensions, including emotion, facial expression, gaze, hand gesture, body posture, and vocal characteristics. Each dimension uses a predefined closed-set label space. The complete annotation schema is provided in Appendix D.

Case-profile extraction. The full consultation transcript is separately parsed using an LLM into a structured patient case profile containing clinical, demographic, and interaction-related attributes.

3.1 Human Validation of Behavioral Annotations

To assess annotation reliability, we conduct blinded human validation. Automated annotations agree with human consensus in over 70% of cases for seven of ten behavioral dimensions. Lower agreement on emotion, body posture, and head movement mirrors low inter-annotator agreement, suggesting inherent ambiguity in these signals (Appendix C).

4 BAPS: Behavior-Aware Patient Simulator

BAPS is an LLM-based SP simulator that jointly models verbal responses, information disclosure, and nonverbal behavior. Instead of relying on a fixed behavioral prompt, BAPS maintains an editable *Skill Library* containing revisable natural-language behavioral guidance. During *context evolution*, the Skill Library is progressively refined by comparing simulated patient responses with matched

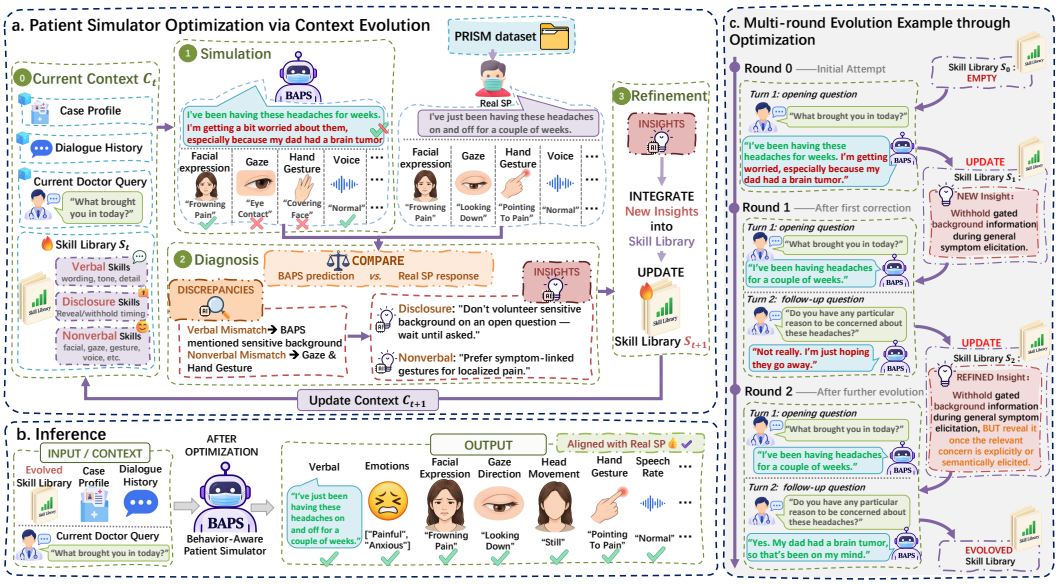


Figure 2: Overview of BAPS. (a) The context evolution framework iteratively refines the Skill Library through Simulation–Diagnosis–Refinement cycle. (b) At inference time, the evolved Skill Library guides patient simulation without further refinement. (c) An example illustrating how behavioral skills progressively evolve across iterations.

real-SP behavior from PRISM. BAPS therefore improves its behavioral guidance through evolving context rather than updating the parameters. Figure 2 provides an overview of the framework.

4.1 Skill Library

The Skill Library consists of three skill sets: $\mathcal{S} = \mathcal{S}_{\text{VERBAL}} \cup \mathcal{S}_{\text{DISCLOSURE}} \cup \mathcal{S}_{\text{NONVERBAL}}$. Here, $\mathcal{S}_{\text{VERBAL}}$ governs how the patient formulates and expresses a response; $\mathcal{S}_{\text{DISCLOSURE}}$ controls when clinical information should be revealed, withheld, or volunteered; and $\mathcal{S}_{\text{NONVERBAL}}$ manages accompanying facial, postural, gestural, and vocal behaviors. This decomposition aligns with the multimodal structure of PRISM (Section 3), allowing each aspect of simulated behavior to be evaluated and refined against its corresponding real-SP reference.

Each skill is represented as a short natural-language rule: $s_i = \langle \text{text}_i, h_i^+, h_i^-, t_i, f_i \rangle$, where text_i contains the behavioral guidance; h_i^+ and h_i^- record how often the skill has been assessed as helpful or harmful during evolution; t_i denotes the evolution step at which the skill was introduced; and f_i, defined only for nonverbal skills, specifies the behavioral dimension governed by the skill, such as gaze, posture, or hand gesture. Together, these fields allow the Skill Library to accumulate behavioral knowledge while retaining the provenance and empirical utility of individual skills.

4.2 Context Evolution

The central mechanism of BAPS is **context evolution**, which transforms turn-level discrepancies between simulated and real-SP behavior into reusable behavioral guidance. Conceptually, the process follows an iterative simulation–diagnosis–refinement procedure. Specifically, BAPS first simulates a patient response when interacting with a physician, diagnoses how that response differs from the corresponding real-SP behavior, and then incorporates the resulting insights into the Skill Library.

At evolution step k, the process is formalized as

$$\mathcal{S}^{(k)} \xrightarrow{\text{SIMULATION}} \hat{\mathbf{y}}_k \xrightarrow[\mathbf{y}_k]{\text{DIAGNOSIS}} d_k \xrightarrow{\text{REFINEMENT}} \mathcal{S}^{(k+1)},$$

where $\mathcal{S}^{(k)}$ denotes the Skill Library at step k, $\hat{\mathbf{y}}_k$ is the simulated multimodal patient response to a physician utterance from PRISM, $\mathbf{y}_k$ is the corresponding real-SP reference, and d_k contains

the behavioral insights distilled from their discrepancies via the backbone LLM. Repeating this cycle allows local supervision from individual interactions to accumulate into general behavioral guidance.

Simulation. Given a patient case profile c, dialogue history h, and the current physician utterance q, BAPS generates $\hat{\mathbf{y}}_k = \mathrm{Simulate}(c, h, q, \mathcal{S}^{(k)})$. The generated output comprises the patient's verbal response together with structured nonverbal behaviors, while disclosure skills regulate which case information is revealed, withheld, or volunteered at the current turn. In this way, all three skill sets jointly condition a single patient response.

Diagnosis. The Diagnosis stage compares $\hat{\mathbf{y}}_k$ with the matched real-SP reference $\mathbf{y}_k$ across verbal expression, information disclosure, and nonverbal behavior. Rather than treating the comparison as a single correctness signal, Diagnosis identifies modality-specific discrepancies and reasons about why the simulated behavior diverges from the reference. Crucially, these discrepancies are abstracted into generalizable behavioral insights rather than case-specific corrections. The resulting diagnosis d_k therefore captures not only *what* was mismatched, but also *how* the simulator should behave in analogous interaction contexts.

Refinement. d_k updates $\mathcal{S}^{(k)}$ into $\mathcal{S}^{(k+1)}$ through one of four operations: a new skill is *added* if no related skill exists; a closely matching skill is *merged* and its track record updated; a related but imprecise skill is *generalized*; and consistently unreliable skills are *pruned*. Because updates accumulate across interactions, knowledge acquired from earlier cases can be refined by later ones, as illustrated in Figure 2(c). Full criteria are given in Appendix B.

4.3 Inference

After context evolution, the resulting Skill Library $\mathcal{S}^*$ is frozen and used solely as behavioral guidance. For a new patient interaction, BAPS generates $\hat{\mathbf{y}} = \mathrm{Simulate}(c, h, q, \mathcal{S}^*)$ using the evolved verbal, disclosure, and nonverbal skills without further diagnosis or refinement. Unlike during context evolution, no matched real-SP response is required at inference time. Thus, the behavioral knowledge distilled from PRISM can be directly reused to guide patient simulation in new interactions.

5 Experiment

We conduct experiments to answer four research questions: **RQ1**: Does BAPS reproduce real standardized-patient (SP) behavior more faithfully than existing patient simulators across verbal response, information disclosure, and nonverbal behavior? **RQ2**: Does BAPS improve as context evolution progresses? **RQ3**: Can an evolved Skill Library transfer to a different backbone LLM without further evolution? **RQ4**: Do human evaluators prefer the patient behavior generated by BAPS over that of a strong baseline?

5.1 Experimental Setup

Data split. We split PRISM at the consultation level into 90 training, 12 validation, and 13 test consultations, ensuring that all turns from the same clinical encounter remain within a single split. Context evolution is performed exclusively on the training split. The validation split is used only to track behavioral performance throughout evolution (Section 6.3).

Baselines. We compare BAPS with three recent LLM-based patient simulators: PWP (Schlager et al., 2026), PatientSim (Kyung et al., 2025), and EasyMED (Zhang et al., 2026a), representing personality-conditioned, persona-controlled, and multi-agent SP simulation, respectively. Since these methods do not natively generate nonverbal behavior, we extend each baseline with the same behavioral output schema used by BAPS to provide a common prediction space, while preserving its original prompting and control mechanisms. For a controlled comparison, all methods use the same backbone LLM.

Implementation. Unless otherwise specified, we use DeepSeek-V4-Flash (DeepSeek-AI, 2026) as the backbone LLM for both context evolution and patient simulation. The Skill Library is initialized as empty and evolved exclusively on the PRISM training split following the procedure described in Section 4.2. Once evolution is complete, the resulting Skill Library is frozen and used unchanged for test-time simulation. To account for residual variability in API-based inference, we independently evaluate each method three times on the test set and report the mean and standard deviation. For cross-backbone transfer (**RQ3**), we directly pair GPT-6 Astra (OpenAI, 2026) with the final Skill Library evolved using DeepSeek-V4-Flash, without any additional context evolution. This setting tests whether the learned behavioral guidance captures transferable simulation knowledge rather than behavior specific to the model used during evolution.

5.2 Evaluation Metrics

We evaluate BAPS along the three components of patient behavior it models: verbal response, nonverbal behavior, and information disclosure. Unless otherwise specified, all automatic metrics are computed at the dialogue-turn level against the matched real-SP reference from PRISM and averaged over the held-out test set.

Verbal response. We evaluate verbal fidelity using SBERT cosine similarity, BLEU-4, and ROUGE-L F1. SBERT captures semantic similarity between the simulated and real-SP responses, while BLEU-1, BLEU-4 and ROUGE-L measure complementary forms of lexical and phrasal overlap. Together, these metrics assess whether the simulator preserves the semantic content of the reference patient response.

Nonverbal behavior. Each dimension is scored according to its label structure. For multi-label dimensions (emotion and facial expression), we report micro- and macro-F1 to capture aggregate performance and performance across labels, respectively. For rest single-label dimensions, we report accuracy. Behavioral intensity, rated on a 1–5 ordinal scale, is evaluated by mean absolute error (MAE).

Information disclosure. We evaluate whether clinical information is disclosed at an appropriate point, using the real SP's disclosure behavior as the reference. *Leak Rate* measures premature disclosure of information that the clinician has not yet elicited, and *Disclosure Score* summarizes overall alignment with the reference disclosure pattern, accounting for both premature and appropriately timed disclosure. Formal definitions are provided in Appendix E.

Human evaluation. To complement automatic evaluation (**RQ4**), we conduct a blinded pairwise comparison between BAPS and EasyMED (Zhang et al., 2026a), the strongest baseline in Table 2. Evaluation criteria are adapted from the Maastricht Assessment of Simulated Patients (MaSP) (Wind et al., 2004), covering verbal and nonverbal behavior, information disclosure, and overall role portrayal. The 13 test consultations are divided into 49 excerpts of approximately 10 turns. Three annotators evaluate anonymized outputs presented in randomized order and indicate a preference for BAPS, EasyMED, or a tie. Full criteria are provided in Appendix A.

6 Results and Analysis

6.1 Overall Comparison

Table 2 reports the overall results on the held-out PRISM test set. With DeepSeek-V4-Flash as the shared backbone, BAPS achieves the best performance across all 12 multimodal behavioral metrics and all three verbal metrics. The gains are particularly pronounced for behavioral dimensions that are weakly captured by existing patient simulators, indicating that the primary benefit of BAPS lies not merely in producing better textual responses and disclosure control, but in modeling how patient behavior accompanies those responses.

Cross-backbone transfer. Results with GPT-6 Astra in Table 2 test whether the evolved Skill Library transfers beyond the DeepSeek-V4-Flash backbone used for context evolution. Without

Table 2: Overall performance on the PRISM test set. ↑ and ↓ indicate whether higher or lower values are better, respectively. Best results within each backbone are shown in **bold**. Baselines: PWP (Schlager et al., 2026), PatientSim (Kyung et al., 2025), and EasyMED (Zhang et al., 2026a). For GPT-6 Astra, BAPS uses the Skill Library evolved on DeepSeek-V4-Flash without further evolution.

	Verbal				**Disclosure Control**	
Method	SBERT↑	BLEU-1↑	BLEU-4↑	ROUGE-L↑	Leak↓	Disc. Score↑
• *Backbone: DeepSeek-V4-Flash*						
PWP	0.327	0.073	0.011	0.144	0.167	0.107
PatientSim	0.340	0.091	0.016	0.166	0.109	0.173
EasyMED	0.367	0.094	0.017	0.183	0.086	0.207
BAPS (Ours)	**0.406**	**0.138**	**0.023**	**0.224**	**0.073**	**0.216**
• *Backbone: GPT-6 Astra*						
GPT-6 Astra (Zero-shot)	0.408	0.129	0.029	0.226	0.099	0.189
GPT-6 Astra + BAPS	**0.470**	0.186	**0.055**	**0.321**	**0.072**	**0.214**

	Multimodal Behavioral Fidelity											
Method	Emo. mF1↑	Emo. MF1↑	Facial mF1↑	Facial MF1↑	Gaze↑	Head↑	Body↑	Hand↑	Voice↑	Speech↑	Traj.↑	Int. MAE↓
• *Backbone: DeepSeek-V4-Flash*												
PWP	0.338	0.246	0.277	0.224	0.412	0.514	0.119	0.554	0.183	0.390	0.857	1.855
PatientSim	0.325	0.252	0.273	0.232	0.516	0.511	0.125	0.630	0.185	0.425	0.778	1.825
EasyMED	0.434	0.300	0.366	0.338	0.472	0.395	0.135	0.538	0.315	0.546	0.855	1.233
BAPS (Ours)	**0.556**	**0.371**	**0.575**	**0.362**	**0.692**	**0.717**	**0.690**	**0.794**	**0.608**	**0.706**	**0.906**	**0.797**
• *Backbone: GPT-6 Astra*												
GPT-6 Astra (Zero-shot)	0.391	**0.287**	0.365	0.236	0.658	0.316	0.269	0.690	0.368	0.697	0.885	0.706
GPT-6 Astra + BAPS	**0.641**	0.279	**0.590**	**0.289**	**0.692**	**0.764**	**0.697**	**0.794**	**0.658**	**0.722**	**0.917**	**0.494**

backbone-specific adaptation, the Skill Library improves all three verbal metrics and 11 of 12 multimodal behavioral metrics. It also reduces Leak Rate from 0.099 to 0.072 and improves Disclosure Score from 0.189 to 0.214. These results suggest that the learned behavioral guidance generalizes across backbone LLMs, supporting **RQ1** and **RQ3**.

6.2 Ablation Study

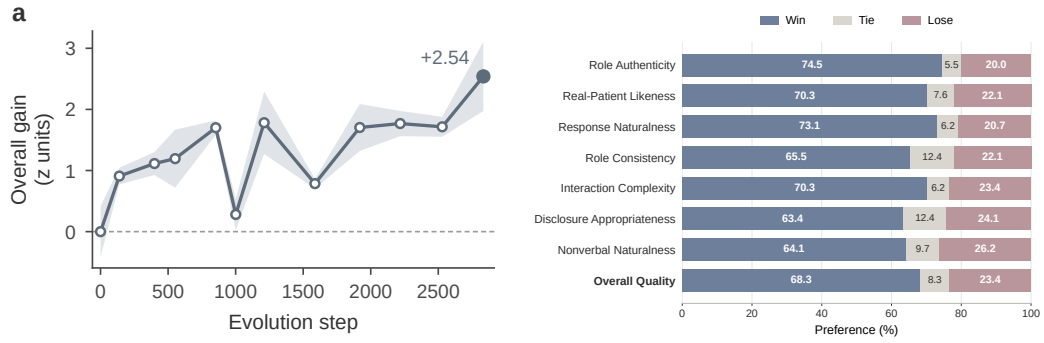
Table 3 examines the contribution of each component of the Skill Library. Removing each skill set primarily degrades its corresponding behavioral dimension. Excluding verbal skills reduces ROUGE-L by 0.023, removing nonverbal skills produces a substantially larger 0.171 drop in nonverbal fidelity, and removing disclosure skills lowers the Disclosure Score by 0.010. In contrast, performance on the other dimensions changes only marginally. This pattern indicates that the three skill sets learn largely specialized behavioral guidance rather than redundantly encoding the same information. Removing the entire Skill Library leads to the largest degradation across all three dimensions, with decreases of 0.042 in verbal fidelity, 0.193 in nonverbal fidelity, and 0.028 in disclosure control. Notably, these drops are larger than those caused by removing any single corresponding skill set, suggesting that the three types of guidance also complement one another when jointly conditioning patient simulation.

6.3 Effect of Context Evolution

We next examine whether the behavioral knowledge accumulated by BAPS improves progressively over context evolution (**RQ2**). We evaluate intermediate Skill Library checkpoints on the validation set, averaging each checkpoint over three independent inference runs. To summarize performance across heterogeneous metrics, we construct a domain-balanced standardized composite score. Each metric is first standardized across checkpoints, with lower-is-better metrics sign-reversed. Scores are then averaged separately within the verbal, nonverbal, and disclosure domains before assigning equal weight to the three domains.

Table 3: Ablation on the Skill Library. Verbal: ROUGE-L; Nonverbal: mean of nine nonverbal metrics; Disclosure: Disclosure Score. Values in parentheses denote changes relative to the full BAPS.

Variant	Verbal↑	Nonverbal↑	Disclosure↑
BAPS (full)	0.252	0.682	0.218
w/o Verbal skills	0.229 (−0.023)	0.686 (+0.005)	0.211 (−0.007)
w/o Nonverbal skills	0.245 (−0.007)	0.511 (−0.171)	0.217 (−0.001)
w/o Disclosure skills	0.248 (−0.004)	0.687 (+0.005)	0.208 (−0.010)
w/o Skill Library	0.210 (−0.042)	0.489 (−0.193)	0.190 (−0.028)



(a) Effect of context evolution. Standardized performance gain over the unevolved checkpoint. Shading denotes ±1 SD over three inference runs.

(b) Human preference evaluation of BAPS against EasyMED across MaSP-inspired dimensions.

Figure 3: Evaluation of BAPS from context evolution and human preference.

Figure 3a shows that every evolved checkpoint performs better than the unevolved simulator with an empty Skill Library, while the final checkpoint achieves the largest standardized gain of $+2.54$. This provides direct evidence that the behavioral guidance accumulated during context evolution improves simulation quality beyond the base prompting configuration.

6.4 Human Evaluation

We further compare BAPS with EasyMED (Zhang et al., 2026a), the strongest baseline in Table 2, through a blinded pairwise human evaluation (**RQ4**). Following MaSP-inspired criteria (Wind et al., 2004), three annotators compare anonymized verbal and nonverbal responses in randomized order and select BAPS, EasyMED, or a tie. Full evaluation criteria are provided in Appendix A.

As shown in Figure 3b, BAPS is preferred across all evaluation dimensions and in 68.3% of comparisons for overall portrayal quality, compared with 23.4% for EasyMED and 8.3% ties.

7 Limitations & Future Work

BAPS produces structured, modality-agnostic behavioral outputs, including emotion, gaze, posture, gesture, and vocal affect. These outputs are intentionally decoupled from any specific rendering pipeline, allowing BAPS to serve as a general behavioral decision layer for interactive avatar systems. Our work therefore focuses on deciding what a patient should say, disclose, and express at each point in a clinical interaction, rather than on how these decisions are rendered into motion and speech. A key limitation is that we do not evaluate BAPS in a fully embodied setting. Our human evaluation assesses verbal responses and structured behavioral descriptions without rendered avatars, so the results demonstrate improved behavioral decision quality but do not establish equivalent gains in perceptual realism after animation and speech synthesis. Future work should integrate BAPS with real-time avatar and text-to-speech systems and evaluate end-to-end interaction quality.

AI USE STATEMENT

Generative AI is the subject of this work: we study LLM-based standardized patient (SP) simulation, and LLMs are used to implement the simulators we propose and evaluate (BAPS). All such uses, including model choices, prompts, and configurations, are described in the main text and appendix. In addition, generative AI tools were used in constructing the PRISM dataset. An LLM pre-screened candidate videos for completeness and doctor–SP interaction density; Qwen3-Omni produced turn-level behavioral annotations from synchronized video, audio, and transcript segments; and [LLM name] extracted structured case profiles and disclosure trajectories from consultation transcripts (Section 3). We have not used generative AI tools to propose research hypotheses, formulate theoretical claims, or interpret experimental results, and the remaining required disclosure tasks (e.g., proofs, translation) are not applicable to this work. Additionally, we used generative AI tools to search for and verify related literature, suggest the organization of the manuscript, and edit the text for clarity and readability.

We have reviewed all AI-assisted work. All videos retained after LLM pre-screening were manually verified by human reviewers for dialogue completeness and audiovisual quality. The model-generated behavioral annotations were validated against independent human annotations on [N] patient turns, with annotators blinded to model labels (Appendix D). Factual claims about prior datasets were checked against the original publications, and all AI-edited text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies virtual standardized-patient (SP) simulation for clinical communication training and does not provide medical diagnosis, treatment recommendations, or decision support for real patients. PRISM is constructed from publicly available videos of simulated clinical encounters involving standardized patients rather than real patient consultations. Nevertheless, audiovisual data may contain identifiable information such as faces and voices. We therefore emphasize responsible use of the dataset and compliance with the terms and access conditions of the original data sources. Any public release of PRISM will follow applicable data-use and privacy requirements and will clearly document its intended research and educational use.

The automated behavioral annotations in PRISM may also reflect biases or errors inherited from the underlying multimodal models, particularly for subjective behavioral attributes such as emotion and posture. We therefore conduct human validation of these annotations and report their limitations rather than treating them as definitive ground truth. Similarly, BAPS is intended as a research system for simulated clinical communication and should not be interpreted as a substitute for human standardized patients, clinical supervision, or professional medical judgment. Future deployment in educational settings should consider potential demographic, cultural, and behavioral biases and include appropriate human oversight.

REPRODUCIBILITY STATEMENT

We provide implementation details, prompt templates, evaluation protocols, and metric definitions in the appendix. The construction and annotation pipeline for PRISM, including data collection, preprocessing, behavioral annotation, and dataset statistics, is described in Section 3 and the supplementary material. Experimental settings, baseline configurations, ablation details, and context-evolution procedures are also documented in the appendix. Code and additional resources for reproducing our experiments are available in our anonymous repository at `https://anonymous.4open.science/r/Multimodal-Behavior-Modeling-for-LLM-based-Virtual-Standardized-Patients--321E/`.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa:

Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026.

Doria Bonzi, Tom Bourgeade, Fabrice Lefèvre, and Irina Illina. A French OSCE dialogue dataset and controllable virtual patient system for clinical training. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 357–374, Atlanta, Georgia, USA, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.sigdial-1.25. URL https://aclanthology.org/2026.sigdial-1.25/.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. Llms can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17278–17306. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.846.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. LLMs can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17278–17306. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.846.

Octavia L. Flanagan and Kristina M. Cummings. Standardized patients in medical education: A review of the literature. *Cureus*, 15(7):e42027, 2023. doi: 10.7759/cureus.42027.

Steffen Hauck, Theresa Schell, Sophie Jörg, and Jens Grubert. An AI-driven virtual patient for breaking bad news: An expert formative study on facial expression intensity. *arXiv preprint arXiv:2607.22118*, 2026.

Stephen G. Henry, Andrea Fuhrel-Forbis, Mary A. M. Rogers, and Susan Eggly. Association between nonverbal communication during clinical interactions and outcomes: a systematic review and meta-analysis. *Patient Education and Counseling*, 86(3):297–315, 2012. doi: 10.1016/j.pec.2011.07.006.

Shijie Jiang, Zefan Zhang, Kehua Zhu, Tian Bai, and Ruihong Zhao. Multi-stage patient role-playing framework for realistic clinical interactions. *arXiv preprint arXiv:2601.10951*, 2026.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, 2023a. doi: 10.13026/5ntk-km72. Version 2.2.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, 2023b. doi: 10.13026/1n74-ne17. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023c. doi: 10.1038/s41597-022-01899-x.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*, 2024.

David Malidze, Lasha Dolidze, Natia Jojua, Tinatin Gognadze, Ana Jabauri, Eka Kokhreidze, Larisa Melia, Tamar Chanturia, Tamar Chkoidze, and Stephen Charles Jones. Standardized patient-based training and clinical diagnostic competence among medical students: a randomized comparative study. *Georgian Medical News*, (375):43–50, Jun 2026.

Pooja A. Nawathe et al. Enhancing healthcare communication education: Standardised patient programmes. *Indian Journal of Anaesthesia*, 2024.

OpenAI. GPT-6 Astra System Card, September 2026. Published September 3, 2026.

Rebecca Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), 2006.

Uri Pollak et al. Training medical students as peer simulated patients in undergraduate medical education: a narrative review and implementation framework. *Postgraduate Medical Journal*, 2026. doi: 10.1093/postmj/qgag100.

Moritz Schlager, Friederike Jungmann, Samuel Schmidgall, Philipp Raffler, Franziska Hartl, Eva Wende, Paula Rossmüller, Conrad Ketzer, Avinatan Hassidim, Dale R. Webster, Yossi Matias, Yun Liu, Daniel Rueckert, Mike Schaekermann, and Paul Hager. Patients with personality: Realistic patient simulation through controlled diversity and selective disclosure. *arXiv preprint arXiv:2606.17441*, 2026.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7080–7106, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.333.

Lidewij A. Wind, Jan Van Dalen, Arno M. M. Muijtjens, and Jan-Joost Rethans. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *Medical Education*, 38(1):39–44, 2004. doi: 10.1111/j.1365-2923.2004.01686.x.

Ping Wu, Yu Han, Jing Zhang, Yunqi Li, Mengna Jiang, Xinyu Lu, Haibin Zhang, Danyang Xu, Hao Ming, Lihong Wang, and Qingping Wen. From simulation to pedagogy: Structured ai standardized patients for clinical communication training validated through multi-model and randomized evaluation. *medRxiv*, 2026. doi: 10.64898/2026.04.26.26351793.

Chen Xu, Yu Ji, Zhenyu Lyu, Yang Yi, Yizhe Yang, Luyao Ji, Chaoyi Chen, Xianyang Wang, Tian Lan, Zhihua Wang, Juan Wang, Xunde Dong, Fuze Tian, Qunxi Dong, and Bin Hu. PUPPET: Neural-symbolic standardized patients for mental health. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2599–2634. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.121.

Haoran Ye, Xuning He, Vincent Arak, Haonan Dong, and Guojie Song. Meta context engineering via agentic skill evolution, 2026. URL `https://arxiv.org/abs/2601.21557`.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10:586, 2023. doi: 10.1038/s41597-023-02487-3.

Bingquan Zhang, Xiaoxiao Liu, Yuchi Wang, Lei Zhou, Qianqian Xie, and Benyou Wang. Human or LLM as standardized patients? a comparative study in medical education. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.1243.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Y. Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026b.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22056–22065, 2025.

Christa Zimmermann, Lidia Del Piccolo, and Arnstein Finset. Cues and concerns by patients in medical consultations: a literature review. *Psychological Bulletin*, 133(3):438–463, 2007. doi: 10.1037/0033-2909.133.3.438.

Table 4: Human evaluation dimensions adapted from MaSP (Wind et al., 2004). The original assessment dimensions are reformulated as pairwise questions for comparing virtual patient simulations.

Dimension	Question shown to annotators
Role Authenticity	Which one sounds more like a real person talking about their own health, rather than a script or an AI?
Real-Patient Likeness	If nobody told you this was a simulation, which one could you more easily believe is a real patient visit?
Response Naturalness	Which one answers in a more natural, conversational way—not stiff or over-explained?
Role Consistency	Which one stays more consistent with the patient's background, mood, personality, and situation without contradicting itself?
Interaction Complexity	Which one better reflects the natural complexity of patient interaction, such as hesitation, uncertainty, emotional reactions, or gradual disclosure?
Disclosure Appropriateness	Which one reveals information more appropriately in response to the doctor's questions, without unnecessarily withholding or volunteering information?
Nonverbal Naturalness	Which patient's body language, emotion, and voice fit better with what they say and with the clinical situation?
Overall Quality	Overall, which one feels more like a real patient?

A Human Evaluation Details

A.1 Evaluation Criteria

We adapt the Maastricht Assessment of Simulated Patients (MaSP) (Wind et al., 2004), which evaluates the quality of human SP portrayals across a broad range of behavioral and interactional dimensions. Because MaSP was originally designed to assess individual human SPs, we reformulate the relevant dimensions as pairwise questions for comparing two virtual patient simulations. Table 4 lists the criteria and the corresponding questions presented to annotators.

B Refinement Criteria

Each piece of diagnosed guidance d_k is compared, via cosine similarity over sentence embeddings (all-MiniLM-L6-v2), against the most similar existing skill within the same section (and, for nonverbal guidance, the same field tag f_i). Let s denote this similarity to the closest matching skill s_i.

Added. If $s < \tau_{\text{add}}$ ($\tau_{\text{add}} = 0.30$), no sufficiently related skill exists, and the guidance is added to the library as a new skill with a fresh identifier and helpful/harmful counters initialized to zero.

Merged or Generalized. If $\tau_{\text{add}} \leq s < \tau_{\text{merge}}$ ($\tau_{\text{merge}} = 0.75$), an LLM judge determines whether the new guidance and the matched skill s_i describe the same underlying behavior expressed differently, in which case s_i is *generalized*: its content is rewritten into a more precise statement (capped at 280 characters) that subsumes both the original skill and the new guidance, while its identifier and counters are retained. If the judge instead determines the guidance is sufficiently distinct, it is added as a separate skill.

If $s \geq \tau_{\text{merge}}$, the guidance is treated as confirming an existing skill and is *merged* without rewriting its content: only the counters are updated.

In both the merged and the generalized case, the matched skill's counters are updated according to the polarity of the diagnosed guidance: guidance arising from a successful interaction (*reinforce*) increments h_i^+; guidance arising from a discrepancy (*correct*) increments h_i^-.

Pruned. After all guidance from a given evolution step has been processed, every skill s_i satisfying

$$h_i^- - h_i^+ \geq 2 \quad \text{and} \quad h_i^+ + h_i^- \geq 3$$

is removed from the library. This condition requires both a sustained imbalance toward harmful outcomes and a minimum amount of evidence, preventing skills from being pruned on the basis of a single early failure.

Every add, merge, generalize, and prune operation is written to an append-only log, allowing each change to the library to be traced back to the diagnosed discrepancy that triggered it.

Algorithm 1 Refinement($\mathcal{S}^{(k)}$, d_k)

for each piece of guidance $g \in d_k$ **do**
  $s_i, s \leftarrow$ most similar skill to g in matching section (and field, if nonverbal)
  if $s < \tau_{\text{add}}$ **then**
    add g as a new skill
  else if $s < \tau_{\text{merge}}$ **then**
    if judge(g, s_i) decides "same behavior" **then**
      generalize s_i: rewrite content to subsume g
    else
      add g as a new skill
    end if
    update (h_i^+, h_i^-) of s_i by polarity of g
  else
    merge: update (h_i^+, h_i^-) of s_i by polarity of g
  end if
end for
prune every skill s_i with $h_i^- - h_i^+ \geq 2$ and $h_i^+ + h_i^- \geq 3$
return updated library $\mathcal{S}^{(k+1)}$

C Human Validation of Behavioral Annotations

To assess the reliability of PRISM's automatically generated behavioral annotations, we conduct a blinded human validation study on 411 patient clips, obtained by sampling 10% of the clips from each video. Two annotators independently label each clip using the same closed-set schema as the automated pipeline, without access to the model predictions. Since these dimensions describe everyday observable behaviors rather than clinical judgments, no specialized medical expertise is required for annotation. For each dimension, we retain the clips on which both annotators agree as a human-consensus reference and measure the agreement between the automated annotations and this reference. Single-label dimensions are evaluated by exact match, intensity by agreement within ± 1 on the 1–5 scale, and multi-label dimensions by exact-set match together with MASI similarity, a Jaccard-based measure that gives partial credit for overlapping label sets (Passonneau, 2006).

Since no universal threshold exists for judging automated annotations, we follow common practice in LLM-based annotation and evaluation and interpret model agreement with reference to human–human agreement (**?**): dimensions on which human annotators themselves rarely agree cannot be expected to reach high agreement with any single reference. We therefore report inter-annotator agreement alongside model agreement as an indicator of the intrinsic difficulty of each dimension. Note that model agreement is computed on the human-consensus subset and is thus not directly comparable in absolute value to inter-annotator agreement.

As shown in Table 5, seven of the ten dimensions achieve over 70% agreement with human consensus, with emotion trajectory and intensity reaching about 90%. The remaining three dimensions (body posture, head movement, and emotion) show lower agreement (51–56%). These dimensions are also among those with the lowest inter-annotator agreement (26.5–47.2%), suggesting that much of the difficulty stems from the inherent ambiguity of these behavioral signals rather than from the automated annotation alone.

Low agreement on emotion is consistent with prior work: categorical emotion annotation is inherently subjective, and inter-annotator agreement in widely used datasets remains moderate at best (e.g., Fleiss' κ of 0.27 for IEMOCAP and 0.43 for MELD) (**??**). Although emotion reaches only 51.4% exact-set agreement, its MASI similarity of 0.610 indicates substantial partial overlap with

the human labels. Most partial mismatches retain the human-consensus emotions but add extra labels, reflecting over-labeling rather than missed core emotions. Head movement exhibits a different limitation: sparse frame sampling during video processing may miss brief motions such as nodding (Zhang et al., 2025), consistent with the model's tendency to predict *Still*.

Table 5: Human validation of PRISM behavioral annotations, grouped by label type. Inter-annotator agreement is computed between the two human annotators; model agreement is computed on the human-consensus subset of each dimension.

Dimension	**Inter-annotator (%)**	**Model (%)**	**Model MASI**
Single-label (exact match)			
Emotion trajectory	80.0	90.9	–
Voice emotion	66.7	76.3	–
Gaze direction	78.3	75.2	–
Hand gesture	61.6	72.3	–
Speech rate	67.4	71.5	–
Body posture	45.0	55.7	–
Head movement	47.2	52.1	–
Ordinal (within ± 1 on a 1–5 scale)			
Intensity	90.0	89.5	–
Multi-label (exact-set match)			
Facial expression	45.5	75.9	0.774
Emotion	26.5	51.4	0.610

D Behavioral Annotation Schema

Table 6 lists the label space of each behavioral dimension. *Emotions* and *facial expressions* are multi-label; all other dimensions are single-label. For visual dimensions (facial expressions, gaze direction, head movement, hand gesture, and body posture), *Not visible* is assigned when the patient is off-frame.

Table 6: Label space of each behavioral dimension. M: multi-label; S: single-label.

Dimension	**Type**	**Label space**
Emotions	M (1–3)	Anxious, Painful, Relieved, Hesitant, Neutral, Fearful, Confused, Trusting
Facial expressions	M (0–3)	Frowning (pain), Grimacing, Guarded eyes, Tearful, Relaxed, Confused, Blank
Gaze direction	S	Eye contact, Looking down, Looking away, Closed eyes, Unfocused
Head movement	S	Nodding, Nodding (agreement), Shaking, Turning away, Tilted, Still
Body posture	S	Leaning forward, Slumped, Guarded, Relaxed, Tense, Open
Hand gesture	S	Pointing to pain, Holding arm, Clasped hands, Fidgeting, Open palms, Covering face, None
Voice emotion	S	Calm, Strained, Trembling, Weak, Hesitant, Clear
Speech rate	S	Fast, Normal, Slow
Emotion trajectory	S	Escalating, De-escalating, Stable
Intensity score	S	1 (very mild) – 5 (very intense)

E Metrics

We evaluate BAPS along three complementary dimensions: verbal fidelity, multimodal behavioral fidelity, and information disclosure. Unless otherwise specified, predictions are compared with the matched real-SP turn in PRISM and aggregated over the evaluation set.

Verbal fidelity. We measure semantic and lexical similarity between the generated response y and the corresponding real-SP response y^*. *SBERT* computes the cosine similarity between their sentence embeddings,

$$\mathrm{SBERT}(y, y^*) = \frac{\mathbf{e}(y)^\top \mathbf{e}(y^*)}{\|\mathbf{e}(y)\| \, \|\mathbf{e}(y^*)\|}.$$

We additionally report *BLEU-1* and *BLEU-4*, which measure modified unigram and up-to-4-gram precision with a brevity penalty, respectively. *ROUGE-L* is reported as the F1 score derived from the longest common subsequence (LCS) between the generated and reference responses. Higher values indicate stronger verbal alignment with the real SP.

Multimodal behavioral fidelity. Metrics are selected according to the label structure of each behavioral output. Emotion and facial expression are multi-label variables, as multiple cues may co-occur within a turn. We therefore report both micro- and macro-F1. Micro-F1 aggregates true positives, false positives, and false negatives across all labels,

$$F_1^{\mathrm{micro}} = \frac{2TP}{2TP + FP + FN},$$

whereas macro-F1 averages the independently computed F1 score over the C behavior classes,

$$F_1^{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} F_{1,c}.$$

Gaze direction, head movement, body posture, hand gesture, voice emotion, speech rate, and emotion trajectory are single-label categorical variables and are evaluated using accuracy,

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i].$$

Behavioral intensity is scored on a 1–5 scale and evaluated using mean absolute error (MAE),

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{s}_i - s_i|.$$

Higher F1 and accuracy indicate better behavioral fidelity, whereas lower MAE is better.

Information disclosure. We evaluate whether clinical information is disclosed at an appropriate point using the real SP's disclosure behavior as the reference. Each information judgment is assigned to one of four states: *leak*, when information is disclosed prematurely; *withheld*, when information that should remain unrevealed is correctly withheld; *revealed*, when information is disclosed when appropriate; and *missed*, when information that should be disclosed remains withheld.

Let N_{leak}, N_{withheld}, N_{revealed}, and N_{missed} denote the numbers of judgments in these four states. *Leak Rate* measures the proportion of premature disclosures among cases in which the information should remain withheld:

$$\mathrm{LeakRate} = \frac{N_{\mathrm{leak}}}{N_{\mathrm{leak}} + N_{\mathrm{withheld}}}.$$

Lower values indicate better control over premature information release.

To capture overall disclosure appropriateness, we further define the *Disclosure Score* by assigning a reward to each state:

$$r(s) = \begin{cases} -1.0, & s = \mathrm{leak}, \\ 0.3, & s = \mathrm{withheld}, \\ 1.0, & s = \mathrm{revealed}, \\ -0.2, & s = \mathrm{missed}. \end{cases}$$

The final score is the mean reward over all disclosure judgments:

$$\mathrm{DisclosureScore} = \frac{-N_{\mathrm{leak}} + 0.3N_{\mathrm{withheld}} + N_{\mathrm{revealed}} - 0.2N_{\mathrm{missed}}}{N_{\mathrm{leak}} + N_{\mathrm{withheld}} + N_{\mathrm{revealed}} + N_{\mathrm{missed}}}.$$

This asymmetric scoring penalizes premature disclosure most strongly, while also rewarding appropriately timed revelation and correctly withheld information. A higher Disclosure Score therefore indicates closer overall alignment with the real SP's disclosure behavior.

Table 7: Distribution of PRISM consultations across medical specialties.

Medical specialty	**# Consultations**
Internal Medicine	48
Psychiatry	20
Neurology	11
Orthopaedic Surgery	7
Urology	5
Dermatology	4
Family Medicine	4
Preventive Medicine	4
Pediatrics	3
Obstetrics & Gynecology	2
Ophthalmology	2
Otolaryngology	2
Palliative Care	2
Anesthesiology	1
Total	**115**

F PRISM DATASET