PREDICTIVELY SUFFICIENT MULTIMODAL SOCIAL WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning latent states that support future evolution remains an open problem for multimodal interactive systems. Existing approaches typically optimize representations for current semantic understanding or direct prediction, without explicitly defining what information a latent state should preserve to remain predictive of future dynamics. In this work, we investigate this problem through the principle of predictive sufficiency, which defines a predictive state as a latent variable that retains future-relevant information from historical multimodal observations while reducing redundant dependence on the observed history. Based on this principle, we propose Predictively Sufficient Multimodal Social World Modeling (PSM-WM), a framework that jointly learns predictive latent states and their stochastic temporal evolution through probabilistic state inference, event-conditioned transitions, and multi-step consistency optimization. Beyond forecasting performance, we introduce state-centric evaluations to measure whether learned representations satisfy predictive state properties, including predictive closure, future semantic alignment, and transition consistency. Experiments on benchmarks demonstrate that PSM-WM improves future multimodal interaction modeling and learns latent states with stronger predictive properties and robustness. These results indicate that predictive sufficiency provides an effective principle for constructing future-oriented multimodal states and highlights the importance of state construction in multimodal world modeling.

1 INTRODUCTION

Building intelligent systems capable of interacting with dynamic environments requires internal states that not only describe the present but also support future reasoning and evolution. Recent advances in world models have demonstrated the importance of learning latent dynamics for predicting future states and enabling long-horizon decision making Hafner et al. (2025). Meanwhile, predictive representation learning has shown that representations optimized for future information can provide more generalizable abstractions beyond conventional discriminative objectives LeCun (2022). These studies collectively suggest that future-oriented representations are essential for intelligent systems. However, a fundamental question remains open: what information should a latent state preserve such that future evolution can be predicted from the state itself?

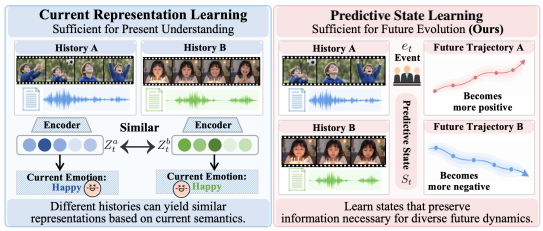


Figure 1: Motivation of predictive state learning. Identical current semantics may correspond to different future dynamics. Predictive state learning aims to construct latent states that preserve information necessary for future evolution.

Multimodal social interaction provides a challenging scenario for studying this question. Recent multimodal representation learning methods have achieved significant progress in modeling human behaviors by exploiting cross-modal interaction, modality-specific information, and semantic alignment Wang et al. (2025); Hazarika et al. (2020); Meng et al. (2026a). Transformer-based approaches learn flexible interactions among heterogeneous modalities Tsai et al. (2019), while representation

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.

disentanglement methods such as MISA Hazarika et al. (2020) and subsequent multimodal learning frameworks Meng et al. (2026b;c); Fang et al. (2025) further improve the separation and integration of complementary information. However, these approaches mainly evaluate representations according to their ability to explain current observations or predict current semantic states. Such representations may preserve discriminative information for the present while discarding historical factors that determine future interaction dynamics. As illustrated in Fig. 1, identical current semantics may correspond to different future interaction dynamics.

Beyond representation learning, future forecasting methods explicitly model temporal evolution by predicting future outcomes from historical observations. Although these methods demonstrate that historical context contains useful future information, they typically formulate forecasting as a direct mapping $H_t \rightarrow F_{t+k}$, without explicitly characterizing whether the learned latent representation itself contains the information required for future evolution. Similarly, predictive representation learning methods optimize future consistency between representations LeCun (2022), but generally focus on learning predictive embeddings rather than defining the principle of a predictive state. Therefore, accurate future prediction does not necessarily imply that a model has discovered a reusable latent state. World models provide another perspective by learning latent dynamics that describe how states evolve over time. Existing latent dynamics models, including PlaNet and Dreamer-style architectures, learn transition mechanisms over latent variables for long-horizon prediction Hafner et al. (2025). Recent social world modeling studies further investigate event-driven human interaction evolution Yu et al. (2026). However, most world models assume that meaningful latent states are already available or focus primarily on transition modeling. Consequently, an important challenge remains unresolved: How should a latent state be constructed such that it preserves exactly the information necessary for future evolution? (More related work in Appendix A).

We investigate this problem from the perspective of predictive sufficiency. Inspired by the principle of sufficient statistics, we define a predictive state as a latent variable that preserves future-relevant information from historical multimodal observations while reducing the additional predictive dependence on the observed history. Formally, an ideal predictive state S_t should satisfy $p(F_{t+1:t+k}|H_t, E_t) \approx p(F_{t+1:t+k}|S_t, E_t)$, which indicates that once the predictive state is known, the remaining historical observations provide limited additional information about future evolution. Equivalently, this objective can be interpreted as minimizing $I(F_{t+1:t+k}; H_t|S_t, E_t)$. This perspective shifts multimodal representation learning from constructing states that explain current observations toward learning states that preserve future evolution dynamics.

Based on this principle, we propose Predictively Sufficient Multimodal Social World Modeling (PSM-WM), a framework for learning future-oriented multimodal latent states in social interactions (Fig. 2). Our contribution is a principled formulation of how future-oriented latent states should be constructed. We introduce predictive sufficiency as a state learning criterion, which separates two fundamental aspects of world models: what information should be preserved in a latent state and how such a state evolves over time. Based on this principle, PSM-WM jointly learns predictive state construction and stochastic latent dynamics. PSM-WM consists of three complementary components: probabilistic belief state inference for modeling uncertainty from partially observed multimodal histories, predictive sufficiency guided state construction for preserving future-relevant information, and event-conditioned stochastic transition modeling for capturing latent state evolution. Furthermore, we introduce state-centric evaluation protocols that go beyond conventional forecasting accuracy and examine whether learned states retain future predictive information. Experiments on Hi-EF, IEMOCAP, and MELD demonstrate that optimizing latent states according to predictive sufficiency improves future multimodal interaction modeling. Our results suggest that predictive sufficiency provides a useful criterion for constructing future-oriented multimodal states. Our contributions are summarized as follows:

- We identify predictive sufficiency as a fundamental criterion for constructing future-oriented multimodal states.
- We develop PSM-WM, a framework that operationalizes this criterion through probabilistic state inference and stochastic latent evolution modeling.
- We establish state-centric evaluations and empirical analyses to examine whether learned representations preserve future-relevant information beyond direct forecasting.

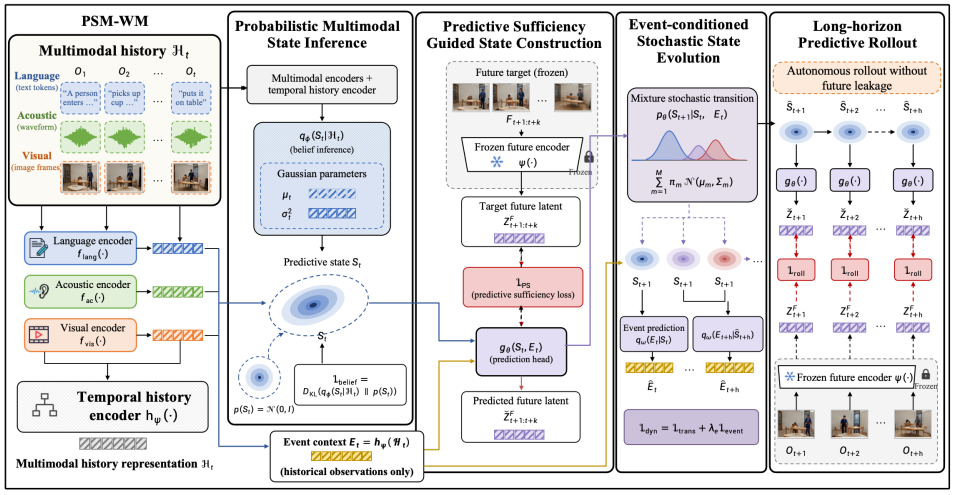


Figure 2: The framework learns predictive multimodal states through probabilistic belief state inference, predictive sufficiency guided state construction, and event-conditioned stochastic transition modeling, enabling long-horizon prediction of multimodal interaction dynamics.

2 METHOD

2.1 PROBLEM FORMULATION AND PREDICTIVE STATE PRINCIPLE

We consider multimodal social interaction sequences consisting of language, acoustic, and visual observations. At time step t, the observed multimodal history is denoted as $H_t = \{O_1, O_2, \cdots, O_t\}$, where each observation contains multiple modalities $O_t = \{X_t^l, X_t^a, X_t^v\}$. Given the historical context H_t, the goal is to learn a latent state S_t that supports future interaction evolution. Different from conventional representation learning that optimizes states according to current prediction objectives, we define a predictive state according to the information required for future dynamics.

Conditional Predictive State Principle. A meaningful latent state should preserve information required for future interaction evolution while reducing the additional predictive dependence on historical observations. Instead of assuming that the latent state alone determines future dynamics, we define predictive states together with an observable event context. Specifically, a predictive state S_t should satisfy, $p(F_{t+1:t+k}|H_t, E_t) \approx p(F_{t+1:t+k}|S_t, E_t)$, where $F_{t+1:t+k}$ denotes the future interaction trajectory within a predefined prediction horizon. This formulation indicates that after observing the learned state and event context, the remaining historical observations provide limited additional information for the evaluated future prediction task.

2.2 PROBABILISTIC MULTIMODAL STATE INFERENCE

The latent social state cannot be directly observed from multimodal interaction signals. Instead, it must be inferred from heterogeneous observations that may contain ambiguity and incomplete evidence. Therefore, we represent the latent state as a probabilistic belief distribution $q_\phi(S_t|H_t) = \mathcal{N}(\mu_t, \sigma_t^2)$, where μ_t represents the estimated latent state and σ_t^2 captures uncertainty induced by incomplete observations. To regularize the latent state space and prevent arbitrary uncertainty estimation, we introduce a prior distribution $p(S_t) = \mathcal{N}(0, I)$, and optimize the belief regularization objective $L_{belief} = D_{KL}(q_\phi(S_t|H_t)||p(S_t))$. The inferred belief state provides the foundation for subsequent predictive state construction and stochastic transition learning.

2.3 PREDICTIVE SUFFICIENCY GUIDED STATE CONSTRUCTION

The central problem of predictive state learning is determining what information should be preserved in the latent state. A representation optimized only for current-state prediction may retain

discriminative features while discarding historical factors that influence future evolution. Therefore, we construct latent states according to predictive sufficiency.

Given the inferred belief state S_t, an ideal predictive state should preserve all future-relevant information contained in the historical context. Following the predictive sufficiency principle introduced in Section 2.1, this requirement can be expressed as $p(F_{t+1:t+k}|H_t, E_t) \approx p(F_{t+1:t+k}|S_t, E_t)$. Equivalently, the residual predictive information from the historical context after observing the latent state should be minimized $I(F_{t+1:t+k}; H_t|S_t, E_t) \to 0$. Directly optimizing this objective is challenging because the complete future distribution is unknown and high-dimensional. Therefore, we derive a practical approximation by maximizing the predictive information retained in the latent state. Specifically, we introduce a future representation encoder $\psi(\cdot)$ that extracts future semantic states $Z^F_{t+1:t+k} = \psi(F_{t+1:t+k})$, and learn a predictive mapping from the current latent state $\hat{Z}^F_{t+1:t+k} = g_\theta(S_t, E_t)$. The predictive sufficiency objective is approximated by minimizing the distance between predicted and target future representations $\mathcal{L}_{pred} = d(\hat{Z}^F_{t+1:t+k}, Z^F_{t+1:t+k})$. This objective encourages the latent state to preserve information that is predictive of future interaction evolution. Importantly, $\mathcal{L}_{pred}$ serves as an empirical surrogate objective for predictive information preservation.

The future encoder ψ is pretrained and frozen during optimization. Therefore, Z^F provides a fixed predictive target rather than an adaptive representation. The objective can be interpreted as predicting a fixed future representation, rather than directly optimizing the unknown future distribution.

2.4 Event-conditioned Stochastic State Evolution

A predictive state should not only preserve future-relevant information, but also support modeling how latent interaction states evolve over time. However, social interactions are inherently stochastic and may contain multiple plausible future trajectories. Therefore, deterministic latent transitions are insufficient for capturing the uncertainty and diversity of future evolution. Given the predictive state S_t, we first extract the current interaction event from available observations $E_t = h_\psi(H_t)$, where $h_\psi(\cdot)$ only receives historical observations available at time t. No future information is involved in event extraction. To model the stochastic nature of social evolution, we formulate the latent transition as a mixture distribution:

$$p_\theta(S_{t+1}|S_t, E_t) = \sum_{m=1}^{M} \pi_m(S_t, E_t)\mathcal{N}(\mu_m(S_t, E_t), \sigma^2_m(S_t, E_t)), \tag{1}$$

where π_m represents the probability of each possible evolution mode. This formulation enables the world model to maintain multiple future hypotheses instead of collapsing uncertain trajectories into a single averaged prediction.

To enable autonomous future rollout, we further learn an event predictor that estimates interaction conditions from latent states. Specifically, the event inference network is optimized by $L_{event} = -\log q_\omega(E_t|S_t)$, where the event label E_t is obtained from the historical context encoder $h_\psi(H_t)$ and does not involve future observations. This objective encourages the latent state to preserve information necessary for predicting future interaction conditions, while preventing future event information from being directly accessed during inference.

During inference, future interaction events are unavailable. Therefore, instead of directly accessing future observations, we infer future event distributions from predicted latent states $q_\omega(E_{t+h}|S_{t+h})$, which enables autonomous latent rollout without future information leakage. The transition process during rollout is therefore, $\hat{E}_{t+h} \sim q_\omega(E|\hat{S}_{t+h})$, followed by, $\hat{S}_{t+h+1} \sim p_\theta(S_{t+h+1}|\hat{S}_{t+h}, \hat{E}_{t+h})$. The transition and event prediction objectives are jointly optimized as:

$$L_{dyn} = -\log p_\theta(S_{t+1}|S_t, E_t) - \lambda_e \log q_\omega(E_t|S_t). \tag{2}$$

This event-conditioned stochastic transition provides two important properties. First, event inference separates observable interaction conditions from latent evolution, enabling leakage-free future rollout. Second, mixture-based stochastic dynamics allow the model to represent diverse future interaction trajectories rather than a single deterministic evolution path.

2.5 Overall Optimization Objective

The objective of PSM-WM is to learn latent states that preserve future-relevant information and support consistent future evolution. Accordingly, the optimization objective consists of five complementary terms corresponding to different properties of predictive states.

First, the belief regularization objective constrains the probabilistic state inference:

$$\mathcal{L}_{belief} = D_{KL}(q_\phi(S_t|H_t)||p(S_t)). \quad \mathcal{L}_{PS} = d(g_\theta(S_t, E_t), Z^F_{t+1:t+k}). \tag{3}$$

This objective regularizes the latent state distribution and enables uncertainty-aware state inference. Second, predictive sufficiency is approximated by future representation prediction $\mathcal{L}_{PS}$.

This objective encourages the latent state to preserve information required for future interaction evolution. Third, the transition objective learns the evolution mechanism between predictive states:

$$\mathcal{L}_{trans} = -\log p_\theta(S_{t+1}|S_t, E_t). \quad \mathcal{L}_{roll} = \sum_{h=1}^{H} \gamma^{h-1} d(\hat{Z}_{t+h}, Z^F_{t+h}). \tag{4}$$

Finally, to improve long-horizon consistency, we recursively apply the transition model and optimize future latent alignment $\mathcal{L}_{roll}$. The final training objective is:

$$L = L_{belief} + \lambda_1 L_{PS} + \lambda_2 L_{trans} + \lambda_3 L_{event} + \lambda_4 L_{roll}. \tag{5}$$

The objective jointly optimizes probabilistic state inference, predictive information preservation, event-aware stochastic transition modeling, and long-horizon latent consistency. Unlike direct forecasting objectives that only optimize future prediction accuracy, PSM-WM explicitly constrains what information should be retained in the latent state and how such states should evolve over time.

3 Experiments

3.1 Experimental Goals

Our experiments aim to evaluate whether predictive sufficiency provides an effective principle for constructing future-oriented multimodal states. Specifically, we investigate four questions: **RQ1:** Can predictive state construction improve future multimodal evolution modeling across different environments? **RQ2:** Do the learned latent states preserve future-relevant information and satisfy the properties of predictive sufficiency? **RQ3:** How do state construction and stochastic transition modeling contribute to stable long-horizon latent evolution? **RQ4:** Can predictive states remain reliable under incomplete observations and uncertain future dynamics?

3.2 Experimental Setup

Datasets. We evaluate PSM-WM on four multimodal temporal benchmarks: Hi-EF Wang et al. (2026), IEMOCAP Busso et al. (2008), MELD Poria et al. (2019), and Ego4D Grauman et al. (2022). Hi-EF, IEMOCAP, and MELD provide temporally evolving multimodal interaction scenarios with aligned language, acoustic, and visual signals, enabling the evaluation of predictive state learning in social environments. To assess generalization beyond social interactions, we additionally evaluate PSM-WM on Ego4D, which captures long-duration human-environment interactions with visual and language observations. **Evaluation Metrics.** We evaluate models from three complementary perspectives. First, future prediction performance is measured using weighted F1 score (W-F1) on social interaction benchmarks and future activity prediction accuracy on Ego4D. Second, predictive state quality is evaluated using state-centric metrics: Predictive Closure Gap (PCG), which measures the performance difference between full-history prediction and state-only prediction; Future Latent Retrieval (FLR), which evaluates whether the learned state contains information predictive of future representations; and Residual History Gain (RHG), which measures the additional predictive benefit obtained by providing the original history after observing the latent state. Third, state evolution quality is evaluated using transition consistency and long-horizon rollout performance. **Implementation Details.** All methods use identical multimodal encoders for fair comparison. Results are averaged over multiple random seeds, and mean and standard deviation are reported.

3.3 RQ1: Predictive States Improve Future Multimodal Evolution Modeling

We first investigate whether constructing latent states according to predictive sufficiency improves future multimodal evolution modeling. To this end, we compare PSM-WM with three categories of approaches: multimodal representation learning methods (MISA Hazarika et al. (2020), EMOE Fang et al. (2025), TSD Meng et al. (2026c), EMART Seong et al. (2026), and HyperEF Qiu et al. (2026)), predictive representation learning methods (CPC Oord et al. (2018), BYOL Grill et al. (2020), and V-JEPA Assran et al. (2025)), and latent dynamics models (RSSM Hafner et al. (2019), DreamerV3 Hafner et al. (2025), LPWM Daniel et al. (2026), HAUWM Wan et al. (2026), and ModularAgent Zhan et al. (2026)). As shown in Table 1, PSM-WM achieves the best performance across Hi-EF, IEMOCAP, and MELD. Compared with multimodal representation learning methods, the improvements indicate that representations optimized for current semantic understanding are less effective for modeling future evolution. Compared with predictive representation learning methods, PSM-WM further benefits from explicitly constructing latent states according to future-relevant information preservation rather than only predicting future representations. Moreover, PSM-WM consistently outperforms latent dynamics baselines, achieving improvements of 0.85, 0.46, and 0.35 percentage points in W-F1 over the strongest world model baseline, ModularAgent, on Hi-EF, IEMOCAP, and MELD, respectively. These results demonstrate that effective future multimodal modeling depends not only on learning transition dynamics, but also on constructing latent states that preserve information required for future evolution.

Table 1: Future interaction prediction performance. Higher W-F1 indicates better future modeling capability.

Method	Hi-EF	IEMOCAP	MELD
Multimodal Representation Learning			
MISA	71.24	65.31	66.52
EMOE	72.18	66.02	67.14
TSD	72.83	66.71	68.21
EMART	73.12	67.03	68.64
HyperEF	73.84	67.45	68.92
Predictive Representation Learning			
CPC	72.96	67.18	67.86
BYOL	73.21	67.42	68.14
V-JEPA	74.10	67.85	68.95
Predictive State and World Models			
RSSM	74.62	67.83	69.39
DreamerV3	74.80	68.02	69.52
LPWM	74.65	68.05	69.42
HAUWM	74.75	68.10	69.55
ModularAgent	74.91	68.15	69.62
PSM-WM	**75.76**	**68.61**	**69.97**

3.4 RQ2: Analysis of Predictive State Properties

Future prediction performance alone does not reveal whether a model learns a predictive state, since accurate forecasting can still rely on implicit access to historical observations. Therefore, we evaluate whether the latent states learned by PSM-WM preserve future-relevant information under a controlled state-level probe protocol. Specifically, we examine three complementary properties of predictive states. PCG measures how closely state-only prediction approaches prediction from complete historical observations. RHG evaluates whether the original history provides additional predictive benefits after observing the learned state. FLR further examines whether the learned state maintains alignment with future interaction semantics. All evaluations use frozen representations and identical prediction probes, ensuring that the measured differences originate from the learned states rather than predictor capacity.

As shown in Table 2, PSM-WM achieves the lowest PCG and RHG while obtaining the strongest future latent retrieval performance. The reduced PCG indicates that predictive sufficiency guided state construction preserves more information required for future prediction. Meanwhile, the low RHG suggests that the learned state leaves less residual predictive information in the observed history under the evaluated probe setting. The retrieval results further demonstrate that the learned states capture latent structures aligned with future interaction semantics rather than only optimizing prediction outputs. These analyses provide empirical evidence that PSM-WM learns future-oriented latent states with stronger predictive properties. These measurements characterize predictive information preservation under the defined prediction tasks and do not claim a strict guarantee of minimal sufficient statistics.

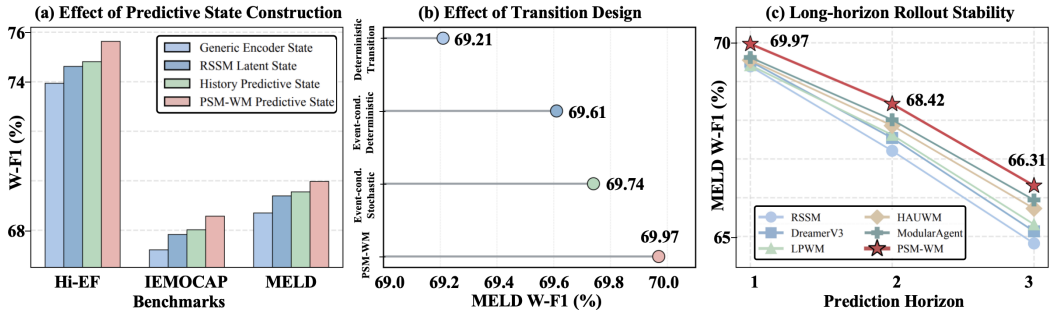


Figure 3: Predictive state evolution analysis. (a) Predictive states improve latent representations under fixed transitions. (b) Event-conditioned stochastic dynamics enhance state evolution. (c) Predictive states enable stable long-horizon rollout.

Table 2: Evaluation of predictive state properties. PCG and RHG measure the remaining predictive gap and residual historical information, respectively. Future latent retrieval evaluates alignment with future semantics. Lower PCG/RHG and higher retrieval indicate stronger predictive sufficiency.

	Predictive Closure Gap (↓)			**Residual History Test**			**Future Latent Retrieval (↑)**	
Method	Hi-EF	IEMOCAP	MELD	State Only	State+History	RHG (↓)	Recall@1	MRR
MISA	8.72	9.15	8.41	65.12	67.47	2.35	32.80	0.440
CPC	6.53	6.91	6.30	-	-	-	40.51	0.543
BYOL	6.18	6.55	5.94	67.12	68.86	1.74	41.73	0.563
RSSM	5.12	5.34	5.00	69.21	70.43	1.22	43.82	0.591
LPWM	4.18	4.42	4.05	69.42	70.24	0.82	45.36	0.624
HAUWM	4.02	4.31	3.92	69.55	70.31	0.76	45.72	0.636
ModularAgent	3.86	4.18	3.75	69.62	70.28	0.66	46.15	0.632
PSM-WM	**2.30**	**2.54**	**2.08**	**69.97**	**70.58**	**0.61**	**47.81**	**0.656**

3.5 Does Predictive State Construction Matter?

We hypothesize that latent state quality is determined by the information preserved during state construction rather than solely by transition modeling. To verify this, we perform a controlled comparison where all components except state construction are fixed, including the multimodal encoder, latent dimension, transition model, event representation, future target, and optimization budget. We compare four state construction strategies: Current Representation, Future Prediction State, Future Prediction with VIB, and PSM-WM. Current Representation optimizes states for present semantic prediction, while Future Prediction State learns future-aware representations without explicit information constraints. Future Prediction with VIB further introduces generic compression through a variational information bottleneck. As shown in Fig. 4(a), future prediction improves over current representation learning, and VIB provides additional compression benefits. However, PSM-WM achieves the lowest PCG and RHG, indicating that predictive state construction is more effective than generic

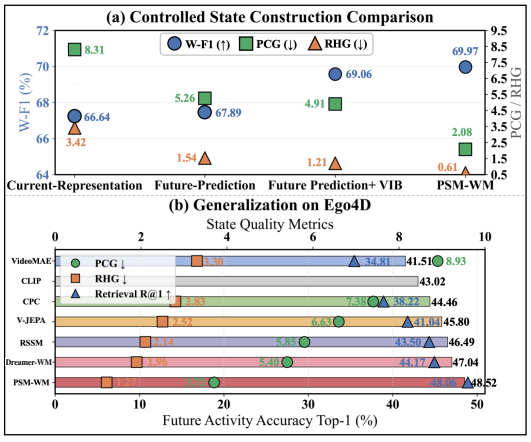


Figure 4: (a) Predictive sufficiency guided state construction improves future modeling under controlled settings with identical architectures and optimization budgets. (b) PSM-WM generalizes to Ego4D by preserving future-relevant information beyond social interaction scenarios.

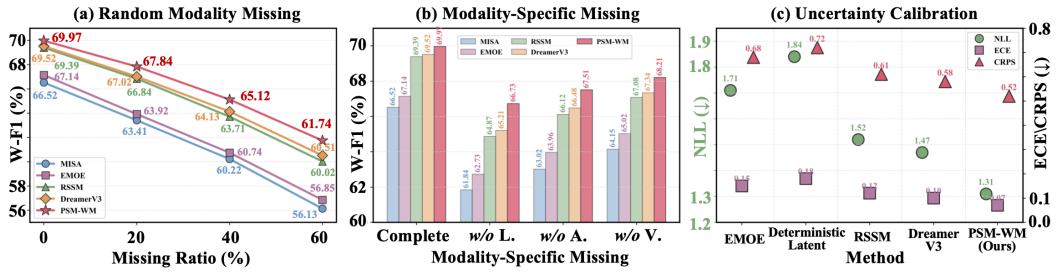


Figure 5: Robustness and uncertainty calibration analysis of PSM-WM. (a) Future prediction performance under varying random modality missing ratios. (b) Robustness under modality-specific missing conditions. (c) Calibration quality of predictive uncertainty using NLL, ECE, and CRPS.

future prediction or compression. These results
isolate the contribution of state construction and demonstrate that the improvement originates from preserving future-relevant information.

3.6 RQ3: Predictive State Evolution Analysis

Predictive states determine what information is preserved, while world models require transition mechanisms to model how such states evolve. Therefore, we analyze the contribution of state construction and latent dynamics from three aspects: predictive state construction, event-conditioned stochastic transitions, and long-horizon rollout stability. **Effect of Predictive State Construction.** We first isolate the impact of state construction by fixing the transition model and only replacing the latent representation. As shown in Fig. 3(a), predictive states consistently outperform generic latent representations, demonstrating that preserving future-relevant information provides a more effective basis for future interaction modeling. **Effect of Event-conditioned Stochastic Transition.** Using the same predictive state representation, we further evaluate the transition mechanism. Event conditioning improves future prediction by incorporating interaction context, while stochastic transitions further enhance performance by capturing diverse evolution trajectories (Fig. 3(b)). This verifies that contextual and uncertainty modeling complement predictive state construction. **Long-horizon Rollout Stability.** We finally evaluate recursive latent rollout under different prediction horizons. PSM-WM maintains slower performance degradation as the rollout horizon increases, indicating that predictive states provide a more stable latent space for modeling future evolution (Fig. 3(c)). These results demonstrate that predictive state construction and latent transition modeling address complementary aspects of world modeling: state construction determines the information retained for future prediction, while transition modeling determines how this information evolves over time.

3.7 RQ4: Robustness under Incomplete Observations and Uncertainty

Multimodal interactions often suffer from missing observations and uncertain future dynamics. Therefore, we evaluate whether PSM-WM maintains reliable future prediction under incomplete evidence and whether its uncertainty estimates are properly calibrated. **Robustness under Incomplete Observations.** We evaluate robustness under both random modality missing ratios and modality-specific missing scenarios. All methods degrade with increasing missing observations, while PSM-WM exhibits slower performance degradation (Fig. 5(a)). The modality-specific results in Fig. 5(b) further show consistent improvements across different missing patterns. These results indicate that probabilistic belief inference helps preserve future-relevant information when multimodal evidence is incomplete. **Uncertainty Calibration Analysis.** We further evaluate whether the estimated uncertainty reflects prediction reliability using negative log-likelihood (NLL), expected calibration error (ECE), and continuous ranked probability score (CRPS). PSM-WM achieves better calibration than deterministic representations and stochastic latent dynamics models (Fig. 5(c)). This demonstrates that the learned uncertainty captures future prediction variability rather than introducing unstructured stochasticity. The results show that probabilistic predictive states improve robustness under incomplete observations and provide more reliable uncertainty estimation for future interaction modeling.

3.8 Ablation Studies

We conduct ablation studies to examine the contribution of each component in PSM-WM, including predictive sufficiency guided state construction, probabilistic belief inference, event-conditioned transitions, stochastic dynamics, and rollout consistency. All variants are trained and evaluated under the same protocol as the full model. We additionally evaluate the ablated models on Ego4D to examine the generality of each component beyond social interaction scenarios. Removing any component consistently degrades performance across both benchmarks (Table 3). The largest degradation is observed when removing the predictive sufficiency objective, highlighting the importance of future-oriented state construction for learning predictive latent representations. Removing probabilistic belief inference also reduces performance, demonstrating the benefit of uncertainty-aware state estimation under incomplete multimodal observations. Event-conditioned transitions, stochastic dynamics, and rollout consistency further contribute to modeling context-dependent and long-horizon state evolution. The results verify that PSM-WM relies on the joint design of state construction and latent dynamics modeling. Predictive state learning determines the information preserved in latent representations, whereas transition modeling governs their evolution over time.

Table 3: Ablation study of PSM-WM.

Variant	Social W-F1	Ego4D Acc.
PSM-WM	**69.97**	**48.5**
w/o Predictive Suff-Obje	68.76	46.9
w/o Belief State	69.12	47.4
w/o Event Conditioning	69.31	47.7
w/o Stochastic Transition	69.42	47.8
w/o Rollout Consistency	69.28	47.3

3.9 Generalization to Egocentric Multimodal World Modeling

To evaluate whether predictive sufficiency extends beyond social interaction, we further test PSM-WM on Ego4D. Compared with conversational scenarios, Ego4D involves different temporal dynamics and partial observability, providing a broader evaluation of future-oriented state construction. We compare PSM-WM with current-state representation methods (VideoMAE and CLIP-based temporal encoders), predictive representation methods (CPC and V-JEPA), and latent world models (RSSM-based and Dreamer-style models). These baselines cover different strategies for learning multimodal temporal representations and modeling future evolution. **Future Evolution Modeling Beyond Social Interaction.** Fig. 4(b) reports future prediction performance and state-centric evaluations on Ego4D. PSM-WM achieves stronger future prediction while obtaining lower predictive closure gap and residual history gain, together with improved future latent retrieval. These results indicate that predictive sufficiency guided state construction preserves future-relevant information under a different interaction domain. The results demonstrate that predictive state learning is not limited to social interactions, but can provide a general criterion for constructing future-oriented multimodal latent states.

4 Conclusion

This work studies how multimodal latent states should be constructed to support future evolution modeling. Different from conventional approaches that optimize representations mainly for current understanding or direct prediction, we investigate predictive states that preserve information necessary for future dynamics. We propose Predictively Sufficient Multimodal Social World Modeling (PSM-WM), which constructs latent states based on predictive sufficiency and models their evolution through probabilistic belief inference and stochastic latent transitions. Extensive experiments demonstrate that PSM-WM improves future interaction modeling across multiple benchmarks and learns latent states with stronger predictive properties, future semantic alignment, and robustness under incomplete observations. Further analyses verify the contribution of future-oriented state construction and latent dynamics modeling, showing that effective multimodal world models require both appropriate state representation and reliable evolution mechanisms. These results highlight that a predictive state is not merely a compact summary of historical observations, but a latent variable that preserves the information required to make future evolution predictable.

AI USE STATEMENT

In this work, we used generative AI tools for Paper polishing and grammar proofreading. We have not used generative AI tools for Method design and Experiment. The grammar of the text polished and proofread using LLM has been verified by the authors for accuracy. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies predictive multimodal state learning using existing research benchmarks. Multimodal social and egocentric data can contain sensitive personal information, and learned predictions may reflect dataset biases. PSM-WM is intended as a research framework; downstream applications should respect privacy, consent, and dataset access conditions, and should not treat its predictions as definitive assessments of individuals.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

Tal Daniel, Carl Qi, Dan Haramati, Amir Zadeh, Chuan Li, Aviv Tamar, Deepak Pathak, and David Held. Latent particle world models: Self-supervised object-centric stochastic dynamics modeling. In *International Conference on Learning Representations*, pp. 127719–127767, 2026.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. EMOE: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14314–14324. IEEE/CVF, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 2555–2565, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pp. 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. Aoept: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning. *arXiv preprint arXiv:2605.24816*, 2026.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL `https://openreview.net/forum?id=BZ5a1r-kVsf`. Version 0.9.2, June 27, 2022.

Chunlei Meng, Pengbin Feng, Rong Fu, Hoi Leong Lee, Xiaojing Du, Zhaolu Kang, Zeyu Zhang, Weilin Zhou, Chun Ouyang, and Zhongxue Gan. Group cognition learning: Making everything better through governed two-stage agents collaboration, 2026a.

Chunlei Meng, Guanhong Huang, Rong Fu, Runmin Jian, Zhongxue Gan, and Chun Ouyang. Clcr: Cross-level semantic collaborative representation for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1606–1615. IEEE/CVF, 2026b.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8791–8800. IEEE/CVF, 2026c.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Xihang Qiu, Yuhao Fang, Qing Zhou, Bin Zhai, Jialong Hong, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Beyond missing modalities: Hypergraph conditioned diffusion for uncertainty-aware multimodal emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22953–22963, 2026.

Eunseon Seong, Harim Lee, Dahye Kim, Changhyun Kim, and Dong-Kyu Chae. Emotion-wheel-guided audio-referred text representation for multimodal emotion recognition in conversation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40392–40403. Association for Computational Linguistics, 2026.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Shenghua Wan, Le Gan, and De-Chuan Zhan. Learning to be uncertain: Pre-training world models with horizon-calibrated uncertainty. In *International Conference on Learning Representations*, volume 2026, pp. 106107–106127, 2026.

Haoran Wang, Xinji Mai, Zeng Tao, Junxiong Lin, Xuan Tong, Ivy Pan, Shaoqi Yan, Yan Wang, and Shuyong Gao. Hi-ef: Benchmarking emotion forecasting in human-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2110–2118, 2026.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 21180–21188. AAAI Press, 2025.

Haofei Yu, Yining Zhao, Guanyu Lin, and Jiaxuan You. Building social world models with large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Wxqr2fkZmX.

Yu-Wei Zhan, Xin Wang, Pengzhe Mao, Tongtong Feng, Ren Wang, and Wenwu Zhu. Modularagent: A task-aware modular framework for joint optimization of multimodal large language models and world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8087–8096, 2026.

A Related Work

A.1 Multimodal Representation Learning

Recent advances in multimodal learning have primarily focused on learning effective representations from heterogeneous modalities for current-state understanding Lang et al. (2026); Grill et al. (2020). Existing approaches investigate cross-modal interaction, modality-specific representation learning, and uncertainty-aware fusion mechanisms. EMART Seong et al. (2026) explores the asymmetric roles of different modalities in conversational emotion recognition and introduces audio-guided text representation learning with emotion-structure-aware contrastive optimization. By incorporating psychological emotion structures into representation learning, EMART improves the geometric organization of multimodal emotion representations. HyperEF Qiu et al. (2026) further studies reliable multimodal learning under incomplete observations. The method introduces hypergraph-conditioned diffusion to recover missing modality representations and incorporates uncertainty-aware evidence fusion for robust prediction. These studies have substantially advanced multimodal perception by improving the quality and reliability of current-state representations. However, their objectives primarily characterize representation quality according to present-state discrimination. Consequently, whether these representations preserve information required for future evolution remains unexplored.

A.2 Multimodal Future Forecasting

Beyond current-state recognition, recent studies have investigated future prediction in human interaction scenarios Oord et al. (2018). Hi-EF Wang et al. (2026) introduces human-interaction-based emotion forecasting and constructs a benchmark for predicting future emotional states from historical multimodal contexts Meng et al. (2026c). The study demonstrates that interaction history contains predictive signals beyond instantaneous observations. Trajectory-based emotion forecasting approaches further emphasize the temporal nature of emotional evolution by modeling historical state transitions. Although these methods demonstrate that future outcomes can be predicted from historical observations, they mainly formulate forecasting as a direct mapping $H_t \rightarrow Y_{t+1}$. Such formulations do not explicitly characterize the latent state that underlies future transitions. A model may obtain accurate future predictions while still relying on historical information implicitly stored in the input sequence. Therefore, future prediction ability does not necessarily imply the discovery of a reusable predictive state.

A.3 World Models and Latent Dynamics

Recent world modeling studies have shifted attention from predicting isolated outputs toward learning latent states and their evolution mechanisms Grill et al. (2020); Hafner et al. (2019). Social World Model investigates event-driven social dynamics and models state transition as $P(S_{t+1}|S_t, E_t)$, highlighting the importance of external events in shaping future social evolution. Other probabilistic world models investigate uncertainty-aware future prediction and demonstrate the importance of modeling stochastic transition processes Daniel et al. (2026). However, existing world models mainly focus on learning transition mechanisms over latent variables, while the explicit construction of future-oriented latent states receives comparatively less attention. In multimodal social in-

teraction, the latent state itself must be inferred from heterogeneous and partially observed signals. Therefore, learning a predictive latent state remains a fundamental challenge.

A.4 Predictive State Learning

Our work connects multimodal representation learning and world modeling by investigating whether learned representations satisfy predictive sufficiency Wan et al. (2026); Hafner et al. (2025). Different from previous studies that define representations according to current discrimination or directly optimize future prediction, we define a meaningful state according to whether it preserves information necessary for future dynamics $P(Future|History, Event) \approx P(Future|State, Event)$. Based on this principle, we develop a multimodal social world model that jointly learns latent state inference and event-conditioned stochastic transition.

B Implementation Details of World Model Baselines

B.1 Unified Multimodal World Modeling Protocol

Existing world modeling approaches have been developed under different application scenarios, including visual control, representation learning, and autonomous interaction. Their original formulations differ in observation spaces, latent state parameterization, and transition objectives. To enable a consistent comparison in multimodal social interaction scenarios, we establish a unified multimodal world modeling protocol.

All approaches receive the same multimodal observations consisting of language, acoustic, and visual signals, $O_t = \{X_t^l, X_t^a, X_t^v\}$. Following the common multimodal preprocessing pipeline, each modality is first encoded using the same modality-specific feature extractors, $z_t^m = f_m(X_t^m), \quad m \in \{l, a, v\}$. The modality representations are then aggregated into a unified observation representation, $z_t^o = f_{agg}(z_t^l, z_t^a, z_t^v)$, which is provided as input to each baseline. Therefore, the comparison focuses on differences in latent state construction and temporal evolution modeling, while controlling the influence of modality encoding capacity.

B.2 Adaptation of Latent Dynamics World Models

Latent dynamics approaches are originally designed to model temporal evolution over latent variables. In our experiments, we adapt these methods to multimodal social interactions while preserving their original state transition principles. For RSSM-based approaches, the deterministic and stochastic latent states are inferred from the shared multimodal observation representation, $z_t^s, z_t^d = q(z_t^o, z_{t-1}^s)$. The transition model predicts future latent states according to the original recurrent state-space formulation. For DreamerV3, we replace the original visual observation encoder with the unified multimodal observation encoder while maintaining its latent dynamics learning objective. The model learns stochastic latent transitions and predicts future latent states through recursive rollout. For LPWM and HAUWM, we follow their original probabilistic transition formulations and replace task-specific observation modules with the shared multimodal interface. For ModularAgent, we adopt the multimodal latent interaction setting and preserve its modular optimization strategy. The only modification is the replacement of task-specific input representations with the unified multimodal observation representation.

B.3 Adaptation of Predictive Representation Learning Methods

Predictive representation learning methods optimize latent representations according to future consistency. To compare their ability to construct future-oriented representations, we adapt CPC, BYOL, and V-JEPA under the same multimodal observation setting. For each method, the multimodal observation representation z_t^o is used as input, and the original predictive objectives are maintained. CPC maximizes the mutual information between current and future latent representations through contrastive prediction. BYOL learns temporally predictive representations by minimizing the distance between online and target networks. V-JEPA predicts future latent representations without reconstructing raw observations. The comparison therefore evaluates the difference between predictive representation objectives and predictive state construction.

B.4 Training and Optimization Settings

All models are trained using the same dataset splits, multimodal features, optimization framework, and evaluation protocol. For each baseline, hyperparameters are selected according to the original implementations and adjusted only when necessary to fit the multimodal social interaction setting. All models are optimized until convergence using the validation set. The reported results are averaged over multiple random seeds, and both mean performance and standard deviation are reported. The parameter number, computational cost, and inference latency are also measured to characterize the relationship between performance and model complexity.

B.5 Unified State-level Evaluation

Beyond future prediction accuracy, we evaluate whether different methods learn latent states that preserve information required for future evolution.

For fair comparison, all learned representations are first frozen after training. A lightweight prediction probe is then independently trained on the training split, with hyperparameters selected on the validation split and evaluated once on the test split. All probes predict the same future target $F_{t+1:t+k}$ under an identical prediction horizon. The predictor architecture, optimization budget, and training procedure are kept identical across different representations. We evaluate three properties: **(1) Predictive closure.** whether the learned state approaches the predictive capability of complete historical observations. **(2) Residual history gain.** whether the original history provides additional predictive improvement after observing the learned state. **(3) Future latent alignment.** We evaluate future latent retrieval to measure whether current latent states preserve information correlated with future interaction semantics. For each test sample, the predicted future representation is compared against a candidate pool consisting of future representations from different temporal positions and different interaction sequences. Future representations from the same timestamp are excluded to avoid trivial matching. The future target encoder is pretrained and frozen before evaluation. Retrieval similarity is measured using cosine similarity, and Recall@1 and mean reciprocal rank (MRR) are reported.

B.6 Predictive Information Decomposition

To further examine whether predictive information is mainly preserved in the learned latent state rather than auxiliary event variables, we perform a predictive information decomposition analysis.

Given the historical observation H_t, PSM-WM produces a latent predictive state S_t and an event context E_t. We evaluate future prediction using different combinations of available information $\{H_t, S_t, E_t, (S_t, E_t), (S_t, E_t, H_t)\}$. Here, H_t represents the complete historical observation, S_t represents the learned predictive state, and E_t represents the observable event context extracted from historical observations. All representations are frozen after model training. For each input configuration, an identical prediction probe is independently trained using the training split, selected on validation data, and evaluated once on the test split. Therefore, performance differences reflect information contained in different inputs rather than differences in optimization capacity. The contribution of each information source is further quantified by:

$$\Delta_S = M(S, E) - M(E), \Delta_E = M(S, E) - M(S), \Delta_H = M(S, E, H) - M(S, E). \quad (6)$$

A predictive state should provide substantial predictive information, while the additional gain from accessing the complete history should remain limited.

The decomposition results provide additional evidence that the predictive state, rather than the event context, carries the majority of future-relevant information. Across both MELD and IEMOCAP, the predictive state alone substantially outperforms the event-only setting, while combining the event context with the state provides only marginal improvements. Furthermore, the performance gap between (S_t, E_t) and the complete history remains small, indicating that the learned state captures most of the predictive information available from historical observations. These results do not imply that the learned state is a minimal sufficient statistic. Instead, they demonstrate that under the evaluated future prediction tasks, PSM-WM successfully transfers future-relevant information from historical observations into a compact latent representation.

Table 4: Predictive information decomposition on MELD and IEMOCAP. All representations are frozen and evaluated with identical prediction probes. Higher W-F1 indicates stronger future prediction capability.

Input Configuration	MELD	IEMOCAP
E_t only	67.85	66.42
S_t only	69.97	68.61
(S_t, E_t)	70.05	68.70
H_t	70.58	69.02
(S_t, E_t, H_t)	70.70	69.15
Δ_S	2.20	2.28
Δ_E	0.08	0.09
Δ_H	0.65	0.45

Table 5: Representation similarity for samples with identical current semantics but different future trajectories. Lower similarity indicates better future trajectory discrimination.

Representation	Cosine Similarity
Current Representation	0.824
Future Prediction State	0.673
PSM-WM	**0.511**

C Case Study: Current Semantic Similarity Does Not Guarantee Future Similarity

Existing representation learning methods mainly optimize current semantic prediction. Therefore, histories with similar current semantics may be mapped to nearby representations even when their future trajectories differ. To examine this phenomenon, we select sample pairs with identical current labels but different future trajectories (Table 5). We compare representation similarity in different latent spaces. A desirable predictive state should distinguish these cases because future evolution depends on historical context beyond current observations.

D Discussion and Limitations

D.1 Predictive Sufficiency as a State Learning Principle

This work investigates predictive sufficiency as a principle for constructing future-oriented multimodal states. Different from conventional representation learning objectives that mainly optimize current semantic understanding, predictive sufficiency emphasizes preserving information required for future evolution. Predictive sufficiency is optimized through empirical variational objectives and therefore provides an approximate criterion rather than an exact sufficient statistic. Therefore, predictive sufficiency should be interpreted as a criterion for guiding state construction rather than a strict guarantee of complete future predictability.

D.2 Relationship to Existing Predictive Learning and World Models

PSM-WM is related to predictive representation learning and latent world models, but differs from both directions in its objective. Predictive representation learning methods, such as CPC and JEPA, aim to learn representations that preserve information useful for predicting future observations. However, they typically focus on representation consistency or future feature prediction without explicitly defining what information a latent state should preserve. Latent world models, such as RSSM-based approaches, focus on learning transition dynamics over latent variables. These methods assume that an appropriate latent state representation is available, while the problem of constructing such states remains less explored.

In contrast, PSM-WM separates state construction from state evolution. It first defines predictive states according to future information preservation and then learns stochastic dynamics over these states.

D.3 Limitations and Future Works

Although PSM-WM demonstrates the effectiveness of predictive state learning for multimodal social interaction modeling, several limitations remain.

First, predictive sufficiency is evaluated through empirical proxies, including predictive closure, future latent retrieval, and residual history analysis. Developing stronger theoretical guarantees for multimodal predictive states remains an important future direction. Second, current experiments focus on social interaction scenarios with relatively structured temporal dynamics. Extending predictive state learning to open-world environments with complex agent interactions and long-horizon planning is an interesting direction. Third, the current formulation models uncertainty through probabilistic latent states. Future work may investigate richer uncertainty representations, including multimodal future distributions and causal state transitions.