PSWS: LEARNING PREDICTIVE SEMANTIC WORLD STATE FOR MULTIMODAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most multimodal understanding models optimize representations for recognizing the semantic label of the current observation. Such supervision, however, does not explicitly require the representation to preserve evidence that is predictive of subsequent semantic changes. We formalize this limitation as Semantic Transition Aliasing (STA): histories with the same current label but different next-label transitions may remain insufficiently separated in the learned state space. We propose Predictive Semantic World State (PSWS) model, a task-grounded and action-free predictive semantic-state model that infers a current state from past-only textual, acoustic, and visual observations and predicts its one-step evolution without accessing future observations at inference. PSWS factorizes current-semantic and transition-oriented evidence through dedicated latent queries, predicts modality- and interaction-specific state increments using an evidence-conditioned router, and aligns the predicted state with an EMA-encoded future target available only during training. A shared decoder grounds current and predicted states in the same label space. PSWS improves over reported current-recognition references and yields gains in future-label prediction and changed-transition performance under a unified causal protocol with supervision- and modality-matched baselines. Additional analyses evaluate transition separation, routing behavior, statistical reliability, and computational cost.

1 INTRODUCTION

Multimodal understanding aims to infer task-relevant semantics from heterogeneous observations, including language, acoustic behavior, visual behavior, and their temporal interactions (Hazarika et al., 2020; Meng et al., 2026b; Wu et al., 2025; Lin et al., 2025). Most existing methods formulate this problem as current-state recognition, $\mathbf{X}_{\leq t}^{T,A,V} \longrightarrow \mathbf{z}_t \longrightarrow y_t$, where the learned representation is optimized to predict the current label (Wang et al., 2025; Meng et al., 2026c; Jin et al., 2025). Cross-modal attention, graph reasoning, representation disentanglement, contrastive learning, and adaptive fusion have substantially improved this paradigm (Yang et al., 2022; Fang et al., 2025; Meng et al., 2026a). Nevertheless, recognizing what the current observation expresses is not equivalent to modeling how the underlying semantic condition may evolve (Zhu et al., 2025; Zhang et al., 2023). Two observations may share the same current label but contain different prosodic, visual, historical, or cross-modal cues that imply different next states (Li et al., 2025b; Zhuang et al., 2025; Li et al., 2023b). This motivates our central question: can multimodal representations be learned as predictive semantic states, rather than only as features for current-label recognition?

The observed task label is only a many-to-one projection of a richer task-relevant state. Let $\mathbf{S}_t$ denote the semantic world state induced by the observed multimodal history $y_t = \pi(\mathbf{S}_t) + \epsilon_t$, where $\pi(\cdot)$ is a task-dependent projection and ϵ_t represents annotation ambiguity or task noise. Since $\pi(\cdot)$ is generally many-to-one, two histories may satisfy $y_t^i = y_t^j, \quad y_{t+1}^i \neq y_{t+1}^j$. An encoder optimized only for y_t has no explicit incentive to retain the evidence required to distinguish these transitions and may produce $\mathbf{z}_t^i \approx \mathbf{z}_t^j$. We term this representation failure **Semantic Transition Aliasing** (STA). Although the resulting representation may remain effective for current recognition, it can be insufficient for anticipatory multimodal understanding.

Several lines of multimodal research partially address the phenomena underlying STA (Zhuang et al., 2025; Chen et al., 2023). Context-aware graph models capture temporal and cross-modal depen-

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

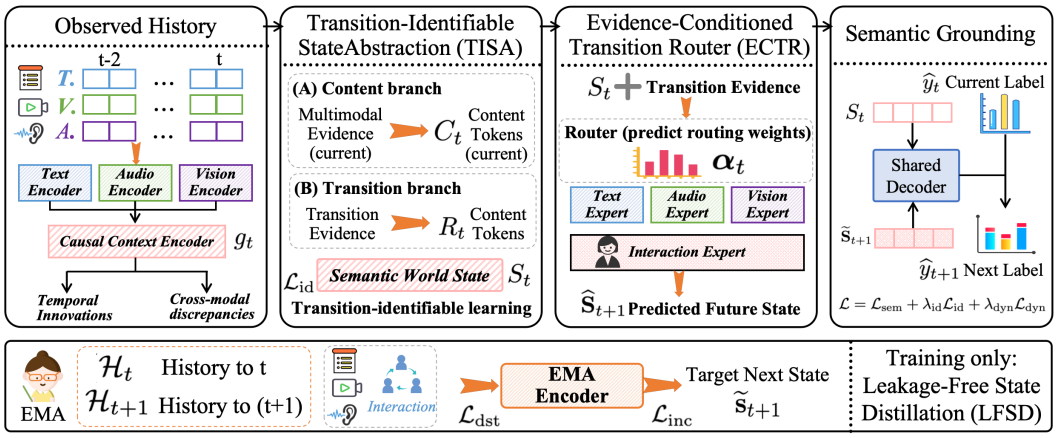


Figure 1: Overview of PSWS. Given past-only multimodal history, TISA separates current semantic content from transition-relevant evidence to construct the semantic world state. ECTR then predicts a routed state increment through modality-specific and interaction experts, producing the one-step future state. LFSD provides leakage-free future-state supervision during training, while a shared decoder grounds both current and predicted states in the same semantic space.

dencies, while GraphSmile further introduces an auxiliary sentiment dynamics perception task to distinguish sentiment shifts between utterances (Li et al., 2025a; 2023a) (More Related Work in Appendix A). Representation-learning methods preserve heterogeneous evidence by decomposing multimodal features into shared and modality-specific components or contrastively learned similarity and dissimilarity components (Hazarika et al., 2020; Yang et al., 2023). Adaptive expert and routing methods further adjust the contributions of modalities or interaction paths according to individual samples (Fang et al., 2025; Chen et al., 2024). These strategies improve contextual sensitivity, feature specialization, and sample-adaptive fusion, but they remain primarily optimized for current-label recognition or explicit shift discrimination. They do not require the current representation to distinguish histories that share the same current label but imply different next states, nor do they jointly model an explicit latent-state increment with future-state supervision under a strictly past-only predictor. Consequently, they can capture aspects of semantic dynamics without directly resolving the representation-level cause of STA. Addressing semantic transition aliasing requires solving three coupled challenges. First, the state representation must preserve transition-relevant distinctions without sacrificing current semantic recognition. Second, semantic evolution is sample-dependent: a transition may be driven primarily by language in one case, by acoustic or visual changes in another, or by cross-modal inconsistency in a third. Third, standard multimodal datasets provide task labels but no explicit latent-state annotations. Future observations can be used as training targets, but they must not be used as predictor inputs under a past-only inference protocol.

We propose PSWS, a Multimodal Semantic World-State Model for one-step predictive multimodal understanding. Given only the observed history H_t, PSWS constructs a task-grounded semantic state S_t, predicts its next state $\widehat{\mathbf{S}}_{t+1}$, and decodes both the current and future labels. PSWS is not intended to reconstruct the physical world or infer unobservable private mental states; instead, it learns a compact semantic state grounded in observable multimodal interaction. The model integrates three components: Transition-Identifiable State Abstraction, Evidence-Conditioned Transition Routing, and Leakage-Free State Distillation. Together, they encourage the state to remain discriminative for the present while preserving information predictive of one-step semantic evolution. Our contributions are summarized as follows:

- We identify Semantic Transition Aliasing, a representation failure where histories with the same current label but different subsequent transitions are collapsed by current-label-oriented learning, and we formalize multimodal understanding as task-grounded predictive semantic-state learning.

- We propose PSWS, which combines transition-identifiable state abstraction, evidence-conditioned residual transition routing, and leakage-free future-state distillation to model one-step semantic evolution from past-only multimodal evidence.
- We introduce transition-oriented evaluation protocols, including supervision-matched causal baselines, transition-signature retrieval, separation margin, and changed-transition prediction, to directly assess whether learned states preserve transition-relevant semantic information.

2 METHOD

2.1 PROBLEM FORMULATION

We consider a temporally ordered multimodal sequence $\mathcal{X} = \left\{\left(\mathbf{X}_t^T, \mathbf{X}_t^A, \mathbf{X}_t^V\right)\right\}_{t=1}^{L}$, where T, A, and V denote the textual, acoustic, and visual modalities. The observable history at time t is $\mathcal{H}_t = \left\{\mathbf{X}_1^{T,A,V}, \ldots, \mathbf{X}_t^{T,A,V}\right\}$. A transition pair $(t, t+1)$ is retained only when both observations belong to the same sequence. Given $\mathcal{H}_t$, PSWS constructs the current semantic world state $\mathbf{S}_t$ and predicts its next state $\mathcal{H}_t \longrightarrow \mathbf{S}_t \longrightarrow \widehat{\mathbf{S}}_{t+1}$, with current and future labels decoded from $\mathbf{S}_t$ and $\widehat{\mathbf{S}}_{t+1}$, respectively. A valid state should retain information for both current and future semantic decoding:

$$p(y_t \mid \mathcal{H}_t) \approx p(y_t \mid \mathbf{S}_t), \qquad p(y_{t+1} \mid \mathcal{H}_t) \approx p(y_{t+1} \mid \mathbf{S}_t). \tag{1}$$

We define potentially aliased sample pairs as $\mathcal{A} = \left\{(i,j) \;\middle|\; y_t^i = y_t^j,\ y_{t+1}^i \neq y_{t+1}^j\right\}$. Semantic transition aliasing occurs when pairs in $\mathcal{A}$ are mapped to excessively similar representations. PSWS addresses this problem through transition-identifiable abstraction, evidence-conditioned transition modeling, and leakage-free future-state supervision.

As shown in Fig. 1, PSWS is an action-free, task-relevant, and one-step semantic world-state model. It summarizes the semantic condition observable from multimodal interaction history and predicts its short-term evolution. It does not attempt to reconstruct the complete physical environment, infer an unobservable private mental state, or model the full distribution of long-term futures.

2.2 MULTIMODAL EVIDENCE ENCODING

For modality $m \in \mathcal{M} = \{T, A, V\}$, we compute $\mathbf{Z}_t^m = \mathrm{LN}\left(\mathrm{Proj}_m\left(f_m(\mathbf{X}_t^m)\right) + \mathbf{P}_t^m\right) \in \mathbb{R}^{L_m \times d}$, where $f_m(\cdot)$ is the modality-specific encoder, $\mathrm{Proj}_m(\cdot)$ maps its output to dimension d, and $\mathbf{P}_t^m$ denotes positional encoding. An attentive pooling operator produces an observation-level descriptor $\mathbf{u}_t^m = \mathrm{Pool}\left(\mathbf{Z}_t^m\right)$. A causal context encoder then summarizes the observed multimodal history $\mathbf{g}_t = \mathrm{CtxEnc}\left([\mathbf{u}_1^T; \mathbf{u}_1^A; \mathbf{u}_1^V], \ldots, [\mathbf{u}_t^T; \mathbf{u}_t^A; \mathbf{u}_t^V]\right)$. We describe modality-specific changes using the temporal innovation $\boldsymbol{\delta}_t^m = \mathbf{u}_t^m - \mathbf{u}_{t-1}^m$, where $\boldsymbol{\delta}_1^m$ is initialized to zero. Cross-modal discrepancy is defined as $\kappa_t^{mn} = 1 - \cos\left(\mathbf{W}_m \mathbf{u}_t^m, \mathbf{W}_n \mathbf{u}_t^n\right), \quad (m,n) \in \{(T,A),(T,V),(A,V)\}$. We collect the three discrepancies as $\boldsymbol{\kappa}_t = [\kappa_t^{TA}, \kappa_t^{TV}, \kappa_t^{AV}]$. The innovations capture modality-wise changes, whereas $\boldsymbol{\kappa}_t$ represents cross-modal agreement or contradiction.

2.3 TRANSITION-IDENTIFIABLE STATE ABSTRACTION (TISA)

TISA separates current semantic content from transition-relevant evidence and explicitly preserves differences among distinct future transitions.

Content and Transition Tokens. We introduce two learnable query banks $\mathbf{Q}^c, \mathbf{Q}^r \in \mathbb{R}^{K \times d}$. The content branch attends to the current multimodal evidence:

$$\mathbf{E}_t = [\mathbf{Z}_t^T; \mathbf{Z}_t^A; \mathbf{Z}_t^V; \mathbf{W}_g \mathbf{g}_t], \qquad \mathbf{C}_t = \mathrm{Attn}\left(\mathbf{Q}^c, \mathbf{E}_t, \mathbf{E}_t\right). \tag{2}$$

For transition modeling, we construct modality descriptors $\mathbf{d}_t^m = \mathbf{W}_d^m[\mathbf{u}_t^m; \boldsymbol{\delta}_t^m], \quad m \in \{T, A, V\}$, and a discrepancy descriptor $\mathbf{d}_t^\kappa = \mathbf{W}_\kappa \boldsymbol{\kappa}_t$. The transition evidence and tokens are:

$$\mathbf{D}_t = [\mathbf{d}_t^T; \mathbf{d}_t^A; \mathbf{d}_t^V; \mathbf{d}_t^\kappa; \mathbf{W}_g^r \mathbf{g}_t], \qquad \mathbf{R}_t = \mathrm{Attn}\left(\mathbf{Q}^r, \mathbf{D}_t, \mathbf{D}_t\right). \tag{3}$$

Since $\mathbf{C}_t$ and $\mathbf{R}_t$ originate from different query banks, we align them through cross-attention rather than direct index-wise addition $\mathbf{F}_t = \mathrm{Attn}\left(\mathbf{C}_t, \mathbf{W}_r\mathbf{R}_t, \mathbf{W}_r\mathbf{R}_t\right)$. The semantic world state is $\mathbf{S}_t = \mathrm{LN}\left(\mathbf{C}_t + \mathbf{F}_t\right) \in \mathbb{R}^{K\times d}$.

Transition-Identifiable Learning.

For each valid transition, we define its semantic transition signature $\boldsymbol{\sigma}_t = (y_t, y_{t+1})$. Samples sharing the same signature form positive pairs, whereas samples with the same current label but different next labels are emphasized as hard negatives. We project the pooled transition tokens as $\mathbf{r}_i = \mathrm{Norm}\left(q_\psi\left(\mathrm{Pool}(\mathbf{R}_i)\right)\right)$, and define $\mathcal{P}(i) = \{p \neq i \mid \boldsymbol{\sigma}_p = \boldsymbol{\sigma}_i\}$. The negative weighting coefficient is $w_{ia} = 1 + \gamma\mathbb{I}\left[y_t^a = y_t^i \wedge y_{t+1}^a \neq y_{t+1}^i\right]$. The transition-identifiable objective is

$$\mathcal{L}_{\mathrm{id}} = -\sum_i \frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)} \log \frac{\exp\left(\mathrm{sim}(\mathbf{r}_i, \mathbf{r}_p)/\rho\right)}{\sum_{a\neq i} w_{ia}\exp\left(\mathrm{sim}(\mathbf{r}_i, \mathbf{r}_a)/\rho\right)}, \tag{4}$$

where ρ is the temperature. A transition-balanced sampler or memory queue ensures that each anchor has at least one positive sample. Unlike conventional class-level contrastive learning, $\mathcal{L}_{\mathrm{id}}$ preserves distinctions that are irrelevant to the current label but predictive of future transitions.

2.4 Evidence-Conditioned Transition Router

Semantic evolution may be driven by different evidence across samples. ECTR therefore models textual, acoustic, visual, and cross-modal state increments.

For each modality, we define $\Delta\mathbf{S}_t^m = \mathcal{T}_m\left(\mathbf{S}_t, \mathbf{d}_t^m\right), m \in \{T, A, V\}$, where $\mathcal{T}_m$ is an evidence-conditioned transition expert. A cross-modal expert models interaction-driven changes $\Delta\mathbf{S}_t^I = \mathcal{T}_I\left(\mathbf{S}_t, [\mathbf{d}_t^T; \mathbf{d}_t^A; \mathbf{d}_t^V; \mathbf{d}_t^\kappa]\right)$. The routing weights are predicted from the current state and transition evidence, $\boldsymbol{\alpha}_t = \mathrm{softmax}\left(\Phi_\alpha[\mathrm{Pool}(\mathbf{S}_t); \boldsymbol{\delta}_t^T; \boldsymbol{\delta}_t^A; \boldsymbol{\delta}_t^V; \boldsymbol{\kappa}_t]\right)$, where $\boldsymbol{\alpha}_t = [\alpha_t^T, \alpha_t^A, \alpha_t^V, \alpha_t^I], \sum_e \alpha_t^e = 1$. The routed state increment is $\Delta\widehat{\mathbf{S}}_t$. We obtain the predicted future state through residual evolution:

$$\Delta\widehat{\mathbf{S}}_t = \sum_{m\in\{T,A,V\}} \alpha_t^m \Delta\mathbf{S}_t^m + \alpha_t^I \Delta\mathbf{S}_t^I. \qquad \widetilde{\mathbf{S}}_{t+1} = \mathbf{S}_t + \Delta\widehat{\mathbf{S}}_t, \qquad \widehat{\mathbf{S}}_{t+1} = \mathrm{LN}\left(\widetilde{\mathbf{S}}_{t+1}\right). \tag{5}$$

This formulation models how the current semantic world state changes, rather than generating an unrelated future representation.

2.5 Leakage-Free State Distillation

Multimodal datasets provide semantic labels but no latent-state annotations. LFSD therefore uses future observations only to construct training targets and never exposes them to the student predictor. Let Φ_θ denote the student state encoder. Its EMA teacher $\Phi_{\bar{\theta}}$ is updated as $\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta$. The teacher produces stop-gradient targets, $\mathbf{S}_t^\star = \mathrm{sg}\left(\Phi_{\bar{\theta}}(\mathcal{H}_t)\right)$, $\mathbf{S}_{t+1}^\star = \mathrm{sg}\left(\Phi_{\bar{\theta}}(\mathcal{H}_{t+1})\right)$. These targets represent slowly evolving semantic encodings of the observed continuation rather than externally annotated latent states. To avoid assuming strict token-wise correspondence, distillation is performed in the pooled state space. $\mathbf{s}_t = \mathrm{Pool}(\mathbf{S}_t)$, $\widetilde{\mathbf{s}}_{t+1} = \mathrm{Pool}(\widetilde{\mathbf{S}}_{t+1})$, $\widehat{\mathbf{s}}_{t+1} = \mathrm{Pool}(\widehat{\mathbf{S}}_{t+1})$, and $\mathbf{s}_t^\star = \mathrm{Pool}(\mathbf{S}_t^\star)$, $\mathbf{s}_{t+1}^\star = \mathrm{Pool}(\mathbf{S}_{t+1}^\star)$. The destination loss aligns the predicted and teacher future states $\mathcal{L}_{\mathrm{dst}}$. We additionally supervise the semantic increment $\mathcal{L}_{\mathrm{inc}}$. The dynamics objective is $\mathcal{L}_{\mathrm{dyn}} = \mathcal{L}_{\mathrm{dst}} + \eta\mathcal{L}_{\mathrm{inc}}$. At inference $\widehat{\mathbf{S}}_{t+1} = \mathcal{F}_\theta\left(\mathcal{H}_t\right)$, and neither the future observation nor the EMA teacher is required.

$$\mathcal{L}_{\mathrm{dst}} = 1 - \cos\left(\widehat{\mathbf{s}}_{t+1}, \mathbf{s}_{t+1}^\star\right), \qquad \mathcal{L}_{\mathrm{inc}} = \frac{1}{d}\left\|(\widetilde{\mathbf{s}}_{t+1} - \mathbf{s}_t) - \mathrm{sg}\left(\mathbf{s}_{t+1}^\star - \mathbf{s}_t^\star\right)\right\|_1. \tag{6}$$

2.6 Semantic Grounding and Overall Objective

A shared decoder $h(\cdot)$ maps the current and predicted states into the same task space, $\widehat{y}_t = h(\mathrm{Pool}(\mathbf{S}_t))$ and $\widehat{y}_{t+1} = h(\mathrm{Pool}(\widehat{\mathbf{S}}_{t+1}))$. The semantic objective is $\mathcal{L}_{\mathrm{sem}}$ ($\mathcal{L}_{\mathrm{sem}} = \ell(\widehat{y}_t, y_t) + \beta\ell(\widehat{y}_{t+1}, y_{t+1})$), where $\ell(\cdot)$ is the task loss and β balances current and future decoding. The complete training objective is $\mathcal{L}$ ($\mathcal{L} = \mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{dyn}}\mathcal{L}_{\mathrm{dyn}}$). Here, $\mathcal{L}_{\mathrm{sem}}$ provides task grounding,

Table 1: Current-state recognition under the GraphSmile MERC setting. † denotes published baseline results reported in GraphSmile (Li et al., 2025a).

Method	IEMOCAP-6 (%)		IEMOCAP-4 (%)		MELD (%)		CMU-MOSEI (%)	
	ACC	WF1	ACC	WF1	ACC	WF1	ACC	WF1
DialogueCRN†	68.70	68.82	81.34	81.28	62.38	63.32	37.88	26.55
MMGCN†	67.10	66.81	79.75	79.72	66.02	64.55	45.67	44.11
MM-DFN†	69.44	68.83	80.91	80.83	67.05	65.21	45.29	42.98
SACL-LSTM†	70.55	70.60	80.70	80.74	64.52	64.55	38.60	25.95
M3Net†	70.92	71.07	83.67	83.57	66.59	65.83	43.67	41.12
GraphSmile†	72.77	72.81	86.53	86.52	67.70	66.71	46.82	44.93
PSWS (Ours)	**74.08±0.26**	**74.19±0.24**	**87.48±0.18**	**87.56±0.17**	**68.34±0.22**	**67.52±0.23**	**47.78±0.31**	**45.98±0.33**

$\mathcal{L}_{\mathrm{id}}$ prevents transition-distinct histories from collapsing, and $\mathcal{L}_{\mathrm{dyn}}$ supervises the future-state destination and increment without future-input leakage. Current-state supervision applies to every valid observation, whereas transition-specific objectives are computed only for observations with a valid within-sequence successor.

3 Experiments

3.1 Experimental setup

Datasets and Evaluation Protocols. We evaluate PSWS on IEMOCAP-6, IEMOCAP-4 (Busso et al., 2008), MELD (Poria et al., 2019), and CMU-MOSEI (Zadeh et al., 2018). IEMOCAP and MELD are used for current recognition and one-step future prediction, while CMU-MOSEI is used for current-state recognition under the same past-only encoding protocol. Dataset splits are performed before transition construction to avoid cross-sequence leakage, and each transition pair $(t, t+1)$ is retained only within the same original sequence. Under the unified causal protocol, predictors access observations up to time t only. For controlled reproductions of non-causal baselines, future context is removed by replacing bidirectional operations with causal counterparts. We further verify the protocol by perturbing future observations and confirming unchanged predictions. Additional dataset statistics are provided in Appendix B.

Baselines. Current-recognition baseline results reported in GraphSmile (Li et al., 2025a) are included for contextual reference. The controlled future-prediction baselines in Table 2 are Persistence, Historical Transformer, GraphSmile+Next with current- and future-label supervision, GraphSmile+Transition with a residual transition head, and GraphSmile+Transition+KD with EMA future-state targets. All learned predictors in this table use the unified causal protocol; the GraphSmile variants are causalized. The external EPC results in Appendix C.2 provide additional context and are not modality-matched comparisons.

Evaluation Metrics. For current recognition, we report accuracy and weighted F1. For future prediction, we use Future Macro-F1 and Future UAR. To evaluate semantic transitions, we report Changed-Transition F1 on samples with $y_{t+1} \neq y_t$, where the Persistence baseline naturally obtains zero score. We further report transition-signature retrieval accuracy and transition-separation margin with class-balanced sampling and dialogue-level candidate exclusion. Unless otherwise stated, reproduced methods are evaluated over five random seeds.

Implementation Details. All controlled models use identical multimodal features and data splits, the same optimizer family, and causal context constraints. PSWS hyperparameters are selected only on the validation set. Unless otherwise specified, results are averaged over five seeds, with checkpoints selected by validation Future Macro-F1 for future prediction and weighted F1 for current recognition. Dataset statistics, the evaluation protocol, and computational cost are provided in the appendix.

Table 2: One-step future-state prediction under the unified causal protocol.

Method	IEMOCAP-4 (%)			MELD (%)		
	Future UAR	Future M-F1	Transition-F1	Future UAR	Future M-F1	Transition-F1
Persistence	61.40	60.72	0.00	38.82	37.94	0.00
Historical Transformer	77.20	76.65	65.30	43.62	43.05	52.40
GraphSmile+Next	79.61	79.18	67.10	45.06	44.52	54.20
GraphSmile+Transition	81.14	80.72	69.40	46.08	45.51	55.90
GraphSmile+Transition+KD	81.76	81.32	70.30	46.68	46.12	57.00
PSWS (Ours)	**82.93±0.36**	**82.55±0.38**	**72.80±0.44**	**48.06±0.48**	**47.42±0.50**	**59.00±0.55**

3.2 Results: Current Semantic Understanding

Table 1 reports the current-state recognition results under the GraphSmile MERC setting. PSWS achieves the best ACC and WF1 results across all four datasets. Compared with the strongest baseline, GraphSmile, PSWS improves WF1 by 1.38, 1.04, 0.81, and 1.05 points on IEMOCAP-6, IEMOCAP-4, MELD, and CMU-MOSEI, respectively. The corresponding ACC improvements are 1.31, 0.95, 0.64, and 0.96 points. These comparisons summarize current-recognition performance relative to published references; the controlled causal results are reported separately in Table 2. Notably, the improvements are obtained across both dyadic and multiparty conversations and across emotion or sentiment label spaces with different granularities, indicating that the benefit is not restricted to a particular dataset or interaction structure. The gains on current-state recognition remain moderate relative to those observed on transition-oriented metrics. This result is consistent with the intended role of PSWS: the model primarily preserves information relevant to future semantic evolution, while improving current recognition as a secondary benefit. The comparatively smaller improvement on MELD should also be interpreted in the context of its pronounced class imbalance and dominant neutral category, which limit the attainable gains in aggregate ACC and WF1.

Under the coarsened MSAC setting (Appendix C.1), PSWS improves GraphSmile's WF1 by 0.69-0.92 points and ACC by 0.64–0.84 points across the four datasets, showing that the benefit persists under the three-class sentiment evaluation.

3.3 One-Step Future Semantic State Prediction

As a contextual external reference, Appendix C.2 compares PSWS with dedicated EPC methods. Because prior EPC methods use Text+Speech whereas PSWS uses Text+Audio+Vision, we do not treat this comparison as a controlled ranking. The supervision-, modality-, and causality-matched evaluation below provides the primary evidence.

Unified Causal Transition Protocol. Table 2 compares PSWS with direct forecasting and latent-transition baselines under the same causal protocol. The Persistence baseline obtains relatively strong future-label performance because consecutive semantic states frequently remain unchanged, but its Transition-F1 is zero by construction. This result shows that semantic continuity alone cannot capture actual state changes. The learned baselines exhibit a clear progression. Historical Transformer improves over Persistence by modeling the observed multimodal history. GraphSmile+Next further benefits from the contextual and sentiment-dynamic representations learned by GraphSmile. Introducing an explicit residual transition function in GraphSmile+Transition produces additional gains, while future-state distillation in GraphSmile+Transition+KD further improves both future-label prediction and transition detection. These comparisons confirm that explicit transition modeling and latent future-state supervision are both useful for one-step semantic forecasting. PSWS consistently outperforms all controlled baselines. Compared with GraphSmile+Transition+KD, the strongest competing variant, PSWS improves Future UAR, Future M-F1, and Transition-F1 by 1.17, 1.23, and 2.50 points on IEMOCAP-4, respectively. On MELD, the corresponding improvements are 1.38, 1.30, and 2.00 points. The larger gains on Transition-F1 highlight improved performance on samples whose semantic labels change. The advantage over GraphSmile+Transition+KD also shows that the performance improvement cannot be fully explained by attaching a generic transition head or introducing future-state distillation alone. PSWS additionally preserves transition-discriminative evidence through its state abstraction and models sample-dependent transition drivers

Table 4: Analysis of semantic transition aliasing. Margin measures the similarity gap between samples sharing the same transition signature and those sharing the same current label but having different next labels. Ret@1 denotes transition-signature retrieval accuracy, and Shift-MF1 denotes future macro-F1 on samples satisfying $y_{t+1} \neq y_t$. Ret@1 and Shift-MF1 are percentages; Margin is unitless.

Representation	IEMOCAP-4			MELD		
	Margin	Ret@1	Shift-MF1	Margin	Ret@1	Shift-MF1
GraphSmile+Next	0.05	57.30	64.70	0.03	42.10	35.10
GraphSmile+Transition	0.10	61.60	66.80	0.07	45.80	36.40
PSWS w/o $\mathcal{L}_{\text{id}}$	0.17	68.40	68.90	0.14	50.60	37.80
PSWS	**0.29**	**74.10**	**70.90**	**0.24**	**56.20**	**39.80**

through evidence-conditioned routing. The consistent gains across IEMOCAP-4 and MELD therefore support the complementarity of transition-identifiable representation learning, evidence-conditioned state evolution, and leakage-free future-state supervision.

3.4 Ablation Study

Table 3 evaluates the contribution of each component. Removing TISA causes the largest degradation on future-oriented metrics, showing that separating content tokens from transition-relevant tokens is important for preserving next-state distinctions. Removing LFSD also substantially reduces Future Macro-F1 and Changed-Transition F1, indicating that future-state destination and increment supervision provide useful training signals beyond label supervision. Removing ECTR consistently hurts transition prediction, suggesting that a homogeneous transition function is less effective than sample-dependent routing over modality-specific and interaction-specific experts. The transition-identifiable objective further improves performance over architectural separation alone. Across ablations, the drops are larger on future and transition metrics than on current WF1, supporting the claim that the proposed components mainly strengthen predictive state modeling rather than merely increasing classifier capacity.

3.5 Analysis of Semantic Transition Aliasing

Table 4 directly evaluates Semantic Transition Aliasing. GraphSmile+Next shows limited separation between samples with the same current label but different next labels, indicating that future-label supervision alone does not necessarily organize the representation by transition signatures. Adding a generic transition head improves separation, but remains weaker than PSWS. TISA increases the transition margin even without $\mathcal{L}_{id}$, while the full objective further improves retrieval and shift-sensitive prediction. These results suggest that architectural separation and transition-identifiable contrastive learning play complementary roles. Across the compared representations, larger separation margins are accompanied by higher retrieval accuracy and Shift-MF1.

Table 3: Ablation Studies. Cu-F1 denotes current-state weighted F1, Fu-F1 denotes future macro-F1, and T-F1 denotes Transition F1.

Variant	IEMOCAP-4 (%)			MELD (%)		
	Cu-F1	Fu-F1	T-F1	Cu-F1	Fu-F1	T-F1
w/o TISA	86.79	79.63	68.10	66.82	44.48	54.20
w/o ECTR	87.12	80.52	69.60	67.06	45.33	55.50
w/o LFSD	86.91	79.88	68.90	66.90	44.72	54.80
w/o $\mathcal{L}_{\text{id}}$	87.28	81.09	70.40	67.21	46.01	56.70
Abs. state only	87.40	81.65	71.30	67.34	46.65	57.80
PSWS	**87.56**	**82.55**	**72.80**	**67.52**	**47.42**	**59.00**

3.6 Analysis of Evidence-Conditioned Transition Routing

3.6.1 Effect of Routing Design

Fig. 2 compares different routing strategies. Uniformly averaging the transition experts improves over a single shared transition function, suggesting that different evidence sources encode complementary transition patterns. Static learned routing further improves performance but remains inferior

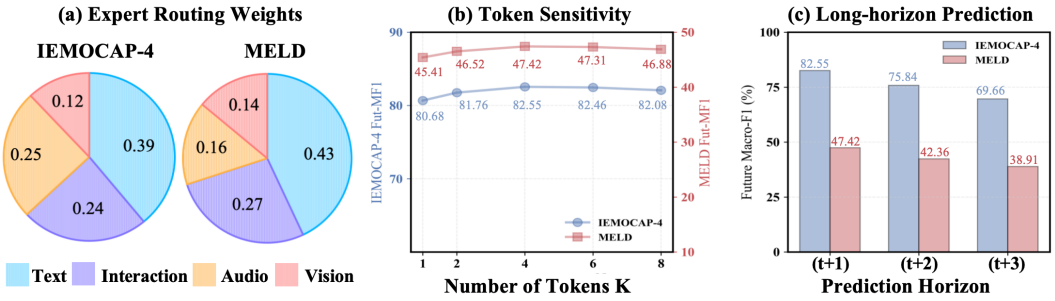


Figure 3: Analysis of dynamic state modeling in PSWS. (a) Distribution of transition expert routing weights across datasets, showing adaptive allocation among different experts. (b) Impact of the shared token number K on future prediction performance. (c) Performance degradation under increasing prediction horizons, evaluating long-horizon predictive capability.

to dynamic routing because it cannot adapt its expert preference to individual samples. Conditioning the router on modality innovations and cross-modal discrepancy yields the strongest performance, supporting the usefulness of sample-dependent evidence selection.

3.6.2 Router-Utility Alignment

We further investigate whether the learned routing weights reflect the actual contribution of each transition branch. For sample i and branch e, we define its counterfactual transition utility as the increase in future-label prediction loss after removing the corresponding innovation evidence, $u_{i,e} = \ell(\hat{\mathbf{p}}_{i,-e}, y_{i,t+1}) - \ell(\hat{\mathbf{p}}_i, y_{i,t+1})$, where $\hat{\mathbf{p}}_i$ and $\hat{\mathbf{p}}_{i,-e}$ denote the future-label distributions predicted with the complete evidence and with branch e removed, respectively. A larger $u_{i,e}$ indicates that branch e contributes more strongly to the prediction of sample i. We compute the Spearman correlation between the learned routing weights $\alpha_{i,e}$ and the corresponding counterfactual utilities $u_{i,e}$. The resulting correlations are 0.61 on IEMOCAP-4 and 0.56 on MELD, indicating that ECTR generally assigns larger weights to branches whose removal causes greater predictive degradation. This alignment provides evidence that the learned gates reflect sample-dependent transition relevance rather than arbitrary expert preferences. The mean routing distributions are reported in Fig. 3(a). Text receives the largest mean weight, while the interaction expert has mean weights 0.24 and 0.27 on IEMOCAP-4 and MELD, respectively; these averages are descriptive and do not replace the sample-level utility analysis above. This analysis should be interpreted as an empirical association, not as a causal identification of modality importance.

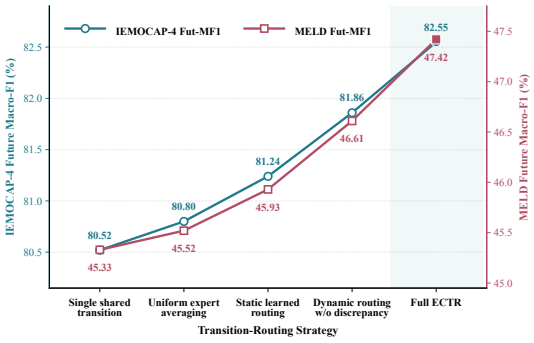


Figure 2: Transition Routing Strategies.

3.7 State Capacity and Predictive Horizon Analysis

We analyze the capacity and temporal generalization of PSWS to understand whether the learned latent state provides sufficient information for future evolution beyond immediate prediction. Specifically, we investigate the effect of state capacity by varying the shared token number K and evaluate long-horizon prediction by extending the prediction horizon beyond the training objective.

State Token Sensitivity. We vary the number of shared tokens K in the content and transition query banks while keeping other components fixed. As shown in Fig. 3(b), increasing K from 1 to 4 consistently improves Future Macro-F1 on both datasets, indicating that overly compressed states cannot fully preserve transition-relevant information. The best performance is achieved at $K = 4$,

while further increasing the token number brings no further improvement. This result suggests that effective future modeling requires an appropriate state capacity rather than simply increasing representation size.

Predictive Horizon Analysis. We further evaluate whether the learned state maintains predictive capability over longer horizons. As shown in Fig. 3(c), Future Macro-F1 gradually decreases as the prediction horizon extends from $t+1$ to $t+3$, which is consistent with the accumulation of future uncertainty. However, PSWS preserves meaningful predictive performance across multiple steps, indicating that the learned state captures temporal dynamics beyond one-step forecasting.

These analyses show that PSWS benefits from a compact yet sufficient latent state capacity and learns representations that remain informative during future state evolution.

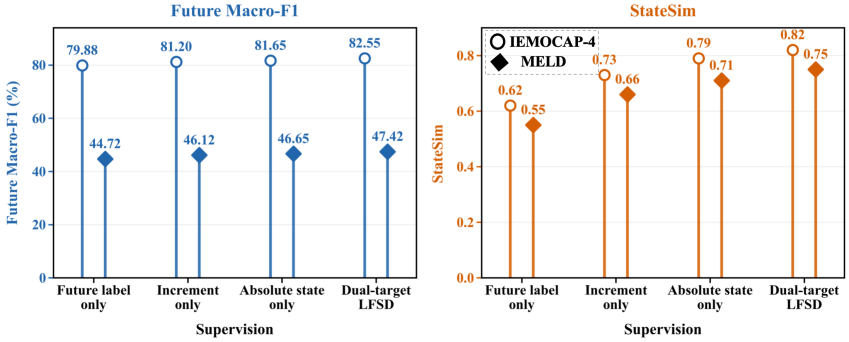


Figure 4: Analysis of leakage-free state distillation. StateSim denotes the cosine similarity between the predicted and teacher future states.

3.8 Analysis of Leakage-Free State Distillation

Fig. 4 analyzes the two supervision signals in LFSD. Future-label supervision alone provides only task-level constraints and yields the weakest future prediction performance. Increment supervision improves performance by regularizing the direction and magnitude of the predicted state change. Absolute-state supervision provides a direct target for the future semantic destination. Combining the two produces the best results, indicating that destination and increment constraints are complementary. Since future observations are used only to construct stop-gradient teacher targets during training, LFSD does not violate the past-only inference protocol.

3.9 Efficiency Analysis

Relative to GraphSmile, PSWS adds 56.1–61.2% per-epoch runtime and 22.2–31.8% GPU memory, with peak usage of about 2.25 GB (Appendix D). The EMA teacher is training-only, so part of this overhead is absent at inference.

4 Conclusions

We studied whether multimodal representations optimized for current recognition also preserve evidence predictive of short-term semantic changes. We formalized Semantic Transition Aliasing as insufficient separation between histories with the same current label but different next-label transitions, and proposed PSWS, a causal predictive semantic-state model combining transition-oriented latent queries, evidence-conditioned state increments, and future-target distillation. PSWS improves over reported current-recognition references and achieves gains in one-step future prediction, changed-transition prediction, and transition-state separation under controlled causal protocols. The model remains limited to task-grounded, deterministic, short-horizon prediction and does not represent the full distribution of plausible futures or action-conditioned physical dynamics. Extending the framework to stochastic transitions and closed-loop multi-step prediction is left for future work.

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 or experiments. 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 multimodal emotion and sentiment understanding and short-term semantic prediction using existing benchmarks. PSWS models task labels grounded in observable interactions and is not intended to infer unobservable private mental states. Any application should account for privacy, dataset bias, and the consequences of prediction errors.

REFERENCES

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. doi: 10.1007/s10579-008-9076-6.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10761–10770, 2023.

Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. D2R: Dual-Branch Dynamic Routing Network for Multimodal Sentiment Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3536–3547, 2024.

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, 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, 2020.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7042–7052, 2021a.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7037–7041, 2022.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10835–10852, 2023.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021b.

Yijie Jin, Junjie Peng, Xuanchao Lin, Haochen Yuan, Lan Wang, and Cangzhi Zheng. Multimodal Transformers are Hierarchical Modal-wise Heterogeneous Graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pp. 2188–2209, 2025.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16051–16069, 2023a.

Jiang Li, Xiaoping Wang, and Zhigang Zeng. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8786–8803, 2025a.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6631–6640, 2023b.

Ziming Li, Yaxin Liu, Chuanpeng Yang, Yan Zhou, and Songlin Hu. ROSA: A robust self-adaptive model for multimodal emotion recognition with uncertain missing modalities. *IEEE Transactions on Multimedia*, 27:6766–6779, 2025b.

Jinhao Lin, Yifei Wang, Yanwu Xu, and Qi Liu. Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 1411–1419, 2025.

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, June 2026a.

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, June 2026b.

Chunlei Meng, Ziyang Zhou, Lucas He, Xiaojing Du, Chun Ouyang, and Zhongxue Gan. Temporal-spatial decouple before act: Disentangled representation learning for multimodal sentiment analysis. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11917–11921, 2026c.

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.

Haoxiang Shi, Ziqi Liang, and Jun Yu. Emotional Cues Extraction and Fusion for Multi-modal Emotion Prediction and Recognition in Conversation. In *Interspeech 2024*, pp. 4074–4078, 2024.

Xiaohan Shi, Sixia Li, and Jianwu Dang. Dimensional Emotion Prediction Based on Interactive Context in Conversation. In *Interspeech 2020*, pp. 4193–4197, 2020.

Xiaohan Shi, Xingfeng Li, and Tomoki Toda. Emotion Awareness in Multi-utterance Turn for Improving Emotion Prediction in Multi-Speaker Conversation. In *Interspeech 2023*, pp. 765–769, 2023.

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 (ACL)*, pp. 6558–6569, 2019.

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)*, volume 39, pp. 21180–21188, 2025.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 1601–1609, 2025.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pp. 1642–1651, 2022.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7617–7630, 2023.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114, 2017.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1208.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 756–767, 2023.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pp. 22123–22138, 2025.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4626–4636, 2025.

A Related Work

A.1 Multimodal Affective Understanding

Multimodal affective understanding integrates textual, acoustic, and visual observations to infer emotion or sentiment. Early fusion methods capture high-order cross-modal interactions or exchange information across heterogeneous sequences through cross-modal attention (Zadeh et al., 2017; Tsai et al., 2019). Representation-oriented methods further separate modality-invariant and modality-specific factors, decompose multimodal features according to shared and complementary information, or coordinate representations across semantic levels (Hazarika et al., 2020; Yang et al., 2022; 2023; Meng et al., 2026b;a). Recent architectures model hierarchical modal-wise interactions or adapt the contributions of modalities and interaction paths to individual samples (Jin et al., 2025; Zhang et al., 2023; Fang et al., 2025; Chen et al., 2024; Wang et al., 2025).

Conversational affective understanding additionally requires temporal, speaker-related, and cross-utterance context modeling. DialogueCRN performs contextual reasoning over dialogue histories, whereas MM-DFN dynamically integrates multimodal contextual evidence (Hu et al., 2021a; 2022). SACL-LSTM improves the discriminability of conversational representations through supervised adversarial contrastive learning (Hu et al., 2023). Graph-based approaches such as MMGCN and M3Net represent multimodal and contextual dependencies through graph propagation (Hu et al., 2021b; Chen et al., 2023). Joyful combines joint modality fusion with graph contrastive learning (Li et al., 2023a). GraphSmile further alternates the propagation of inter-modal and intra-modal cues and introduces sentiment-shift classification as an auxiliary objective (Li et al., 2025a). These methods cover the principal contextual reasoning, dynamic fusion, graph propagation, and contrastive learning paradigms represented by the current-recognition baselines in Table 1.

These approaches substantially improve contextual modeling and current-state recognition. GraphSmile additionally encourages sensitivity to whether sentiment changes across adjacent utterances. Nevertheless, recognizing a sentiment shift does not uniquely determine the subsequent semantic state, because multiple next labels may correspond to a change from the same current label. Existing objectives therefore do not explicitly require histories with the same current label but different next labels to remain separated in the current representation. They also do not model semantic evolution as a sample-dependent residual transition under a strictly past-only inference protocol. PSWS addresses these limitations by learning a current semantic state that remains discriminative for present recognition while preserving evidence predictive of its one-step evolution.

A.2 Emotion Prediction in Conversation

Emotion Prediction in Conversation predicts a forthcoming emotional state from preceding dialogue context. DEP incorporates the interactive context of both interlocutors (Shi et al., 2020). EAMT extends context modeling to complete multi-utterance turns and multi-speaker conversations (Shi et al., 2023). Shi et al. subsequently combine word-level textual relations, prosodic enhancement, and multimodal fusion for categorical emotion prediction (Shi et al., 2024). These studies establish that historical, speaker-specific, and modality-specific cues provide useful evidence for future emotion prediction.

The principal output of these methods is the future emotion label. They do not explicitly construct a current latent state whose internal organization is evaluated according to subsequent semantic transitions. PSWS instead constructs a task-grounded state $\mathbf{S}_t$, predicts its residual evolution to $\widehat{\mathbf{S}}_{t+1}$, and examines whether histories sharing y_t but differing in y_{t+1} remain distinguishable. The external EPC results in Table 7 are therefore reported only as contextual evidence. Existing EPC methods use Text+Speech inputs, whereas PSWS uses Text+Audio+Vision, so this comparison is not treated as a modality-matched ranking.

A.3 Predictive Representation Learning and Distillation

Multimodal distillation transfers complementary information across modality-specific or fused representations to improve affective recognition (Li et al., 2023b). Such methods commonly supervise features or predictions associated with the same observed instance. Predictive representation learning instead constrains a current representation using a target derived from an unobserved or subsequent observation. This principle provides a means of learning future-relevant information without reconstructing the raw future input.

PSWS applies predictive supervision to task-grounded semantic-state learning. Its EMA teacher encodes $\mathcal{H}_t$ and $\mathcal{H}_{t+1}$ only to construct stop-gradient training targets, whereas the student predictor receives only $\mathcal{H}_t$. LFSD supervises both the destination of the predicted future state and the corresponding semantic increment. Unlike conventional feature distillation, each target is constructed from a valid within-sequence transition and is used to constrain an explicit state-evolution process. The resulting representation is evaluated through future-label prediction, changed-transition recognition, and transition-state separation, while future observations remain unavailable to the student at inference.

B Experimental Protocol and Dataset Statistics

We evaluate PSWS on IEMOCAP-6, IEMOCAP-4, MELD, and CMU-MOSEI. IEMOCAP and MELD provide dialogue-level affective sequences and support both current recognition and one-step future prediction. CMU-MOSEI provides video-level multimodal sentiment sequences and is used for current-state recognition under the same past-only encoding protocol. Dataset partitions are established at the dialogue or video level before transition construction. A transition pair $(t, t+1)$ is retained only when both labeled units belong to the same original sequence. Consequently, no pair crosses a training, validation, or test boundary.

Every predictor receives observations up to time t only. For a baseline with bidirectional context, the contextual operator is replaced by its causal counterpart. For graph-based baselines, the future context window is set to zero. We additionally perturb observations after t and verify that predictions

Table 5: Dataset statistics and eligible within-sequence adjacent pairs. A labeled unit is an utterance for IEMOCAP and MELD and a labeled video segment for CMU-MOSEI. Train+Val counts are recorded before validation extraction.

Dataset	Sequences		Labeled Units		Adjacent Pairs	
	Train+Val	Test	Train+Val	Test	Train+Val	Test
IEMOCAP-6	120	31	5,810	1,623	5,690	1,592
IEMOCAP-4	120	31	3,600	943	3,480	912
MELD	1,153	280	11,098	2,610	9,945	2,330
CMU-MOSEI	2,549	646	18,198	4,662	15,649	4,016

conditioned on $\mathcal{H}_t$ remain unchanged. For a partition containing sequence lengths $\{L_d\}_{d=1}^{D}$, the number of eligible adjacent pairs is $\sum_{d=1}^{D}(L_d - 1)$. The counts in Table 5 satisfy this identity for every dataset and partition. Train+Val counts are reported before the validation sequences are separated from that partition.

B.1 Reproducibility and Statistical Protocol

All controlled methods use the same pre-extracted multimodal features, sequence partitions, optimizer family, and causal context constraints. Hyperparameters introduced by PSWS are selected only on the validation set. Unless otherwise stated, reproduced models are trained with five random seeds (42,1110,1111,1112,1150). Checkpoints are selected by validation Future Macro-F1 for future prediction and validation weighted F1 for current recognition. Reported PSWS results use the mean and standard deviation across the five seeds (42,1110,1111,1112,1150).

C Secondary Recognition and External Prediction Context

C.1 Coarsened Multimodal Sentiment Recognition

Table 6 evaluates current-state recognition under the GraphSmile MSAC setting, where the original emotion or sentiment labels are mapped to negative, neutral, and positive sentiment classes. PSWS obtains the highest ACC and WF1 values on all four datasets. Relative to GraphSmile, the WF1 gains are 0.84, 0.72, 0.69, and 0.92 points on IEMOCAP-6, IEMOCAP-4, MELD, and CMU-MOSEI, respectively. The corresponding ACC gains are 0.83, 0.73, 0.64, and 0.84 points.

The gains are generally smaller than those under the fine-grained MERC setting. Coarsening merges several original label categories into one sentiment class, so some transition distinctions evaluated under MERC are no longer visible in the mapped labels. The consistent improvements nevertheless show that the benefit of predictive semantic-state learning persists after label coarsening and is not limited to fine-grained emotion recognition.

C.2 Contextual Comparison with Emotion Prediction in Conversation

Table 7 provides an external reference to representative EPC methods. The first three rows are the Text+Speech results reported by Shi et al. (2024). PSWS obtains numerically higher M-F1 values by 1.02 points on IEMOCAP-4 and 1.69 points on MELD relative to the strongest external row. Its UAR values are higher by 1.01 and 1.96 points, respectively.

This table is not a controlled ranking. The prior methods use Text+Speech, whereas PSWS uses Text+Audio+Vision and predicts an explicit latent-state transition. Differences in modality availability and model objectives prevent a direct attribution of the numerical gaps to the proposed components. The supervision-, modality-, split-, and causality-matched comparisons in the main manuscript provide the primary evidence for PSWS.

Table 6: Current-state recognition under the GraphSmile MSAC setting. All values are percentages. † denotes published baseline results reported in GraphSmile; ‡ denotes PSWS mean ± standard deviation over five runs. DialogueCRN (Hu et al., 2021a), MMGCN (Hu et al., 2021b), MM-DFN (Hu et al., 2022), SACL-LSTM (Hu et al., 2023), M3Net (Chen et al., 2023), GraphSmile (Li et al., 2025a).

Method	IEMOCAP-6		IEMOCAP-4		MELD		CMU-MOSEI	
	ACC	WF1	ACC	WF1	ACC	WF1	ACC	WF1
DialogueCRN†	83.43	83.52	80.70	80.34	74.18	73.96	54.11	48.50
MMGCN†	82.75	82.94	81.87	81.56	73.56	73.46	66.66	65.89
MM-DFN†	84.60	84.69	81.34	81.03	73.30	73.27	66.35	64.62
SACL-LSTM†	85.15	85.35	82.61	82.47	73.56	73.44	55.28	44.50
M3Net†	84.29	84.19	84.52	84.44	73.33	73.32	64.72	61.66
GraphSmile†	84.97	85.09	87.49	87.48	74.44	74.31	67.73	66.73
PSWS‡	**85.80**	**85.93**	**88.22**	**88.20**	**75.08**	**75.00**	**68.57**	**67.65**
±Std. Dev.	±0.21	±0.20	±0.16	±0.16	±0.18	±0.19	±0.24	±0.25

Table 7: Contextual external comparison with EPC methods. All values are percentages. Prior rows use Text+Speech. PSWS uses Text+Audio+Vision. This comparison is not modality matched. DEP (Shi et al., 2020), EAMT (Shi et al., 2023). Shi et al. (Shi et al., 2024).

Method	IEMOCAP-4		MELD	
	UAR	M-F1	UAR	M-F1
DEP	76.31	75.50	42.39	42.51
EAMT	80.18	80.01	45.21	44.36
Shi et al.	81.92	81.53	46.10	45.73
PSWS	**82.93±0.36**	**82.55±0.38**	**48.06±0.48**	**47.42±0.50**

D Computational Efficiency

Table 8 compares PSWS with GraphSmile under the same experimental environment. Relative to GraphSmile, the per-epoch runtime increases by 58.9%, 61.2%, 56.4%, and 56.1% on IEMOCAP-6, IEMOCAP-4, MELD, and CMU-MOSEI, respectively. Peak GPU memory increases by 22.2%, 27.6%, 31.5%, and 31.8%. These values match the ranges summarized in the main manuscript. The largest PSWS memory usage is $2{,}250$ MB on IEMOCAP-6. The additional training cost arises from the separate content and transition abstraction branches, four transition experts, and future-state distillation. The EMA teacher is active only during training and is removed at inference. The student TISA, ECTR, and shared decoder remain active at inference. Accordingly, the per-epoch measurements below quantify the reported experimental cost and should not be interpreted as inference-latency measurements.

Table 8: Per-epoch runtime and peak GPU memory under the GraphSmile experimental environment. The EMA teacher is used only during training.

Dataset	Method	Time (s/epoch)	Peak Mem. (MB)
IEMOCAP-6	GraphSmile	1.12	1840.89
	PSWS	1.78	2250.07
IEMOCAP-4	GraphSmile	0.67	509.88
	PSWS	1.08	650.65
MELD	GraphSmile	5.21	513.33
	PSWS	8.15	675.18
CMU-MOSEI	GraphSmile	7.38	467.13
	PSWS	11.52	615.76