EVIDENCE-LATTICE FLOW: TASK-POSTERIOR TRANSPORT AND INTEGRATION FOR INCOMPLETE MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multimodal learning (IML) aims to make reliable predictions when some input modalities are unavailable. Many existing methods reconstruct missing inputs or map incomplete observations to a single complete-like representation. This **completion-centric strategy** collapses ambiguous evidence into a single representation, which may obscure variation among plausible task outcomes and lead to overconfident predictions. To address this issue, we propose *Evidence-Lattice Flow* (ELF), a general framework that recasts IML as **task-posterior inference over an evidence lattice**. Each lattice node represents an available modality subset and carries the posterior supported by that evidence. ELF implements this view through three stages. *Full-Evidence Task Anchoring* establishes task semantics from complete observations, *Lattice-Conditioned Flow Transport* moves stochastic particles toward subset-specific task-latent distributions, and *Posterior Task Integration* converts transported samples into predictions and uncertainty. ELF thus learns evidence-conditioned task-latent distributions and integrates predictions from transported samples, **without reconstructing missing modalities**. We characterize consistency relations among underlying task posteriors and relate transport error to downstream risk. Experiments on five datasets spanning emotion recognition, sentiment analysis, and action quality assessment demonstrate ELF's effectiveness. Relative to the strongest baseline, ELF reduces MSE averaged over six incomplete modality subsets by 8.7% on FS1000 and 4.9% on Fis-V.

1 INTRODUCTION

Multimodal learning combines images, audio, text, and sensor signals to capture complementary evidence about complex environments. Yet acquisition failures, privacy constraints, and transmission errors often leave some modalities unavailable. This challenge spans medical analysis (Shi et al., 2024), autonomous perception (Ma et al., 2026), remote sensing (Chen et al., 2024b), audiovisual understanding (Park et al., 2024), affective computing (Xu et al., 2024), and behavior analysis (Xu et al., 2026b). It has motivated incomplete multimodal learning (IML), which seeks reliable predictions from any available modality subset (Wu et al., 2026).

Existing IML methods rely on missing-aware fusion, complete-data teachers, or generative reconstruction of missing modalities and features (Woo et al., 2023; Zhao et al., 2021; Li et al., 2025; Dai et al., 2025; Qiu et al., 2026; Xu et al., 2026b). Despite this progress, many approaches construct a complete proxy before solving the task. Even without raw-data generation, they usually map incomplete samples to one complete-like representation.

This pattern is problematic because observed modalities seldom determine missing information uniquely. In a visual-only action clip, the same movement can be accompanied by different sounds, forces, or contexts that change its class or score. A single predicted feature or mean representation collapses these alternatives and hides unresolved uncertainty. Faithful modality recovery also does not guarantee a better decision. The central question for downstream prediction is thus not what the missing modality looks like, but **which outcomes remain plausible given the observed evidence**.

We instead treat each available modality subset as an *evidence state*. Ordered by set inclusion, modality subsets form an *evidence lattice*, whose nonempty nodes range from sparse to complete

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.

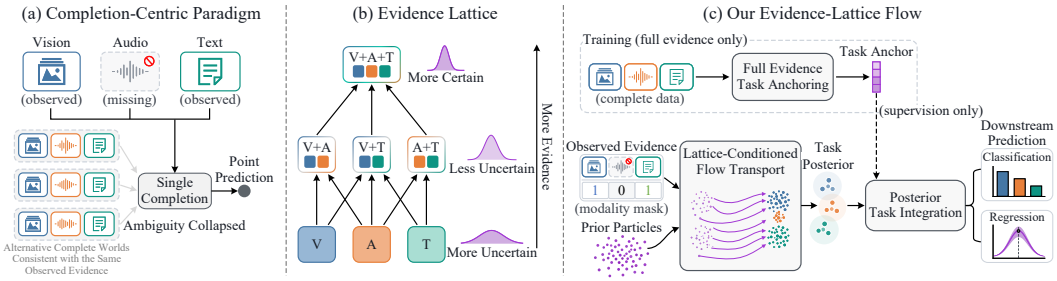


Figure 1: From missing-modality completion to evidence-conditioned task-posterior transport. (a) A single complete proxy can collapse plausible interpretations. (b) Schematic of the evidence lattice formed by modality subsets. (c) Evidence-Lattice Flow (ELF) transports particles to evidence-conditioned task-latent distributions. A shared head yields downstream predictions and uncertainty.

evidence. Each node carries a task posterior matched to its available information. A lower-node posterior should average over plausible richer states rather than imitate one complete example. When new evidence arrives, the posterior should update toward the corresponding upper node. This view echoes Bayesian accounts of multisensory perception, where cues are integrated by reliability and uncertainty is represented explicitly (Ernst & Banks, 2002; Knill & Pouget, 2004). Missing-modality degradation therefore combines **irreducible loss** from unavailable information, which no model can remove, and **reducible modeling error** from an inaccurate evidence-conditioned posterior.

How can we turn this view into a trainable model? The key is to represent the **one-to-many relation from limited evidence to plausible task interpretations**. Standard predictors return one result, while generation in the original modality space is costly and often models task-irrelevant detail. We therefore need a mechanism that preserves a conditional distribution while focusing on task semantics. This requirement naturally leads us to flow matching, which learns continuous transport from a simple source to a target distribution (Lipman et al., 2023). Conditioned on observed modalities and their mask, particles move toward plausible regions of a low-dimensional task space, preserving alternative interpretations while filtering irrelevant variation. Recent diffusion- and flow-based methods mainly target missing-data imputation or feature recovery (Wang et al., 2023; 2026; Simkus & Gutmann, 2025; Kan et al., 2026). In contrast, we use flow matching to directly learn evidence-conditioned distributions over task anchors, bypassing the need to reconstruct missing modalities.

To this end, we propose **E**vidence-**L**attice **F**low (**ELF**), a general framework for incomplete multimodal learning built on three stages. As shown in Figure 1, *Full-Evidence Task Anchoring* learns a task anchor from complete labeled observations so that the transport target retains task-relevant semantics. *Lattice-Conditioned Flow Transport* samples evidence nodes during training, aggregates only the available modalities, explicitly encodes the missingness mask, and uses a shared conditional flow to learn the corresponding task-latent distributions. *Posterior Task Integration* maps multiple transported samples through a shared task head and aggregates them into class probabilities or a continuous score, while their variation measures particle disagreement. Subset supervision and task consistency preserve task alignment. ELF therefore replaces missing-input recovery with prediction from evidence-conditioned task-latent distributions.

We characterize posterior consistency across evidence states and show that adding evidence cannot increase Bayes risk in expectation. Our analysis separates irreducible information loss from modeling error and links transport error to task risk. Across five datasets spanning emotion recognition, sentiment analysis, and action quality assessment, ELF achieves the best incomplete-condition averages and full-modality results. For example, ELF reduces average FS1000 MSE by 8.7% over the strongest competing baseline. Integrating particle predictions also reduces average Fis-V MSE by 6.1% compared with decoding the mean latent, supporting prediction from a transported distribution rather than a single representation. Our main contributions are as follows.

- We reformulate incomplete multimodal learning as task-posterior inference over an evidence lattice, where predictions update with evidence and retain unresolved uncertainty.
- Evidence-Lattice Flow (ELF) realizes this formulation through Full-Evidence Task Anchoring, Lattice-Conditioned Flow Transport, and Posterior Task Integration, jointly learn-

ing subset-specific task-latent distributions and supporting uncertainty-aware classification and regression from any nonempty modality subset.

- We provide a risk-based analysis and extensive evaluation across five datasets and three downstream tasks, assessing predictive performance, missing-subset robustness, and uncertainty quality to establish the broad applicability of the proposed ELF.

2 Related Work

Incomplete Multimodal Learning. IML methods are commonly organized by how they compensate for absent evidence (Wu et al., 2026). Early studies show that multimodal Transformers are not inherently robust to missing inputs and that the most effective fusion strategy depends on the data (Ma et al., 2022). Reconstruction-based approaches synthesize missing inputs or latent features using imagination networks and diffusion models (Zhao et al., 2021; Wang et al., 2023; Dai et al., 2025; Qiu et al., 2026). Distillation-based approaches transfer correlations or probability distributions from complete-modality teachers to incomplete students (Li et al., 2024; Chen et al., 2024a; Zhuang et al., 2025). Missing-aware fusion methods instead adapt aggregation, expert routing, or optimization to the available modalities (Woo et al., 2023; Li et al., 2025; Nguyen et al., 2026). A newer direction introduces structural or external priors, including retrieved modality prototypes for multimodal large language models (Pipoli et al., 2025). Although probabilistic distillation and diffusion relax deterministic point recovery, their targets remain missing modalities or representations defined by complete evidence. *In contrast, our ELF learns a task-latent distribution conditioned on each observed subset and explicitly retains ambiguity that the available evidence cannot resolve.*

Flow Matching and Missing Data. Flow matching learns a continuous vector field that transports a simple source distribution to a target distribution without simulating trajectories during training (Lipman et al., 2023). Subsequent work improves source-target couplings and extends deterministic transport to stochastic bridges (Pooladian et al., 2023; Tong et al., 2024). Flow matching has also been developed as a likelihood-free posterior sampler, highlighting its ability to represent conditional uncertainty (Zhai et al., 2026). Its conditional form has recently been adapted to missing data. CFMI conditions flows on observed entries for multiple imputation (Simkus & Gutmann, 2025), while Missing-Data Flow Matching marginalizes unobserved coordinates in the training objective (Khan et al., 2026). In multimodal settings, PAMF couples flow-based imputation with prediction for incomplete time series, and IM-PFR recovers complete probability flows for emotion recognition (Kan et al., 2026; Wang et al., 2026). These studies establish flow matching as a flexible conditional imputer, but they still transport toward missing data or a complete proxy before prediction. *In contrast, our proposed ELF changes both the structure and destination of transport by learning a shared flow over an evidence lattice that targets subset-conditioned distributions over task anchors.*

3 Approach

3.1 Problem Formulation

Let $X = (X^{(1)}, \ldots, X^{(M)})$ denote a complete observation with label Y and modality index set $[M] = \{1, \ldots, M\}$. Training observations contain all modalities. At test time, a nonempty subset $S \subseteq [M]$ specifies the available evidence $X_S = \{X^{(m)} : m \in S\}$ and its binary mask $\boldsymbol{r}_S \in \{0,1\}^M$. Modality subsets ordered by inclusion form the evidence lattice $\mathcal{L}_M = 2^{[M]}$. ELF predicts at its nonempty nodes $\mathcal{L}_M^+ = \mathcal{L}_M \setminus \{\emptyset\}$, each associated with a task posterior

$$\Pi_S(\cdot) = P(Y \in \cdot \mid X_S), \qquad S \in \mathcal{L}_M^+. \tag{1}$$

ELF learns an evidence-conditioned distribution over task anchors and integrates predictions from a shared task head to estimate class probabilities or regression conditional means. Variation across particle predictions measures disagreement among task interpretations. The task posterior is the inference target, and Theorem 1 establishes recovery under the stated sufficiency and task-head assumptions and ideal transport. As illustrated in Figure 2, the proposed ELF comprises three stages. *Full-Evidence Task Anchoring* learns a task anchor from complete labeled observations. *Lattice-Conditioned Flow Transport* moves Gaussian particles to task-latent distributions conditioned on the available modalities and their mask. *Posterior Task Integration* decodes each particle and aggregates the predictions. Tasks use their own encoders and task heads within this shared framework.

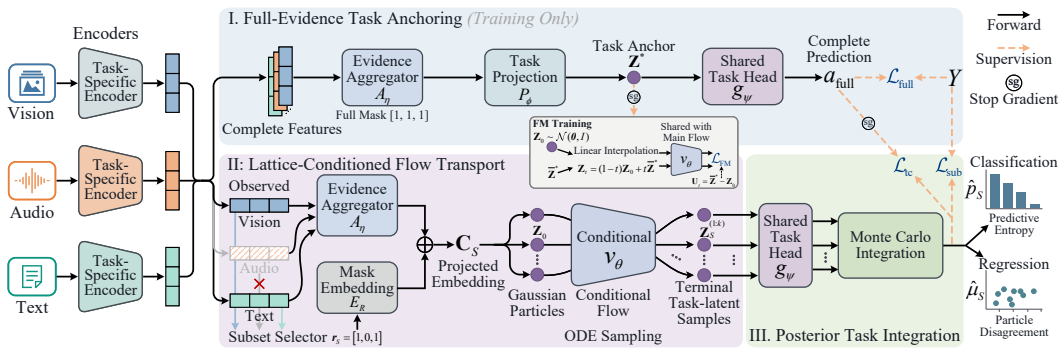


Figure 2: Overview of Evidence-Lattice Flow. (I) Full-Evidence Task Anchoring learns a supervised anchor from complete observations. (II) Lattice-Conditioned Flow Transport generates task-latent samples conditioned on available modalities and their mask. (III) Posterior Task Integration decodes each sample with a shared task head and averages the predictions. The upper branch and FM inset are used during training. Conditional sampling supports both training and inference.

3.2 Full-Evidence Task Anchoring

Complete-evidence supervision establishes a task-relevant target space for transport. Modality encoders produce sequences $\boldsymbol{H}_m = E_m(X^{(m)}) \in \mathbb{R}^{T_m \times d}$, where T_m is the sequence length and d is the projected feature dimension. Mean pooling and a learnable modality embedding $\boldsymbol{e}_m$ yield $\boldsymbol{h}_m = \mathrm{Pool}(\boldsymbol{H}_m) + \boldsymbol{e}_m$. An evidence aggregator A_η combines all features under the complete mask, and a task projection P_ϕ produces the anchor

$$\mathbf{Z}^* = P_\phi\big(A_\eta(\{\boldsymbol{h}_m\}_{m=1}^M, E_R(\boldsymbol{r}_{[M]}))\big) \in \mathbb{R}^{d_z}, \tag{2}$$

where E_R embeds the modality mask. A task head g_ψ predicts the label from $\mathbf{Z}^*$ using

$$\mathcal{L}_{\text{full}} = \mathbb{E}_{X,Y}\left[\ell_{\text{task}}(g_\psi(\mathbf{Z}^*), Y)\right]. \tag{3}$$

We use squared loss for regression and cross-entropy for classification. The head outputs a scalar score or a probability vector, whose class-y component is denoted by $p_\psi(y \mid \boldsymbol{z})$. This supervision makes the anchor task-relevant. Flow matching uses $\bar{\mathbf{Z}}^* = \mathrm{sg}(\mathbf{Z}^*)$ as its endpoint, where sg stops gradients through the target branch.

3.3 Lattice-Conditioned Flow Transport

Evidence conditioning. During training, we sample $S \sim \rho$ independently of the observation, with ρ covering all evaluated nonempty subsets. The subset branch reuses the encoders and aggregator of the complete branch:

$$\mathbf{C}_S = A_\eta(\{\boldsymbol{h}_m : m \in S\}, E_R(\boldsymbol{r}_S)). \tag{4}$$

A learnable query attends to available modality features, with a padding mask excluding missing keys and values. The attended representation is concatenated with the mask embedding and projected to $\mathbf{C}_S$. Modality embeddings identify feature sources, while the mask embedding specifies the available combination.

Conditional flow matching. We independently sample $\mathbf{Z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_z})$ and $t \sim \mathcal{U}(0,1)$, then connect the prior particle to the task anchor of the same observation:

$$\mathbf{Z}_t = (1-t)\mathbf{Z}_0 + t\bar{\mathbf{Z}}^*, \qquad \mathbf{U}_t = \bar{\mathbf{Z}}^* - \mathbf{Z}_0. \tag{5}$$

A shared velocity field predicts transport velocity from the path state, time, and evidence condition:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{X,S,t,\mathbf{Z}_0}\left[\left\|v_\theta(\mathbf{Z}_t, t, \mathbf{C}_S, \boldsymbol{r}_S) - \mathbf{U}_t\right\|_2^2\right]. \tag{6}$$

Conditional velocity regression learns transport to the anchor distribution $p(\mathbf{Z}^* \mid \mathbf{C}_S, \boldsymbol{r}_S)$ (Lipman et al., 2023). Different complete observations sharing the same subset evidence can yield different anchors, represented by stochastic particles. The velocity network projects $\mathbf{Z}_t$ and a Fourier time encoding into a hidden space, then applies FiLM residual blocks modulated by $\mathbf{C}_S$. Mask conditioning is implemented through the embedding in $\mathbf{C}_S$. All evidence nodes share the velocity network.

Task-latent sampling. Given X_S, we draw K independent Gaussian particles and solve

$$\frac{d\widetilde{\mathbf{Z}}_t^{(k)}}{dt} = v_\theta(\widetilde{\mathbf{Z}}_t^{(k)}, t, \mathbf{C}_S, \boldsymbol{r}_S), \qquad \widetilde{\mathbf{Z}}_0^{(k)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_z}), \quad k = 1, \dots, K. \tag{7}$$

Heun integration yields terminal samples $\widehat{\mathbf{Z}}_S^{(k)}$. We denote the conditional flow endpoint distribution by $q_\theta(d\boldsymbol{z} \mid \mathbf{C}_S, \boldsymbol{r}_S)$, which numerical sampling approximates. The differentiable solver also generates training particles for task supervision.

3.4 Posterior Task Integration

Each latent sample passes through the task head shared with the complete branch, followed by Monte Carlo integration in prediction space. Nonlinear decoding generally does not commute with expectation, so decoding each particle preserves its contribution to the task output. With regression scores $\hat{y}_k = g_\psi(\widehat{\mathbf{Z}}_S^{(k)})$ and class probabilities $p_k(y) = p_\psi(y \mid \widehat{\mathbf{Z}}_S^{(k)})$, the predictions are

$$\widehat{\mu}_S = \frac{1}{K}\sum_{k=1}^{K} \hat{y}_k, \qquad \widehat{p}_S(y) = \frac{1}{K}\sum_{k=1}^{K} p_k(y). \tag{8}$$

Regression uses $\widehat{\mu}_S$ as the score and the empirical variance $\widehat{v}_S = K^{-1}\sum_k (\hat{y}_k - \widehat{\mu}_S)^2$ to measure disagreement among task interpretations. This variance does not capture residual label noise (Appendix D.3). For classification, predictive entropy $H(\widehat{p}_S)$ measures uncertainty, and $H(\widehat{p}_S) - K^{-1}\sum_k H(p_k)$ measures particle disagreement.

3.5 Learning Objective

We supervise integrated predictions at incomplete nodes with labels and use complete-evidence predictions for task consistency. Let $\widehat{a}_S$ be the regression score or class probabilities in Eq. (8), and let $a_{\text{full}} = g_\psi(\mathbf{Z}^*)$. The objectives are

$$\mathcal{L}_{\text{sub}} = \mathbb{E}[\ell_{\text{task}}(\widehat{a}_S, Y) \mid S \neq [M]], \quad \mathcal{L}_{\text{tc}} = \mathbb{E}[D_{\text{task}}(\widehat{a}_S, \text{sg}(a_{\text{full}}))], \tag{9}$$

$$\mathcal{L}_{\text{ELF}} = \mathcal{L}_{\text{full}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{sub}}\mathcal{L}_{\text{sub}} + \lambda_{\text{tc}}\mathcal{L}_{\text{tc}}. \tag{10}$$

Here, $\lambda_{\text{FM}}, \lambda_{\text{sub}}, \lambda_{\text{tc}} \geq 0$ balance the loss terms, and expectations are over training observations, sampled subsets, and prior particles. We jointly optimize these objectives, using mean squared error between integrated predictions and complete-evidence targets for D_{task}. For action quality assessment (AQA), scoring queries are further regularized with a triplet loss (Appendix A.2). Inference consists of available-evidence encoding, conditional sampling, and task integration.

4 Theoretical Analysis

We analyze posterior relations across evidence nodes, ideal recovery by conditional transport, and the effect of transport error on task risk. Masks are sampled independently of the data, and every evaluated node has $\rho(S) > 0$. Regression assumes $\mathbb{E}[Y^2] < \infty$, and classification has finitely many classes. Detailed assumptions and proofs appear in Appendix D.

Posterior relations over the evidence lattice. Let $\mathcal{G}_S = \sigma(X_S)$ and $\mu_S = \mathbb{E}[Y \mid \mathcal{G}_S]$.

Proposition 1 (Evidence consistency and information value)**.** *For nonempty nodes $S \subseteq T$ and any measurable label set B, the true posteriors and Bayes risks satisfy*

$$\begin{aligned} \mathbb{E}[\Pi_T(B) \mid \mathcal{G}_S] &= \Pi_S(B), \\ R_S^{*,\text{sq}} - R_T^{*,\text{sq}} &= \mathbb{E}[(\mu_T - \mu_S)^2] \geq 0, \\ R_S^{*,\log} - R_T^{*,\log} &= I(Y, X_{T\setminus S} \mid X_S) \geq 0, \end{aligned} \tag{11}$$

where sq *and* log *denote squared loss and classification log-loss, and I denotes conditional mutual information.*

These identities follow from the tower property, conditional-expectation orthogonality, and the conditional mutual information identity (Cover & Thomas, 2006). A lower-evidence posterior averages compatible richer-evidence posteriors, and additional evidence cannot increase Bayes risk in expectation. For any predictor a_S based on X_S, its expected risk $R(a_S)$ decomposes as

$$R(a_S) - R^*_{[M]} = \underbrace{R^*_S - R^*_{[M]}}_{\text{irreducible information loss}} + \underbrace{R(a_S) - R^*_S}_{\text{reducible modeling error}} . \tag{12}$$

Here R^*_S is the Bayes risk at node S. ELF targets the modeling error within each node.

From conditional flow to task prediction. Fix deterministic encoders, the aggregator, and the anchor mapping, and assume $\mathbb{E}\|\mathbf{Z}^*\|^2 < \infty$. At the population optimum of Eq. (6), conditional squared regression gives the optimal velocity for each node with $\rho(S) > 0$:

$$v^*_S(\boldsymbol{z}, t, \boldsymbol{c}, \boldsymbol{r}_S) = \mathbb{E}[\bar{\mathbf{Z}}^* - \mathbf{Z}_0 \mid \mathbf{Z}_t = \boldsymbol{z}, \mathbf{C}_S = \boldsymbol{c}, \boldsymbol{r}_S]. \tag{13}$$

This is the population conditional flow matching solution (Lipman et al., 2023). Assume the shared field can realize these optima simultaneously and that the conditional paths satisfy the continuity equation with a unique probability flow for $t < 1$. Taking the distributional limit as $t \uparrow 1$ defines the endpoint and recovers $p(\mathbf{Z}^* \mid \mathbf{C}_S, \boldsymbol{r}_S)$.

Theorem 1 (Task prediction recovery under ideal transport)**.** *Under the conditions above, suppose $p(\mathbf{Z}^* \mid \mathbf{C}_S, \boldsymbol{r}_S) = p(\mathbf{Z}^* \mid X_S, S)$ and $Y \perp X \mid \mathbf{Z}^*$. If the classification head satisfies $p_\psi(y \mid \boldsymbol{z}) = p(y \mid \mathbf{Z}^* = \boldsymbol{z})$ and the regression head satisfies $g_\psi(\boldsymbol{z}) = \mathbb{E}[Y \mid \mathbf{Z}^* = \boldsymbol{z}]$, exact integration under the ideal flow gives*

$$\begin{aligned} p^{\mathrm{ELF}}_S(y) &= \int p_\psi(y \mid \boldsymbol{z})\, q_\theta(d\boldsymbol{z} \mid \mathbf{C}_S, \boldsymbol{r}_S) = p(y \mid X_S), \\ \mu^{\mathrm{ELF}}_S &= \int g_\psi(\boldsymbol{z})\, q_\theta(d\boldsymbol{z} \mid \mathbf{C}_S, \boldsymbol{r}_S) = \mathbb{E}[Y \mid X_S]. \end{aligned} \tag{14}$$

These sufficient conditions identify ideal task recovery. The resulting class posteriors satisfy Proposition 1, and the regression means satisfy the corresponding tower relation. Equation (8) approximates these integrals with particle averages. For a fixed model and independent exact samples, the averages are unbiased, with conditional variance decreasing as $1/K$ when second moments are finite. Appendix D.4 separates the effects of imperfect representations, transport, and task heads.

Transport error and task risk. Let $p_S = p(\mathbf{Z}^* \mid X_S, S)$ and $q_S = q_\theta(\cdot \mid \mathbf{C}_S, \boldsymbol{r}_S)$, with integrated predictions $a_p = \int g_\psi(\boldsymbol{z}) p_S(d\boldsymbol{z})$ and $a_q = \int g_\psi(\boldsymbol{z}) q_S(d\boldsymbol{z})$.

Proposition 2 (Transport risk bound)**.** *Suppose p_S and q_S have finite first moments, g_ψ is L_g-Lipschitz, and $\ell(\cdot, y)$ is uniformly L_ℓ-Lipschitz on the prediction domain for all labels. If the risks and the expectation below are finite, then*

$$|R_q - R_p| \le L_\ell L_g\, \mathbb{E}_{X,S}[W_1(q_S, p_S)], \tag{15}$$

where W_1 is the first Wasserstein distance, $R_q = \mathbb{E}[\ell(a_q, Y)]$, and $R_p = \mathbb{E}[\ell(a_p, Y)]$.

The bound follows from Wasserstein coupling and the Lipschitz property of the loss (Villani, 2009). Squared loss satisfies this condition for bounded predictions and labels. Cross-entropy requires a common positive lower bound on all compared class probabilities. The bound compares risks under the same task head without requiring task sufficiency. Under Theorem 1, $R_p = \mathbb{E}_{S\sim\rho}[R^*_S]$.

5 Experiments

5.1 Experimental Setup

Datasets and metrics. We evaluate ELF on five datasets spanning emotion recognition, sentiment analysis, and action quality assessment (AQA). IEMOCAP (Busso et al., 2008) uses four emotion classes, evaluated by weighted accuracy (WA) and unweighted accuracy (UA). CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018) use sentiment regression, with binary accuracy (ACC) and weighted F1 evaluated on nonzero labels. These datasets provide audio, text, and visual inputs. FS1000 (Xia et al., 2023) and Fis-V (Xu et al., 2020) provide visual, optical-flow, and audio inputs for action quality score regression. We report Spearman's Rank Correlation (ρ) and mean squared error (MSE), aggregated over seven and two scoring targets, respectively. Dataset protocols, preprocessing, and metric aggregation are detailed in Appendix A.1.

Table 1: Emotion recognition and sentiment analysis results. Inputs are audio (a), text (t), and visual (v). Average covers six incomplete subsets. Unstarred baselines are from TMDC (Zhuang et al., 2026), and * marks our reimplementation. **Bold**/underline indicate best/second-best results. Δ_{SOTA} reports relative change from the best baseline, with blue/gray indicating improvement/degradation.

Datasets	Methods	Testing Condition							
		$\{a\}$	$\{t\}$	$\{v\}$	$\{a,v\}$	$\{a,t\}$	$\{t,v\}$	**Average**	$\{a,t,v\}$
		WA / UA	WA / UA	WA / UA	WA / UA	WA / UA	WA / UA	WA / UA	WA / UA
IEMOCAP	MCTN (AAAI'19)	49.75 / 51.62	62.42 / 63.78	48.92 / 45.73	56.34 / 55.84	68.34 / 69.46	67.84 / 68.34	58.94 / 59.13	–
	MMIN (ACL'21)	56.58 / 59.00	66.57 / 68.02	52.52 / 51.60	63.99 / 65.43	72.94 / 75.14	72.67 / 73.61	64.21 / 65.47	–
	IF-MMIN (ICASSP'23)	55.03 / 53.20	67.02 / 68.20	51.97 / 50.41	65.33 / 66.52	74.05 / 75.44	72.68 / 73.62	64.35 / 64.57	–
	MRAN (MMM'23)	55.44 / 57.01	65.31 / 66.42	53.23 / 49.80	64.70 / 64.46	73.00 / 74.58	72.11 / 72.24	63.97 / 64.08	–
	MoMKE (ACMMM'24)	70.32 / 71.38	77.82 / 78.37	58.60 / 54.70	68.85 / 67.65	79.89 / 79.53	77.87 / 77.84	72.23 / 71.58	80.13 / 79.99
	MCMoE* (AAAI'26)	69.69 / 69.93	77.45 / 78.39	58.90 / 54.38	70.86 / 69.80	79.48 / 80.36	77.65 / 78.18	72.34 / 71.84	80.41 / 80.98
	TMDC (AAAI'26)	70.45 / 71.40	77.88 / 78.44	59.18 / 55.33	70.21 / 69.91	79.99 / 81.45	78.20 / 78.29	72.65 / 72.47	80.48 / 80.69
	ELF (This Paper)	**71.16 / 72.49**	**79.25 / 79.97**	**59.65 / 56.84**	**72.53 / 72.41**	**82.26 / 83.11**	**80.23 / 80.46**	**74.18 / 74.21**	**82.47 / 82.74**
	Δ_{SOTA}	↑1.0% / ↑1.5%	↑1.8% / ↑2.0%	↑0.8% / ↑2.7%	↑2.4% / ↑3.6%	↑2.8% / ↑2.0%	↑2.6% / ↑2.8%	↑2.1% / ↑2.4%	↑2.5% / ↑2.2%
		ACC / F1	ACC / F1	ACC / F1	ACC / F1	ACC / F1	ACC / F1	ACC / F1	ACC / F1
MOSI	MCTN (AAAI'19)	56.10 / 54.50	79.10 / 79.20	55.00 / 54.40	57.50 / 57.40	81.00 / 81.00	81.10 / 81.20	68.30 / 67.95	81.40 / 81.50
	MMIN (ACL'21)	55.30 / 51.50	83.80 / 83.80	57.00 / 54.00	60.40 / 58.50	84.00 / 84.00	83.80 / 83.90	70.72 / 69.28	84.60 / 84.40
	GCNet (TPAMI'23)	56.10 / 54.50	83.70 / 83.60	56.10 / 55.70	62.00 / 61.90	84.50 / 84.40	84.30 / 84.20	71.12 / 70.72	85.20 / 85.10
	IMDer (NeurIPS'23)	62.00 / **62.20**	84.80 / 84.70	61.30 / 60.80	63.60 / 63.40	85.40 / 85.30	85.50 / 85.40	73.77 / 73.63	85.70 / 85.60
	MoMKE (ACMMM'24)	63.19 / 58.61	86.59 / 86.52	63.35 / 63.34	64.04 / 64.66	87.20 / 87.17	87.04 / 87.00	75.24 / 74.55	87.96 / 87.89
	MCMoE* (AAAI'26)	62.20 / 60.04	86.13 / 85.98	64.48 / 64.62	65.55 / 64.94	87.04 / 87.03	86.43 / 86.41	75.31 / 74.84	86.13 / 86.09
	TMDC (AAAI'26)	62.35 / 60.24	87.35 / 87.27	64.63 / **64.82**	65.40 / 65.60	87.50 / 87.45	87.96 / 87.87	75.87 / 75.54	88.26 / 88.19
	ELF (This Paper)	**63.87** / 61.72	**88.11 / 88.12**	**64.94** / 64.66	**66.16 / 65.98**	**88.26 / 88.28**	**88.11 / 88.13**	**76.58 / 76.15**	**88.41 / 88.38**
	Δ_{SOTA}	↑1.1% / ↓0.8%	↑0.9% / ↑1.0%	↑0.5% / ↓0.2%	↑0.9% / ↑0.6%	↑0.9% / ↑0.9%	↑0.2% / ↑0.3%	↑0.9% / ↑0.8%	↑0.2% / ↑0.2%
MOSEI	MCTN (AAAI'19)	62.70 / 54.50	82.60 / 82.80	62.60 / 57.10	63.70 / 62.70	83.50 / 83.30	83.20 / 83.20	73.05 / 70.60	84.20 / 84.20
	MMIN (ACL'21)	58.90 / 59.50	82.30 / 82.40	59.30 / 60.00	63.50 / 61.90	83.70 / 83.30	83.80 / 83.40	71.92 / 71.75	84.30 / 84.20
	GCNet (TPAMI'23)	60.20 / 60.30	83.00 / 83.20	61.90 / 61.60	64.10 / 57.20	84.30 / 84.40	84.30 / 84.40	72.97 / 71.85	85.20 / 85.10
	IMDer (NeurIPS'23)	63.80 / 60.60	84.50 / 84.50	63.90 / 63.60	64.90 / 63.50	85.10 / 85.10	85.00 / 85.00	74.53 / 73.72	85.10 / 85.10
	MoMKE (ACMMM'24)	72.56 / 71.03	86.46 / 86.43	70.12 / 70.23	73.34 / 71.82	86.68 / 86.61	86.79 / 86.69	79.33 / 78.80	87.12 / 87.03
	MCMoE* (AAAI'26)	72.98 / 72.37	86.82 / 86.86	71.63 / 70.32	73.56 / 72.49	87.15 / 87.04	87.45 / 87.31	79.93 / 79.40	87.42 / 87.34
	TMDC (AAAI'26)	73.64 / 72.23	86.87 / 86.82	71.60 / 70.65	74.13 / 73.41	87.15 / 87.14	87.48 / 87.43	80.15 / 79.61	87.67 / 87.62
	ELF (This Paper)	**74.49 / 73.34**	**87.61 / 87.58**	**71.88 / 71.54**	**74.46 / 73.64**	**87.78 / 87.74**	**87.97 / 87.85**	**80.70 / 80.28**	**88.06 / 87.97**
	Δ_{SOTA}	↑1.2% / ↑1.3%	↑0.9% / ↑0.8%	↑0.3% / ↑1.3%	↑0.4% / ↑0.3%	↑0.7% / ↑0.7%	↑0.6% / ↑0.5%	↑0.7% / ↑0.8%	↑0.4% / ↑0.4%

Implementation details. All experiments are performed on an NVIDIA RTX 3090 GPU and a 2.40 GHz CPU using PyTorch 2.4.1. Training uses complete observations and uniformly sampled nonempty modality subsets. Testing covers all six incomplete subsets and the complete set. ELF uses a 256-dimensional task latent and a shared conditional flow with three residual blocks. We use two particles during training and $K = 8$ at inference, integrated with four Heun steps. The loss weights in Eq. (10) are $\lambda_{\text{FM}} = \lambda_{\text{tc}} = 1$ and $\lambda_{\text{sub}} = 0.1$. Emotion and sentiment models use AdamW for 300 epochs, while AQA models use Adam with cosine learning-rate decay. All runs use batch size 32. We follow the experimental settings and evaluation protocols of the corresponding existing methods (Lian et al., 2023; Xu et al., 2024; Zhuang et al., 2026) for fair comparison. More details about the architecture, optimization, and hyperparameter selection are provided in Appendix A.2.

5.2 Comparison with State-of-the-Art Methods

We compare our ELF framework with the state-of-the-art (SOTA) methods listed in Tables 1 and 2, including MCTN (Pham et al., 2019), MMIN (Zhao et al., 2021), IF-MMIN (Zuo et al., 2023), MRAN (Luo et al., 2023), GCNet (Lian et al., 2023), IMDer (Wang et al., 2023), MoMKE (Xu et al., 2024), TMDC (Zhuang et al., 2026), and MCMoE (Xu et al., 2026b). The AQA comparison also includes ActionMAE (Woo et al., 2023), Skating-Mixer (Xia et al., 2023), PAMFN (Zeng & Zheng, 2024), and SDR-GNN (Fu et al., 2025).

Emotion recognition and sentiment analysis. Our ELF achieves the highest incomplete-condition averages on all three datasets in Table 1. On IEMOCAP, it reaches 74.18 WA and 74.21 UA, exceeding TMDC by 1.53 and 1.74 percentage points. On MOSI and MOSEI, average ACC improves by 0.71 and 0.55 points, with F1 gains of 0.61 and 0.67 points. ELF also improves both full-modality metrics on all three datasets. Across the 36 incomplete-condition metric comparisons, ELF achieves the best performance in 34 and second-best performance in the remaining two.

Action quality assessment. In Table 2, ELF raises average ρ from 0.782 to 0.793 on FS1000 and from 0.734 to 0.746 on Fis-V, relative to the SOTA method MCMoE. Average MSE decreases from 15.37 to 14.03 and from 17.02 to 16.18, corresponding to reductions of 8.7% and 4.9%. ELF improves both full-modality metrics on each dataset. Across both AQA datasets, ELF ranks first or second on every reported metric under every modality combination.

Table 2: AQA results aggregated over seven FS1000 targets and two Fis-V targets. Inputs are visual (v), optical flow (f), and audio (a). Average covers six incomplete subsets. Unstarred baselines are from MCMoE (Xu et al., 2026b), and * marks our reimplementation. **Bold**/underline indicate best/second-best results. Δ_{SOTA} reports relative change from the best baseline, with blue/gray indicating improvement/degradation.

Datasets	Methods	Testing Condition (Spearman Correlation (↑) / Mean Squared Error (↓))							
		$\{v,f\}$	$\{v,a\}$	$\{f,a\}$	$\{v\}$	$\{f\}$	$\{a\}$	**Average**	$\{v,f,a\}$
FS1000 (7 targets)	ActionMAE (AAAI'23)	0.775 / 24.66	0.766 / 64.13	0.556 / 26.51	0.761 / 50.64	0.462 / 21.47	0.458 / 41.66	0.651 / 38.18	0.809 / 17.96
	GCNet (TPAMI'23)	0.730 / 25.56	0.740 / 23.86	0.507 / 24.97	0.696 / 26.67	0.447 / 31.27	0.442 / 39.40	0.610 / 28.62	0.764 / 21.82
	IMDer (NeurIPS'23)	0.760 / 22.34	0.745 / 28.46	0.573 / 24.86	0.724 / 35.99	0.424 / 22.92	0.488 / 32.56	0.636 / 27.86	0.788 / 25.95
	Skating-Mixer (AAAI'23)	0.722 / 25.56	0.542 / 60.57	0.474 / 87.20	0.623 / 101.35	0.472 / 87.59	0.177 / 68.43	0.520 / 71.78	0.819 / 14.56
	PAMFN (TIP'24)	0.727 / 59.96	0.644 / 62.80	0.561 / 56.36	0.713 / 92.10	0.486 / 117.63	0.145 / 62.13	0.571 / 75.16	0.855 / 13.02
	MoMKE (ACMMM'24)	0.798 / 18.86	0.805 / 23.88	0.541 / 24.96	0.785 / 37.96	0.398 / 23.31	0.499 / 27.53	0.668 / 26.08	0.819 / 16.85
	SDR-GNN (KBS'25)	0.789 / 17.50	0.785 / 25.08	0.564 / 22.29	0.749 / 28.47	0.504 / 29.96	0.477 / 25.46	0.665 / 24.79	0.817 / 15.91
	TMDC* (AAAI'26)	0.806 / 15.09	0.830 / 13.95	0.575 / 18.53	0.805 / 17.71	0.338 / 24.72	0.576 / 19.12	0.689 / 18.19	0.834 / 13.58
	MCMoE (AAAI'26)	0.845 / 12.66	**0.882** / 11.85	0.738 / 14.88	0.845 / 13.64	0.650 / 22.47	0.615 / **16.72**	0.782 / 15.37	0.881 / 11.53
	ELF (This Paper)	**0.851 / 12.52**	0.878 / **10.81**	**0.765 / 13.00**	**0.847 / 13.17**	**0.667 / 17.94**	**0.657 / 16.72**	**0.793 / 14.03**	**0.883 / 10.75**
	Δ_{SOTA}	↑0.7% / ↓1.1%	↓0.5% / ↓8.8%	↑3.7% / ↓12.6%	↑0.2% / ↓3.4%	↑2.6% / ↓16.4%	↑6.8% / ↓0.0%	↑1.4% / ↓8.7%	↑0.2% / ↓6.8%
Fis-V (2 targets)	ActionMAE (AAAI'23)	0.704 / 33.54	0.678 / 27.61	0.575 / 25.55	0.616 / 40.07	0.484 / 24.87	0.486 / 29.29	0.597 / 30.16	0.698 / 17.34
	GCNet (TPAMI'23)	0.738 / 19.86	0.656 / 21.32	0.594 / 21.72	0.667 / 20.87	0.602 / 19.61	0.455 / 34.77	0.626 / 23.03	0.698 / 16.93
	IMDer (NeurIPS'23)	0.748 / 15.19	0.658 / 22.66	0.568 / 23.99	0.675 / 25.45	0.618 / 26.38	0.405 / 31.96	0.622 / 24.27	0.703 / 17.02
	Skating-Mixer (AAAI'23)	0.732 / 30.34	0.651 / 46.70	0.572 / 26.96	0.618 / 48.46	0.546 / 27.18	0.325 / 67.25	0.586 / 41.15	0.772 / 13.97
	PAMFN (TIP'24)	0.801 / 33.49	0.661 / 54.34	0.622 / 110.50	0.644 / 84.93	0.616 / 110.42	0.141 / 86.16	0.610 / 79.97	0.822 / 15.33
	MoMKE (ACMMM'24)	0.754 / 14.84	0.689 / 20.60	0.646 / 19.62	0.684 / 23.16	0.654 / 22.46	0.497 / 29.09	0.660 / 21.63	0.747 / 17.30
	SDR-GNN (KBS'25)	0.752 / 14.99	0.680 / 20.62	0.619 / 20.89	0.689 / 20.26	0.648 / 21.50	0.479 / 32.13	0.651 / 21.73	0.733 / 16.45
	TMDC* (AAAI'26)	0.740 / 14.04	0.737 / 15.24	0.586 / 21.27	0.721 / 17.05	0.533 / 20.14	0.512 / 27.69	0.649 / 19.24	0.746 / 13.60
	MCMoE (AAAI'26)	0.813 / **11.02**	**0.787** / 14.64	0.727 / 17.41	**0.765 / 15.14**	0.698 / 15.39	0.557 / 28.54	0.734 / 17.02	0.829 / 12.15
	ELF (This Paper)	**0.823** / 11.55	0.786 / **13.73**	**0.746 / 14.71**	0.763 / 16.57	**0.730 / 14.56**	**0.580 / 25.94**	**0.746 / 16.18**	**0.830 / 11.33**
	Δ_{SOTA}	↑1.2% / ↑4.8%	↓0.1% / ↓6.2%	↑2.6% / ↓15.5%	↓0.3% / ↑9.4%	↑4.6% / ↓5.4%	↑4.1% / ↓6.3%	↑1.6% / ↓4.9%	↑0.1% / ↓6.7%

Table 3: (a) Component and loss ablations on IEMOCAP (WA/UA, %) and Fis-V (ρ/MSE). Average covers six incomplete subsets. (b) Computational cost and ρ/MSE on FS1000.

(a) Components and losses

Setting	IEMOCAP		Fis-V	
	Average	$\{a,t,v\}$	Average	$\{v,f,a\}$
Baseline	70.10 / 69.77	76.36 / 77.02	0.649 / 34.47	0.754 / 14.38
+ FETA	71.94 / 72.04	78.51 / 79.09	0.681 / 21.07	0.783 / 13.16
+ LCFT (w/o PTI)	73.32 / 73.17	81.62 / 82.02	0.718 / 17.82	0.811 / 12.20
+ PTI (Ours)	**74.18 / 74.21**	**82.47 / 82.74**	**0.746 / 16.18**	**0.830 / 11.33**
w/o FETA	72.45 / 72.37	80.48 / 80.86	0.706 / 20.36	0.802 / 13.05
w/o LCFT (point)	71.94 / 72.04	78.51 / 79.09	0.681 / 21.07	0.783 / 13.16
w/o $\mathcal{L}_{\text{FM}}$	72.72 / 72.63	81.18 / 81.54	0.700 / 20.85	0.805 / 12.79
w/o $\mathcal{L}_{\text{sub}}$	73.40 / 73.36	82.21 / 82.46	0.731 / 17.48	0.828 / 11.52
w/o $\mathcal{L}_{\text{tc}}$	73.62 / 73.58	82.04 / 82.35	0.738 / 16.96	0.820 / 11.96

(b) Computational cost on FS1000

Method	#Params	#FLOPs	Average	$\{v,f,a\}$
Skating-Mixer (AAAI'23)	14.32M	49.90G	0.520 / 71.78	0.819 / 14.56
PAMFN (TIP'24)	18.06M	2.56G	0.571 / 75.16	0.855 / 13.02
ActionMAE (AAAI'23)	14.05M	62.12G	0.651 / 38.18	0.809 / 17.96
GCNet (TPAMI'23)	8.78M	1191.39G	0.610 / 28.62	0.764 / 21.82
IMDer (NeurIPS'23)	7.97M	23.53G	0.636 / 27.86	0.788 / 25.95
MoMKE (ACMMM'24)	5.39M	2.60G	0.668 / 26.08	0.819 / 16.85
SDR-GNN (KBS'25)	22.63M	24.35G	0.665 / 24.79	0.817 / 15.91
TMDC (AAAI'26)	5.78M	1.80G	0.689 / 18.19	0.834 / 13.58
MCMoE (AAAI'26)	4.90M	1.34G	0.782 / 15.37	0.881 / 11.53
ELF (This Paper)	5.66M	1.22G	**0.793 / 14.03**	**0.883 / 10.75**

Results across tasks. ELF achieves the best incomplete-condition averages and full-modality results across all five datasets, demonstrating its effectiveness across classification and regression. Full-evidence anchoring establishes task semantics, while conditional transport and posterior integration adapt predictions to available evidence. These results support a shared posterior formulation across distinct modality and label spaces. The ablations below quantify each stage's contributions.

5.3 Ablation Study

Effect of components and loss terms. To validate our components, we build a **Baseline** with modality encoders, a mask-aware aggregator, a deterministic latent projector, and a shared task head. The top half of Table 3(a) adds Full-Evidence Task Anchoring (FETA), Lattice-Conditioned Flow Transport (LCFT), and Posterior Task Integration (PTI) sequentially, while the bottom half removes individual components and loss terms. FETA raises average IEMOCAP WA by 1.84 points and reduces average Fis-V MSE by 38.9%, establishing a task-aligned target. LCFT adds 1.38 WA points by modeling alternative task interpretations, while PTI adds 0.86 points and reduces MSE by 9.2% through prediction integration. Complete-modality performance also improves at each stage. Loss ablations confirm the contributions of flow matching, subset supervision, and task consistency, with $\mathcal{L}_{\text{full}}$ retained as basic task supervision. Appendix B.1 specifies the controls.

Computational cost. Table 3(b) compares all methods with the same single-sample input dimensions on FS1000. The full ELF inference configuration, including eight-particle readout and four Heun steps, uses 5.66M parameters and the fewest FLOPs (1.22G). With 9.0% fewer FLOPs than MCMoE and the best results, ELF offers a favorable balance between accuracy and efficiency.

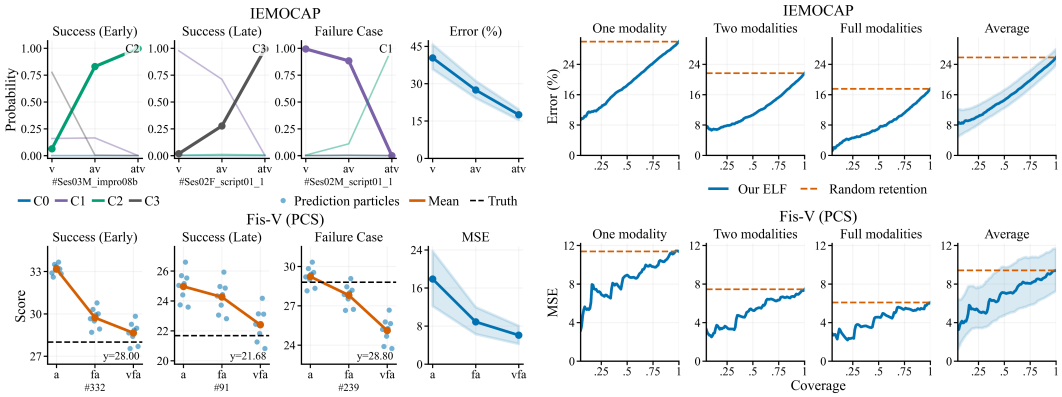
Table 4: (a) Particle count and readout at four Heun steps. (b) Heun steps H at $K = 8$, with $2H$ velocity evaluations in parentheses. Bold marks the default configuration.

(a) Posterior integration

Setting	IEMOCAP		Fis-V	
	Average	$\{a,t,v\}$	Average	$\{v,f,a\}$
$K = 1$	73.32 / 73.17	81.62 / 82.02	0.718 / 17.82	0.811 / 12.20
$K = 2$	73.79 / 73.80	82.07 / 82.40	0.727 / 17.14	0.821 / 11.62
$K = 4$	74.07 / 74.10	82.26 / 82.53	0.733 / 16.52	0.829 / 11.45
$K = 8$	**74.18 / 74.21**	**82.47 / 82.74**	**0.746 / 16.18**	**0.830 / 11.33**
$K = 16$	74.20 / 74.24	82.50 / 82.78	0.747 / 16.15	0.832 / 11.31
Mean latent, 8	73.53 / 73.55	82.11 / 82.43	0.734 / 17.24	0.824 / 11.91

(b) Numerical integration

Setting	IEMOCAP		Fis-V	
	Average	$\{a,t,v\}$	Average	$\{v,f,a\}$
$H = 1$ (2)	72.86 / 72.71	81.12 / 81.47	0.712 / 18.63	0.808 / 12.85
$H = 2$ (4)	73.67 / 73.62	82.02 / 82.31	0.736 / 16.79	0.823 / 11.71
$H = 3$ (6)	74.03 / 74.06	82.38 / 82.63	0.743 / 16.34	0.828 / 11.42
$H = 4$ (8)	**74.18 / 74.21**	**82.47 / 82.74**	**0.746 / 16.18**	**0.830 / 11.33**
$H = 6$ (12)	74.20 / 74.25	82.49 / 82.76	0.747 / 16.13	0.831 / 11.29
$H = 8$ (16)	74.21 / 74.23	82.50 / 82.75	0.747 / 16.15	0.830 / 11.31



(a) Predictions along evidence paths (b) Uncertainty-based selective prediction

Figure 3: Evidence paths and selective prediction on IEMOCAP and Fis-V (Program Component Score, PCS). (a) Individual cases and all-sample performance. (b) Risk-coverage curves by modality count and averaged over incomplete subsets. Shaded bands show 95% confidence intervals.

Effect of posterior integration and solver resolution. Table 4 shows diminishing gains beyond eight particles and four Heun steps. Our prediction averaging improves average IEMOCAP WA by 0.65 points and reduces Fis-V MSE from 17.24 to 16.18 over mean-latent decoding, supporting integration in prediction space after nonlinear decoding. With weights and prior particles fixed, four steps capture most gains from finer integration. We thus adopt $K = 8$ and $H = 4$ to balance predictive gains against sampling cost.

5.4 Visualization Analysis

To examine evidence-dependent predictions and uncertainty, we visualize ELF on IEMOCAP and Fis-V (PCS) in Figure 3. Panel (a) shows clear gains as more evidence becomes available, with both classification error and regression error decreasing substantially from early to late evidence settings. Failure cases further show a confident class reversal and score drift with increasing particle disagreement. In panel (b), uncertainty-based retention reduces the incomplete-average area under the risk-coverage curve (AURC) by 40.7% and 28.7% versus random retention. These results link posterior uncertainty to prediction reliability across classification and regression, supporting selective prediction. More visualizations appear in Appendix C.

6 Conclusion

In this paper, we present Evidence-Lattice Flow (ELF) for incomplete multimodal learning, which combines task anchoring, evidence-conditioned flow transport, and posterior integration to predict from available modalities without reconstructing missing inputs. Our analysis characterizes posterior consistency and relates transport error to task risk. Experiments on five datasets spanning emotion recognition, sentiment analysis, and action quality assessment demonstrate the effectiveness of ELF across classification and regression under both incomplete and complete evidence. Future work will explore adaptive sampling budgets and larger modality sets.

AI USE STATEMENT

We used LLMs solely for language polishing, error correction, and as an auxiliary tool in cross-checking theoretical derivations. All changes were carefully reviewed and validated by the authors, who assume full responsibility for the accuracy and scientific integrity of the manuscript.

ETHICS STATEMENT

We evaluate ELF on existing benchmarks. We do not collect additional personal data. All data usage complies with the original datasets' licensing terms. However, emotion and sentiment predictions may reflect dataset biases and should not be treated as definitive assessments of individuals. Meanwhile, AQA predictions may contain scoring errors and should be used only as a reference for evaluation and cannot replace human assessment.

REPRODUCIBILITY STATEMENT

Section 3 describes the model and training objectives. Appendix D provides assumptions and complete proofs for Section 4. Section 5.1 and Appendices A.1 and A.2 document dataset protocols, preprocessing, metrics, architecture, optimization, and checkpoint selection, with hyperparameters in Tables 5 and 6. Appendices B and C specify the ablation and visualization protocols, respectively. Code will be released upon acceptance.

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.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *Adv. Neural Inform. Process. Syst.*, pp. 36218–36242, 2024a.

Yuxing Chen, Maofan Zhao, and Lorenzo Bruzzone. A novel approach to incomplete multimodal learning for remote sensing data fusion. *IEEE Trans. Geosci. Remote Sens.*, 62:1–14, 2024b.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Int. Conf. Comput. Vis.*, pp. 24507–24517, 2025.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Fangze Fu, Wei Ai, Fan Yang, Yuntao Shou, Tao Meng, and Keqin Li. SDR-GNN: spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *Knowl. Based Syst.*, 309:112825, 2025.

Ziwen Kan, Wugeng Zheng, Tianlong Chen, and Song Wang. PAMF: Prior-aware multimodal fusion for incomplete time series data. *arXiv preprint arXiv:2606.06328*, 2026.

Fairoz Nower Khan, Nabuat Zaman Nahim, and Peizhong Ju. Flow matching with missing data. *arXiv preprint arXiv:2607.28698*, 2026.

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12458–12468, 2024.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Int. Conf. Comput. Vis.*, pp. 24068–24077, 2025.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7): 8419–8432, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Int. Conf. Learn. Represent.*, 2023.

Wei Luo, Mengying Xu, and Hanjiang Lai. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *Int. Conf. Multimedia Modeling*, pp. 411–422, 2023.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18156–18165, 2022.

Xin Ma, Xuqi Cai, Yuansheng Song, Xiaoheng Deng, and Shaohua Wan. Incomseg: Incomplete multi-modal 3d semantic segmentation for robust autonomous driving. *Expert Syst. Appl.*, 324: 132891, 2026.

Phuong-Anh Nguyen, Tien Anh Pham, Duc-Trong Le, and Cam-Van Thi Nguyen. Balm: A model-agnostic framework for balanced multimodal learning under imbalanced missing rates. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 30246–30256, 2026.

Kyu Ri Park, Hong Joo Lee, and Jung Uk Kim. Learning trimodal relation for audio-visual question answering with missing modality. In *Eur. Conf. Comput. Vis.*, pp. 42–59, 2024.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, pp. 6892–6899, 2019.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Int. Conf. Comput. Vis.*, pp. 3215–3224, 2025.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Int. Conf. Mach. Learn.*, pp. 28100–28127, 2023.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22953–22963, 2026.

Junjie Shi, Caozhi Shang, Zhaobin Sun, Li Yu, Xin Yang, and Zengqiang Yan. Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates. In *ACM Int. Conf. Multimedia*, pp. 456–465, 2024.

Vaidotas Simkus and Michael U Gutmann. CFMI: flow matching for missing data imputation. *arXiv preprint arXiv:2506.09258*, 2025.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. In *Int. Conf. Artif. Intell. Statist.*, pp. 1279–1287, 2024.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Adv. Neural Inform. Process. Syst.*, pp. 17117–17128, 2023.

Yuanzhi Wang, Zhen Cui, Mengyi Liu, Xiaoya Zhang, and Zechao Li. Incomplete multimodal probability flow recovery for emotion recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2026.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *AAAI*, pp. 2776–2784, 2023.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Trans. Mach. Learn Res.*, 2026, 2026.

Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *AAAI*, pp. 2901–2909, 2023.

Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3232–3241, 2022.

Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Trans. Circuit Syst. Video Technol.*, 30(12):4578–4590, 2020.

Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 23967–23977, 2025a.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Yuezhou Li, Rui Xu, and Wenzhong Guo. Quality-guided vision-language learning for long-term action quality assessment. *IEEE Trans. Multimedia*, 27: 7326–7339, 2025b.

Huangbiao Xu, Huanqi Wu, Xiao Ke, and Yuxin Peng. Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. In *Int. Conf. Mach. Learn.*, 2026a.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Junyi Wu, Rui Xu, and Jinglin Xu. Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In *AAAI*, pp. 11241–11249, 2026b.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *ACM Int. Conf. Multimedia*, pp. 438–446, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE intelligent systems*, 31(6): 82–88, 2016.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Assoc. Comput. Linguist.*, pp. 2236–2246, 2018.

Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Trans. Image Process.*, 33:1600–1613, 2024.

Percy S Zhai, So Won Jeong, and Veronika Ročková. Conditional flow matching for bayesian posterior inference. In *Int. Conf. Artif. Intell. Statist.*, pp. 2044–2052, 2026.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Assoc. Comput. Linguist.*, pp. 2608–2618, 2021.

Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *IJCAI*, pp. 1771–1779, 2024.

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 *Int. Conf. Comput. Vis.*, pp. 4626–4636, 2025.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *AAAI*, pp. 2281–2289, 2026.

Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP*, pp. 1–5, 2023.

APPENDIX OVERVIEW

The appendix is organized as follows.

- Appendix A describes datasets, evaluation protocols, and implementation details.
- Appendix B presents ablation configurations and additional studies.
- Appendix C analyzes evidence-dependent predictions and uncertainty across datasets.
- Appendix D provides assumptions and proofs for the theoretical results.
- Appendix E discusses limitations and future work.

A EXPERIMENTAL DETAILS

A.1 DATASETS, PREPROCESSING, AND METRICS

To assess the performance of our ELF across classification and regression, we evaluate three tasks: emotion recognition, sentiment analysis, and action quality assessment (AQA). We first describe their multimodal challenges, then introduce the five datasets (IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), FS1000 (Xia et al., 2023), and Fis-V (Xu et al., 2020)) and the evaluation protocols.

Task characteristics. Emotion recognition combines facial expressions, vocal prosody, and spoken content, with conversational context and speaker differences complicating interpretation (Lian et al., 2023). Missing facial or acoustic observations remove cues needed to distinguish similar utterances. Sentiment analysis estimates opinion polarity and intensity from language and its audiovisual delivery (Zadeh et al., 2018; Zhuang et al., 2026). Unavailable transcripts or recording channels require prediction from less explicit evidence. Long-term AQA evaluates subtle execution differences over entire routines, combining visual appearance, motion, and musical context (Xia et al., 2023; Zeng & Zheng, 2024). Missing audio or unavailable motion features reduce evidence for timing and coordination, motivating incomplete-modality scoring (Xu et al., 2026b). Together, established affective-computing benchmarks and emerging multimodal AQA benchmarks provide challenging incomplete multimodal learning (IML) settings with different temporal structures and prediction targets.

IEMOCAP. IEMOCAP (Busso et al., 2008) records conversations between pairs of actors across five sessions, combining scripted exchanges with improvised interactions. The recordings provide speech, video, and transcripts, with emotion annotations assigned to individual utterances. Following the four-class protocol used by prior work (Lian et al., 2023; Xu et al., 2024; Zhuang et al., 2026), we recognize happy, sad, neutral, and angry emotions.

CMU-MOSI. CMU-MOSI (Zadeh et al., 2016) contains 2,199 opinion-bearing segments from YouTube videos. Each segment pairs spoken content with visual and acoustic cues and has a sentiment score between -3 (strongly negative) and $+3$ (strongly positive). Following prior work (Lian et al., 2023; Xu et al., 2024; Zhuang et al., 2026), we use the benchmark split of 1,284 training, 229 validation, and 686 test segments.

CMU-MOSEI. CMU-MOSEI (Zadeh et al., 2018) extends multimodal sentiment analysis to a broader range of speakers and topics, drawing on YouTube monologues from around 1,000 speakers. Following the feature-release protocol adopted by prior work (Lian et al., 2023; Xu et al., 2024; Zhuang et al., 2026), we use 22,856 labeled segments, partitioned into 16,326 training, 1,871 validation, and 4,659 test examples. Sentiment annotations use the same $[-3, 3]$ scale as CMU-MOSI.

FS1000. FS1000 (Xia et al., 2023) evaluates long-form audiovisual understanding through figure skating scores. Following prior work (Xia et al., 2023; Xu et al., 2026b;a), we use its 1,247-video training/validation partition, with 1,000 videos for training and 247 for evaluation. Eight program categories cover short and free programs for men's singles, women's singles, and pairs, plus rhythm and free dance. Videos contain roughly 5,000 frames at 25 fps. The seven prediction targets are

technical element score (TES), program component score (PCS), and five component scores: skating skills (SS), transitions (TR), performance (PE), composition (CO), and interpretation (IN).

Fis-V. Fis-V (Xu et al., 2020) focuses on women's singles short programs and contains 500 competition videos. Following prior work (Zeng & Zheng, 2024; Xu et al., 2026b;a), we retain the split of 400 training and 100 test examples. Each routine lasts roughly three minutes and is recorded at 25 fps. Official TES and PCS annotations provide two regression targets, assessing technical execution and program components, respectively.

Evaluation protocol. All datasets use complete training observations and evaluation under every nonempty modality subset. For each dataset, we select a single checkpoint with the best performance under complete multimodal input and use it for all complete and incomplete modality subsets throughout the experiments. This rule is applied separately within each cross-validation fold, random-seed run, and AQA scoring target, as applicable. IEMOCAP uses five-fold leave-one-session-out evaluation on the four-class release, with WA and UA averaged equally over held-out sessions. MOSI and MOSEI retain the feature release's training, validation, and test partitions. For FS1000 and Fis-V, we train a separate regressor for each scoring target using the splits above, following prior work (Xia et al., 2023; Xu et al., 2026b;a).

Benchmark sources and comparison fairness. Our comparisons combine results reported in the benchmarks established by prior work (Xu et al., 2024; Zhuang et al., 2026; Xu et al., 2026b) with our reproductions following the same standard benchmark protocols. All compared existing methods and ELF use the same pre-extracted features for each dataset, ensuring a common input representation across methods.

Feature extraction. We reuse features extracted by pretrained backbones under the established benchmark protocols. IEMOCAP, MOSI, and MOSEI use the GCNet/MoMKE feature releases (Lian et al., 2023; Xu et al., 2024), with wav2vec-large for audio, DeBERTa-large for text, and MA-Net for faces. For FS1000, visual and audio features follow Skating-Mixer (Xia et al., 2023), while optical-flow features follow MCMoE (Xu et al., 2026b). Fis-V uses the Video Swin Transformer, Audio Spectrogram Transformer, and I3D features released with PAMFN (Zeng & Zheng, 2024) for visual, audio, and optical-flow inputs, respectively. The AQA input dimensions are 768/768/1024 on FS1000 and 1024/768/1024 on Fis-V, matching the feature settings used by the state-of-the-art baseline MCMoE.

Temporal preprocessing and padding. Following prior work (Lian et al., 2023; Xu et al., 2024; Zhuang et al., 2026; Xu et al., 2026b), we mean-pool emotion and sentiment features within each utterance before dialogue encoding, and average FS1000 visual features over their within-segment axis. IEMOCAP visual features correspond to the active speaker, with a zero vector used when no matching face is available. Dialogues are padded within batches, with padding excluded from attention and losses. Context for an observed modality comes only from that modality. AQA sequences are aligned and cropped to 95 segments on FS1000 and 124 on Fis-V. Training uses random contiguous crops, evaluation uses centered crops, and short sequences are zero-padded.

Label processing. We adopt the label-processing protocol used in prior work (Lian et al., 2023; Xu et al., 2024; Zhuang et al., 2026; Xu et al., 2026b). We regress sentiment scores on MOSI and MOSEI, then evaluate binary sentiment after excluding zero-valued ground-truth labels. A predicted score above zero indicates positive sentiment. For AQA, scores are divided by the target-specific constants in Table 6 during training and restored to their scoring units for MSE evaluation. FS1000 PCS is first divided by its sample-specific factor from the label file.

Classification metrics. Let c_i and $\hat{c}_i$ be the ground-truth and predicted classes for N evaluated examples, and let n_c, TP_c, FP_c, and FN_c denote class support, true positives, false positives, and false negatives. IEMOCAP uses $C = 4$ classes, with

$$\mathrm{WA} = \frac{100}{N}\sum_{i=1}^{N}\mathbf{1}\{\hat{c}_i = c_i\}, \qquad \mathrm{UA} = \frac{100}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{n_c}. \tag{16}$$

For MOSI and MOSEI, let $\mathcal{I}_0 = \{i : y_i \neq 0\}$ and $N_0 = |\mathcal{I}_0|$. Binary labels are $c_i = \mathbf{1}\{y_i > 0\}$ and $\hat{c}_i = \mathbf{1}\{\hat{y}_i > 0\}$. Accuracy and support-weighted F1 are

$$\begin{aligned} \mathrm{ACC} &= \frac{100}{N_0} \sum_{i \in \mathcal{I}_0} \mathbf{1}\{\hat{c}_i = c_i\}, \\ \mathrm{F1} &= 100 \sum_{c \in \{0,1\}} \frac{n_c}{N_0} \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \end{aligned} \tag{17}$$

All binary counts are computed on $\mathcal{I}_0$. A class term with a zero denominator is set to zero. WA, UA, ACC, and F1 are reported as percentages.

AQA metrics. For a scoring target with N evaluation videos, let y_i and $\hat{y}_i$ denote the ground-truth and predicted scores in the evaluation scoring units. Define their ranks as $q_i = \mathrm{rank}(y_i)$ and $\hat{q}_i = \mathrm{rank}(\hat{y}_i)$, assigning average ranks to tied values. Spearman correlation is

$$\rho = \frac{\sum_{i=1}^{N}(q_i - \bar{q})(\hat{q}_i - \bar{\hat{q}})}{\sqrt{\sum_{i=1}^{N}(q_i - \bar{q})^2}\sqrt{\sum_{i=1}^{N}(\hat{q}_i - \bar{\hat{q}})^2}}, \qquad \bar{q} = \frac{1}{N}\sum_{i=1}^{N} q_i, \quad \bar{\hat{q}} = \frac{1}{N}\sum_{i=1}^{N} \hat{q}_i. \tag{18}$$

Mean squared error measures score error after reversing the training normalization:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{19}$$

Higher ρ and lower MSE indicate better performance.

Aggregation. Within an AQA condition S, correlations are averaged over J scoring targets using the Fisher transformation. The incomplete-condition average then aggregates the six strict subsets, denoted by $\mathcal{S}_{\mathrm{inc}}$:

$$\begin{aligned} \rho_S &= \tanh\left(\frac{1}{J}\sum_{j=1}^{J}\operatorname{arctanh}(\rho_{S,j})\right), \\ \rho_{\mathrm{avg}} &= \tanh\left(\frac{1}{6}\sum_{S \in \mathcal{S}_{\mathrm{inc}}}\operatorname{arctanh}(\rho_S)\right). \end{aligned} \tag{20}$$

Here $J = 7$ for FS1000 and $J = 2$ for Fis-V. MSE is averaged arithmetically over targets and then over the six incomplete conditions. IEMOCAP metrics are averaged equally over folds before condition averaging. All classification metrics use arithmetic condition averages. Full-modality results are reported separately. Relative changes in the main tables compare ELF with the strongest competing value for each metric, whereas main-text accuracy gains are absolute percentage-point differences.

A.2 Architecture and Optimization

Shared posterior model. The evidence aggregator uses one learned query, two attention heads, modality embeddings, and a separate embedding of the three-bit availability mask. Unavailable modalities are excluded from attention. The anchor mapping has widths $256 \to 512 \to 256$ followed by layer normalization. The velocity network uses a 32-dimensional Fourier time embedding and three residual blocks of width 256, each modulated by the evidence condition. The flow transports standard Gaussian particles with four uniform Heun steps, requiring eight velocity evaluations. At evaluation, a fixed set of eight prior particles is reused across examples and evidence subsets.

Task adapters. We use task-specific adapters based on prior work (Xu et al., 2024; Zhuang et al., 2026; Xu et al., 2026b). Emotion and sentiment models use a separate two-layer Transformer encoder for each modality, with two heads and feed-forward width 768. A residual multilayer perceptron maps the task latent to four class probabilities on IEMOCAP or a scalar on MOSI and MOSEI. AQA uses modality-specific two-layer pointwise projections followed by a shared three-layer temporal Transformer encoder. Its three-layer decoder attends from four learned scoring queries to the

Table 5: Optimization settings. E/S denotes emotion recognition and sentiment analysis.

Setting	E/S	FS1000	Fis-V
Optimizer	AdamW	Adam	Adam
Initial learning rate	10^{-4}	10^{-4}	2×10^{-4}
Weight decay	10^{-5}	10^{-4}	10^{-4}
Learning-rate schedule	Constant	Cosine	Cosine
Final/initial learning rate	1	0.01	0.01
Batch size	32 dialogues	32 videos	32 videos
Encoder/readout dropout	0.5	0.3	0.3
Aggregator/anchor/flow dropout	0	0.3	0.3
Gradient clipping norm	5	None	None
Training epochs	300	Table 6	Table 6

Table 6: AQA training epochs and score normalization divisors.

Dataset	Target	Epochs	Score divisor
FS1000	TES	280	130
	PCS	410	60
	SS	300	10
	TR	290	10
	PE	400	10
	CO	420	10
	IN	310	10
Fis-V	TES	300	45
	PCS	390	40

task latent. Normalized positive query weights combine four equally spaced score bins in $[0, 1]$. All Transformer layers have width 256, two heads, and feed-forward width 768. The four AQA queries are distinct from the $K = 8$ posterior particles.

Losses and training. We jointly optimize the anchor, conditional flow, and task head, sampling one nonempty evidence subset uniformly per example. Full-evidence supervision uses cross-entropy on IEMOCAP, MSE on MOSI and MOSEI, and ten times normalized-score MSE on AQA, following prior work (Xu et al., 2024; Zhuang et al., 2026; Xu et al., 2026b). Incomplete-node label supervision uses the same task loss with weight 0.1 and excludes the complete node. Flow matching and task consistency both have weight 1 and include all sampled nodes. Consistency uses mean squared error between scalar predictions or class-probability vectors after particle integration, with the complete-evidence target detached. The flow-matching target is also detached. MSE reductions average over all output coordinates.

For AQA, we employ a grade-based regressor (Xu et al., 2022; Zhou et al., 2024; Xu et al., 2025a;b; 2026b) and apply a batch-hard triplet loss to the full-evidence query embeddings and the particle-averaged subset query embeddings. Query indices define positive and negative pairs, and the distance is one minus cosine similarity. The margin is 1. The sum of the two triplet terms has weight 1 on FS1000 and 0.5 on Fis-V. This auxiliary regularizer is omitted for emotion and sentiment tasks.

Optimization. Tables 5 and 6 summarize the optimization settings. Following prior work (Xu et al., 2024; Zhuang et al., 2026; Xu et al., 2026b), AQA uses cosine learning-rate decay without warm-up. Emotion recognition and sentiment analysis use a constant learning rate and a fixed 300-epoch training schedule.

B Ablation Studies

B.1 Ablation Configuration

Datasets and evaluation. IEMOCAP and Fis-V represent classification and regression. The evaluation includes all five IEMOCAP session folds and both Fis-V scoring targets. The component, integration, and sensitivity tables report the average over six incomplete subsets and the complete-modality result. The component definitions below specify the architectural and objective changes relative to ELF.

Sequential component additions. The top half of Table 3(a) follows the three stages in Section 3.

1. **Baseline.** The baseline consists of modality encoders, a mask-aware evidence aggregator, a deterministic latent projector, and a shared task head. The same projector processes complete and incomplete evidence, with task supervision at both levels.
2. **+ FETA.** A separate full-evidence projection P_ϕ defines the supervised task anchor. The subset projector regresses the detached anchor with mean squared error, and task consistency aligns its prediction with the detached complete-evidence prediction.

Table 7: Performance variability under complete and incomplete evidence. Entries report mean ± standard deviation across five IEMOCAP session folds and five independent random seeds for each remaining dataset. TMDC and MCMoE are our reimplementations (∗) on all five datasets. Average covers six incomplete subsets. Full denotes $\{a, t, v\}$ for affective tasks and $\{v, f, a\}$ for AQA.

Dataset	Metric	ELF (Ours)		TMDC*		MCMoE*	
		Average	Full	Average	Full	Average	Full
IEMOCAP	WA (%) ↑	74.18 ± 0.23	82.47 ± 0.18	72.41 ± 0.26	80.51 ± 0.23	72.34 ± 0.21	80.41 ± 0.25
	UA (%) ↑	74.21 ± 0.25	82.74 ± 0.21	72.04 ± 0.28	80.67 ± 0.27	71.84 ± 0.26	80.98 ± 0.24
MOSI	ACC (%) ↑	76.58 ± 0.36	88.41 ± 0.11	75.74 ± 0.34	88.11 ± 0.14	75.31 ± 0.39	86.13 ± 0.16
	F1 (%) ↑	76.15 ± 0.27	88.38 ± 0.13	75.24 ± 0.25	88.11 ± 0.16	74.84 ± 0.30	86.09 ± 0.19
MOSEI	ACC (%) ↑	80.70 ± 0.21	88.06 ± 0.26	80.19 ± 0.26	87.56 ± 0.25	79.93 ± 0.18	87.42 ± 0.28
	F1 (%) ↑	80.28 ± 0.30	87.97 ± 0.27	79.60 ± 0.27	87.48 ± 0.30	79.40 ± 0.30	87.34 ± 0.24
FS1000	ρ ↑	0.793 ± 0.009	0.883 ± 0.006	0.689 ± 0.012	0.834 ± 0.015	0.778 ± 0.008	0.874 ± 0.005
	MSE ↓	14.03 ± 0.33	10.75 ± 0.29	18.19 ± 0.31	13.58 ± 0.33	15.49 ± 0.41	12.07 ± 0.32
Fis-V	ρ ↑	0.746 ± 0.010	0.830 ± 0.007	0.649 ± 0.017	0.746 ± 0.028	0.729 ± 0.015	0.821 ± 0.007
	MSE ↓	16.18 ± 0.37	11.33 ± 0.28	19.24 ± 0.36	13.60 ± 0.32	17.05 ± 0.42	12.29 ± 0.27

3. **+ LCFT.** A conditional flow replaces the deterministic subset projector, and $\mathcal{L}_{\mathrm{FM}}$ replaces anchor regression. Training uses the joint ELF objective with two particles. Inference decodes one transported particle.
4. **+ PTI.** The same trained flow generates eight particles, and the task head decodes each particle before averaging their predictions.

Component removal controls. The **w/o FETA** control removes complete-evidence label supervision and replaces P_ϕ with the identity, making $\mathrm{sg}(\mathbf{C}_{[M]})$ the flow endpoint. Subset supervision trains the prediction branch, and task consistency uses the detached prediction from the fused full-evidence representation.

The **w/o LCFT** control replaces the flow with a deterministic projector trained by anchor regression, retaining task supervision and consistency. It is equivalent to Baseline + FETA. The **w/o PTI** control uses one particle from the complete model and appears as + LCFT (w/o PTI) in Table 3(a) and $K = 1$ in Table 4(a).

Task supervision and auxiliary losses. The loss controls retain $\mathcal{L}_{\mathrm{full}}$ as the task-dependent supervision that gives anchors their label semantics. Flow matching and task consistency use detached targets, so neither trains P_ϕ through the anchor branch. The auxiliary-loss controls isolate $\mathcal{L}_{\mathrm{FM}}$, $\mathcal{L}_{\mathrm{sub}}$, and $\mathcal{L}_{\mathrm{tc}}$, whereas w/o FETA removes the anchoring stage as a whole.

Each auxiliary-loss control sets one coefficient to zero and retains the remaining objectives. Without $\mathcal{L}_{\mathrm{FM}}$, the flow still receives gradients through subset task supervision and consistency. Without $\mathcal{L}_{\mathrm{sub}}$, complete-evidence anchoring and consistency supply the label information. Without $\mathcal{L}_{\mathrm{tc}}$, anchor matching and direct subset supervision remain. The Fis-V controls retain the same scoring head and query-level triplet regularizer.

B.2 Additional Ablation Studies

Error bar analysis. Table 7 compares ELF with our reimplementations of TMDC and MCMoE on all five datasets. All methods use the same five IEMOCAP session folds and identical random-seed values for the five runs on each remaining dataset, which vary weight initialization and data shuffling. Data splits, pre-extracted features, hardware and software environments, metric aggregation, and the full-modality checkpoint-selection rule are shared across methods. Standard deviations quantify variation across held-out sessions for IEMOCAP and across random seeds for the other datasets. All baseline results in Table 7 come from our reproductions under this shared evaluation protocol. Tables 1 and 2 additionally include results reported in prior work, as indicated in their captions.

ELF achieves the best mean performance for every reported metric under both complete and incomplete-average evaluation. Relative to the stronger reproduced baseline, it improves

Table 8: Evidence-conditioning ablations. IEMOCAP reports WA/UA (%), and Fis-V reports ρ/MSE. Average covers six incomplete subsets.

Setting	IEMOCAP		Fis-V	
	Average	$\{a, t, v\}$	Average	$\{v, f, a\}$
ELF (Ours)	**74.18 / 74.21**	**82.47 / 82.74**	**0.746 / 16.18**	**0.830 / 11.33**
w/o mask embedding	73.65 / 73.71	82.14 / 82.50	0.735 / 17.02	0.826 / 11.68
w/o modality embedding	73.96 / 74.00	82.27 / 82.53	0.742 / 16.50	0.829 / 11.44
w/o both embeddings	73.34 / 73.05	81.92 / 82.19	0.727 / 17.88	0.821 / 12.05
Mean evidence pooling	73.48 / 73.55	81.83 / 82.02	0.733 / 17.31	0.822 / 11.93

Table 9: Training-node sampling ablations. IEMOCAP reports WA/UA (%), and Fis-V reports ρ/MSE. Average covers six incomplete subsets.

Setting	IEMOCAP		Fis-V	
	Average	$\{a, t, v\}$	Average	$\{v, f, a\}$
Uniform subsets (Ours)	**74.18 / 74.21**	82.47 / 82.74	**0.746 / 16.18**	0.830 / 11.33
Uniform cardinalities	74.02 / 74.08	82.55 / 82.81	0.744 / 16.30	0.831 / 11.29
Singleton-biased	74.06 / 74.10	82.10 / 82.34	0.745 / 16.25	0.825 / 11.66
Bimodal-biased	74.11 / 74.16	82.33 / 82.60	0.744 / 16.34	0.828 / 11.46
Full-biased	73.62 / 73.69	**82.64 / 82.90**	0.735 / 17.03	**0.833 / 11.19**

incomplete-average IEMOCAP WA by 1.77 percentage points and reduces incomplete-average MSE by 9.4% on FS1000 and 5.1% on Fis-V. These gains are accompanied by standard deviations of a similar order to those of the baselines. For AQA, ELF standard deviations are at most 0.010 for ρ and 0.37 for MSE, indicating limited sensitivity to training randomness across the evaluated seeds.

Effect of evidence conditioning. To examine how ELF represents available evidence, we ablate the mask and modality embeddings and replace query-based aggregation with mean pooling, as summarized in Table 8. All variants continue to mask unavailable inputs. Removing both embeddings causes a larger drop than removing either alone, supporting their complementary roles in identifying evidence states. Mean pooling also underperforms query-based aggregation, showing the benefit of learning how to combine the available sources. The smaller effect of removing modality embeddings alone is consistent with modality-specific encoders retaining source information.

Effect of training-node sampling. Beyond the conditioning architecture, we examine how training coverage affects prediction by comparing uniform subset sampling with uniform cardinalities and three biased distributions in Table 9. The singleton-, bimodal-, and full-biased variants assign cardinality probabilities $(0.6, 0.3, 0.1)$, $(0.3, 0.6, 0.1)$, and $(0.2, 0.2, 0.6)$, respectively, with equal probabilities within each cardinality. Uniform subset sampling achieves the best incomplete averages on both datasets. Emphasizing full evidence improves complete prediction but degrades incomplete performance, illustrating the importance of training coverage across evidence states.

Effect of loss balancing. We next assess sensitivity to the joint objective by varying each nonzero loss weight around $(\lambda_{\mathrm{FM}}, \lambda_{\mathrm{sub}}, \lambda_{\mathrm{tc}}) = (1, 0.1, 1)$ in Table 10. Unlike the loss-removal controls, this study examines the relative strength of objectives that remain active. Performance varies only modestly across the tested weights, with incomplete-average WA spanning 73.88–74.21% on IEMOCAP and MSE spanning 16.12–16.68 on Fis-V. This local stability supports the shared loss balance used for classification and regression.

Effect of transport capacity. Finally, we examine transport capacity by varying the velocity-network width and residual-block count in Table 11, keeping the 256-dimensional task latent and the particle and solver budgets fixed. The results show that narrower or shallower networks reduce performance, while increasing the width to 512 or the depth to four yields only small gains. A fifth residual block provides no further improvement. The default width of 256 and three blocks therefore offer a compact configuration with performance close to the best observed results.

Table 10: Loss-weight sensitivity on IEMOCAP (WA/UA, %) and Fis-V (ρ/MSE). Average covers six incomplete subsets. Bold marks the default weights.

λ_{FM}	λ_{sub}	λ_{tc}	IEMOCAP		Fis-V	
			Average	$\{a,t,v\}$	Average	$\{v,f,a\}$
1	**0.1**	**1**	74.18 / 74.21	82.47 / 82.74	0.746 / 16.18	0.830 / 11.33
0.5	0.1	1	73.88 / 73.93	82.31 / 82.57	0.740 / 16.68	0.827 / 11.53
2	0.1	1	74.02 / 74.06	82.39 / 82.65	0.743 / 16.43	0.829 / 11.41
1	0.05	1	73.97 / 74.01	82.48 / 82.74	0.741 / 16.59	0.830 / 11.35
1	0.2	1	74.21 / 74.24	82.38 / 82.65	0.747 / 16.12	0.828 / 11.49
1	0.1	0.5	73.95 / 74.01	82.27 / 82.53	0.742 / 16.50	0.826 / 11.65
1	0.1	2	74.10 / 74.15	82.50 / 82.77	0.744 / 16.36	0.831 / 11.30

Table 11: Velocity-network capacity on IEMOCAP and Fis-V. Depth denotes the number of residual blocks. Average covers six incomplete subsets. Bold marks the default capacity.

Width	Depth	IEMOCAP		Fis-V	
		Average	$\{a,t,v\}$	Average	$\{v,f,a\}$
256	**3**	74.18 / 74.21	82.47 / 82.74	0.746 / 16.18	0.830 / 11.33
128	3	73.67 / 73.74	82.12 / 82.40	0.735 / 17.07	0.825 / 11.74
512	3	74.23 / 74.25	82.51 / 82.79	0.747 / 16.09	0.831 / 11.27
256	2	73.91 / 73.97	82.28 / 82.55	0.739 / 16.63	0.827 / 11.57
256	4	74.22 / 74.24	82.54 / 82.81	0.747 / 16.13	0.831 / 11.29
256	5	74.16 / 74.20	82.45 / 82.71	0.746 / 16.24	0.830 / 11.38

B.3 Comparison with Simpler Conditional Models

Conditional model controls. To assess the contribution of conditional flow beyond deterministic prediction and simpler latent distributions, we compare six models in Tables 12 and 13. All variants retain the modality encoders, evidence aggregator, task head, full-evidence supervision, and prediction consistency. Direct predicts from the evidence condition without anchor matching. Point regresses the detached full-evidence anchor and corresponds to + FETA in Table 3(a). We replace flow transport with a conditional diagonal Gaussian (Gaussian-8) or a four-component diagonal Gaussian mixture model (GMM, denoted GMM4-8). Both are trained by anchor negative log-likelihood. ELF-8 and ELF-2 use flow matching with eight and two training particles. The suffix denotes the training integration budget, while all stochastic models use $K_{\mathrm{test}} = 8$ and flow models use four Heun steps at evaluation. GMM4-8 averages two predictions per component and weights the component averages by their mixture probabilities, preserving gradients to the mixture weights.

We allocate the same validation search budget to each distribution objective, using weights in $\{0.1, 1, 10\}$, with IEMOCAP selection performed within each outer training fold. We report averages over the six incomplete subsets, with equal weights across IEMOCAP folds and the two Fis-V targets. Classification NLL is computed from the integrated true-class probability, and the Brier score sums squared probability errors over classes. These predictive metrics are distinct from the latent anchor NLL used to train Gaussian models. Entropy ranks IEMOCAP predictions, while particle score variance ranks stochastic Fis-V predictions. Deterministic regressors do not provide particle variance. Inference latency includes evidence encoding, sampling, and task readout on one RTX 3090, using batch size one and identical input shapes within each dataset.

Effect of conditional modeling. Table 12 shows that Gaussian and mixture models improve on Point, supporting distributional modeling beyond a single task anchor. ELF further improves both task performance and predictive probability quality. Compared with GMM4-8 at the same training and evaluation particle budgets, ELF-8 reduces IEMOCAP NLL from 0.885 to 0.842 and Fis-V MSE from 17.36 to 16.06, relative reductions of 4.9% and 7.5%. Its lower Brier score complements the accuracy gains, while AURC decreases from 16.36 to 15.10 on IEMOCAP and from 12.50 to

Table 12: Predictive performance of conditional models. GMM denotes Gaussian mixture model. K_{train} denotes the training integration budget. All stochastic models use $K_{\text{test}} = 8$, with four Heun steps for ELF. Metrics average six incomplete subsets and, for Fis-V, both TES and PCS. Bold identifies the default configuration.

Model	K_{train}	IEMOCAP				Fis-V	
		NLL ↓	WA (%) ↑	UA (%) ↑	Brier score ↓	MSE ↓	ρ ↑
Direct	1	0.978	71.34	71.75	0.384	23.63	0.672
Point	1	0.952	71.94	72.04	0.376	21.07	0.681
Gaussian-8	8	0.900	72.89	72.93	0.361	18.83	0.710
GMM4-8	8	0.885	73.52	73.59	0.355	17.36	0.734
ELF-8	8	0.842	74.27	74.32	0.337	16.06	0.748
ELF-2 (default)	**2**	0.851	74.18	74.21	0.340	16.18	0.746

Table 13: Selective prediction and inference latency. AURC averages six incomplete subsets, using classification error (%) for IEMOCAP and score MSE averaged over TES and PCS for Fis-V. A dash denotes unavailable particle-variance ranking for deterministic regressors. Latency is measured per input at batch size one. ELF-8 and ELF-2 differ in training particle count and both use eight particles and four Heun steps at inference.

Model	AURC ↓		Latency (ms) ↓	
	IEMOCAP	Fis-V	IEMOCAP	Fis-V
Direct	19.51	—	3.8	8.2
Point	18.90	—	4.0	8.5
Gaussian-8	17.54	14.35	4.9	9.7
GMM4-8	16.36	12.50	5.4	10.6
ELF-8	15.10	11.49	11.6	16.8
ELF-2 (default)	15.31	11.62	11.6	16.8

11.49 on Fis-V. These results support the practical value of conditional transport over the tested Gaussian families for prediction and uncertainty ranking.

Training budget and inference cost. ELF-8 improves over ELF-2 by only 0.09 IEMOCAP WA points and 0.12 Fis-V MSE, supporting the smaller default training budget. The two variants differ only in training particle count and share the same eight-particle, four-step Heun inference configuration, yielding latencies of 11.6 ms on IEMOCAP and 16.8 ms on Fis-V at the reported precision. Gaussian and mixture models offer faster direct sampling, while ELF achieves the lowest predictive losses and AURC among the evaluated models. This comparison identifies the accuracy and uncertainty benefits associated with iterative transport.

C Additional Visualizations

Predictions along an evidence path. To illustrate how posterior predictions respond to additional evidence, we follow fixed samples through nested modality subsets in Figure 3(a). On IEMOCAP, the path $\{v\} \subset \{a, v\} \subset \{a, t, v\}$ displays integrated class probabilities, with markers highlighting the true class. On Fis-V PCS, the path $\{a\} \subset \{f, a\} \subset \{v, f, a\}$ displays the eight particle scores, their mean, and the actual ground-truth score. Particle dispersion measures disagreement among task interpretations. These paths illustrate evidence expansion, while the risk-coverage analysis includes all seven subsets.

Risk aggregation and uncertainty intervals. Figures 3(b) and 6 use the same risk-coverage protocol, showing IEMOCAP and Fis-V (PCS) in the main text and all five datasets in the appendix. We rank samples by increasing predictive entropy for IEMOCAP or particle score variance for the score-regression datasets, then measure risk as coverage increases. AURC averages empirical risk over all nonempty retention sizes, with tied uncertainties evaluated by expected risk under random

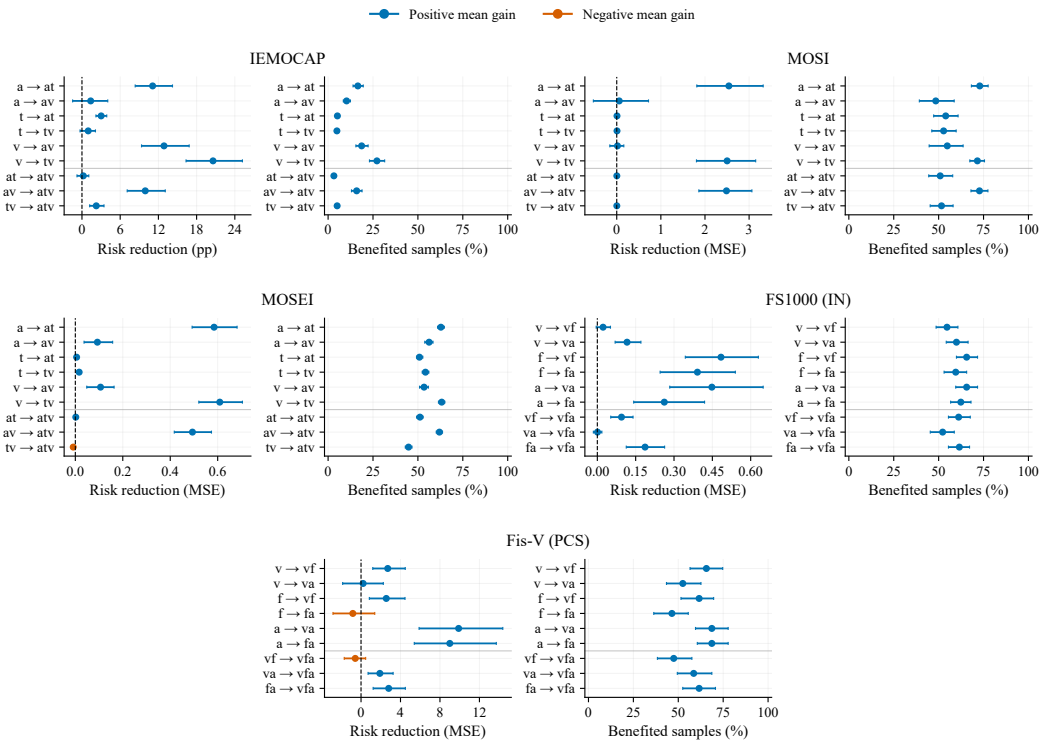


Figure 4: Single-step evidence gains across five datasets. Each dataset pairs mean risk reduction (left) with the percentage of samples with reduced loss (right), with error bars following the bootstrap protocol in Appendix C. Orange points indicate negative mean gains. IEMOCAP uses classification error, with reductions in percentage points. The other datasets use score MSE. Samples with unchanged losses remain in the denominator.

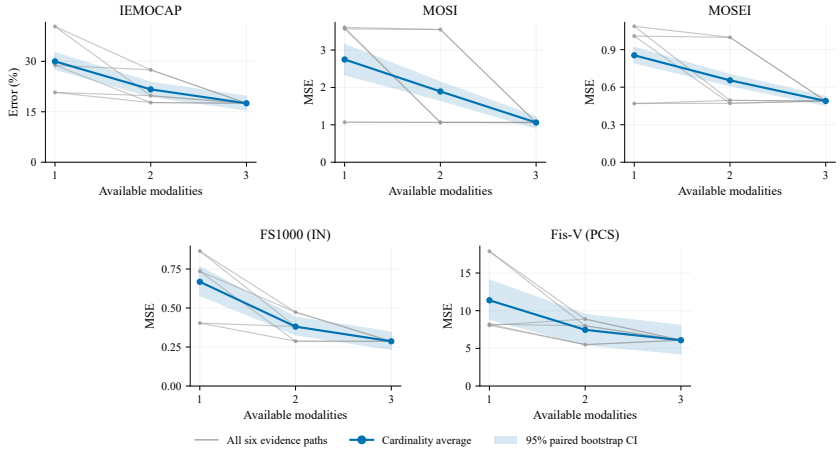


Figure 5: Empirical risk along six evidence paths on five datasets. Each node uses all evaluation samples. Thin gray curves show individual modality-addition orders, and the blue curve averages subsets of each cardinality. Bands follow the bootstrap protocol in Appendix C.

tie-breaking. In both figures, the four columns summarize singleton subsets, two-modality subsets, complete evidence, and the six-subset incomplete average. We average curves and AURC values across the corresponding subsets and IEMOCAP folds after ranking within each subset and fold. Across Figures 3–6, error bars and shaded bands denote pointwise 95% intervals from 1,000 paired cluster-bootstrap draws. The reported AURC differences use the same procedure. We resample di-

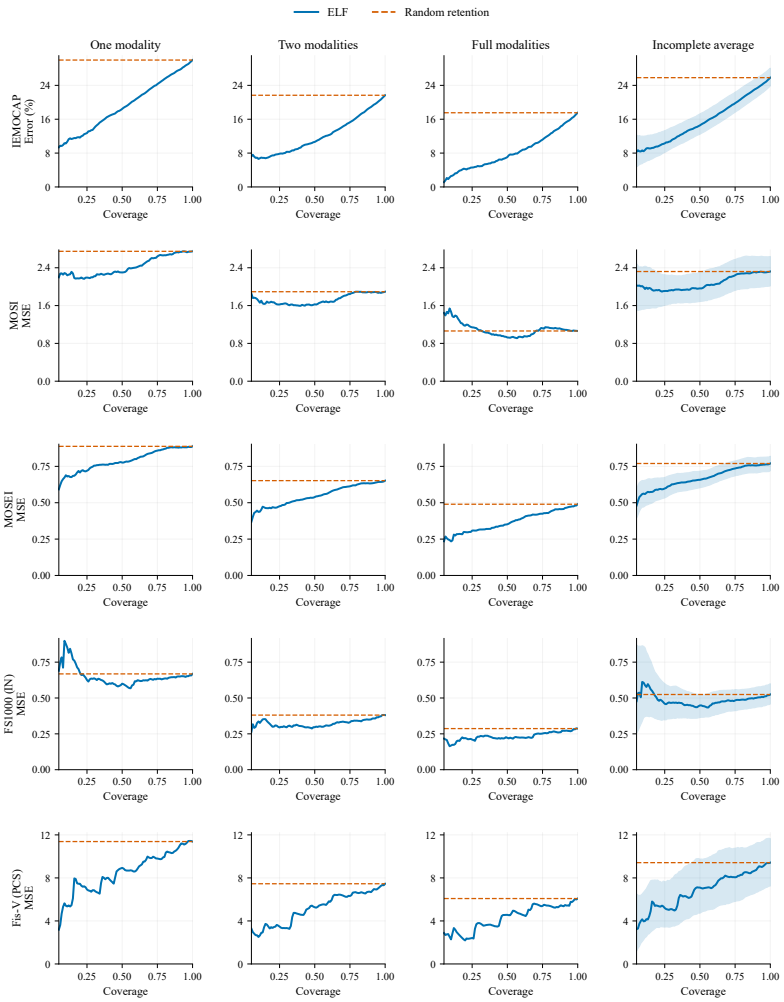


Figure 6: Risk-coverage curves across five datasets, with FS1000 evaluated on IN and Fis-V on PCS. Columns show singleton subsets, two-modality subsets, complete evidence, and the six-subset incomplete average. Dashed lines denote random retention, and shaded bands follow the bootstrap protocol above.

alogues within IEMOCAP folds, source videos for MOSI and MOSEI, and performance videos for FS1000 and Fis-V. Paired draws across subsets preserve their dependence, and the intervals quantify evaluation-sample variability for fixed models.

Single-step evidence gains. To isolate the effect of adding one modality, we evaluate all nine nonempty lattice edges $S \subset T$ with $|T| = |S| + 1$ in Figure 4. Each edge uses matched samples and the same selected checkpoint at both nodes. We report the mean risk reduction $R_S - R_T$ and the fraction of samples with strictly lower loss, using paired cluster-bootstrap intervals and equal IEMOCAP fold weights. We use classification error for IEMOCAP and score MSE for the other datasets, consistent with Figures 5 and 6. For MOSI and MOSEI, MSE captures changes in sentiment-score accuracy even when predicted polarity remains unchanged, using all samples including zero-valued labels. Mean risk decreases on 42 of the 45 dataset-edge pairs, with the largest affective-task gains arising from text additions. All three negative point estimates occur when audio is added. On MOSEI, $\{t, v\} \to \{a, t, v\}$ increases MSE by 0.0089, with its risk-reduction interval below zero. The two Fis-V reversals have intervals that include zero. A possible reason is that acoustic cues provide limited additional information beyond text and vision for sentiment or motion for scoring, allowing conditional estimation error to outweigh their marginal benefit. These results

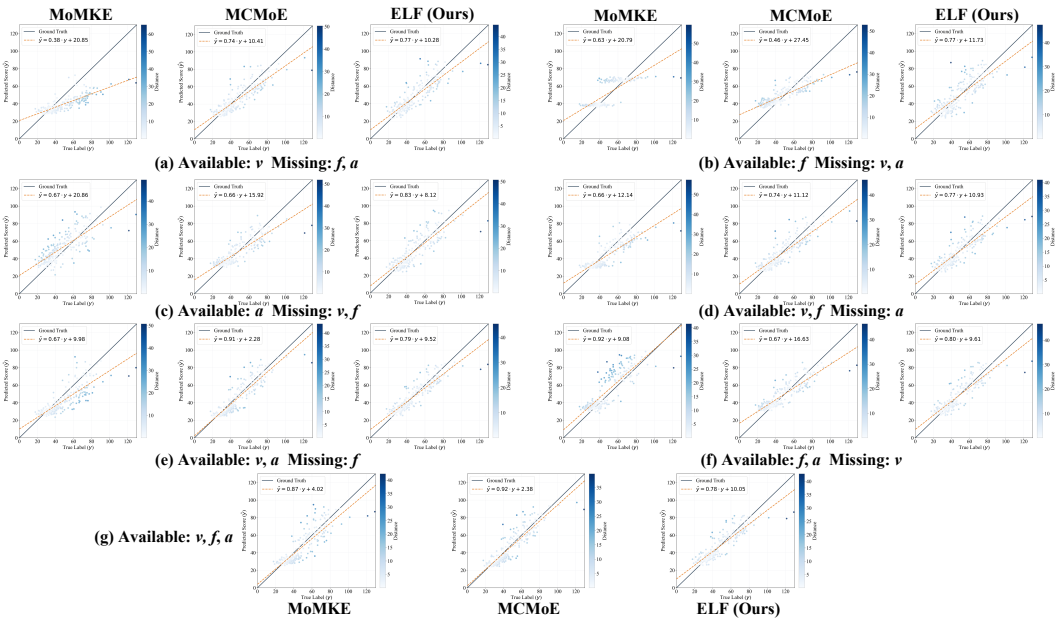


Figure 7: Predicted versus ground-truth TES scores on FS1000 for MoMKE (Xu et al., 2024), MCMoE (Xu et al., 2026b), and our ELF under all seven modality subsets. Each point represents one video score. Solid diagonals indicate ideal predictions, and dashed lines show linear fits. Colors encode error magnitude on the scale shown in each panel.

show that evidence gains depend on the available sources and distinguish learned-model behavior from the population Bayes-risk relation in Proposition 1.

Dataset-level evidence paths. To complement the single-step analysis with complete evidence paths, we evaluate all six modality-addition orders using all evaluation samples in Figure 5. Thin gray curves trace individual orders, while the thick blue curve averages subsets with the same cardinality. The cardinality-average risk decreases at both transitions on all five datasets. From singleton to complete evidence, IEMOCAP error falls from 29.98% to 17.53%, and MSE decreases from 0.668 to 0.287 on FS1000 (IN) and from 11.38 to 6.08 on Fis-V (PCS). The sentiment datasets show larger gains along paths that introduce text. These dataset-level summaries show lower aggregate risk with more evidence while retaining the variation across specific modality combinations.

Selective prediction across datasets. To assess selective prediction across tasks, we extend the IEMOCAP and Fis-V (PCS) results in Figure 3(b) to all five datasets in Figure 6, using the protocol above. Relative to random retention, incomplete-average AURC decreases by 40.7% on IEMOCAP, 10.3% on MOSI, 14.5% on MOSEI, 7.2% on FS1000 (IN), and 28.7% on Fis-V (PCS). These results connect posterior uncertainty to prediction reliability across the three tasks. On FS1000 (IN), the bimodal and full-modality AURC reductions reach 16.6% and 19.8%, with gains sustained across most coverage levels. The singleton curve is less well ordered at low coverage, and the incomplete-average AURC difference has a paired bootstrap interval of $[-0.111, 0.035]$. MOSI likewise shows stronger uncertainty ordering under incomplete evidence than with complete inputs.

Visualization of predicted scores. To examine score predictions under missing modalities, we compare ELF with MoMKE (Xu et al., 2024) and MCMoE (Xu et al., 2026b) on FS1000 (TES) in Figure 7. The scatter plots complement the aggregate error metrics in Table 2. AQA requires preserving score differences between performances, making both prediction errors and compression of the score range informative. Under each singleton subset, ELF has a fitted slope closer to one than either baseline, indicating less compression toward intermediate scores. Its fitted slopes also remain within 0.77–0.83 across all seven subsets, showing a consistent response to ground-truth score variation as evidence changes. These fitted slopes describe score-range variation across evidence conditions and should be interpreted together with prediction errors.

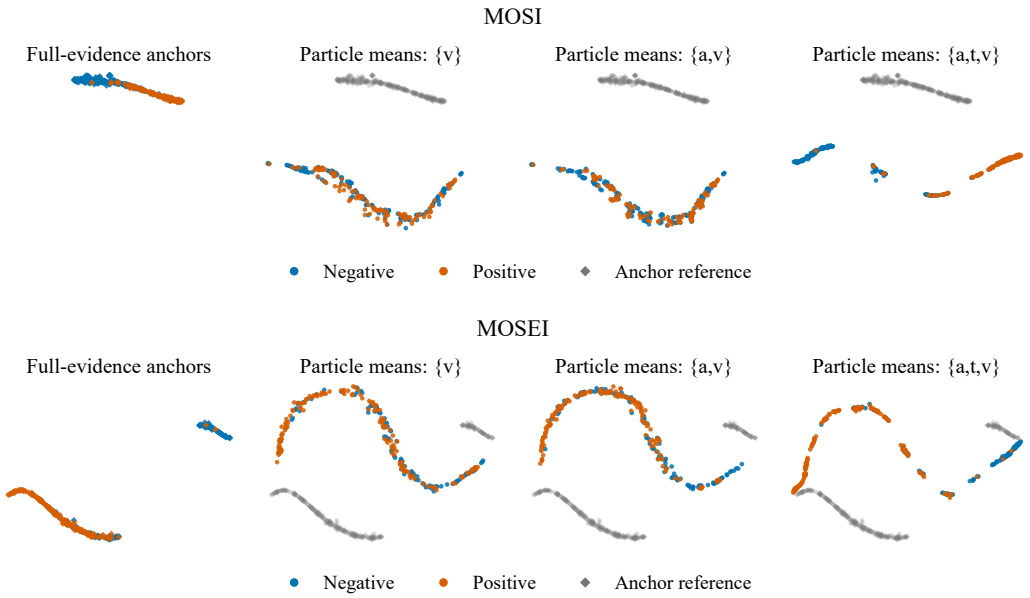


Figure 8: t-SNE of full-evidence task anchors and mean transported latents on MOSI (top) and MOSEI (bottom). Colors indicate ground-truth sentiment polarity, and gray diamonds mark anchor references. Each row uses a shared embedding across evidence conditions.

Task representations under increasing evidence. To examine how available evidence shapes task representations, Figure 8 visualizes full-evidence task anchors and mean transported latents on MOSI and MOSEI. Positive and negative samples partly overlap under visual-only and audiovisual inputs, while adding text produces clearer polarity-specific regions on both datasets. This pattern agrees with the decline in binary classification error from 41.46% to 11.59% on MOSI and from 30.90% to 11.94% on MOSEI between visual-only and complete evidence with a fixed checkpoint per dataset. These observations support ELF's ability to incorporate complementary evidence into more discriminative task representations.

D Detailed Theoretical Analysis

This appendix proves the results in Section 4 and specifies the regularity conditions for conditional transport. All random variables are defined on a common probability space, and observation and label spaces are standard Borel spaces so that regular conditional distributions exist. Conditional identities hold almost surely. Latent and prediction spaces use Euclidean norms. We analyze a fixed population distribution with an independently sampled mask $S \sim \rho$ and $\rho(S) > 0$ at every evaluated node.

D.1 Evidence Relations and Bayes Risk

For each fixed nonempty subset S, let $\mathcal{G}_S = \sigma(X_S)$. Independence of the mask gives $p(Y \mid X_S, S) = p(Y \mid X_S)$. Thus the posterior at a node depends on its observed evidence, without an additional selection effect from the masking mechanism.

Proof of Proposition 1. For $S \subseteq T$, we have $\mathcal{G}_S \subseteq \mathcal{G}_T$. Applying the tower property to the indicator of any measurable label set B gives

$$\mathbb{E}[\Pi_T(B) \mid \mathcal{G}_S] = \mathbb{E}\big[\mathbb{E}[\mathbf{1}_{\{Y \in B\}} \mid \mathcal{G}_T] \mid \mathcal{G}_S\big] = \mathbb{E}[\mathbf{1}_{\{Y \in B\}} \mid \mathcal{G}_S] = \Pi_S(B). \tag{21}$$

For regression, $\mathbb{E}[Y^2] < \infty$ ensures that $\mu_S = \mathbb{E}[Y \mid \mathcal{G}_S]$ and μ_T are square-integrable. Conditional expectation minimizes squared risk, so $R_S^{*,\mathrm{sq}} = \mathbb{E}[(Y - \mu_S)^2]$. Writing $Y - \mu_S = (Y - \mu_T) +$

$(\mu_T - \mu_S)$, the cross term vanishes because

$$\mathbb{E}[(Y - \mu_T)(\mu_T - \mu_S)] = \mathbb{E}[(\mu_T - \mu_S)\mathbb{E}[Y - \mu_T \mid \mathcal{G}_T]] = 0. \tag{22}$$

Consequently,

$$R_S^{*,\mathrm{sq}} = R_T^{*,\mathrm{sq}} + \mathbb{E}[(\mu_T - \mu_S)^2]. \tag{23}$$

The same orthogonality argument gives, for any $\mathcal{G}_S$-measurable predictor a_S with finite squared risk,

$$R^{\mathrm{sq}}(a_S) - R_S^{*,\mathrm{sq}} = \mathbb{E}[(a_S - \mu_S)^2]. \tag{24}$$

For classification with J classes, write $\pi_S(j) = P(Y = j \mid \mathcal{G}_S)$ and let a_S be a $\mathcal{G}_S$-measurable probability vector. Conditional log-loss decomposes as

$$\mathbb{E}[-\log a_S(Y) \mid \mathcal{G}_S] = H(\pi_S) + \mathrm{KL}(\pi_S \| a_S), \tag{25}$$

where $0 \log 0 = 0$ and a positive probability divided by zero gives infinite KL divergence. The Bayes predictor is π_S, and $R_S^{*,\log} = H(Y \mid X_S) \le \log J$. Hence

$$\begin{aligned} R_S^{*,\log} - R_T^{*,\log} &= H(Y \mid X_S) - H(Y \mid X_T) \\ &= \mathbb{E}[\mathrm{KL}(\pi_T \| \pi_S)] \\ &= I(Y, X_{T \setminus S} \mid X_S) \ge 0. \end{aligned} \tag{26}$$

The second equality follows by conditioning the term involving $\log \pi_S(Y)$ on $\mathcal{G}_S$. The last equality is the conditional mutual information identity (Cover & Thomas, 2006). Taking expectations in Eq. (25) also gives $R^{\log}(a_S) - R_S^{*,\log} = \mathbb{E}[\mathrm{KL}(\pi_S \| a_S)]$. □

Risk decomposition. Adding and subtracting R_S^* proves Eq. (12). Equations (24) and (25) identify the reducible term as a squared prediction error or an expected KL divergence at the observed node. The irreducible term is determined by the information in the missing modalities. Proposition 1 compares population Bayes risks. It does not require the loss or posterior entropy of every individual observation to decrease along an evidence path. With informative missingness, the relevant posterior would instead condition on the observation mechanism as well as the available values.

D.2 Conditional Flow and Endpoint Recovery

We fix deterministic encoders, the evidence aggregator, and the anchor mapping. Assume $\mathbb{E}\|\mathbf{Z}^*\|_2^2 < \infty$. The prior $\mathbf{Z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_z})$ is independent of (X, Y, S), and $t \sim \mathcal{U}(0, 1)$ is independent of all these variables. With fixed representations, $\bar{\mathbf{Z}}^*$ and $\mathbf{Z}^*$ have the same value and distribution. Stop-gradient affects parameter derivatives, not the probability path.

Population velocity regression. Let $\mathbf{U} = \mathbf{Z}^* - \mathbf{Z}_0$ and $\mathcal{H} = \sigma(\mathbf{Z}_t, t, \mathbf{C}_S, S)$. The binary mask uniquely identifies S. Since $\mathbf{U}$ is square-integrable, its conditional expectation $v^* = \mathbb{E}[\mathbf{U} \mid \mathcal{H}]$ is the orthogonal projection onto the space of $\mathcal{H}$-measurable square-integrable vector fields. For any such field v,

$$\mathcal{L}_{\mathrm{FM}}(v) = \mathbb{E}\|\mathbf{U} - v^*\|_2^2 + \mathbb{E}\|v(\mathbf{Z}_t, t, \mathbf{C}_S, \boldsymbol{r}_S) - v^*\|_2^2. \tag{27}$$

To verify the identity, expand $\|v - \mathbf{U}\|_2^2$ around v^* and condition the cross term on $\mathcal{H}$. It vanishes because $\mathbb{E}[\mathbf{U} - v^* \mid \mathcal{H}] = 0$. The first term is independent of v and can be positive.

The second term in Eq. (27) is a sum over nodes weighted by $\rho(S)$. If a shared field realizes all conditional optima, every population minimizer agrees with Eq. (13) at each evaluated node for almost every condition, time, and path state. Positive node weights ensure that no evaluated node is omitted from this conclusion.

Conditional path and continuity equation. Fix a node S and a condition $\boldsymbol{c}$ for which the conditional second moment of $\mathbf{Z}^*$ is finite. Such conditions have probability one. Write

$$\nu_{S,\boldsymbol{c}} = \mathcal{L}(\mathbf{Z}^* \mid \mathbf{C}_S = \boldsymbol{c}, S), \qquad \nu_{t,S,\boldsymbol{c}} = \mathcal{L}((1-t)\mathbf{Z}_0 + t\mathbf{Z}^* \mid \mathbf{C}_S = \boldsymbol{c}, S), \tag{28}$$

where $\mathcal{L}(\cdot \mid \cdot)$ denotes a conditional law. For $t < 1$, $\nu_{t,S,\boldsymbol{c}}$ has the strictly positive Gaussian-mixture density

$$f_{t,S,\boldsymbol{c}}(\boldsymbol{z}) = \int \frac{\exp\!\big(-\|\boldsymbol{z} - t\boldsymbol{z}^*\|_2^2 / [2(1-t)^2]\big)}{(2\pi)^{d_z/2}(1-t)^{d_z}} \, \nu_{S,\boldsymbol{c}}(d\boldsymbol{z}^*). \tag{29}$$

We abbreviate conditioning on $(\mathbf{C}_S = \boldsymbol{c}, S)$ by $(\boldsymbol{c}, S)$. For any $\varphi \in C_c^1(\mathbb{R}^{d_z})$, the chain rule and conditional expectation yield

$$\begin{aligned}\frac{d}{dt}\int \varphi(\boldsymbol{z})\,\nu_{t,S,\boldsymbol{c}}(d\boldsymbol{z}) &= \mathbb{E}[\nabla\varphi(\mathbf{Z}_t)^\top \mathbf{U} \mid \boldsymbol{c}, S] \\ &= \int \nabla\varphi(\boldsymbol{z})^\top v_S^*(\boldsymbol{z}, t, \boldsymbol{c}, \boldsymbol{r}_S)\,\nu_{t,S,\boldsymbol{c}}(d\boldsymbol{z}).\end{aligned} \tag{30}$$

Indeed, the absolute derivative inside the expectation is bounded by $\|\nabla\varphi\|_\infty\|\mathbf{U}\|_2$, which is conditionally integrable. Dominated differentiation and Fubini therefore justify both the differential identity and its time-integrated form. Equation (30) is the weak continuity equation $\partial_t\nu_{t,S,\boldsymbol{c}} + \nabla\cdot(v_S^*\nu_{t,S,\boldsymbol{c}}) = 0$, with initial law $\nu_{0,S,\boldsymbol{c}} = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_z})$. This gives the conditional flow matching relation used in the main text (Lipman et al., 2023).

Sufficient regularity for a unique probability flow. For almost every $(\boldsymbol{c}, S)$ and every $\tau < 1$, assume that v_S^* admits a version continuous in $(\boldsymbol{z}, t)$ with continuous spatial derivatives on $\mathbb{R}^{d_z} \times [0, \tau]$. Suppose that finite constants $L_{\tau,S,\boldsymbol{c}}$ and $b_{\tau,S,\boldsymbol{c}}$ satisfy

$$\begin{aligned}\|v_S^*(\boldsymbol{z}, t, \boldsymbol{c}, \boldsymbol{r}_S) - v_S^*(\boldsymbol{z}', t, \boldsymbol{c}, \boldsymbol{r}_S)\|_2 &\le L_{\tau,S,\boldsymbol{c}}\|\boldsymbol{z} - \boldsymbol{z}'\|_2, \\ \|v_S^*(\mathbf{0}, t, \boldsymbol{c}, \boldsymbol{r}_S)\|_2 &\le b_{\tau,S,\boldsymbol{c}}, \qquad 0 \le t \le \tau.\end{aligned} \tag{31}$$

These conditions imply linear growth and existence of unique forward and backward ODE solutions on $[0, \tau]$. They are sufficient regularity assumptions in addition to the moment condition. Their constants may grow as $\tau \uparrow 1$.

For completeness, let $\Phi_{s,t}$ be the characteristic map of this field with $(\boldsymbol{c}, S)$ fixed. Given $\varphi \in C_c^1(\mathbb{R}^{d_z})$, define $\chi_s(\boldsymbol{z}) = \varphi(\Phi_{s,\tau}(\boldsymbol{z}))$. The chain rule gives

$$\partial_s\chi_s(\boldsymbol{z}) + \nabla\chi_s(\boldsymbol{z})^\top v_S^*(\boldsymbol{z}, s, \boldsymbol{c}, \boldsymbol{r}_S) = 0, \qquad \chi_\tau = \varphi. \tag{32}$$

The supports of χ_s lie in a common compact set because the backward flow has linear growth on $[0, \tau]$. Applying the time-dependent weak continuity equation to any narrowly continuous probability solution η_s gives $\frac{d}{ds}\int \chi_s\, d\eta_s = 0$. Thus

$$\int \varphi\, d\eta_\tau = \int \varphi(\Phi_{0,\tau}(\boldsymbol{z}))\,\eta_0(d\boldsymbol{z}). \tag{33}$$

This identifies η_τ uniquely as the pushforward of η_0. The path in Eq. (28) is continuous in the second Wasserstein distance and satisfies Eq. (30). It therefore equals the law of the ideal ODE solution at every $t < 1$. The interpolated pairs and ODE trajectories share these marginals, although their individual trajectories need not coincide.

One concrete sufficient case is a bounded conditional anchor distribution, $\|\mathbf{Z}^*\|_2 \le R_{S,\boldsymbol{c}}$ almost surely given $(\boldsymbol{c}, S)$. Writing $\mathbf{m}_t(\boldsymbol{z}) = \mathbb{E}[\mathbf{Z}^* \mid \mathbf{Z}_t = \boldsymbol{z}, \boldsymbol{c}, S]$, Gaussian conditioning gives

$$v_S^*(\boldsymbol{z}, t, \boldsymbol{c}, \boldsymbol{r}_S) = \frac{\mathbf{m}_t(\boldsymbol{z}) - \boldsymbol{z}}{1 - t}, \qquad D_{\boldsymbol{z}}\mathbf{m}_t(\boldsymbol{z}) = \frac{t}{(1-t)^2}\,\mathrm{Cov}(\mathbf{Z}^* \mid \mathbf{Z}_t = \boldsymbol{z}, \boldsymbol{c}, S). \tag{34}$$

The derivative formula follows by differentiating the numerator and denominator of the conditional mean under the Gaussian density in Eq. (29). Bounded support justifies this differentiation and bounds the covariance operator by $R_{S,\boldsymbol{c}}^2$. Hence, on $[0, \tau]$,

$$\|D_{\boldsymbol{z}}v_S^*\|_{\mathrm{op}} \le \frac{1}{1-\tau} + \frac{\tau R_{S,\boldsymbol{c}}^2}{(1-\tau)^3}, \qquad \|v_S^*(\mathbf{0}, t, \boldsymbol{c}, \boldsymbol{r}_S)\|_2 \le \frac{R_{S,\boldsymbol{c}}}{1-\tau}. \tag{35}$$

This example permits discrete or lower-dimensional anchor distributions.

Endpoint recovery. Use the same pair $(\mathbf{Z}_0, \mathbf{Z}^*)$ to couple $\nu_{t,S,\boldsymbol{c}}$ with $\nu_{S,\boldsymbol{c}}$. By the definition of the second Wasserstein distance,

$$\begin{aligned}W_2^2(\nu_{t,S,\boldsymbol{c}}, \nu_{S,\boldsymbol{c}}) &\le \mathbb{E}[\|\mathbf{Z}_t - \mathbf{Z}^*\|_2^2 \mid \boldsymbol{c}, S] \\ &= (1-t)^2\mathbb{E}[\|\mathbf{Z}_0 - \mathbf{Z}^*\|_2^2 \mid \boldsymbol{c}, S] \longrightarrow 0.\end{aligned} \tag{36}$$

The conditional moment is finite. Therefore the ideal probability flow has the endpoint law $q_\theta(\cdot \mid \boldsymbol{c}, \boldsymbol{r}_S) = \nu_{S,\boldsymbol{c}}$ when its endpoint is defined by $t \uparrow 1$. This conclusion allows a singular target law and requires no density at $t = 1$.

At the complete node, $\mathbf{Z}^* = P_\phi(\mathbf{C}_{[M]})$ is deterministic given the condition, so $\nu_{[M],\boldsymbol{c}} = \delta_{P_\phi(\boldsymbol{c})}$. For a point target $\boldsymbol{z}^*$, the optimal velocity and its solution are

$$v^*(\boldsymbol{z}, t) = \frac{\boldsymbol{z}^* - \boldsymbol{z}}{1-t}, \qquad \Phi_{0,t}(\boldsymbol{z}_0) = (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}^*, \quad t < 1. \tag{37}$$

The limiting law is $\delta_{\boldsymbol{z}^*}$, while the spatial Lipschitz constant diverges as $t \uparrow 1$. This example explains why regularity is imposed on truncated intervals and endpoint recovery is stated as distributional convergence.

D.3 Task Integration and Predictive Dispersion

Proof of Theorem 1. Appendix D.2 establishes $q_\theta(\cdot \mid \mathbf{C}_S, \boldsymbol{r}_S) = p(\mathbf{Z}^* \mid \mathbf{C}_S, \boldsymbol{r}_S)$. The evidence-encoding assumption identifies this conditional law with $p_S = p(\mathbf{Z}^* \mid X_S, S)$. Since $\mathbf{Z}^*$ is a deterministic function of X, task sufficiency implies, for any bounded measurable h,

$$\mathbb{E}[h(Y) \mid \mathbf{Z}^*, X_S] = \mathbb{E}[h(Y) \mid \mathbf{Z}^*]. \tag{38}$$

This follows by conditioning the identity $\mathbb{E}[h(Y) \mid X] = \mathbb{E}[h(Y) \mid \mathbf{Z}^*]$ on $\sigma(\mathbf{Z}^*, X_S)$.

For classification, take $h(Y) = \mathbf{1}_{\{Y=j\}}$ and use the assumed head probabilities. The tower property gives

$$\begin{aligned} \int p_\psi(j \mid \boldsymbol{z})\, q_\theta(d\boldsymbol{z} \mid \mathbf{C}_S, \boldsymbol{r}_S) &= \mathbb{E}[P(Y = j \mid \mathbf{Z}^*) \mid X_S] \\ &= \mathbb{E}[P(Y = j \mid \mathbf{Z}^*, X_S) \mid X_S] \\ &= P(Y = j \mid X_S). \end{aligned} \tag{39}$$

Mask independence allows conditioning on S to be omitted. For regression, $\mathbb{E}[Y^2] < \infty$ permits the same argument with $h(Y) = Y$ by integrable approximation. With $g_\psi(\mathbf{Z}^*) = \mathbb{E}[Y \mid \mathbf{Z}^*]$,

$$\int g_\psi(\boldsymbol{z})\, q_\theta(d\boldsymbol{z} \mid \mathbf{C}_S, \boldsymbol{r}_S) = \mathbb{E}[\mathbb{E}[Y \mid \mathbf{Z}^*, X_S] \mid X_S] = \mathbb{E}[Y \mid X_S]. \tag{40}$$

These are Eq. (14). Substituting the recovered posteriors or means into the tower property proves their consistency across nested evidence nodes. □

Monte Carlo integration. Fix the model and observed evidence. Let $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(K)}$ be independent samples from a conditional endpoint law q_S, and set $\mathbf{G}_k = g_\psi(\mathbf{Z}^{(k)})$. This notation includes scalar regression and vector-valued classification. Define

$$a_q = \mathbb{E}_{q_S}[g_\psi(\mathbf{Z})], \qquad \boldsymbol{\Sigma}_q = \mathrm{Cov}_{q_S}(g_\psi(\mathbf{Z})), \qquad \widehat{a}_K = \frac{1}{K}\sum_{k=1}^{K} \mathbf{G}_k. \tag{41}$$

For finite second moments, linearity and independence give

$$\mathbb{E}_{\mathrm{MC}}[\widehat{a}_K] = a_q, \qquad \mathrm{Cov}_{\mathrm{MC}}(\widehat{a}_K) = \frac{\boldsymbol{\Sigma}_q}{K}, \qquad \mathbb{E}_{\mathrm{MC}}\|\widehat{a}_K - a_q\|_2^2 = \frac{\operatorname{tr} \boldsymbol{\Sigma}_q}{K}. \tag{42}$$

Classification probabilities are bounded, so their second moments are finite. Under Theorem 1, regression second moments follow from conditional Jensen's inequality and $\mathbb{E}[Y^2] < \infty$.

A deterministic numerical solver applied separately to independent prior particles also produces independent samples, now from its numerical endpoint law. Equation (42) then holds for that law. Reusing a sampled particle bank makes evaluation reproducible. Unbiasedness refers to expectation over draws of the bank, whereas the reported predictions condition on the realized bank.

Regression dispersion. Under Theorem 1, conditional total variance and task sufficiency give

$$\mathrm{Var}(Y \mid X_S) = \mathbb{E}[\mathrm{Var}(Y \mid \mathbf{Z}^*) \mid X_S] + \mathrm{Var}(g_\psi(\mathbf{Z}^*) \mid X_S). \tag{43}$$

To see this directly, decompose $Y - \mu_S$ into $Y - g_\psi(\mathbf{Z}^*)$ and $g_\psi(\mathbf{Z}^*) - \mu_S$. The first term has zero conditional mean given $(\mathbf{Z}^*, X_S)$, so the cross term vanishes. Task sufficiency identifies its conditional second moment with $\mathrm{Var}(Y \mid \mathbf{Z}^*)$. The second term gives the variance of decoded anchors.

The particle disagreement in Section 3 estimates the second term in Eq. (43) under ideal recovery. Its $1/K$ normalization satisfies

$$\mathbb{E}_{\mathrm{MC}}[\widehat{v}_S] = \frac{K-1}{K}\,\mathrm{Var}_{q_S}(g_\psi(\mathbf{Z})). \tag{44}$$

Indeed, writing the sum of squared deviations as $\sum_k (\mathbf{G}_k - a_q)^2 - K(\widehat{a}_K - a_q)^2$ and using Eq. (42) proves the identity for scalar outputs. At a deterministic complete-evidence anchor, ideal particle disagreement is zero. The first term in Eq. (43) can still contain label variability.

Classification dispersion. Let $\pi_{\boldsymbol{z}}$ be the class-probability vector produced at $\boldsymbol{z}$ and $\bar{\pi} = \mathbb{E}_{q_S}[\pi_{\mathbf{Z}}]$. Expanding entropy gives

$$H(\bar{\pi}) - \mathbb{E}_{q_S}[H(\pi_{\mathbf{Z}})] = \mathbb{E}_{q_S}[\mathrm{KL}(\pi_{\mathbf{Z}} \| \bar{\pi})] \geq 0. \tag{45}$$

Under ideal recovery and at fixed evidence $X_S = x_S$, this equals $I(Y, \mathbf{Z}^* \mid X_S = x_S)$. The empirical quantity used by ELF obeys the corresponding exact identity

$$H(\widehat{p}_S) - \frac{1}{K}\sum_{k=1}^{K} H(p_k) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{KL}(p_k \| \widehat{p}_S) \geq 0. \tag{46}$$

Thus predictive entropy describes uncertainty in the integrated class distribution, while particle disagreement describes variation among its decoded components.

Finite-particle training objectives. For a fixed observation, a fixed target b, and particles sampled independently of that target, expanding around a_q gives

$$\mathbb{E}_{\mathrm{MC}}\|\widehat{a}_K - b\|_2^2 = \|a_q - b\|_2^2 + \frac{\mathrm{tr}\,\boldsymbol{\Sigma}_q}{K}. \tag{47}$$

The cross term vanishes by unbiasedness. This formula applies to regression label supervision and to squared task consistency with $b = \mathrm{sg}(a_{\mathrm{full}})$. For class-probability consistency, the implemented mean over the J output coordinates divides both sides by J. For classification label supervision, convexity instead gives $\mathbb{E}_{\mathrm{MC}}[-\log \widehat{a}_K(Y)] \geq -\log a_q(Y)$.

Equation (47) shows how the training particle count and task-loss weights affect the preference for output dispersion. Theorem 1 characterizes the population flow-matching optimum with fixed representations, the two sufficiency assumptions, and the stated task-head identities. The joint objective in Eq. (10) also learns these representations and readouts and includes the finite-particle task terms above. Its practical accuracy therefore depends on their joint approximation and optimization, as well as numerical sampling.

D.4 Wasserstein Risk Bound

Proof of Proposition 2. Fix (X_S, S) and let $\Gamma(q_S, p_S)$ be the set of couplings of the two conditional endpoint laws. For any $\gamma \in \Gamma(q_S, p_S)$, finite first moments and the Lipschitz property of g_ψ ensure that the following integrals exist:

$$\begin{aligned} \|a_q - a_p\|_2 &= \left\| \int \big(g_\psi(\boldsymbol{z}_q) - g_\psi(\boldsymbol{z}_p)\big)\, \gamma(d\boldsymbol{z}_q, d\boldsymbol{z}_p) \right\|_2 \\ &\leq \int \|g_\psi(\boldsymbol{z}_q) - g_\psi(\boldsymbol{z}_p)\|_2\, \gamma(d\boldsymbol{z}_q, d\boldsymbol{z}_p) \\ &\leq L_g \int \|\boldsymbol{z}_q - \boldsymbol{z}_p\|_2\, \gamma(d\boldsymbol{z}_q, d\boldsymbol{z}_p). \end{aligned} \tag{48}$$

Taking the infimum over couplings gives

$$\|a_q - a_p\|_2 \le L_g W_1(q_S, p_S). \tag{49}$$

Here W_1 uses Euclidean transport cost in latent space (Villani, 2009). Uniform Lipschitz continuity of the loss yields, for each label, $|\ell(a_q, Y) - \ell(a_p, Y)| \le L_\ell \|a_q - a_p\|_2$. Taking expectations and using the assumed integrability proves

$$|R_q - R_p| \le \mathbb{E}|\ell(a_q, Y) - \ell(a_p, Y)| \le L_\ell L_g \, \mathbb{E}_{X,S}[W_1(q_S, p_S)]. \tag{50}$$

□

Loss domains and constants. For scalar squared loss, if $|a|, |b| \le B_a$ and $|y| \le B_y$, then

$$|(a-y)^2 - (b-y)^2| = |a-b|\,|a+b-2y| \le 2(B_a + B_y)|a-b|. \tag{51}$$

For normalized predictions and labels in $[0,1]$, the sharper constant is $L_\ell = 2$. Multiplying normalized MSE by ten multiplies this constant by ten. These constants apply when both predictions and labels lie in the stated domain.

For the Brier loss $\ell(a, y) = \|a - \boldsymbol{e}_y\|_2^2$ on the probability simplex, where $\boldsymbol{e}_y$ is the class basis vector,

$$|\ell(a,y) - \ell(b,y)| = |(a-b)^\top (a+b-2\boldsymbol{e}_y)| \le 2\sqrt{2}\,\|a-b\|_2. \tag{52}$$

A mean over J class coordinates scales this constant by $1/J$. For classification log-loss, a common probability floor $a_j, b_j \ge \varepsilon > 0$ implies

$$|-\log a_y + \log b_y| \le \frac{|a_y - b_y|}{\varepsilon} \le \frac{\|a-b\|_2}{\varepsilon}. \tag{53}$$

Thus $L_\ell = 1/\varepsilon$. The probability floor is an assumption on the compared predictions. Positivity at individual inputs alone does not supply a uniform constant. Proposition 2 controls these task losses directly. Accuracy, F1, and rank correlation require separate metric-specific arguments.

Bayes risk and numerical sampling. Under the task sufficiency and head assumptions of Theorem 1, integration under p_S gives the node's Bayes predictor. For either squared or log-loss, let $\overline{R}^* = \mathbb{E}_{S\sim\rho}[R_S^*]$. Combining Propositions 1 and 2 gives

$$0 \le R_q - \overline{R}^* \le L_\ell L_g \, \mathbb{E}_{X,S}[W_1(q_S, p_S)], \qquad R_q - R_{[M]}^* = (\overline{R}^* - R_{[M]}^*) + (R_q - \overline{R}^*). \tag{54}$$

The first term in the final decomposition is the average information loss from missing evidence, and the second is the excess risk of the transported prediction.

Let $q_{S,H}$ denote the endpoint law produced by H numerical integration steps. The same proof applies with $q_{S,H}$ in place of q_S whenever its moment and loss-domain assumptions hold. The resulting Wasserstein distance measures the total endpoint discrepancy. A bound on empirical flow-matching loss or a convergence rate for a finite-step solver would additionally require stability and numerical regularity assumptions.

Approximation errors beyond transport. The recovery assumptions in Theorem 1 are sufficient conditions, whereas Proposition 2 also applies to imperfect representations and heads. To separate their effects, let $\widetilde{p}_S = p(\mathbf{Z}^* \mid \mathbf{C}_S, \boldsymbol{r}_S)$ be the conditional anchor law targeted by flow matching and retain $p_S = p(\mathbf{Z}^* \mid X_S, S)$. Define

$$\begin{aligned} \varepsilon_{\text{flow}} &= \mathbb{E}_{X,S}[W_1(q_S, \widetilde{p}_S)], \\ \varepsilon_{\text{enc}} &= \mathbb{E}_{X,S}[W_1(\widetilde{p}_S, p_S)], \\ \varepsilon_{\text{task}} &= R_p - \overline{R}^*. \end{aligned} \tag{55}$$

For squared or classification log-loss, under the moment, integrability, and Lipschitz conditions of Proposition 2, including finite first moments for $\widetilde{p}_S$ and finite discrepancies above, the Wasserstein triangle inequality gives

$$0 \le R_q - \overline{R}^* \le L_\ell L_g(\varepsilon_{\text{flow}} + \varepsilon_{\text{enc}}) + \varepsilon_{\text{task}}. \tag{56}$$

Indeed, $R_q - \overline{R}^* = (R_q - R_p) + (R_p - \overline{R}^*)$, and Eq. (15) bounds the first difference. The three terms distinguish flow approximation, information lost by evidence encoding, and the excess risk of the decoded anchor distribution. The last term is nonnegative because a_p is an evidence-measurable predictor. It includes errors from insufficient anchor semantics and an imperfect task head. Exact flow fitting removes only $\varepsilon_{\text{flow}}$, so it cannot by itself correct the other two sources of error. This bound quantifies the contributions of representation, transport, and task-head errors to excess risk.

E LIMITATIONS AND FUTURE WORK

Theorem 1 characterizes population recovery when the anchor is task-sufficient, evidence encoding preserves its conditional law, and the task head is correct. Learned representations can lose relevant information, and an imperfect full-evidence head can bias decoded predictions even when the flow matches its anchor target. Equation (56) separates these residual errors from transport error. Full-evidence supervision and subset labels encourage task alignment, but do not guarantee the recovery conditions. Evaluating sensitivity to encoder compression and head calibration is a useful direction for further study.

Proposition 1 characterizes true posteriors independently of these representation assumptions. ELF couples evidence nodes through shared parameters and prediction consistency with the full-evidence head. This objective does not explicitly enforce conditional marginalization between arbitrary pairs of nodes, so learned predictions can depart from the tower relation. The evidence-path analyses characterize predictive changes across subsets rather than verify that identity. Future work could investigate cross-node marginalization constraints and direct diagnostics of their approximation error.

Our experiments use three modalities per dataset and fixed pretrained features. Larger modality sets introduce more evidence subsets, making efficient coverage during training an important extension. Future work could combine selective node sampling with end-to-end feature adaptation to extend ELF to richer inputs and additional domains.

ELF currently uses a fixed particle count and solver resolution for every example. Although the default configuration balances performance and computation on the evaluated benchmarks, different evidence states may require different sampling budgets. Adapting the particle count or integration steps to predictive uncertainty could reduce computation on easier examples while preserving more detailed posterior estimates for ambiguous inputs.