FROM AGGREGATION TO IMAGINATION: LATENT WORLD MODELING FOR MULTIMODAL FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In multimodal federated learning (MMFL) over nonstationary environments, modality availability couples observation geometry with staleness, producing modality-conditioned temporal misalignment: same-round client evidence may describe different semantic times, and scalar staleness correction cannot remove the resulting directional temporal bias. The remedy is not a better weighting scheme but a different estimand. Federated Latent World Modeling (FedLWM) shifts federation from client-parameter aggregation to a shared latent world state and re-times state inference from server-processing rounds to the semantic times to which the evidence pertains. A fixed hyperspherical semantic reference frame establishes common coordinates, a structured state preserves the geometry on which the dynamics act, and a chart-wise structure-preserving transition supports both retrospective assimilation of delayed evidence and prospective latent imagination at each client's use time. Our theoretical analysis establishes an impossibility result for dynamical prediction from aggregate first- and second-moment summaries: a positive non-closure floor independent of sample size and predictor capacity. It further establishes finite-sample identification guarantees for the structured linear transition blocks and explicit conditions under which retrospective assimilation and prospective imagination improve upon persistence under a shared rollout-error budget. Across four benchmarks, FedLWM consistently outperforms strong baselines under modality-conditioned staleness, with gains widening as modality–staleness coupling intensifies. Anonymized code is available here.

1 INTRODUCTION

Multimodal federated learning (MMFL) over nonstationary environments must reconcile client evidence whose modality availability and observation time vary across clients and rounds. In cooperative V2X perception, for example, vehicles and roadside units observe an evolving traffic scene through heterogeneous sensor suites, while modality-specific sensing and processing pipelines make these observations available at different times (Wu et al., 2026). However, conventional MMFL remains round-indexed, implicitly treating evidence processed in the same server round as contemporaneous (Feng et al., 2023). Under modality-dependent staleness, this assumption no longer holds: evidence processed in the same round may correspond to different *semantic times*—the times at which the underlying observations were valid. The question therefore shifts from how distributed evidence should be aggregated to which state the evidence describes and at what time.

The core difficulty is not that updates arrive late, but that what arrives late depends on what is observed. In MMFL, this dependence couples staleness with observation geometry: different modality subsets constrain distinct directions of the latent state and are subject to systematically different delays. As the latent state evolves, this mismatch induces anisotropic rather than direction-independent bias: directions primarily informed by slower modalities are represented by older evidence. We term this **modality-conditioned temporal misalignment**. A scalar staleness weight can only rescale stale evidence but cannot correct the direction of the resulting bias, which is determined by modality-dependent information geometry. Nor does synchronous aggregation remove the mismatch: a round barrier aligns processing time, not evidence timestamps.

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.

The remedy is not a better aggregation weight but a different estimand: a time-indexed shared latent world state. Such a state, however, exposes two structural obstacles to a well-defined transition. First, heterogeneous representations are identifiable only up to client- and time-varying orthogonal gauges; temporal transitions lack a common coordinate domain, confounding semantic evolution with representational drift—*representational gauge indeterminacy*. Second, the state must retain what its dynamics act on: point prototypes and aggregate moment summaries need not be closed under admissible evolution, yielding an irreducible non-closure floor—*predictive state insufficiency*.

We introduce **FedLWM** (Federated Latent World Modeling) to realize this estimand. Aggregation is not abandoned; FedLWM pools distributed likelihood information into a time-indexed shared latent world state at the semantic times to which the evidence pertains. A fixed hyperspherical semantic reference frame establishes a common coordinate domain for cross-client likelihood-information pooling and temporal dynamics; a structured latent world state retains the primitive-level structure on which those dynamics act; and a chart-wise structure-preserving parameterization enables transition identification from finitely many temporal pairs.

The learned transition supports two asymmetric temporal uses. Retrospectively, delayed evidence is assimilated at its semantic time and propagated forward to revise the present state. Prospectively, the current posterior is rolled open-loop to each client's use time and returned as a predictive prior for local learning. This asymmetry is fundamental: the former corrects an error already present in the estimate, whereas the latter accepts model risk in exchange for a future prior. Together, they shift federation from round-indexed aggregation to time-indexed latent world inference, with the observation-free prospective rollout constituting *latent imagination*.

Contributions. (i) **Multimodality makes staleness directional.** At the likelihood level, we formalize modality-conditioned temporal misalignment: observation age interacts with modality-dependent information geometry to induce anisotropic temporal bias that age-only correction cannot eliminate. (ii) **Relocation is forced, not preferred.** We show that likelihood information cannot be pooled consistently without a fixed representational gauge and, to the best of our knowledge, establish the first impossibility result for dynamical prediction from federated summaries containing only aggregate first and second moments: an irreducible non-closure floor independent of sample size and predictor capacity. (iii) **One operator, two temporal uses.** FedLWM combines a fixed hyperspherical semantic reference frame, a structured latent world state, and chart-wise structure-preserving dynamics. The same learned transition admits finite-sample identification guarantees for its linear blocks and supports both retrospective semantic assimilation and prospective latent imagination. Explicit sufficient conditions characterize when each use improves over persistence under a shared rollout-error budget.

2 Related Work

Federation Without a Temporal State. Classical FL aggregates locally updated model parameters (McMahan et al., 2017), with later variants addressing statistical heterogeneity and server-side optimization (Li et al., 2020; Lee et al., 2025). Asynchronous FL mitigates update staleness through delay-aware or semi-asynchronous aggregation (Wang et al., 2024; Li et al., 2026). MMFL extends federation to the multimodal setting, where clients have heterogeneous modality availability (Nguyen et al., 2026; Tan et al., 2026). Beyond parameter aggregation, other approaches federate semantic anchors (Fu et al., 2025), sufficient statistics (Dieuleveut et al., 2021), or Gaussian primitives (Zhu et al., 2026a). These lines differ in what they aggregate and how they handle heterogeneity, but generally treat delayed multimodal evidence as arriving updates or shared summaries rather than observations of an evolving state indexed by semantic time.

Temporal State Modeling Beyond Federation. Latent world models learn compact state dynamics for prediction and planning (Hafner et al., 2020; Zhu et al., 2026b), whereas data assimilation combines dynamical priors with distributed observations to infer evolving states (Chen et al., 2026). These approaches generally presuppose a well-defined state domain and observation model. In MMFL, however, neither a common state domain nor a shared observation geometry is available a priori: client representations may be identifiable only up to client- and time-dependent gauges, while modality subsets provide partial views of the latent process. FedLWM addresses this gap with a coordinate-consistent structured latent state that retains the primitive-level structure acted on by temporal dynamics, enabling semantic-time inference and prediction under partial observations.

3 MOTIVATION

3.1 PROBLEM FORMULATION

Let $[K] := \{1, \dots, K\}$ index clients, $\mathcal{M}$ the modality set, $\mathcal{Y}$ the fixed label space, and $[T] := \{1, \dots, T\}$ a discrete time index. Server round $t \in [T]$ indexes processing on the shared logical clock, whereas $s \in [T]$ indexes semantic time; z_s denotes the latent state at s. Client k observes a time-varying subset $\mathcal{M}_{k,s} \subseteq \mathcal{M}$. For client k at server round t, let $\tau_{k,t}, \Delta_{k,t} \in \mathbb{Z}_{\geq 0}$ denote the staleness of the upload processed at t and the consumption horizon of the prior broadcast at t, respectively, with $s^{\text{obs}}_{k,t} := t - \tau_{k,t}$ and $s^{\text{use}}_{k,t} := t + \Delta_{k,t}$. The bounds $\tau_{k,t} \leq t-1$ and $\Delta_{k,t} \leq T-t$ keep both times in $[T]$. Given evidence processed through round t, $(\widehat{z}_{s|t}, P_{s|t})$ denote the estimate and uncertainty of z_s. The offsets are known or estimated from protocol metadata, while the states $z_{s^{\text{obs}}_{k,t}}$ and $z_{s^{\text{use}}_{k,t}}$ remain latent. Staleness may depend on the observed modality subset: $\tau_{k,t} \not\perp \mathcal{M}_{k,s^{\text{obs}}_{k,t}}$.

We model the shared object as an observer-independent latent semantic process $\{z_s\}_{s\in[T]}$, with client models as observers. Because the latent state may change in cardinality and topology over time, $[T]$ is partitioned into maximal fixed-topology regimes $\{\mathcal{I}_p\}_{p=1}^{P}$. For consecutive $s, s+1 \in \mathcal{I}_p$, $z_s \in \mathcal{Z}_p \subseteq \mathbb{R}^{D_p}$ evolves as $z_{s+1} = F(z_s) + \zeta_s$, where $F : \mathcal{Z}_p \to \mathcal{Z}_p$ is unknown and ζ_s is an almost surely bounded martingale-difference disturbance with proxy scale σ_ζ. The trajectory remains in $\mathcal{Z}_p$ almost surely; structural changes occur only at regime boundaries, making the analysis regime-local.

For each local sample j and observed modality $m \in \mathcal{M}_{k,s}$, let $u^m_{k,s,j} \in \mathbb{S}^{d-1}$ denote the ℓ_2-normalized local encoder representation. We assume no common representational gauge across clients or time. Available modalities are modeled jointly without conditional-independence assumptions, so $\mathcal{M}_{k,s}$ specifies the observation mechanism rather than merely the amount of evidence. Let $\mathcal{H}^{\text{srv}}_{k,s}$ denote the server history available when the local likelihood is formed. Without assuming an unbiased linear observation model, we quantify the resulting evidence about the shared state by

$$\ell_{k,s}(z) := \log p\Big(\big\{\big(\{u^m_{k,s,j}\}_{m\in\mathcal{M}_{k,s}}, y_{k,j}\big)\big\}_j \,\Big|\, z, \mathcal{M}_{k,s}, \mathcal{H}^{\text{srv}}_{k,s}\Big). \tag{1}$$

3.2 THE CORE FAILURE: MODALITY-CONDITIONED TEMPORAL MISALIGNMENT

Same-round uploads can describe different semantic times, so their aggregate need not correspond to a single-time state (Fig. 1(a)). This temporal misalignment also has a likelihood-level consequence: stale evidence is generated under modality-dependent observation geometries. For $s = s^{\text{obs}}_{k,t}$, define the conditional likelihood curvature $\mathcal{J}_{k,s} := -\mathbb{E}\Big[\nabla^2_z \ell_{k,s}(z_s) \mid z_s, \mathcal{M}_{k,s}, \mathcal{H}^{\text{srv}}_{k,s}\Big]$. Under correct local specification, conditional independence between the observation and future state given z_s, and local Hessian regularity, a first-order expansion yields

$$\mathbb{E}\big[\nabla_z \ell_{k,s}(z_t) \mid z_s, z_t, \mathcal{M}_{k,s}, \mathcal{H}^{\text{srv}}_{k,s}\big] = -\mathcal{J}_{k,s}(z_t - z_s) + O\big(\|z_t - z_s\|_2^2\big). \tag{2}$$

Interpreting evidence generated at z_s as evidence for z_t therefore induces a systematic score bias. Because $\mathcal{J}_{k,s}$ depends on $\mathcal{M}_{k,s}$, scalar age weighting can only rescale this bias, whereas modality-conditioned curvature reorients it relative to $z_t - z_s$ (Fig. 1(b)). Likelihood transport is generally unavailable: stochastic or noninvertible dynamics require a backward state kernel that may be intractable and need not preserve the local information form. The mismatch is bidirectional (Fig. 1(c)): observation delay $\tau_{k,t}$ can bias present-state inference, while the use-time gap $\Delta_{k,t}$ can render the shared target stale before use. This target drift is distinct from, and can compound, classical client drift: even exact local optimization may converge to an obsolete target at consumption time. The former motivates retrospective assimilation; the latter motivates prospective prediction to the client's use time. Slow evolution attenuates both; eliminating either removes only its mismatch. Discarding stale uploads is not neutral: under modality-dependent staleness, age thresholding implicitly selects modalities and can disproportionately suppress paired-modality evidence with cross-modal information.

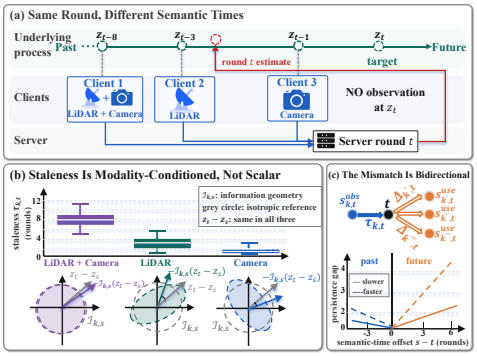

Figure 1: Modality-conditioned temporal misalignment: CooperScene measurements in (b,c).

3.3 Two Obstacles: No Common Domain, No Sufficient State

Resolving this mismatch requires a shared-state transition. Such a transition is well defined only when it acts on a common coordinate domain and a dynamically sufficient state representation, neither of which is guaranteed in federated settings (Appendix B).

Representational Gauge Indeterminacy. Federated representations need not share a common coordinate system across clients or over time. Under the assumed inner-product/geodesic invariance of the representation model, normalized representations are identifiable only up to an orthogonal gauge: if $u_{k,s} = R_{k,s}\widetilde{u}_{k,s}$, then $u_{k,s} = (R_{k,s}Q_s)(Q_s^\top \widetilde{u}_{k,s})$ for any time-varying $Q_s \in O(d)$. Semantic evolution and gauge rotation are therefore observationally indistinguishable (Fig. 2(a))—a structural non-identifiability that makes cross-round differences gauge dependent. Likelihood curvatures likewise transform by congruence, so summing them across unaligned clients remains gauge dependent. Thus, both likelihood-information pooling and temporal dynamics require a common gauge.

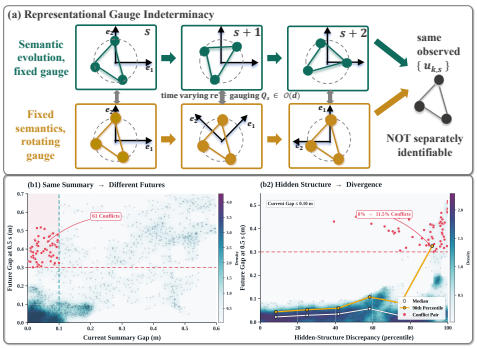

Figure 2: Two structural obstacles, with CooperScene evidence for predictive state insufficiency.

Predictive State Insufficiency. Even on a common domain, prediction from a shared summary is well defined only if the dynamics close on it. For a point-summary operator Π_{pt}, closure requires some $\overline{F}$ with $\Pi_{\mathrm{pt}} \circ F = \overline{F} \circ \Pi_{\mathrm{pt}}$. If $\Pi_{\mathrm{pt}}(z) = \Pi_{\mathrm{pt}}(z')$ but $\Pi_{\mathrm{pt}}(F(z)) \neq \Pi_{\mathrm{pt}}(F(z'))$, no such $\overline{F}$ exists. Fig. 2(b) shows both signatures on CooperScene: near-identical summaries yield divergent futures, while among matched pairs divergence grows with hidden-structure discrepancy. Class prototypes illustrate this: a class mean can erase within-class structure and entity-level variation that governs future evolution. Neither more samples nor a richer predictor can recover information already discarded by the summary. The obstruction is therefore dynamical insufficiency, not a limitation of any particular prototype construction. Overcoming these obstacles therefore requires a common gauge, a dynamically sufficient shared state, and a transition class learnable from finite trajectories. Finite-sample learnability is demanding: each time step yields at most one transition pair regardless of client count, whereas a dense linear map on D_p dimensions already has $O(D_p^2)$ parameters.

4 FedLWM: Federated Latent World Modeling

FedLWM treats federation as inference over a time-evolving latent world whose state is the time-indexed estimand. In each round, clients upload timestamped semantic evidence alongside any backbone update; the server assimilates this evidence at its observation times, rolls the inferred state to client-specific use times, and returns predictive priors. Figure 3 summarizes this inference–prediction loop, and Algorithms C.1–C.2 specify one round.

4.1 Hyperspherical Semantic Reference Frame

FedLWM establishes a common semantic coordinate domain for heterogeneous client representations, likelihood-information pooling, and temporal dynamics through a *Hyperspherical Semantic Reference Frame* (HSRF). Shared across clients and time, the HSRF anchors representations to fixed semantic references and fixes the modeled orthogonal gauge without trainable shared parameters.

Let $\mathcal{C}_{\mathrm{sem}} = \{c_y \in \mathbb{S}^{d-1} : y \in \mathcal{Y}\}$ denote the semantic fiducials, and let $\mathcal{C} = \mathcal{C}_{\mathrm{sem}} \cup \mathcal{C}_{\mathrm{cal}}$ include fixed calibration fiducials that enlarge the reference span. Define $\mathcal{V}_{\mathcal{C}} := \mathrm{span}(\mathcal{C})$ and $r := \dim(\mathcal{V}_{\mathcal{C}})$. For each class y, a shared deterministic orthonormalization constructs $B_y \in \mathbb{R}^{d \times d_v}$ satisfying $B_y^\top B_y = I$ and $\mathrm{col}(B_y) \subseteq \mathcal{V}_{\mathcal{C}} \cap T_{c_y}\mathbb{S}^{d-1}$, hence $d_v \leq r - 1$. A normalized representation assigned to class y is then mapped to the canonical coordinate $v_{k,s,j}^m = B_y^\top \mathrm{Log}_{c_y}(u_{k,s,j}^m)$, defining a class-centered semantic chart around c_y. Here $T_{c_y}\mathbb{S}^{d-1}$ and Log_{c_y} denote the spherical tangent space and Riemannian logarithm, respectively. At (k, s), let a index semantic and calibration frame targets, with $c_a \in \mathcal{C}$, count $n_{k,s,a}$, and normalized mean $\bar{u}_{k,s,a}^m$. With $d_g(c, u) = \arccos(c^\top u)$ and chart

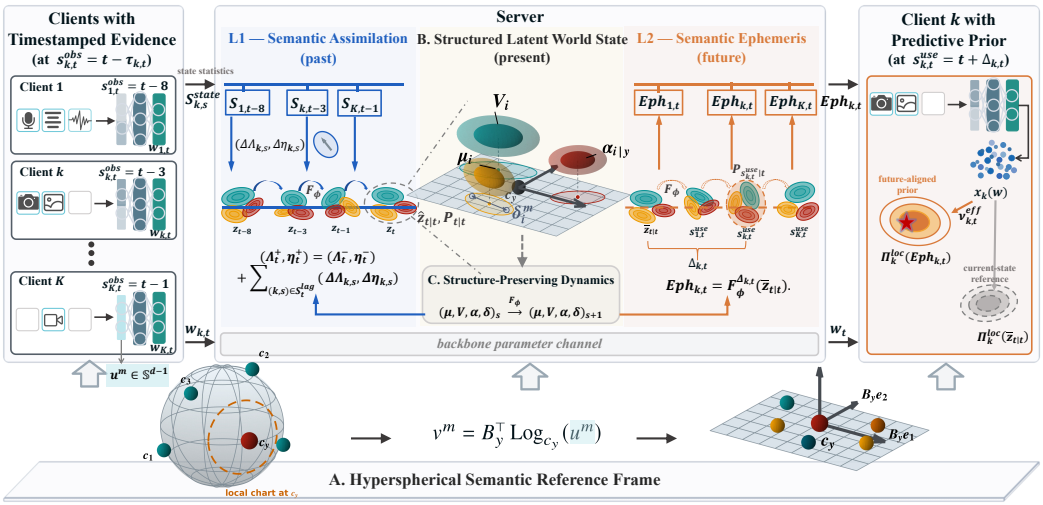

Figure 3: Overview of the FedLWM semantic inference–prediction loop. The HSRF (A) canonicalizes heterogeneous client representations for coordinate-consistent evidence pooling into the structured latent state (B); structure-preserving dynamics (C) evolve this state across semantic time, enabling delayed-evidence assimilation (L1) and use-time predictive priors (L2).

radius θ_0, each client aligns its representations to the HSRF through

$$\mathcal{L}_{k,s}^{\text{frame}} = \sum_{m\in\mathcal{M}_{k,s}} \sum_{a:n_{k,s,a}>0} n_{k,s,a}\, d_g^2(\bar{u}_{k,s,a}^m, c_a) + \sum_j \sum_{m\in\mathcal{M}_{k,s}} \left[d_g\left(u_{k,s,j}^m, c_{y_{k,j}}\right) - \theta_0\right]_+^2. \quad (3)$$

The frame term aligns modality-wise means with fixed fiducials, while the hinge penalizes chart violations without suppressing the within-chart variation modeled by the latent state.

Proposition 1 (Gauge-Consistent Canonical Pooling)**.** *Under the orthogonal gauge model, every residual gauge fixing $\mathcal{C}$ pointwise leaves the canonical coordinates invariant. Hence likelihood-information increments are residual-gauge invariant and, under conditional independence given their state slices and recorded server histories, enter additively into the trajectory information form. Conversely, if* $\mathrm{col}(B_y) \not\subseteq \mathcal{V}_{\mathcal{C}}$*, some residual gauge alters the projected coordinates.*

4.2 Structured Latent World State

FedLWM fixes the reference geometry through the HSRF and represents temporal variation through a structured latent world state comprising persistent primitives. Within each fixed-topology regime p, these primitives parameterize the canonical class-conditional distribution. Primitive i in chart y_i carries a mean $\mu_{i,s}$ and raw second moment $M_{i,s}$, which determine its location and covariance $V_{i,s} = M_{i,s} - \mu_{i,s}\mu_{i,s}^\top$, together with mixture weight $\alpha_{i|y_i,s}$ and modality offsets $\{\delta_{i,s}^m\}_{m\in\mathcal{M}}$ describing its mass and modality-specific shifts. Collecting these quantities, the latent state is

$$z_s = \mathrm{vec}\left(\mu_{i,s}, \mathrm{vech}(M_{i,s}), \alpha_{i|y_i,s}, \{\delta_{i,s}^m\}_{m\in\mathcal{M}}\right)_{i=1}^{N_p} \in \mathcal{Z}_p. \quad (4)$$

Primitive identities are maintained within a regime, with adjacent-slice matching when needed, while their locations are inferred as latent variables rather than client-centroid averages, preserving the component structure required by the dynamics. Raw second moments support additive responsibility-weighted accumulation, and the closed convex admissible set $\mathcal{K}_p \subseteq \mathcal{Z}_p$ enforces $V_{i,s} \succeq \sigma_{\min}^2 I$, bounded chart coordinates, feasible mixture weights, and centered modality offsets $\sum_m w_m \delta_{i,s}^m = 0$, with fixed $w_m \geq 0$ and $\sum_m w_m = 1$.

Partial multimodal observations are related to this state through a joint emission model. For class y, let $i \sim \mathrm{Cat}(\alpha_{\cdot|y})$, $v \sim \mathcal{N}(\mu_i, V_i)$, and $v^m = B_y^\top \mathrm{Log}_{c_y}(u^m) = v + \delta_i^m + \epsilon_i^m$, where residual cross-modal dependence is captured by the joint emission covariance. For any observed modality subset, inference uses the corresponding marginal of the joint emission model, so modality availability changes the likelihood-information geometry without altering the latent state or requiring explicit imputation. Conditioned on the recorded server prior, posterior responsibilities yield sufficient statistics $\mathsf{S}_{k,s}^{\text{state}}$ as likelihood evidence for server-side state inference.

4.3 Temporal State Inference and Prediction

Within each regime, FedLWM fits a structure-preserving transition F_ϕ from consecutive state estimates, evolving the structured state for retrospective assimilation and prospective prediction.

Structure-Preserving Dynamics. Rather than parameterizing a dense full-state transition, F_ϕ factorizes the evolution into chart-wise low-dimensional blocks. For chart y, the primitive means, modality offsets, covariances, and mixture weights evolve as

$$\begin{aligned} \mu_{i,s+1} &= A_y^\mu \mu_{i,s} + b_{y,i}, & \delta_{i,s+1}^m &= A_y^\delta \delta_{i,s}^m + b_y^{\delta,m}, \\ V_{i,s+1} &= \sigma_{\min}^2 I + A_y^V (V_{i,s} - \sigma_{\min}^2 I)(A_y^V)^\top + Q_{y,i}, & \alpha_{\cdot|y,s+1} &= S_y \alpha_{\cdot|y,s}. \end{aligned} \tag{5}$$

Here $Q_{y,i} \succeq 0$ is a deterministic covariance-injection term, distinct from the process disturbance ζ_s. With $\sum_m w_m b_y^{\delta,m} = 0$ and an $\alpha_{\min}$-preserving column-stochastic S_y, the unprojected transition preserves the covariance floor, mixture feasibility, and centered-offset constraint. It evolves $V_{i,s}$ directly to preserve covariance structure and recovers the stored raw moment as $M_{i,s+1} = V_{i,s+1} + \mu_{i,s+1}\mu_{i,s+1}^\top$. The full transition is $F_\phi = \Pi_{\mathcal{K}_p} \circ \overline{F}_\phi$, where projection onto $\mathcal{K}_p$ enforces the remaining admissibility constraints and is non-expansive. Separate A_y^μ and A_y^V let location and dispersion evolve independently, while chart-wise factorization reduces transition dimension relative to a dense full-state map and underpins finite-sample identification of the linear blocks.

Semantic Assimilation (L1). At round t, client k's upload carries semantic time $s = s_{k,t}^{\mathrm{obs}}$ and sufficient statistics $\mathsf{S}_{k,s}^{\mathrm{state}}$ formed under $\widehat{z}_{k,s}^-$, the server prior recorded when evidence at s was formed. The server reconstructs the score $g_{k,s} = \nabla_z \ell_{k,s}(\widehat{z}_{k,s}^-)$ and a calibrated positive-semidefinite curvature surrogate $\widetilde{J}_{k,s} \succeq 0$, yielding a quadratic approximation (Appendix C.2). Unlike scalar age weighting, $\widetilde{J}_{k,s}$ preserves modality-conditioned information directions of the observation.

To place each contribution at its semantic time, the server maintains the fixed-lag trajectory $X_t = (z_{t-L_{\mathrm{lag}}}, \dots, z_t)$, whose slices are coupled by the learned dynamics F_ϕ. For tractable smoothing, its dynamic prior is approximated as $q_t^-(X_t) \approx \mathcal{N}(\widehat{X}_t^-, P_t^-)$, with information parameters $\Lambda_t^- = (P_t^-)^{-1}$ and $\eta_t^- = \Lambda_t^- \widehat{X}_t^-$. At slice s, $\iota_{s,t}$ extracts $z_s = \iota_{s,t} X_t$, while $\iota_{s,t}^\top$ embeds local information into the trajectory. Let $\mathcal{S}_t^{\mathrm{lag}}$ collect retained-lag pairs processed at t, and $(\Delta\Lambda_t^{\mathrm{old}}, \Delta\eta_t^{\mathrm{old}})$ the boundary contribution of older evidence. Each retained factor then contributes:

$$\begin{aligned} \Delta\Lambda_{k,s} &= \iota_{s,t}^\top \widetilde{J}_{k,s} \iota_{s,t}, & \Delta\eta_{k,s} &= \iota_{s,t}^\top \big(g_{k,s} + \widetilde{J}_{k,s} \widehat{z}_{k,s}^-\big), \\ \Lambda_t^+ &= \Lambda_t^- + \sum_{(k,s)\in\mathcal{S}_t^{\mathrm{lag}}} \Delta\Lambda_{k,s} + \Delta\Lambda_t^{\mathrm{old}}, & \eta_t^+ &= \eta_t^- + \sum_{(k,s)\in\mathcal{S}_t^{\mathrm{lag}}} \Delta\eta_{k,s} + \Delta\eta_t^{\mathrm{old}}. \end{aligned} \tag{6}$$

The resulting posterior is $\widehat{X}_t^+ = (\Lambda_t^+)^{-1}\eta_t^+$, with $P_t^+ = (\Lambda_t^+)^{-1}$ and current marginal $(\widehat{z}_{t|t}, P_{t|t})$. Within the retained lag, each likelihood factor is inserted at its semantic time; older evidence is folded into the boundary slice, and F_ϕ-coupling propagates its posterior effect to the present. Thus, L1 assimilates delayed evidence at its semantic time rather than correcting it only on arrival. Conditional independence under the recorded server history makes the factors additive; the HSRF makes their cross-client sum coordinate-consistent.

Semantic Ephemeris (L2). Starting from the admissible current estimate $\bar{z}_{t|t} := \Pi_{\mathcal{K}_p}(\widehat{z}_{t|t})$ and uncertainty $P_{t|t}$, FedLWM rolls open-loop to each client's use time $s_{k,t}^{\mathrm{use}}$, yielding $q_{k,t}^{\mathrm{use}} \approx \mathcal{N}(\mathrm{Eph}_{k,t}, P_{s_{k,t}^{\mathrm{use}}|t})$, where $P_{s_{k,t}^{\mathrm{use}}|t}$ is the propagated covariance and $\mathrm{Eph}_{k,t} := F_\phi^{\Delta_{k,t}}(\bar{z}_{t|t})$ the *semantic ephemeris*, *i.e.*, the predicted shared state at use time. A regularized client maps the shared-state ephemeris to compatible local coordinates through

$$\mathcal{L}_k(w) = \mathcal{L}_k^{\mathrm{base}}(w) + \frac{\nu_{k,t}^{\mathrm{eff}}}{2} \left\| x_k(w) - \Pi_k^{\mathrm{loc}}(\mathrm{Eph}_{k,t}) \right\|_2^2 . \tag{7}$$

Here $\mathcal{L}_k^{\mathrm{base}}$ is the backbone objective before ephemeris regularization. Let $x_k(w) \in \mathbb{R}^{d_k^{\mathrm{loc}}}$ denote the client-side state readout induced by w, and $\Pi_k^{\mathrm{loc}} : \mathbb{R}^{D_p} \to \mathbb{R}^{d_k^{\mathrm{loc}}}$ a linear shared-to-local map. Predictive uncertainty gates the prior strength via $\nu_{k,t}^{\mathrm{eff}} = \nu_0 \chi\Big(\Pi_k^{\mathrm{loc}} P_{s_{k,t}^{\mathrm{use}}|t} (\Pi_k^{\mathrm{loc}})^\top\Big)$, where $\chi : \mathbb{S}_+ \to [0,1]$ is nonincreasing in PSD order. Thus, L2 turns observation-free rollout into an uncertainty-gated use-time prior, grounding latent imagination in local learning.

5 THEORETICAL ANALYSIS

The analysis establishes dynamic non-closure of aggregate moment summaries, finite-sample identifiability of structured linear transition blocks, and sufficient conditions for learned dynamics to outperform persistence in retrospective propagation and prospective prediction. Coordinate consistency follows from Prop. 1; all results are regime-local, with proofs and assumptions in Appendices A–D.

Theorem 1 (Dynamic Non-Closure of Aggregate Moment States)**.** *Let Π_{mom} retain chart-wise aggregate mean and second raw moment under the Euclidean–Frobenius norm. There exist $z^{(1)}, z^{(2)} \in \mathcal{K}_p$ and an admissible transition F of the form in Eq. 5 such that $\Pi_{\mathrm{mom}}(z^{(1)}) = \Pi_{\mathrm{mom}}(z^{(2)})$ but $\Pi_{\mathrm{mom}}(F(z^{(1)})) \neq \Pi_{\mathrm{mom}}(F(z^{(2)}))$. Consequently, every summary-space predictor G satisfies $\max_{r\in\{1,2\}} \|G(\Pi_{\mathrm{mom}}(z^{(r)})) - \Pi_{\mathrm{mom}}(F(z^{(r)}))\| \geq c_{\mathrm{cl}}$, where $c_{\mathrm{cl}} \coloneqq \frac{1}{2}\|\Pi_{\mathrm{mom}}(F(z^{(1)})) - \Pi_{\mathrm{mom}}(F(z^{(2)}))\| > 0$. This bound is independent of sample size and predictor capacity.*

Theorem 2 (Finite-Sample Identification of Linear Transition Blocks)**.** *Let N_{trans} be the number of distinct transition pairs and $\vartheta_y^{\mathrm{lin}} \in \mathbb{R}^{p_y^{\mathrm{lin}}}$ collect the free parameters of the vectorized affine-linear blocks in chart y. Suppose the corresponding design is persistently exciting with level σ_{PE}, the disturbances are almost surely bounded martingale differences with proxy scale σ_ζ, both source- and target-state estimation errors are bounded by $\epsilon_{\mathrm{state}}$, and the residual confounding error by $\epsilon_{\mathrm{conf},y}$. There exists $C > 0$, depending only on fixed regularity bounds, such that for every $\xi \in (0,1)$, with probability at least $1-\xi$,*

$$\left\|\widehat{\vartheta}_y^{\mathrm{lin}} - \vartheta_y^{\mathrm{lin}}\right\|_2 \leq C\left[\frac{\sigma_\zeta}{\sigma_{\mathrm{PE}}}\sqrt{\frac{p_y^{\mathrm{lin}} + \log(1/\xi)}{N_{\mathrm{trans}}}} + \frac{(1+\|\vartheta_y^{\mathrm{lin}}\|_2)\epsilon_{\mathrm{state}}}{\sigma_{\mathrm{PE}}} + \epsilon_{\mathrm{conf},y}\right]. \tag{8}$$

Only the sampling term decays with temporal pairs; state-estimation and residual-confounding errors form non-vanishing floors. Accordingly, the rate depends on the structured dimension p_y^{lin} rather than the $O(D_p^2)$ dimension of a dense full-state operator.

Theorem 3 (Retrospective and Prospective Utility of Learned Dynamics)**.** *For the deterministic drift map F within a fixed-topology regime, let $\epsilon_F \coloneqq \sup_{u\in\mathcal{K}_p} \|F_\phi(u) - F(u)\|_2$ and $L_\phi \coloneqq \mathrm{Lip}(F_\phi; \mathcal{K}_p)$. Set $E_0 \coloneqq 0$ and $E_h \coloneqq \epsilon_F \sum_{n=0}^{h-1} L_\phi^n$ for $h \geq 1$. Then $\|F_\phi^h(u) - F^h(u)\|_2 \leq E_h$ for any $u \in \mathcal{K}_p$ whose true rollout remains in $\mathcal{K}_p$.* (i) Retrospective assimilation. *For $\tau \geq 1$, let $\mathrm{L1}^0$ denote timestamped assimilation with identity dynamics, conditioned on the same delayed-state estimate $z \in \mathcal{K}_p$ at semantic time $t-\tau$. If $E_\tau < \|F^\tau(z) - z\|_2$, learned dynamics have strictly smaller drift-propagation error than persistence. This comparison concerns temporal propagation rather than total filtering error.* (ii) Prospective prediction. *For $\Delta \geq 1$, suppose $\|\widehat{z}_{t|t} - z_t\|_2 \leq \epsilon_{\mathrm{state}}$ and $\|\Pi_k^{\mathrm{loc}}\|_2 \leq 1$. Let $x_k^\star$ denote the ideal client coordinate at use time, and define $D_\Delta \coloneqq \|\Pi_k^{\mathrm{loc}}(F^\Delta(z_t) - z_t)\|_2$, $\epsilon_{\mathrm{loc}} \coloneqq \|x_k^\star - \Pi_k^{\mathrm{loc}} F^\Delta(z_t)\|_2$, and $\Gamma_\Delta^\star \coloneqq \mathrm{Lip}(F^\Delta; \mathcal{K}_p)$. If $E_\Delta < D_\Delta - (1+\Gamma_\Delta^\star)\epsilon_{\mathrm{state}} - 2\epsilon_{\mathrm{loc}}$, the semantic-ephemeris center $\Pi_k^{\mathrm{loc}} F_\phi^\Delta(\bar{z}_{t|t})$ is strictly closer to $x_k^\star$ than the persistence center $\Pi_k^{\mathrm{loc}}\bar{z}_{t|t}$. If the regularized consumer satisfies the local strong-convexity/sensitivity condition and $\nu_{k,t}^{\mathrm{eff}} > 0$, this improvement yields a strictly smaller certified upper bound on the post-update coordinate error.*

6 EXPERIMENTS

Setup. We evaluate FedLWM on MELD (Poria et al., 2019), CrisisMMD (Alam et al., 2018), CREMA-D (Cao et al., 2014), and CooperScene (Wu et al., 2026). We follow the FedMultimodal protocol (Feng et al., 2023) for multimodal recognition and the original CooperScene architecture for cooperative driving. We construct federations with heterogeneous client data, modality access, and encoder backbones, and impose temporal stress through fixed service traces governing modality availability, latent evolution, and modality-conditioned arrival delay. We compare centralized, conventional FL, asynchronous FL, and multimodal FL baselines under identical client partitions, service traces, participation schedules, and local optimization budgets.

Overall performance. Table 1 demonstrates broad effectiveness across contemporaneous and temporally misaligned federation. Even under S0, FedLWM leads global evaluation on both datasets

Table 1: Quantitative comparison. **S0** denotes the contemporaneous setting with cross-client heterogeneity, whereas **S1** denotes the medium-severity target regime with dynamic modality availability, latent drift, and modality-conditioned staleness. Results are averaged over five matched seeds. E-P and E-G1 evaluate personalized and global models, respectively; E-G1′ applies 50% test-time modality dropout, and E-G2 evaluates an unseen heterogeneous client with eight support samples. Centralized is a pooled-data reference under the same evaluation constraints. † marks parameter-sharing methods with matched homogeneous encoders.

	MELD – Weighted-F1 (%)						CrisisMMD – Macro-F1 (%)					
	S0		S1		Missing	Unseen	S0		S1		Missing	Unseen
Method	**E-P**	**E-G1**	**E-P**	**E-G1**	**E-G1′**	**E-G2**	**E-P**	**E-G1**	**E-P**	**E-G1**	**E-G1′**	**E-G2**
Centralized	48.32	53.44	44.61	50.39	52.54	41.37	31.04	28.14	30.20	22.20	22.60	14.20
FedAvg† (McMahan et al., 2017)	34.74	49.14	32.37	49.01	47.89	29.24	23.53	22.81	20.96	18.55	17.95	10.01
FedProx† (Li et al., 2020)	35.30	49.70	32.90	49.45	48.35	29.60	24.99	22.90	20.75	18.43	18.20	10.65
FedAsync† (Xie et al., 2019)	38.52	50.24	34.90	50.39	50.01	**31.29**	26.17	24.35	25.34	19.50	20.19	11.78
AdaMasFL† (Yan et al., 2025)	33.74	44.44	34.77	49.15	48.95	29.24	24.12	23.13	21.81	19.76	21.97	11.88
HAMFL (Qi & Li, 2024)	41.40	47.96	32.71	49.77	47.62	27.80	28.54	25.02	29.66	20.94	20.28	12.47
FedCola (Sun et al., 2024)	48.00	55.39	33.68	49.66	50.12	29.24	28.63	25.14	29.46	20.12	18.85	11.74
FedAFD (Tan et al., 2026)	**49.98**	52.40	31.70	49.12	49.47	28.86	26.21	23.31	21.01	20.05	19.93	10.70
FedMSplit (Chen & Zhang, 2022)	45.63	55.29	32.54	49.92	51.90	27.09	28.50	25.25	30.63	20.21	19.85	12.50
PEPSY (Nguyen et al., 2026)	47.21	56.22	33.61	50.12	52.21	28.86	29.56	26.05	31.08	19.10	19.39	10.87
FedTaste (Liang et al., 2026)	45.80	56.43	33.89	50.05	52.30	28.90	28.64	24.87	30.19	20.03	19.44	12.79
FedLWM	48.27	**58.62**	**35.97**	**54.32**	**56.05**	30.99	**30.51**	**27.74**	**32.97**	**22.87**	**27.30**	**13.98**

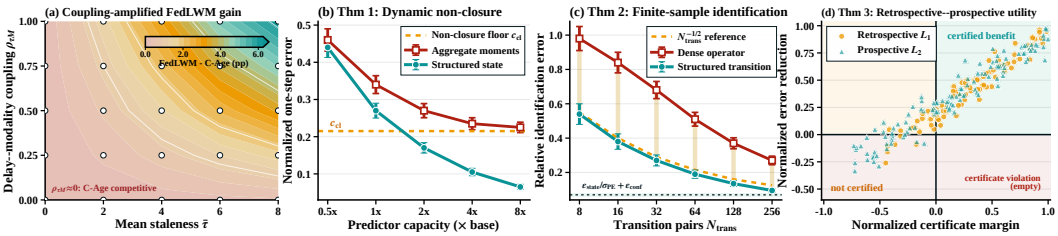

Figure 4: Empirical signatures predicted by FedLWM. **(a)** The margin over scalar age weighting grows with staleness and delay–modality coupling, and vanishes with the coupling. **(b)** Aggregate-moment states saturate at the non-closure floor under increasing capacity, while the structured state keeps improving (Thm. 1). **(c)** The structured transition attains the $N_{\text{trans}}^{-1/2}$ rate down to its estimation floor, whereas a dense operator attains neither (Thm. 2). **(d)** Error reduction occurs whenever the certificate margin is positive, leaving the lower-right quadrant empty (Thm. 3). Panel (a) uses MELD; (b–d) use transition dynamics calibrated to CREMA-D.

and remains among the strongest personalized methods, showing that its structured shared state remains useful beyond temporal asynchrony. Its distinctive advantage becomes more pronounced as temporal–modality misalignment intensifies: on MELD, the E-G1 margin over the strongest baseline grows by about 80% from S0 to S1, while on CrisisMMD the missing-modality margin exceeds three times its S0 E-G1 margin. This amplification directly supports the problem formulation—semantic-time state inference becomes increasingly valuable when partial observations refer to different underlying times—and the controlled sweep in Fig. 4(a) independently confirms the trend with increasing staleness and delay–modality coupling. Across the more demanding evaluations, gains appear jointly in E-P and E-G1, persist under severe test-time modality dropout, while unseen-client performance remains competitive, indicating stronger shared prediction while preserving personalized utility. Finally, Fig. 5 extends the evidence beyond multimodal recognition to cooperative 3D detection, where FedLWM recovers objects missed by FedAvg on identical frames.

Ablation analysis. Table 2 validates the three structural requirements behind FedLWM rather than interchangeable architectural capacity. Removing the fixed HSRF (**A**) preserves class–fiducial alignment but increases residual-gauge pooling drift by roughly 2–3×, showing that semantic alignment alone does not define a coordinate domain for cross-client likelihood pooling (Prop. 1). Replacing the structured state with aggregate moments (**B**) exposes the second limitation: although *B-Mom-XL* improves one-step prediction with 4× predictor capacity, its four-step error remains 0.18 above FedLWM, indicating that discarded state structure creates a dynamic non-closure floor that cannot be recovered by a larger predictor (Thm. 1). Both A and B ablations also reduce E-P and E-G1′, linking these structural deficiencies to downstream task performance. These ablations separate coordinate consistency from dynamic state sufficiency, showing that neither substitutes for the other.

Table 2: Ablation study on MELD under medium **S1** stress. ϵ_{frame} measures class–fiducial alignment; ϵ_{pool} measures residual-gauge pooling drift; NLL_1, NLL_4, and NLL_{use} are one-step, four-step, and use-time prediction NLLs, respectively. σ_{use} is the across-seed standard deviation of NLL_{use}. Δ_{L1} and Δ_{L2} are paired task gains (pp) over identity L1 transport and current-state broadcasting.

	Task		**A: Reference Frame**		**B: World State**		**C: Inference–Prediction**			
Variant	**E-P**↑	**E-G1′**↑	$\epsilon_{\text{frame}}\downarrow$	$\epsilon_{\text{pool}}\downarrow$	$\text{NLL}_1\downarrow$	$\text{NLL}_4\downarrow$	$\text{NLL}_{\text{use}}\downarrow$	$\sigma_{\text{use}}\downarrow$	$\Delta_{\text{L1}}\uparrow$	$\Delta_{\text{L2}}\uparrow$
A-EMA: w/o HSRF (EMA)	33.82	50.84	**0.390**	0.250	1.150	1.430	1.590	0.061	0.82	0.54
A-VIC: w/o HSRF (VICReg)	34.41	52.18	0.440	0.180	1.100	1.370	1.540	0.055	1.08	0.72
B-Mom: w/o structured state	34.18	50.96	0.421	0.092	1.130	1.450	1.580	0.058	0.68	0.61
B-Mom-XL: 4× predictor capacity	34.57	52.04	0.420	0.091	**0.990**	1.360	1.520	0.052	0.94	0.76
C-Dense: w/o structured transition	34.76	52.63	0.422	0.092	1.000	1.250	1.500	0.104	1.12	0.79
C-Age: Scalar age weighting	34.48	54.18	0.419	0.091	1.060	1.320	1.480	0.043	0.56	0.43
C-L1^0: Identity L1 transport	34.10	53.95	0.421	0.091	1.080	1.350	1.510	0.047	0.00	0.91
C-L2: Current-state prior	34.62	54.70	0.419	0.091	1.020	1.210	1.460	0.036	1.63	0.00
FedLWM	**35.97**	**56.05**	0.420	**0.090**	1.010	**1.180**	**1.400**	**0.029**	**2.10**	**1.35**



Figure 5: Qualitative results on CooperScene. FedAvg (top, blue) and FedLWM (bottom, red) are compared on identical cooperative 3D-detection frames; green dashed boxes denote ground truth, yellow insets highlight FedAvg misses recovered by FedLWM, and the right insets show the corresponding BEV detections.

The dynamics ablations (**C**) further verify both the learnability and utility of structure-preserving temporal transport. *C-Dense* remains competitive at short horizons but exhibits more than 3.5× higher across-seed use-time standard deviation, consistent with the finite-sample identification advantage of structured transitions (Thm. 2). Scalar age weighting (*C-Age*) recovers only about one quarter of the retrospective gain, showing that modality-conditioned temporal transport cannot be reduced to scalar staleness correction. Finally, removing either temporal direction eliminates its corresponding benefit: identity L1 transport removes Δ_{L1}, whereas current-state broadcasting removes Δ_{L2}, consistent with the sufficient conditions of Thm. 3. Together, these results support the full inference–prediction loop: **A** establishes a common coordinate domain for pooling, **B** maintains a dynamically sufficient state, and **C** enables learnable temporal transport with both retrospective and prospective utility.

Appendix E details the experimental protocol and additional results, while Appendix C.3 examines communication, computation, and privacy.

7 Conclusion

FedLWM reframes multimodal federated learning under temporal misalignment as inference over an evolving latent world, moving beyond the assumption that round-indexed updates alone provide a consistent basis for federation. By establishing a shared semantic space and a structured temporal state, FedLWM enables delayed observations to be assimilated according to the times they describe and transforms the inferred state into predictive priors for future client use without additional observations. Together, the theoretical analysis and empirical evaluation demonstrate that this latent-state perspective provides a principled solution to representational gauge indeterminacy, predictive state insufficiency, and modality-conditioned temporal misalignment. The present work develops latent imagination for use-time prediction, while action-conditioned planning provides a natural extension of the structured world transition.

AI USE STATEMENT

We used generative AI tools for language refinement, proofreading, and improving the clarity and readability of the manuscript. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions to facilitate reproducibility. The main paper specifies the problem formulation, model components, and theoretical assumptions, while proofs and technical details are provided in the appendix. Experimental protocols, heterogeneity construction, baseline configurations, hyperparameters, and additional analyses are documented in the appendix. An anonymous implementation is provided with the submission to support verification and future research.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Jiayi Chen and Aidong Zhang. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 87–96, 2022.

Siyi Chen, Yixuan Jia, Qing Qu, He Sun, and Jeffrey Fessler. Flowdas: A stochastic interpolant-based framework for data assimilation. *Advances in Neural Information Processing Systems*, 38: 103061–103102, 2026.

Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Geneviève Robin. Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34:29553–29566, 2021.

Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4035–4045, 2023.

Lele Fu, Sheng Huang, Yanyi Lai, Tianchi Liao, Chuanfu Zhang, and Chuan Chen. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16648–16656, 2025.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*, 2020. URL https://mlanthology.org/iclr/2020/hafner2020iclr-dream/.

Su Hyeong Lee, Sidharth Sharma, Manzil Zaheer, and Tian Li. Efficient adaptive federated optimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 171750–171809. Curran Associates, Inc., 2025. doi: 10.52202/085713-5717. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fab71c08b15508fd6c435e59b0e82b68-Paper-Conference.pdf.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Yunbo Li, Jiaping Gui, Zhihang Deng, Fanchao Meng, and Yue Wu. Fedqs: Optimizing gradient and model aggregation for semi-asynchronous federated learning. *Advances in Neural Information Processing Systems*, 38:86502–86550, 2026.

Haochen Liang, Jie Zhang, and Hideya Ochiai. Fedtaste: Topology-aware structural transfer for multimodal federated learning with missing modalities. *arXiv preprint arXiv:2607.23245*, 2026.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Duong Nguyen, Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. *Advances in Neural Information Processing Systems*, 38:28808–28845, 2026.

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

Fan Qi and Shuai Li. Adaptive hyper-graph aggregation for modality-agnostic federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12312–12321, 2024.

Guangyu Sun, Matias Mendieta, Aritra Dutta, Xin Li, and Chen Chen. Towards multi-modal transformers in federated learning. In *European Conference on Computer Vision*, pp. 229–246. Springer, 2024.

Min Tan, Junchao Ma, Yinfu Feng, Jiajun Ding, Wenwen Pan, Tingting Han, Qian Zheng, Zhenzhong Kuang, and Zhou Yu. Fedafd: Multimodal federated learning via adversarial fusion and distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3400–3409, June 2026.

Yujia Wang, Shiqiang Wang, Songtao Lu, and Jinghui Chen. Fadas: towards federated adaptive asynchronous optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Bo Wu, Ruoshen Mo, Justin Yue, Yanyu Zhang, Janice Nguyen, Guoyuan Wu, Amit Roy-Chowdhury, Matthew J. Barth, and Hang Qiu. Cooperscene: Multi-modal cooperative autonomy benchmark with c-v2x communication characterization. In *European Conference on Computer Vision*, 2026.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

Wenjing Yan, Xiangyu Zhong, Xiaolu Wang, and Ying-Jun Angela Zhang. Momentum-driven adaptivity: Towards tuning-free asynchronous federated learning. In *Forty-second International Conference on Machine Learning*, 2025.

Morui Zhu, Mohammad Dehghani Tezerjani, Mátyás Szántó, Márton Vaitkus, Song Fu, and Qing Yang. F3dgs: Federated 3d gaussian splatting for decentralized multi-agent world modeling. *arXiv preprint arXiv:2604.01605*, 2026a.

Yiyao Zhu, Ying Xue, Haiming Zhang, Guangfeng Jiang, Wending Zhou, Xu Yan, Jiantao Gao, Yingjie Cai, Bingbing Liu, Zhen Li, and Shaojie Shen. Dlwm: Dual latent world models enable holistic gaussian-centric pre-training in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39713–39723, June 2026b.

A NOTATION AND ASSUMPTIONS

All continuous-state statements are regime-local, within a fixed-topology regime $\mathcal{I}_p$ where the primitive set, chart assignment $i \mapsto y_i$, and state space $\mathcal{Z}_p$ are fixed; constants may depend on p. We equip $\mathcal{Z}_p$ with the Euclidean–Frobenius product geometry (equivalently, Euclidean after representing symmetric blocks by svec) and use $\|\cdot\|_2$ for the induced state norm, operator norm, or Euclidean parameter norm as appropriate. Unless stated otherwise, expectations of local quantities are conditional on the server history $\mathcal{H}^{\mathrm{srv}}_{k,s}$ under which they were formed, and hats denote estimates.

A.1 NOTATION

Table 3: Core notation used across the main text and appendices.

Symbol	Meaning	Symbol	Meaning
Indices, sets, and semantic time			
$[K], k$	Client set and client index	$\mathcal{M}, m, \mathcal{A}$	Modality set, modality index, and observed mask $\emptyset \neq \mathcal{A} \subseteq \mathcal{M}$
$\mathcal{M}_{k,s}$	Modalities observed by client k at semantic time s	$\mathcal{Y}, y$	Fixed label space; class/chart index
$\mathcal{I}_p, p$	Fixed-topology regime and regime index	$[T], t, s$	Shared logical time set; server round; semantic time
$\tau_{k,t}, \Delta_{k,t}$	Upload staleness and broadcast-prior consumption horizon	$s^{\mathrm{obs}}_{k,t}, s^{\mathrm{use}}_{k,t}$	$t - \tau_{k,t}$ and $t + \Delta_{k,t}$
N_p, N_y	Numbers of primitives in regime p and chart y	L_{lag}, X_t	Fixed-lag length and trajectory $(z_{t-L_{\mathrm{lag}}}, \dots, z_t)$
Representation and semantic reference frame			
$d, u^m_{k,s,j}$	Representation dimension and normalized modality representation	$\mathcal{C}_{\mathrm{sem}}, \mathcal{C}_{\mathrm{cal}}$	Semantic and calibration fiducials
$\mathcal{C}, \mathcal{V}_{\mathcal{C}}, r$	$\mathcal{C}_{\mathrm{sem}} \cup \mathcal{C}_{\mathrm{cal}}$; $\mathrm{span}(\mathcal{C})$; its dimension	$\mathrm{Stab}(\mathcal{C})$	Residual orthogonal gauges fixing $\mathcal{C}$ pointwise
B_y, d_v	Shared tangent basis for chart y and canonical chart dimension	Log_{c_y}, d_g	Spherical logarithm at c_y and geodesic distance
$v^m_{k,s,j}, v$	Canonical modality coordinate $B_y^\top \mathrm{Log}_{c_y}(u^m_{k,s,j})$ and latent chart coordinate	w	Client-model parameters
Latent world state and emission model			
$z_s, \mathcal{Z}_p$	Latent world state at semantic time s and regime-p state space	D_p, D_y	State dimension and chart-block dimension
$\mu_{i,s}, M_{i,s}, V_{i,s}$	Primitive mean, raw second moment, and $V_{i,s} = M_{i,s} - \mu_{i,s}\mu_{i,s}^\top$	$\alpha_{i\|y,s}, \delta^m_{i,s}$	Within-chart mixture weight and modality offset
$\Psi^{\mathcal{A}}_i, \Sigma^{\mathcal{A}}_i$	Joint emission-noise covariance and induced observed covariance on mask $\mathcal{A}$	$\mathcal{K}_p, \Pi_{\mathcal{K}_p}$	Admissible state set and metric projection
$\sigma_{\min}, \theta_\mu, \theta_\delta$	Dispersion floor and mean/offset radii	$M_{\max}, \alpha_{\min}, w_m$	Moment bound, mixture-weight floor, and offset-centering weights
$\Pi_{\mathrm{pt}}, \Pi_{\mathrm{mom}}$	Chart-aggregate mean summary and mean–raw-second-moment summary		
Dynamics			
F, ζ_s, σ_ζ	Nominal transition, process disturbance $z_{s+1} - F(z_s)$, and sub-Gaussian scale	$F_\phi, \overline{F}_\phi, \Phi_{\mathrm{str}}$	Learned projected transition, its unprojected map, and structured family
$A^\mu_y, A^\delta_y, A^V_y$	Chart-wise location, offset, and dispersion operators	$b_{y,i}, b^{\delta,m}_y, Q_{y,i}$	Primitive drift, centred modality intercept, and covariance injection
$S_y, \widetilde{S}_y$	Mixture-weight transition and its column-stochastic core	ϵ_F, L_ϕ	One-step transition error and Lipschitz constant of F_ϕ
E_h	Rollout-error budget $\epsilon_F \sum_{n=0}^{h-1} L^n_\phi$	$\Gamma^\star_\Delta$	$\mathrm{Lip}(F^\Delta; \mathcal{K}_p)$
Likelihood evidence and server inference			
$\mathcal{D}_{k,s}, \ell_{k,s}$	Local multimodal data and its log-likelihood	$\mathcal{J}_{k,s}$	Conditional likelihood curvature
$\mathsf{S}^{\mathrm{state}}_{k,s}$	Uploaded sufficient statistics for latent-state inference	$g_{k,s}, \widetilde{J}_{k,s}$	Score at the recorded prior and calibrated PSD curvature
$\widehat{z}_{s\|t}, P_{s\|t}$	Mean and covariance of the belief about z_s given evidence processed through round t	$\bar{z}_{t\|t}$	$\Pi_{\mathcal{K}_p}(\widehat{z}_{t\|t})$
$\widehat{z}^-_{k,s}$	Server prior recorded for client k when evidence at semantic time s was formed	$\widehat{X}^\pm_t, P^\pm_t$	Fixed-lag trajectory mean and covariance before/after round-t assimilation
$\Lambda^\pm_t, \eta^\pm_t$	$(P^\pm_t)^{-1}$ and $\Lambda^\pm_t \widehat{X}^\pm_t$	$\Delta\Lambda_{k,s}, \Delta\eta_{k,s}$	Likelihood-information increments placed at semantic slice s
$\mathrm{Eph}_{k,t}$	$F_\phi^{\Delta_{k,t}}(\bar{z}_{t\|t})$	$\nu_0, \nu^{\mathrm{eff}}_{k,t}, \chi$	Base/effective ephemeris strength and uncertainty attenuation
$x_k(w), \Pi^{\mathrm{loc}}_k$	Client-side state readout induced by w and shared-to-local map	$x^\star_k, D_\Delta, \epsilon_{\mathrm{loc}}$	Ideal use-time readout, use-time displacement, and local mismatch in Theorem 3
Transition identification			

continued on next page

Table 3 (continued)

Symbol	Meaning	Symbol	Meaning
$\mathcal{T}_y$, N_{trans}	Chart-y transition-pair set and its size	$\vartheta_y^{\mathrm{lin}}, p_y^{\mathrm{lin}}$	Structured linear-block parameter and its dimension
Ξ_y, o_y, m_y	Design matrix, known offset, and output dimension in Eq. 11	σ_{PE}	Persistent-excitation level
$\epsilon_{\mathrm{state}}$	Uniform state-estimation error	$\epsilon_{\mathrm{conf},y}$	Residual feedback/confounding floor for chart y

A.2 Assumptions

Admissible set and structured transition family. Fix modality weights $w_m > 0$ with $\sum_{m\in\mathcal{M}} w_m = 1$ and constants $\sigma_{\min}, \theta_\mu, \theta_\delta, M_{\max} > 0$ and $\alpha_{\min} \geq 0$ such that $\sigma_{\min}^2 d_v < M_{\max}$ and $N_y \alpha_{\min} \leq 1$ for every y. The admissible set is

$$\mathcal{K}_p \coloneqq \big\{ z \in \mathcal{Z}_p : \text{ (K1)–(K5) hold for every } i,\ m \text{ and } y \big\}, \tag{9}$$

$$\begin{aligned}
&\text{(K1)} \quad M_i - \mu_i \mu_i^\top \succeq \sigma_{\min}^2 I, && \text{(K2)} \quad \|\mu_i\|_2 \leq \theta_\mu,\ \ \|\delta_i^m\|_2 \leq \theta_\delta,\\
&\text{(K3)} \quad \operatorname{tr} M_i \leq M_{\max}, && \text{(K4)} \quad \textstyle\sum_{m\in\mathcal{M}} w_m \delta_i^m = 0,\\
&\text{(K5)} \quad \alpha_{i|y} \geq \alpha_{\min},\ \ \textstyle\sum_{i:\, y_i = y} \alpha_{i|y} = 1.
\end{aligned}$$

Constraint (K1) is equivalent to the linear matrix inequality $\left(\begin{smallmatrix} M_i - \sigma_{\min}^2 I & \mu_i \\ \mu_i^\top & 1 \end{smallmatrix}\right) \succeq 0$ in (μ_i, M_i), and (K2)–(K5) are norm-ball, linear and affine constraints; hence $\mathcal{K}_p$ is nonempty, closed, bounded and convex, and $\Pi_{\mathcal{K}_p}$ is single-valued and non-expansive. The structured family Φ_{str} consists of the maps $F_\phi = \Pi_{\mathcal{K}_p} \circ \overline{F}_\phi$, where $\overline{F}_\phi$ acts by Eq. 5, returns $M_{i,s+1} = V_{i,s+1} + \mu_{i,s+1}\mu_{i,s+1}^\top$, and has parameters satisfying

$$\sum_{m\in\mathcal{M}} w_m b_y^{\delta,m} = 0, \qquad Q_{y,i} \succeq 0, \qquad S_y = (1 - N_y\alpha_{\min})\,\widetilde{S}_y + \alpha_{\min}\mathbf{1}\mathbf{1}^\top, \quad \widetilde{S}_y \text{ column stochastic.} \tag{10}$$

Let x denote the chart-y block of a state. The location, offset and weight components of $\overline{F}_\phi$ in chart y are affine in the linear-block parameter $\vartheta_y^{\mathrm{lin}}$, which collects A_y^μ, A_y^δ, $\{b_{y,i}\}_i$, $\{b_y^{\delta,m}\}_m$ modulo the centring constraint and $\widetilde{S}_y$ modulo its column sums:

$$\big[\overline{F}_\phi(x)\big]_{\mu,\delta,\alpha} = \Xi_y(x)\,\vartheta_y^{\mathrm{lin}} + o_y(x), \qquad p_y^{\mathrm{lin}} = 2d_v^2 + N_y d_v + (|\mathcal{M}| - 1)d_v + N_y(N_y - 1), \tag{11}$$

where $\Xi_y(x) \in \mathbb{R}^{m_y \times p_y^{\mathrm{lin}}}$ and $o_y(x)$ are known and affine in x, and $m_y = N_y\big((1 + |\mathcal{M}|)d_v + 1\big)$. Because $\mathcal{K}_p$ is bounded, $B_\Xi \coloneqq \sup_x \|\Xi_y(x)\|_2$ and the Lipschitz modulus L_Ξ of Ξ_y on $\mathcal{K}_p$ are finite. Properties that follow from these definitions are not assumed below: for every $F_\phi \in \Phi_{\mathrm{str}}$, $\overline{F}_\phi$ preserves (K1), (K4) and (K5), and L_ϕ is finite (Lemma 5). The local strong convexity of the client coordinate subproblem is stated as a hypothesis in Theorem 3(ii).

Setting and observation.

Assumption 1 (Observer-Independent Latent Process)**.** *There exists a map $F : \mathcal{K}_p \to \mathcal{K}_p$, not depending on the server history, such that $\mathbb{E}[z_{s+1} \mid \mathcal{F}_s] = F(z_s)$ for all $s, s+1 \in \mathcal{I}_p$. Server messages may alter the observation mechanism, through the client parameters and hence the representations $u_{k,s,j}^m$, but do not enter the drift of $\{z_s\}$.*

Assumption 2 (Disturbance Regularity)**.** *For $s, s+1 \in \mathcal{I}_p$, the disturbance $\zeta_s = z_{s+1} - F(z_s)$ is $\mathcal{F}_{s+1}$-measurable, $\mathbb{E}[\zeta_s \mid \mathcal{F}_s] = 0$, $\|\zeta_s\|_2 \leq b_\zeta$ almost surely, and $\mathbb{E}\big[\exp\langle v, \zeta_s\rangle \mid \mathcal{F}_s\big] \leq \exp\big(\sigma_\zeta^2 \|v\|_2^2/2\big)$ for every $v \in \mathcal{Z}_p$. Moreover, $z_s \in \mathcal{K}_p$ almost surely for every $s \in \mathcal{I}_p$.*

Representation and reference frame.

Assumption 3 (Isometry-Only Reparameterization)**.** *For every client, the task loss depends on the normalized representations only through inner products, among themselves and with task-head parameters that transform covariantly with them. Consequently, the admissible reparameterizations of the normalized representation are the isometries of $\mathbb{S}^{d-1}$, and the representational gauge group is $O(d)$.*

Assumption 4 (Shared Configuration and Shared Semantics). *(i) For every (k, s, j, m), there exist a representation $\widetilde{u}^m_{k,s,j} \in \mathbb{S}^{d-1}$ that depends on the underlying sample and modality but not on the client, and a gauge $R_{k,s} \in \mathrm{Stab}(\mathcal{C})$ such that $\|u^m_{k,s,j} - R_{k,s}\widetilde{u}^m_{k,s,j}\|_2 \le \epsilon_{\mathrm{enc},k}$ almost surely.*

(ii) The label space $\mathcal{Y}$ is shared, and each $y \in \mathcal{Y}$ denotes the same semantic category at every client.

Assumption 5 (Chart Geometry and Membership). *(i) $\theta_0 \le \theta_{\max} < \pi/2$, and $d_g(c_y, c_{y'}) > 2\theta_{\max}$ for all distinct $y, y' \in \mathcal{Y}$.*

(ii) Whenever the latent state lies in $\mathcal{K}_p$, a sample of class y satisfies $d_g(u^m, c_y) \le \theta_{\max}$ for every observed modality m with probability at least $1 - \delta_{\mathrm{chart}}$.

Local likelihood and pooling.

Assumption 6 (Local Likelihood Regularity). *Let $s = s^{\mathrm{obs}}_{k,t}$. Conditionally on $(z_s, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s})$:*

(i) $\mathcal{D}_{k,s}$ follows the mixture–emission model of Sec. 4.2 at z_s, and $\ell_{k,s}$ is its log-likelihood;

(ii) $\mathcal{D}_{k,s}$ is independent of $(z_{s'})_{s'>s}$;

(iii) on a convex neighbourhood containing z_s and z_t, $\ell_{k,s}$ is twice continuously differentiable, its Hessian is Lipschitz with a modulus of finite conditional expectation, and differentiation under the integral sign is permitted, so that $\mathbb{E}[\nabla_z \ell_{k,s}(z_s) \mid z_s, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s}] = 0$.

Assumption 7 (Conditional Independence). *For every finite set $\mathcal{S}$ of pairs (k, s),*

$$p\big(\{\mathcal{D}_{k,s}\}_{(k,s)\in\mathcal{S}} \,\big|\, z_{1:T}, \{\mathcal{H}^{\mathrm{srv}}_{k,s}\}_{(k,s)\in\mathcal{S}}\big) = \prod_{(k,s)\in\mathcal{S}} p\big(\mathcal{D}_{k,s} \,\big|\, z_s, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s}\big).$$

Dynamics and identification.

Assumption 8 (Lipschitz Nominal Dynamics). *F is Lipschitz on $\mathcal{K}_p$; consequently $\Gamma^\star_\Delta \le \mathrm{Lip}(F; \mathcal{K}_p)^\Delta < \infty$ for every $\Delta \ge 1$.*

Assumption 9 (Structural Realizability). *For every chart y, there exist parameters $(\vartheta^{\mathrm{lin}}_y, A^V_y, \{Q_{y,i}\}_i)$ satisfying Eq. 10, with A^V_y invertible, whose unprojected map $\overline{F}^\circ$ satisfies, for every $j \in \mathcal{T}_y$: (i) the chart-y block of $F(z_{s_j})$ equals $\overline{F}^\circ(x_j)$; and (ii) $\overline{F}^\circ(x_j)$ satisfies (K2) and (K3).*

Assumption 10 (Persistent Excitation). *$\lambda_{\min}\big(\sum_{j\in\mathcal{T}_y} \Xi_y(x_j)^\top \Xi_y(x_j)\big) \ge N_{\mathrm{trans}}\, \sigma^2_{\mathrm{PE}}$ with $\sigma_{\mathrm{PE}} > 0$.*

Assumption 11 (Trivial Congruence Stabilizer). *Let $\mathcal{W}_y := \{W_{i,s_j} - W_{i,s_{j'}} : j, j' \in \mathcal{T}_y,\ y_i = y\}$. Then $\mathcal{G}_y := \big\{U \in GL(d_v) : U\,\Delta W\,U^\top = \Delta W \ \ \forall\, \Delta W \in \mathcal{W}_y\big\} = \{\pm I\}$.*

Assumption 12 (State-Estimation Error and Residual Confounding). *For every $j \in \mathcal{T}_y$:*

(i) $\widehat{x}_j$ is $\mathcal{H}^{\mathrm{srv}}_{s_j}$-measurable and $\|\widehat{x}_j - x_j\|_2 \le \epsilon_{\mathrm{state}}$;

(ii) $\widehat{x}^+_j - x^+_j = \varepsilon^+_j + b^{\mathrm{fb}}_j$ with $\|\varepsilon^+_j\|_2 \le \epsilon_{\mathrm{state}}$, where b^{fb}_j is the shift induced by the forecasts in force at time $s_j + 1$;

(iii) with $\widehat{\Xi}_j := \Xi_y(\widehat{x}_j)$, $\big\|\big(\sum_{j'\in\mathcal{T}_y} \widehat{\Xi}^\top_{j'}\widehat{\Xi}_{j'}\big)^{-1} \sum_{j'\in\mathcal{T}_y} \widehat{\Xi}^\top_{j'} \big[b^{\mathrm{fb}}_{j'}\big]_{\mu,\delta,\alpha}\big\|_2 \le \epsilon_{\mathrm{conf},y}$;

(iv) $2B_\Xi L_\Xi\, \epsilon_{\mathrm{state}} \le \sigma^2_{\mathrm{PE}}/2$.

B Formal Analysis of the Core Failure and the Structural Obstacles

B.1 The Core Failure: Modality-Conditioned Temporal Misalignment

Lemma 1 (Temporal Score Bias). *Let the upload of client k processed at round t have semantic time $s = s^{\mathrm{obs}}_{k,t}$, let $\Delta z_{s,t} \coloneqq z_t - z_s$, and suppose Assumption 6 holds. Let $\bar{L}_H$ be the conditional expectation of the Lipschitz modulus of $\nabla^2_z \ell_{k,s}$ given $(z_s, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s})$.*

(i) The conditional score at the present state satisfies

$$\mathbb{E}\big[\nabla_z \ell_{k,s}(z_t) \mid z_s, z_t, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s}\big] = -\mathcal{J}_{k,s}\Delta z_{s,t} + \mathsf{e}_{k,s}, \quad \|\mathsf{e}_{k,s}\|_2 \le \tfrac{1}{2}\bar{L}_H \|\Delta z_{s,t}\|_2^2. \tag{12}$$

(ii) Writing $b_{k,s} \coloneqq -\mathcal{J}_{k,s}\Delta z_{s,t}$, $\mathcal{J}_{k,s} \succeq 0$, and the leading bias $b_{k,s}$ satisfies $\langle b_{k,s}, \Delta z_{s,t}\rangle \le 0$. For $\Delta z_{s,t} \neq 0$, the leading bias is collinear with $\Delta z_{s,t}$ if and only if $\Delta z_{s,t}$ is an eigenvector of $\mathcal{J}_{k,s}$.

Proof. (i) The segment from z_s to z_t lies in the convex neighbourhood of Assumption 6(iii), so

$$\nabla_z \ell_{k,s}(z_t) = \nabla_z \ell_{k,s}(z_s) + \nabla^2_z \ell_{k,s}(z_s)\,\Delta z_{s,t} + \mathsf{R}, \qquad \mathsf{R} \coloneqq \int_0^1 \big[\nabla^2_z \ell_{k,s}(z_s + u\Delta z_{s,t}) - \nabla^2_z \ell_{k,s}(z_s)\big]\Delta z_{s,t}\,\mathrm{d}u,$$

with $\|\mathsf{R}\|_2 \le \frac{1}{2}L_H\|\Delta z_{s,t}\|_2^2$, where L_H is the Lipschitz modulus of the Hessian. By Assumption 6(ii), the conditional law of $\mathcal{D}_{k,s}$ given $(z_s, z_t, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s})$ coincides with its law given $(z_s, \mathcal{M}_{k,s}, \mathcal{H}^{\mathrm{srv}}_{k,s})$. Under this conditional expectation, the first term vanishes by Assumption 6(iii), the second equals $b_{k,s}$ by the definition of $\mathcal{J}_{k,s}$, and the third is bounded by $\frac{1}{2}\bar{L}_H\|\Delta z_{s,t}\|_2^2$. (ii) Under the interchange of differentiation and integration in Assumption 6(iii), $\mathcal{J}_{k,s}$ equals the conditional covariance of $\nabla_z \ell_{k,s}(z_s)$ and is therefore positive semidefinite. Hence $\langle b_{k,s}, \Delta z_{s,t}\rangle = -\Delta z_{s,t}^\top \mathcal{J}_{k,s}\Delta z_{s,t} \le 0$, and $b_{k,s} \in \mathrm{range}(\mathcal{J}_{k,s})$. Collinearity of $\mathcal{J}_{k,s}\Delta z_{s,t}$ and $\Delta z_{s,t}$ is the definition of an eigenvector. □

For a mask $\mathcal{A}$ and a primitive i in chart y, the emission model of Sec. 4.2 gives

$$v^{\mathcal{A}} \mid i \;\sim\; \mathcal{N}\big(\bar{v}_i^{\mathcal{A}}, \Sigma_i^{\mathcal{A}}\big), \qquad \bar{v}_i^{\mathcal{A}} = \mathbf{1}_{\mathcal{A}}\mu_i + \delta_i^{\mathcal{A}}, \qquad \Sigma_i^{\mathcal{A}} = \mathbf{1}_{\mathcal{A}} V_i \mathbf{1}_{\mathcal{A}}^\top + \Psi_i^{\mathcal{A}}. \tag{13}$$

Lemma 2 (Support–Mask Factorization). *Let $\mathcal{A} = \mathcal{M}_{k,s}$ and $\Psi_i^{\mathcal{A}} \succ 0$.*

(i) The mean block of the complete-data curvature, with the primitive index completed, factorizes as

$$\mathcal{J}^{\mathrm{cd}}_{k,s}[\mu_i] = n_{k,s,i}\,\Omega_i(\mathcal{A}), \qquad \Omega_i(\mathcal{A}) = \mathbf{1}_{\mathcal{A}}^\top (\Sigma_i^{\mathcal{A}})^{-1}\mathbf{1}_{\mathcal{A}} \succ 0.$$

(ii) $\ell_{k,s}$ does not depend on δ_i^m for $m \notin \mathcal{A}$, nor on the emission blocks $C_i^{m,m'}$ with $\{m, m'\} \not\subseteq \mathcal{A}$. Consequently, the coordinates $\{\delta_i^m\}_{m\notin\mathcal{A}}$ lie in the kernel of $\nabla^2_z \ell_{k,s}$ at every point, and hence in $\ker \mathcal{J}_{k,s}$, $\ker \mathcal{J}^{\mathrm{cd}}_{k,s}$ *and* $\ker \mathcal{J}^{\mathrm{mis}}_{k,s}$.

Proof. (i) With the primitive index completed, the complete-data log-likelihood is

$$\sum_j \sum_i \mathbf{1}\{i_j = i\}\Big[\log \alpha_{i|y_j} + \log \mathcal{N}\big(v_j^{\mathcal{A}}; \bar{v}_i^{\mathcal{A}}, \Sigma_i^{\mathcal{A}}\big)\Big].$$

Since $\bar{v}_i^{\mathcal{A}}$ is affine in μ_i with Jacobian $\mathbf{1}_{\mathcal{A}}$ and $\Sigma_i^{\mathcal{A}}$ does not depend on μ_i, the negative Hessian of each Gaussian term in μ_i equals $\Omega_i(\mathcal{A})$. Taking the posterior expectation over i_j replaces the indicator by $\rho_{j,i}$, and summing over j gives $n_{k,s,i}\Omega_i(\mathcal{A})$. Positive definiteness follows from $\Sigma_i^{\mathcal{A}} \succ 0$ and from $\mathbf{1}_{\mathcal{A}}$ having full column rank d_v for $\mathcal{A} \neq \emptyset$. (ii) The law in Eq. 13 involves only the offsets and emission blocks indexed by $\mathcal{A}$, so $\ell_{k,s}$ is constant in the remaining ones. All partial derivatives of $\ell_{k,s}$ in $\{\delta_i^m\}_{m\notin\mathcal{A}}$ therefore vanish identically, and so do the corresponding rows and columns of $\nabla^2_z \ell_{k,s}$ and of its conditional expectation $-\mathcal{J}_{k,s}$. The same holds for Louis's decomposition $-\nabla^2_z \ell_{k,s} = \mathcal{J}^{\mathrm{cd}}_{k,s} - \mathcal{J}^{\mathrm{mis}}_{k,s}$, whose two terms are formed from complete-data derivatives in the same coordinates. □

Proposition 2 (Scalar Age Correction and Selection). *Let $\omega(\tau) > 0$ be a scalar age weight.*

(i) Scalar attenuation. *For $\widetilde{\mathcal{J}}_{k,s} = \omega(\tau_{k,t})\mathcal{J}_{k,s}$, $\ker \widetilde{\mathcal{J}}_{k,s} = \ker \mathcal{J}_{k,s}$, $-\widetilde{\mathcal{J}}_{k,s}\Delta z_{s,t} = \omega(\tau_{k,t})\, b_{k,s}$. Hence scalar age weighting changes the magnitude but not the direction or information support of the leading temporal bias.*

(ii) Selection by age. *For $E_{\bar{\tau}} \coloneqq \{\tau_{k,t} \le \bar{\tau}\}$, age thresholding changes the mask law from $p(\mathcal{M}_{k,s} = \mathcal{A})$ to $p(\mathcal{M}_{k,s} = \mathcal{A} \mid E_{\bar{\tau}})$. If $\tau_{k,t}$ and $\mathcal{M}_{k,s}$ are dependent, these laws differ for some threshold; by Lemma 2, the resulting selection need not preserve likelihood-information geometry.*

(iii) Paired-modality evidence. *Let $\mathcal{P}_{m,m'} \coloneqq \{(k,s) : \{m,m'\} \subseteq \mathcal{M}_{k,s}\}$, $\mathcal{P} \coloneqq \bigcup_{m \neq m'} \mathcal{P}_{m,m'}$, and let F and $\mathrm{F}_{\mathcal{P}}$ denote the delay distribution functions over all uploads and over $\mathcal{P}$. If $\mathrm{F}_{\mathcal{P}} \le \mathrm{F}$ pointwise, then $\Pr(\mathcal{P} \mid \tau_{k,t} \le \bar{\tau}) \le \Pr(\mathcal{P})$ for every threshold $\bar{\tau}$. If no retained upload belongs to $\mathcal{P}_{m,m'}$, the pooled likelihood is constant in $C_i^{m,m'}$, so this cross-modal block is not identifiable from the retained evidence.*

Proof. Part (i) follows because multiplication by a positive scalar preserves both the kernel and direction of a linear map applied to a fixed displacement. For (ii), thresholding conditions the modality mask on $E_{\bar{\tau}}$; dependence between delay and mask implies a change for some threshold, while Lemma 2 shows that different masks constrain different state directions. For (iii), Bayes' rule gives $\Pr(\mathcal{P} \mid \tau_{k,t} \le \bar{\tau}) = \Pr(\mathcal{P})\mathrm{F}_{\mathcal{P}}(\bar{\tau})/\mathrm{F}(\bar{\tau})$. The final claim follows from Lemma 2(ii) and conditional independence of the retained likelihood factors. □

Proposition 3 (Likelihood Transport). *For $t > s$, exact reinterpretation of evidence generated at semantic time s as evidence for z_t requires*

$$p\big(\mathcal{D}_{k,s} \mid z_t\big) = \int p\big(\mathcal{D}_{k,s} \mid z_s\big)\, p\big(z_s \mid z_t\big)\, \mathrm{d}z_s. \tag{14}$$

The backward kernel $p(z_s \mid z_t)$ depends on the state law and is therefore not determined by the forward transition alone. Moreover, even under deterministic invertible dynamics $G \coloneqq F^{t-s}$, the transported log-likelihood becomes $\ell_{k,s} \circ G^{-1}$, which generally need not preserve a quadratic information form when G^{-1} is nonlinear.

Proof. Eq. 14 follows by marginalizing the latent state z_s under Assumption 6(ii). By Bayes' rule, $p(z_s \mid z_t) \propto p(z_t \mid z_s)\, p(z_s)$, so the backward kernel depends on the marginal law of z_s in addition to the forward transition. If the dynamics are deterministic and invertible, then $z_t = G(z_s)$ implies $z_s = G^{-1}(z_t)$, and transporting the local log-likelihood gives $\ell_{k,s}\big(G^{-1}(z_t)\big)$. For nonlinear G^{-1}, composition generally does not preserve quadratic dependence on z_t; hence the local quadratic information form need not be retained after transport. □

B.2 The Two Structural Obstacles

Representational gauge indeterminacy.

Proposition 4 (Gauge Indeterminacy of Trajectories and Pooled Curvature). *Let $u_{k,s} = R_{k,s}\widetilde{u}_{k,s}$ with $R_{k,s} \in O(d)$, and let $\{Q_s\}_{s\in[T]} \subset O(d)$ be arbitrary.*

(i) The reparameterization $R_{k,s} \mapsto R_{k,s}Q_s$, $\widetilde{u}_{k,s} \mapsto Q_s^\top \widetilde{u}_{k,s}$ leaves every observation unchanged.

(ii) If $\widetilde{u}_{k,s+1} = \widetilde{F}_s(\widetilde{u}_{k,s})$ for all k, the reparameterized representations satisfy the transition $Q_{s+1}^\top \circ \widetilde{F}_s \circ Q_s$.

(iii) Under a time-varying gauge, even the identity transition and any fixed $Q \in O(d)$ admit observationally equivalent representations.

(iv) Pooled curvature is gauge-dependent before alignment: in general, summing client curvatures in different gauges is not congruent to summing them in a common gauge.

Proof. (i) $(R_{k,s}Q_s)(Q_s^\top \widetilde{u}_{k,s}) = R_{k,s}\widetilde{u}_{k,s}$. (ii) $Q_{s+1}^\top \widetilde{u}_{k,s+1} = Q_{s+1}^\top \widetilde{F}_s\big(Q_s\, Q_s^\top \widetilde{u}_{k,s}\big)$. (iii) Take $Q_s = (Q^\top)^s$; then $Q_{s+1}^\top \,\mathrm{id}\, Q_s = Q^{s+1}(Q^\top)^s = Q$. (iv) If a log-likelihood ℓ is expressed in the coordinates $u' = \Upsilon_k u$, then $\ell'(u') = \ell(\Upsilon_k^\top u')$, with Hessian $\Upsilon_k \nabla^2 \ell\, \Upsilon_k^\top$. Take $\mathcal{J}_1 = \mathcal{J}_2 = e_1 e_1^\top$, $\Upsilon_1 = I$ and Υ_2 a rotation with $\Upsilon_2 e_1 = e_2$. The unaligned sum $e_1 e_1^\top + e_2 e_2^\top$ has rank two, whereas $\mathcal{J}_1 + \mathcal{J}_2 = 2e_1 e_1^\top$ has rank one, and congruence by an orthogonal matrix preserves rank. □

Predictive state insufficiency.

Proposition 5 (Non-Closure of the Point Summary)**.** *Let chart y carry $N_y = 2$ primitives and let $\alpha_{\min} < 1/2$. There exist $\delta \neq 0$, $z^{(1)}, z^{(2)} \in \mathcal{K}_p$ and $F \in \Phi_{\mathrm{str}}$ such that $\Pi_{\mathrm{pt}}(z^{(1)}) = \Pi_{\mathrm{pt}}(z^{(2)})$ and $\Pi_{\mathrm{mom}}(z^{(1)}) = \Pi_{\mathrm{mom}}(z^{(2)})$, whereas $\|\Pi_{\mathrm{pt}}(F(z^{(1)})) - \Pi_{\mathrm{pt}}(F(z^{(2)}))\|_2 = (1 - 2\alpha_{\min})\|\delta\|_2$. Consequently, every map G on the point-summary space satisfies*

$$\max_{a\in\{1,2\}} \big\|G(\Pi_{\mathrm{pt}}(z^{(a)})) - \Pi_{\mathrm{pt}}(F(z^{(a)}))\big\|_2 \geq \tfrac{1}{2}(1 - 2\alpha_{\min})\|\delta\|_2.$$

Proof. Choose $\delta \in \mathbb{R}^{d_v} \setminus \{0\}$ and $\sigma_0^2 \geq \sigma_{\min}^2$ with $\|\delta\|_2 \leq \theta_\mu$ and $\sigma_0^2 d_v + \|\delta\|_2^2 \leq M_{\max}$, and set, with $\alpha_{1|y} = \alpha_{2|y} = \frac{1}{2}$ and zero offsets,

$$z^{(1)}:\ \mu_1 = \delta,\ \mu_2 = -\delta,\ V_1 = V_2 = \sigma_0^2 I; \qquad z^{(2)}:\ \mu_1 = \mu_2 = 0,\ V_1 = V_2 = \sigma_0^2 I + \delta\delta^\top.$$

Both satisfy (K1)–(K5) and have chart-aggregate mean 0 and chart-aggregate raw second moment $\sigma_0^2 I + \delta\delta^\top$. Let F have $A_y^\mu = A_y^\delta = A_y^V = I$, zero drifts, intercepts and injections, and $S_y = (1-2\alpha_{\min})\widetilde{S}_y + \alpha_{\min}\mathbf{1}\mathbf{1}^\top$ with the column-stochastic $\widetilde{S}_y = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, so $F \in \Phi_{\mathrm{str}}$. Then $S_y(\frac{1}{2}, \frac{1}{2})^\top = (1-\alpha_{\min}, \alpha_{\min})^\top$, all other components are unchanged, and the unprojected images lie in $\mathcal{K}_p$, so the projection is inactive. The aggregate means of $F(z^{(1)})$ and $F(z^{(2)})$ are $(1 - 2\alpha_{\min})\delta$ and 0. Since G takes a single value on the common summary, the triangle inequality gives the lower bound. □

Finite-sample learnability.

Proposition 6 (Temporal Rank Budget)**.** *(i) Within $\mathcal{I}_p$, the number of distinct transition pairs satisfies $N_{\mathrm{trans}} \leq |\mathcal{I}_p| - 1$, independently of the number of clients.*

(ii) Let a transition class be affine in a parameter $\theta \in \mathbb{R}^p$, with each transition pair supplying m scalar equations of the form $x^+ = \Xi(x)\theta + o(x)$. Persistent excitation therefore requires $N_{\mathrm{trans}} m \geq p$.

(iii) Consequently, a dense affine transition on $\mathcal{Z}_p$ is not identifiable unless $|\mathcal{I}_p| \geq D_p + 2$, and a dense affine transition on a single chart block is not identifiable unless $|\mathcal{I}_p| \geq D_y + 2$, whereas the linear blocks of Φ_{str} require only $|\mathcal{I}_p| \geq \lceil p_y^{\mathrm{lin}}/m_y \rceil + 1$.

Proof. (i) A transition pair is a pair of consecutive semantic times $(s, s+1)$ in $\mathcal{I}_p$; uploads from different clients at the same semantic time refer to the same latent state z_s and therefore to the same pair. (ii) $\mathrm{rank}\big(\sum_j \Xi(x_j)^\top \Xi(x_j)\big) \leq \sum_j \mathrm{rank}\,\Xi(x_j) \leq N_{\mathrm{trans}} m$, and positive definiteness requires rank p. (iii) A dense affine map on $\mathcal{Z}_p$ has $p = D_p^2 + D_p$ and $m = D_p$, so $N_{\mathrm{trans}} \geq D_p + 1$; on a chart block, $p = D_y^2 + D_y$ and $m = D_y$. For Φ_{str}, Eq. 11 gives $p = p_y^{\mathrm{lin}}$ and $m = m_y$. Combining with (i) yields the three bounds. □

C Algorithm and Implementation Details

C.1 Federated Protocol

Initialization. Semantic fiducials and bases are fixed before training; calibration fiducials come from shared public probes under a frozen encoder and enter Eq. 3 as client alignment targets.

Algorithm C.1 FedLWM server round t

Require: (Λ_t^-, η_t^-); F_ϕ; $\{\Psi_i\}$; $\{\widehat{\gamma}_{y,\mathcal{A}}\}$; ν_0; χ; L_{lag}

L1: semantic assimilation

$(\Delta\Lambda_t^{\mathrm{old}}, \Delta\eta_t^{\mathrm{old}}) \leftarrow (0, 0)$

for each $(k, s, \mathcal{A}, \mathsf{S}_{k,s}^{\mathrm{state}})$ processed at t **do**

  assemble $(g_{k,s}, \widetilde{J}_{k,s})$ at $\widehat{z}_{k,s}^-$ ▷ Eqs. 18, 21

  if $s \geq t - L_{\mathrm{lag}}$ **then**

    add $(\Delta\Lambda_{k,s}, \Delta\eta_{k,s})$ at slice s ▷ Eq. 6

  else

    accumulate its boundary increment into $(\Delta\Lambda_t^{\mathrm{old}}, \Delta\eta_t^{\mathrm{old}})$ ▷ Eq. 22

  end if

end for

add $(\Delta\Lambda_t^{\mathrm{old}}, \Delta\eta_t^{\mathrm{old}})$ at $t - L_{\mathrm{lag}}$

solve $\Lambda_t^+ \widehat{X}_t^+ = \eta_t^+$; extract $(\widehat{z}_{t|t}, P_{t|t})$ and set $\bar{z}_{t|t} \leftarrow \Pi_{\mathcal{K}_p}(\widehat{z}_{t|t})$

L2: semantic ephemeris

for each scheduled client k **do**

  $h \leftarrow 0$ on anchor rounds, else $\Delta_{k,t}$; $s \leftarrow s_{k,t}^{\mathrm{use}}$

  propagate $(\bar{z}_{t|t}, P_{t|t})$ to $(\widehat{z}_{k,s}^-, P_{t+h|t})$ ▷ Sec. 4.3 (L2)

  $\nu_{k,t}^{\mathrm{eff}} \leftarrow \nu_0 \chi\big(\Pi_k^{\mathrm{loc}} P_{t+h|t} (\Pi_k^{\mathrm{loc}})^\top\big)$, or 0 on anchor rounds

  broadcast $(\widehat{z}_{k,s}^-, \nu_{k,t}^{\mathrm{eff}})$ and record $\widehat{z}_{k,s}^-$

end for

Estimation and optional parameter channel

if t is an anchor round **then**

  refit ϕ, $\{\Psi_i\}$, and $\{\widehat{\gamma}_{y,\mathcal{A}}\}$ ▷ Eqs. 24, 23, 20

end if

optionally aggregate backbone updates with the chosen FL rule

Schur-marginalize $t - L_{\mathrm{lag}}$ and append the $t+1$ prior, yielding $(\Lambda_{t+1}^-, \eta_{t+1}^-)$

Algorithm C.2 Client k using the broadcast from round t'

Require: w; $(\mathcal{C}, \{B_y\})$; $(\widehat{z}_{k,s}^-, \nu_{k,t'}^{\mathrm{eff}})$ with $s = s_{k,t'}^{\mathrm{use}}$; $\{\Psi_i\}$

update w on $\mathcal{L}_k^{\mathrm{base}}(w) + \frac{1}{2}\nu_{k,t'}^{\mathrm{eff}} \|x_k(w) - \Pi_k^{\mathrm{loc}} \widehat{z}_{k,s}^-\|_2^2$ ▷ Eq. 7

compute $v_j^m \leftarrow B_{y_j}^\top \mathrm{Log}_{c_{y_j}}(u_j^m)$ for labelled j and $m \in \mathcal{A} = \mathcal{M}_{k,s}$ ▷ Sec. 4.1

compute $\rho_{j,i}$, $\widetilde{v}_{j,i}$, and $\widetilde{M}_{j,i}$ under $\widehat{z}_{k,s}^-$ ▷ Eq. 15

form and upload $\mathsf{S}_{k,s}^{\mathrm{state}}$ with $(s, \mathcal{A})$ ▷ Eq. 16

if t' is an anchor round **then**

  also upload emission-residual statistics $\{\mathsf{R}_{k,s,i}^{m,m'}\}$ and $\widehat{\jmath}_{k,s,y}^{\mathrm{mis}}$ ▷ Eq. 17

end if

optionally upload the backbone update through the parameter channel

C.2 Local Statistics and Server-Side Assembly

Local statistics. Under the recorded prior, whose components carry the superscript $-$, and with $\bar{v}_i^{\mathcal{A}}$, $\Sigma_i^{\mathcal{A}}$ as in Eq. 13, the joint E-step gives

$$\begin{aligned} \rho_{j,i} &\propto \alpha_{i|y_j}^- \, \mathcal{N}\Big(v_j^{\mathcal{A}}; \bar{v}_i^{\mathcal{A},-}, \Sigma_i^{\mathcal{A},-}\Big), \\ \widetilde{v}_{j,i} &= \mu_i^- + V_i^- \mathbf{1}_{\mathcal{A}}^\top (\Sigma_i^{\mathcal{A},-})^{-1} \big(v_j^{\mathcal{A}} - \bar{v}_i^{\mathcal{A},-}\big), \\ \widetilde{M}_{j,i} &= \Sigma_i^{v|\mathcal{A}} + \widetilde{v}_{j,i} \widetilde{v}_{j,i}^\top. \end{aligned} \tag{15}$$

where $\Sigma_i^{v|\mathcal{A}} = V_i^- - V_i^- \mathbf{1}_{\mathcal{A}}^\top (\Sigma_i^{\mathcal{A},-})^{-1} \mathbf{1}_{\mathcal{A}} V_i^-$. The state stream is

$$n_{k,s,i} = \sum_j \rho_{j,i}, \qquad \mathsf{m}_{k,s,i} = \sum_j \rho_{j,i} \widetilde{v}_{j,i}, \qquad \mathsf{M}_{k,s,i} = \sum_j \rho_{j,i} \widetilde{M}_{j,i}, \qquad \mathsf{r}^m_{k,s,i} = \sum_j \rho_{j,i} \big(v_j^m - \widetilde{v}_{j,i}\big), \tag{16}$$

On anchor rounds, the emission stream additionally carries

$$\mathsf{R}^{m,m'}_{k,s,i} = \sum_j \rho_{j,i} \big(v_j^m - \widetilde{v}_{j,i} - \delta_i^{m,-}\big)\big(v_j^{m'} - \widetilde{v}_{j,i} - \delta_i^{m',-}\big)^\top, \tag{17}$$

for $m, m' \in \mathcal{A}$.

Score. By the Fisher identity with (i, v) completed, the score is affine in $\mathsf{S}^{\mathrm{state}}_{k,s}$; in the stored (μ, M) coordinates,

$$\begin{aligned}
&G_i = \tfrac{1}{2}(V_i^-)^{-1}\Big[\mathsf{M}_{k,s,i} - \mathsf{m}_{k,s,i}(\mu_i^-)^\top - \mu_i^- \mathsf{m}_{k,s,i}^\top + n_{k,s,i}\big(\mu_i^-(\mu_i^-)^\top - V_i^-\big)\Big](V_i^-)^{-1}, \\
&g_{k,s}[M_i] = G_i, \qquad g_{k,s}[\mu_i] = (V_i^-)^{-1}(\mathsf{m}_{k,s,i} - n_{k,s,i}\mu_i^-) - 2G_i\mu_i^-, \\
&g_{k,s}[\delta_i^{\mathcal{A}}] = (\Psi_i^{\mathcal{A}})^{-1}(\mathsf{r}^{\mathcal{A}}_{k,s,i} - n_{k,s,i}\delta_i^{\mathcal{A},-}), \qquad g_{k,s}[\alpha_{i|y}] = \frac{n_{k,s,i}}{\alpha^-_{i|y}}.
\end{aligned} \tag{18}$$

where $-2G_i\mu_i^-$ is the chain rule for $V = M - \mu\mu^\top$.

Curvature and calibration. Using the mask-dependent factor $\Omega_i(\mathcal{A})$ from Lemma 2, the complete-data Fisher curvature is positive semidefinite on every block. For a perturbation $(\mathrm{d}\mu_i, \mathrm{d}V_i, \mathrm{d}\delta_i^{\mathcal{A}}, \mathrm{d}\alpha_{i|y})$,

$$\mathrm{d}^\top \mathcal{J}^{\mathrm{cd}}_{k,s} \mathrm{d} = \sum_i n_{k,s,i}\Big[\big\|(\Sigma_i^{\mathcal{A}})^{-1/2}(\mathbf{1}_{\mathcal{A}}\mathrm{d}\mu_i + \mathrm{d}\delta_i^{\mathcal{A}})\big\|_2^2 + \tfrac{1}{2}\operatorname{tr}\big(\Omega_i(\mathcal{A})\,\mathrm{d}V_i\,\Omega_i(\mathcal{A})\,\mathrm{d}V_i\big) + \big(\mathrm{d}\alpha_{i|y}/\alpha_{i|y}\big)^2\Big]. \tag{19}$$

The curvature is evaluated at $\widehat{z}^-_{k,s}$ and transformed to the stored (μ, M) coordinates. By Louis's identity, $\mathcal{J}_{k,s} = \mathcal{J}^{\mathrm{cd}}_{k,s} - \mathcal{J}^{\mathrm{mis}}_{k,s}$, where $\mathcal{J}^{\mathrm{mis}}_{k,s} \succeq 0$ is the posterior covariance, summed over samples, of the complete-data scores $\mathcal{U}_{j,i} = \nabla_z \log p(v_j^{\mathcal{A}}, i \mid z)$. Let $\widehat{\jmath}^{\mathrm{mis}}_{k,s,y}$ denote its trace on chart y.

For each chart–mask pair $(y, \mathcal{A})$, let $\mathcal{B}_{y,\mathcal{A}}$ denote the anchor-round uploads with mask $\mathcal{A}$ and nonzero support in chart y. For pairs with positive pooled complete-data trace, define

$$\widehat{\gamma}_{y,\mathcal{A}} = \min\left\{1, \left[1 - \frac{\sum_{(k,s)\in\mathcal{B}_{y,\mathcal{A}}} \widehat{\jmath}^{\mathrm{mis}}_{k,s,y}}{\sum_{(k,s)\in\mathcal{B}_{y,\mathcal{A}}} \operatorname{tr}\Big(\mathcal{J}^{\mathrm{cd}}_{k,s}\big|_y\Big)}\right]_+\right\}. \tag{20}$$

The calibrated curvature is applied chartwise:

$$\widetilde{J}_{k,s}\Big|_y = \widehat{\gamma}_{y,\mathcal{A}}\; \mathcal{J}^{\mathrm{cd}}_{k,s}\big|_y. \tag{21}$$

Hence $\widetilde{J}_{k,s} \succeq 0$, while the scalar calibration rescales each chart block without rotating its modality-conditioned information directions.

Out-of-window evidence. When slice r leaves the window, the server stores the conditional $z_r = \beta_r + G_r' z_{r+1} + \varpi_r$, $\varpi_r \sim \mathcal{N}(0, W_r)$, of the window posterior; composing from s to $r_0 = t - L_{\mathrm{lag}}$ gives $z_s = \bar{\beta}_s + H_s z_{r_0} + \bar{\varpi}_s$ with $\bar{\varpi}_s \sim \mathcal{N}(0, \bar{W}_s)$, and marginalizing $\bar{\varpi}_s$ from the factor of Eq. 6 gives

$$\Delta\Lambda = H_s^\top K_s H_s, \quad \Delta\eta = H_s^\top\big[(I + \widetilde{J}_{k,s}\bar{W}_s)^{-1}(g_{k,s} + \widetilde{J}_{k,s}\widehat{z}^-_{k,s}) - K_s\bar{\beta}_s\big], \quad K_s = \widetilde{J}_{k,s}(I + \bar{W}_s\widetilde{J}_{k,s})^{-1}. \tag{22}$$

Table 4: Regular-round communication per active client (FP32), with $P_{\text{fed}} = 2.0$M backbone parameters, $d_v = 4$, $N^{\text{act}}_{k,s} = 8$, and $|\mathcal{A}| = 2$. P_{blk} denotes the parameters of the shared modality blocks. †Parameter-sharing methods use homogeneous encoder shapes. The FedLWM world channel adds only approximately 0.01% to the backbone payload.

Method	Exchanged object	Scales with	Uplink	Downlink	Parameter-channel scalars
FedAvg†, FedAsync†	Full model	P_{fed}	7.63 MiB	7.63 MiB	$2P_{\text{fed}}$
FedMSplit	Modality blocks	P_{blk}	$4P_{\text{blk}}$ B	$4P_{\text{blk}}$ B	$2P_{\text{blk}}$
FedLWM: world channel	Statistics / ephemeris	$N^{\text{act}}_{k,s} d_v^2$	0.72 KiB	0.72 KiB	0
FedLWM + parameter channel	Parameters + world	$P_{\text{fed}} + N^{\text{act}}_{k,s} d_v^2$	7.63 MiB + 0.72 KiB	7.63 MiB + 0.72 KiB	$2P_{\text{fed}}$

Estimation on anchor rounds. Write $[\Psi_i]_{m,m'}$ for the (m, m') block of the joint emission covariance, with $[\Psi_i]_{m,m} = \Psi_i^m$ and $[\Psi_i]_{m,m'} = C_i^{m,m'}$ for $m \neq m'$. Let $\mathcal{P}^{\text{anch}}_{m,m'}$ denote the anchor-round uploads in the estimation pool satisfying $\{m, m'\} \subseteq \mathcal{M}_{k,s}$. For every block with nonzero total support, the emission update is

$$[\Psi_i]_{m,m'} \leftarrow \frac{\sum_{(k,s)\in\mathcal{P}^{\text{anch}}_{m,m'}} \left(\mathsf{R}^{m,m'}_{k,s,i} + n_{k,s,i}\Sigma_i^{v|\mathcal{M}_{k,s}}\right)}{\sum_{(k,s)\in\mathcal{P}^{\text{anch}}_{m,m'}} n_{k,s,i}}. \tag{23}$$

Within each regime, primitive indices remain consistent across adjacent estimates, so every fitted transition pair refers to the same primitive. The transition is fitted from consecutive filtered estimates $(\widehat{x}_j, \widehat{x}_j^+)$ within the regime. The linear blocks solve a constrained quadratic program, while the dispersion operator eliminates $Q_{y,i}$ through differences between transitions of the same primitive:

$$\begin{aligned}
\widehat{\vartheta}_y^{\text{lin}} &\in \arg\min_{\vartheta:\, \widetilde{S}_y \text{ col. stoch.}} \sum_{j\in\mathcal{T}_y} \left\| [\widehat{x}_j^+]_{\mu,\delta,\alpha} - \Xi_y(\widehat{x}_j)\vartheta - o_y(\widehat{x}_j) \right\|_2^2, \\
\widehat{A}_y^V &\in \arg\min_A \sum_{i:\, y_i = y} \sum_{\substack{j,j'\in\mathcal{T}_y \\ j<j'}} \left\| \Delta\widehat{W}^+_{i,jj'} - A\Delta\widehat{W}_{i,jj'}A^\top \right\|_{\text{F}}^2, \\
\widehat{Q}_{y,i} &= \Pi_{\mathbb{S}_+}\left[\frac{1}{|\mathcal{T}_y|} \sum_{j\in\mathcal{T}_y} \left(\widehat{W}_{i,s_j+1} - \widehat{A}_y^V \widehat{W}_{i,s_j} (\widehat{A}_y^V)^\top \right) \right].
\end{aligned} \tag{24}$$

Here $\Delta\widehat{W}_{i,jj'} \coloneqq \widehat{W}_{i,s_j} - \widehat{W}_{i,s_{j'}}$ and $\Delta\widehat{W}^+_{i,jj'} \coloneqq \widehat{W}_{i,s_j+1} - \widehat{W}_{i,s_{j'}+1}$. The dispersion problem is solved by Gauss–Newton over the d_v^2 entries of A, initialized at $\widehat{A}_y^\mu$.

C.3 Efficiency and Privacy of the World Channel

Communication. For each retained primitive, a regular upload contains $1 + d_v + \frac{d_v(d_v+1)}{2} + |\mathcal{A}|d_v$ scalars, and the ephemeris broadcast contains the corresponding active state block plus one client-level prior strength. Thus the world channel scales as $O(N^{\text{act}}_{k,s} d_v^2)$, independently of client-model size. Anchor rounds additionally transmit emission and calibration statistics, amortized over n_{anch} rounds. Table 4 reports representative regular-round payloads.

Computation. Clients perform one joint E-step, reusing each primitive's factorization of $\Sigma_i^{\mathcal{A}}$ across samples. The server window is block tridiagonal and solves in $O((L_{\text{lag}} + 1)(N_p d_{\text{prim}}^3 + \sum_y N_y^3))$ operations, where $d_{\text{prim}} = d_v + d_v(d_v+1)/2 + |\mathcal{M}|d_v$. Anchor rounds additionally solve one quadratic program in p_y^{lin} variables and one d_v^2-variable Gauss–Newton problem per chart.

Privacy considerations. The FedLWM world channel transmits neither raw observations nor per-example features; uploads consist of responsibility-weighted aggregate sufficient statistics in canonical coordinates. Statistics are released only for sufficiently supported primitives, suppressing low-support messages. When the optional parameter channel is enabled, it exchanges the same model updates as the underlying FL rule and does not alter the world-channel protocol.

D PROOFS

D.1 PRELIMINARIES

Lemma 3 (Spherical Log Equivariance). *For $c, u \in \mathbb{S}^{d-1}$ with $d_g(c,u) < \pi$ and every $\Upsilon \in O(d)$,*

$$\mathrm{Log}_{\Upsilon c}(\Upsilon u) = \Upsilon \,\mathrm{Log}_c(u).$$

Proof. Let $\theta = \arccos(c^\top u) \in [0, \pi)$. For $\theta > 0$, $\mathrm{Log}_c(u) = \frac{\theta}{\sin\theta}\big(u - (\cos\theta)c\big)$, with $\mathrm{Log}_c(c) = 0$. Since orthogonal maps preserve inner products and hence θ,

$$\mathrm{Log}_{\Upsilon c}(\Upsilon u) = \frac{\theta}{\sin\theta}\big(\Upsilon u - (\cos\theta)\Upsilon c\big) = \Upsilon\,\mathrm{Log}_c(u).$$

□

Lemma 4 (Admissible Set). *$\mathcal{K}_p$ is nonempty, closed, bounded, and convex, and $\Pi_{\mathcal{K}_p}$ is single-valued and non-expansive.*

Proof. By the Schur complement with respect to the positive entry 1, (K1) is equivalent to the linear matrix inequality

$$\begin{pmatrix} M_i - \sigma_{\min}^2 I & \mu_i \\ \mu_i^\top & 1 \end{pmatrix} \succeq 0$$

in (μ_i, M_i); (K2) consists of norm balls, and (K3)–(K5) are linear constraints. Hence $\mathcal{K}_p$ is closed and convex. It is bounded because $M_i \succeq \sigma_{\min}^2 I \succ 0$ gives $\|M_i\|_{\mathrm{F}} \le \mathrm{tr}\, M_i \le M_{\max}$, and the remaining blocks are bounded by (K2) and (K5). The state with $\mu_i = 0$, $M_i = \sigma_{\min}^2 I$, $\delta_i^m = 0$ and uniform weights satisfies (K1)–(K5) since $\sigma_{\min}^2 d_v < M_{\max}$ and $N_y \alpha_{\min} \le 1$. The metric projection onto a nonempty closed convex subset of a Hilbert space is single-valued and non-expansive. □

Lemma 5 (Structural Feasibility). *Let $F_\phi \in \Phi_{\mathrm{str}}$ and $z \in \mathcal{K}_p$. Then $\overline{F}_\phi(z)$ satisfies (K1), (K4), and (K5), so projection onto $\mathcal{K}_p$ is needed only to enforce (K2) and (K3). Moreover,*

$$L_\phi \le \sup_{z \in \mathcal{K}_p} \|D\overline{F}_\phi(z)\|_2 < \infty.$$

Proof. With $W_i = V_i - \sigma_{\min}^2 I \succeq 0$, the dispersion update of Eq. 5 reads $W_i' = A_y^V W_i (A_y^V)^\top + Q_{y,i}$, which is positive semidefinite because congruence preserves the cone and $Q_{y,i} \succeq 0$; hence $M_i' - \mu_i'\mu_i'^\top = V_i' \succeq \sigma_{\min}^2 I$, which is (K1). For (K4), $\sum_m w_m \delta_i^{m\prime} = A_y^\delta \sum_m w_m \delta_i^m + \sum_m w_m b_y^{\delta,m} = 0$. For (K5), by Eq. 10, S_y is column stochastic, so $\mathbf{1}^\top S_y \alpha = \mathbf{1}^\top \alpha = 1$, and $(S_y\alpha)_i \ge \alpha_{\min} \mathbf{1}^\top \alpha = \alpha_{\min}$ because $(1 - N_y\alpha_{\min})\widetilde{S}_y \alpha \ge 0$. For the Lipschitz statement, $\overline{F}_\phi$ is a polynomial map of degree two in μ_i and affine in the remaining blocks, $\mathcal{K}_p$ is bounded and convex by Lemma 4, and $\Pi_{\mathcal{K}_p}$ is 1-Lipschitz; the mean value inequality along segments in $\mathcal{K}_p$ gives the bound. Each primitive block of $\overline{F}_\phi$ depends only on the same primitive, and the weight block only on $\alpha_{\cdot|y}$, which gives the block structure. □

Lemma 6 (Self-Normalized Bound). *Let $\{\mathcal{G}_j\}_{j\ge 0}$ be a filtration, let $\Xi_j \in \mathbb{R}^{m\times p}$ be $\mathcal{G}_{j-1}$-measurable with $\|\Xi_j\|_2 \le \widehat{B}$, and let $\zeta_j \in \mathbb{R}^m$ be $\mathcal{G}_j$-measurable with $\mathbb{E}[\exp\langle v, \zeta_j\rangle \mid \mathcal{G}_{j-1}] \le \exp(\sigma^2\|v\|_2^2/2)$ for every v. For a deterministic $\lambda > 0$, let $\Lambda_N^\lambda := \sum_{j\le N} \Xi_j^\top \Xi_j + \lambda I$ and $\mathsf{S}_N := \sum_{j\le N} \Xi_j^\top \zeta_j$. Then, for every $\xi \in (0,1)$, with probability at least $1-\xi$,*

$$\big\|(\Lambda_N^\lambda)^{-1/2}\mathsf{S}_N\big\|_2^2 \le 2\sigma^2\Big[\tfrac{1}{2}\log\det\big(\Lambda_N^\lambda/\lambda\big) + \log(1/\xi)\Big] \le 2\sigma^2\Big[\tfrac{p}{2}\log\Big(1 + \frac{N\min\{m,p\}\widehat{B}^2}{p\lambda}\Big) + \log(1/\xi)\Big]. \tag{25}$$

Proof. For $x \in \mathbb{R}^p$, the sub-Gaussian condition applied with $v = \Xi_j x/\sigma$ shows that $\exp\big(\langle x, \mathsf{S}_N\rangle/\sigma - \frac{1}{2}\sum_{j\le N}\|\Xi_j x\|_2^2\big)$ is a nonnegative supermartingale. This is the only property used in the method-of-mixtures argument of Abbasi-Yadkori et al. (2011, Theorem 1), which, with the Gaussian mixture $x \sim \mathcal{N}(0, \lambda^{-1}I)$, yields the first inequality with $\sum_j \Xi_j^\top \Xi_j$ in place of $\sum_j \psi_j \psi_j^\top$. The second follows from the arithmetic–geometric mean inequality on the eigenvalues and $\mathrm{tr}(\Xi_j^\top \Xi_j) \le \min\{m,p\}\widehat{B}^2$. □

D.2 Proof of Proposition 1

Proof of Proposition 1. Invariance. Let $\Upsilon \in \mathrm{Stab}(\mathcal{C})$. Since Υ fixes every element of $\mathcal{C}$, it fixes $\mathcal{V}_{\mathcal{C}}$ pointwise. Thus $\Upsilon B_y = B_y$ and, by orthogonality, $\Upsilon^\top B_y = B_y$. Moreover $\Upsilon c_y = c_y$, so Lemma 3 gives

$$B_y^\top \operatorname{Log}_{c_y}(\Upsilon u) = B_y^\top \Upsilon \operatorname{Log}_{c_y}(u) = B_y^\top \operatorname{Log}_{c_y}(u).$$

Hence the canonical coordinates are invariant under every residual gauge in $\mathrm{Stab}(\mathcal{C})$. Under the emission model of Sec. 4.2, the local likelihood depends on representations only through these coordinates, up to terms independent of z. Therefore $\ell_{k,s}$, $g_{k,s}$, $\widetilde{J}_{k,s}$, $\Delta\Lambda_{k,s}$, and $\Delta\eta_{k,s}$ are residual-gauge invariant.

Additivity. By Assumption 7, for any finite collection $\mathcal{S}$ of client–time pairs,

$$\log p\big(\{\mathcal{D}_{k,s}\}_{(k,s)\in\mathcal{S}} \mid z_{1:T}, \{\mathcal{H}^{\mathrm{srv}}_{k,s}\}\big) = \sum_{(k,s)\in\mathcal{S}} \ell_{k,s}(z_s).$$

Writing $z_s = \iota_{s,t} X_t$, the corresponding score and curvature in trajectory coordinates are sums of $\iota_{s,t}^\top \nabla_z \ell_{k,s}$ and $\iota_{s,t}^\top \nabla_z^2 \ell_{k,s} \iota_{s,t}$, respectively. Because every term is expressed in the same canonical coordinates, the likelihood-information increments enter additively into the trajectory information form of Eq. 6.

Converse. Suppose $\mathrm{col}(B_y) \not\subseteq \mathcal{V}_{\mathcal{C}}$. Then some column b of B_y has $w := P_{\mathcal{V}_{\mathcal{C}}^\perp} b \neq 0$. Let Υ act as the identity on $\mathcal{V}_{\mathcal{C}}$ and as $-I$ on $\mathcal{V}_{\mathcal{C}}^\perp$; then $\Upsilon \in \mathrm{Stab}(\mathcal{C})$. Since $c_y \in \mathcal{V}_{\mathcal{C}}$, we have $w \perp c_y$ and hence $w \in T_{c_y}\mathbb{S}^{d-1}$. For sufficiently small $\varepsilon > 0$, let $x = \varepsilon w$ and $u = \operatorname{Exp}_{c_y}(x)$, so u lies in chart y and $\operatorname{Log}_{c_y}(u) = x$. Lemma 3 gives $\operatorname{Log}_{c_y}(\Upsilon u) = \Upsilon x = -x$. Therefore the coordinate along b changes by

$$b^\top \Big[\operatorname{Log}_{c_y}(\Upsilon u) - \operatorname{Log}_{c_y}(u)\Big] = -2\varepsilon\, b^\top w = -2\varepsilon \|w\|_2^2 \neq 0.$$

Thus a residual gauge alters the projected coordinates whenever $\mathrm{col}(B_y) \not\subseteq \mathcal{V}_{\mathcal{C}}$. □

D.3 Proof of Theorem 1

Proof of Theorem 1. Fix a chart y with $N := N_y \geq 2$ primitives and set $\alpha_{i|y} = 1/N$ for all i. This is feasible because $N\alpha_{\min} \leq 1$. Set all modality offsets to zero.

Since $\sigma_{\min}^2 d_v < M_{\max}$, choose $\sigma_0^2 \geq \sigma_{\min}^2$ and $\delta \in \mathbb{R}^{d_v} \setminus \{0\}$ such that

$$2\|\delta\|_2 \leq \theta_\mu, \qquad \sigma_0^2 d_v + 4\|\delta\|_2^2 \leq M_{\max}.$$

Define two states that differ only in chart y by

$$\begin{aligned} z^{(1)}: &\quad \mu_1 = \delta, \quad \mu_2 = -\delta, \quad \mu_i = 0 \ (i \geq 3), \quad V_i = \sigma_0^2 I \text{ for all } i, \\ z^{(2)}: &\quad \mu_i = 0 \text{ for all } i, \quad V_1 = V_2 = \sigma_0^2 I + \delta\delta^\top, \quad V_i = \sigma_0^2 I \ (i \geq 3). \end{aligned} \tag{26}$$

Both states belong to $\mathcal{K}_p$. Their aggregate means are zero, while their aggregate raw second moments coincide:

$$\bar{M}^{(1)} = \bar{M}^{(2)} = \sigma_0^2 I + \frac{2}{N}\delta\delta^\top.$$

Hence $\Pi_{\mathrm{mom}}(z^{(1)}) = \Pi_{\mathrm{mom}}(z^{(2)}) =: \varpi$.

Choose an admissible transition F with $A_y^\mu = 2I$, $A_y^V = A_y^\delta = I$, zero drifts, offset intercepts, and covariance injections, and $S_y = (1 - N\alpha_{\min})I + \alpha_{\min}\mathbf{1}\mathbf{1}^\top$. This is of the form defining Φ_{str} and preserves the uniform weight vector. All other charts evolve identically in the two states.

For $z^{(1)}$, the first two primitive means become $\pm 2\delta$, while their dispersions remain $\sigma_0^2 I$. For $z^{(2)}$, the first two means remain zero and their dispersions remain $\sigma_0^2 I + \delta\delta^\top$. The choice of σ_0 and δ keeps both unprojected images in $\mathcal{K}_p$, so the projection is inactive. Their aggregate means remain equal to zero, whereas the aggregate raw second moments differ by

$$\bar{M}^{(1)\prime} - \bar{M}^{(2)\prime} = \frac{6}{N}\,\delta\delta^\top.$$

Therefore

$$\left\|\Pi_{\mathrm{mom}}(F(z^{(1)})) - \Pi_{\mathrm{mom}}(F(z^{(2)}))\right\| = \frac{6}{N}\|\delta\|_2^2 > 0.$$

For any summary-space predictor G, the common input ϖ produces one prediction $G(\varpi)$. The triangle inequality therefore gives

$$\max_{r\in\{1,2\}} \left\|G(\varpi) - \Pi_{\mathrm{mom}}(F(z^{(r)}))\right\| \geq \frac{1}{2}\left\|\Pi_{\mathrm{mom}}(F(z^{(1)})) - \Pi_{\mathrm{mom}}(F(z^{(2)}))\right\| = \frac{3}{N}\|\delta\|_2^2 =: c_{\mathrm{cl}} > 0.$$

The floor depends only on the admissible witness and not on sample size or predictor capacity. □

D.4 Proof of Theorem 2

Regression form. Fix a chart y and order the pairs of $\mathcal{T}_y$ by their earlier times $s_1 < s_2 < \cdots < s_N$, $N = N_{\mathrm{trans}}$, set $s_{N+1} := s_N + 1$, and let $\mathcal{G}_j := \mathcal{F}_{s_{j+1}}$ for $0 \leq j \leq N$, so that $\mathcal{G}_{j-1} = \mathcal{F}_{s_j}$ and $\mathcal{F}_{s_j+1} \subseteq \mathcal{G}_j$. Write $\Xi_j := \Xi_y(x_j)$, $\widehat{\Xi}_j := \Xi_y(\widehat{x}_j)$, $o_j := o_y(x_j)$, $\widehat{o}_j := o_y(\widehat{x}_j)$, and let P select the location, offset and weight components. By Assumptions 1, 2 and 9,

$$Px_j^+ = \Xi_j\vartheta + o_j + \zeta_j, \qquad \zeta_j := P[\zeta_{s_j}]_y, \tag{27}$$

where $\vartheta = \vartheta_y^{\mathrm{lin}}$, ζ_j is $\mathcal{G}_j$-measurable, $\mathbb{E}[\zeta_j \mid \mathcal{G}_{j-1}] = 0$, and ζ_j is conditionally σ_ζ-sub-Gaussian because $\|P^\top v\|_2 \leq \|v\|_2$. The estimator regresses $\widehat{y}_j := P\widehat{x}_j^+ - \widehat{o}_j$ on $\widehat{\Xi}_j$; by Assumption 12(i), $\widehat{\Xi}_j$ is $\mathcal{G}_{j-1}$-measurable. Substituting Eq. 27 and Assumption 12(ii),

$$\widehat{y}_j - \widehat{\Xi}_j\vartheta = \zeta_j + P\varepsilon_j^+ + Pb_j^{\mathrm{fb}} - \varrho_j, \qquad \varrho_j := (\widehat{\Xi}_j - \Xi_j)\vartheta + (\widehat{o}_j - o_j), \tag{28}$$

which separates the disturbance, the target error, the forecast-induced shift and the regressor error. Let L_o be the Lipschitz modulus of the affine map o_y and

$$\widehat{B} := B_\Xi + L_\Xi\epsilon_{\mathrm{state}} \geq \max_j \|\widehat{\Xi}_j\|_2.$$

By Assumption 12(iv),

$$L_\Xi\epsilon_{\mathrm{state}} \leq \frac{\sigma_{\mathrm{PE}}^2}{4B_\Xi},$$

and hence

$$\widehat{B} \leq B_\Xi + \frac{\sigma_{\mathrm{PE}}^2}{4B_\Xi} =: \bar{B}.$$

Thus $\bar{B}$ depends only on fixed regularity quantities and not on N_{trans} or $\epsilon_{\mathrm{state}}$. Because $\mathcal{K}_p$ is bounded and $F(\mathcal{K}_p) \subseteq \mathcal{K}_p$, the quantities

$$B_{F,y} := \sup_{z\in\mathcal{K}_p} \left\|[F(z)]_{\mu,\delta,\alpha,y}\right\|_2, \qquad B_{o,y} := \sup_{x\in\mathcal{K}_p} \|o_y(x)\|_2$$

are finite. Structural realizability and persistent excitation then imply

$$\begin{aligned} N_{\mathrm{trans}}\sigma_{\mathrm{PE}}^2\|\vartheta_y^{\mathrm{lin}}\|_2^2 &\leq \sum_{j\in\mathcal{T}_y} \|\Xi_y(x_j)\vartheta_y^{\mathrm{lin}}\|_2^2 \\ &\leq N_{\mathrm{trans}}(B_{F,y} + B_{o,y})^2. \end{aligned}$$

Hence

$$\|\vartheta_y^{\mathrm{lin}}\|_2 \leq \frac{B_{F,y} + B_{o,y}}{\sigma_{\mathrm{PE}}} =: B_{\vartheta,y},$$

where $B_{\vartheta,y}$ is a fixed regularity bound.

Proof of Theorem 2. Gram bound. With $E_j := \widehat{\Xi}_j - \Xi_j$, $\|E_j\|_2 \leq L_\Xi\epsilon_{\mathrm{state}}$ and $\widehat{\Lambda} := \sum_j \widehat{\Xi}_j^\top\widehat{\Xi}_j \succeq \sum_j \Xi_j^\top\Xi_j + \sum_j(\Xi_j^\top E_j + E_j^\top\Xi_j)$. By Assumption 10 and Weyl's inequality,

$$\lambda_{\min}(\widehat{\Lambda}) \geq N\big(\sigma_{\mathrm{PE}}^2 - 2B_\Xi L_\Xi\epsilon_{\mathrm{state}}\big) \geq \tfrac{1}{2}N\sigma_{\mathrm{PE}}^2, \qquad \lambda_{\max}(\widehat{\Lambda}) \leq N\widehat{B}^2, \tag{29}$$

the second inequality by Assumption 12(iv).

Error decomposition. The least-squares estimate $\bar{\vartheta} := \widehat{\Lambda}^{-1} \sum_j \widehat{\Xi}_j^\top \widehat{y}_j$ satisfies, by Eq. 28, $\widehat{\Lambda}^{1/2}(\bar{\vartheta} - \vartheta) = \widehat{\Lambda}^{-1/2} \sum_j \widehat{\Xi}_j^\top (\zeta_j + P\varepsilon_j^+ + Pb_j^{\mathrm{fb}} - \varrho_j)$. The operator $\mathsf{T} : (e_j)_j \mapsto \widehat{\Lambda}^{-1/2} \sum_j \widehat{\Xi}_j^\top e_j$ satisfies $\mathsf{T}\mathsf{T}^* = \widehat{\Lambda}^{-1/2} \widehat{\Lambda} \widehat{\Lambda}^{-1/2} = I$, so $\|\mathsf{T}\|_2 = 1$ and, for any vectors e_j,

$$\Big\| \widehat{\Lambda}^{-1/2} \sum_j \widehat{\Xi}_j^\top e_j \Big\|_2 \le \Big(\sum_j \|e_j\|_2^2 \Big)^{1/2}. \tag{30}$$

Disturbance. Apply Lemma 6 with $\lambda = \frac{1}{2} N \sigma_{\mathrm{PE}}^2$, which is deterministic. By Eq. 29, $\widehat{\Lambda} + \lambda I \preceq 2\widehat{\Lambda}$, so $\|\widehat{\Lambda}^{-1/2} \mathsf{S}_N\|_2 \le \sqrt{2}\, \|(\widehat{\Lambda} + \lambda I)^{-1/2} \mathsf{S}_N\|_2$. Using $\widehat{B} \le \bar{B}$, define the fixed logarithmic factor

$$\bar{c}_0 := \log\left(1 + \frac{2 \min\{m_y, p_y^{\mathrm{lin}}\} \bar{B}^2}{p_y^{\mathrm{lin}} \sigma_{\mathrm{PE}}^2} \right).$$

It is independent of N_{trans} and $\epsilon_{\mathrm{state}}$. Consequently, with probability at least $1 - \xi$,

$$\left\| \widehat{\Lambda}^{-1/2} \sum_j \widehat{\Xi}_j^\top \zeta_j \right\|_2 \le 2\sigma_\zeta \sqrt{\tfrac{1}{2} p_y^{\mathrm{lin}} \bar{c}_0 + \log(1/\xi)}.$$

Target and regressor errors. By Eq. 30, $\|P\varepsilon_j^+\|_2 \le \epsilon_{\mathrm{state}}$ and $\|\varrho_j\|_2 \le (L_\Xi \|\vartheta\|_2 + L_o)\epsilon_{\mathrm{state}}$, these two terms contribute at most $\sqrt{N}\big(1 + L_o + L_\Xi B_{\vartheta,y}\big)\epsilon_{\mathrm{state}}$.

Residual feedback/confounding. In Euclidean norm this term equals $\|\widehat{\Lambda}^{-1} \sum_j \widehat{\Xi}_j^\top P b_j^{\mathrm{fb}}\|_2 \le \epsilon_{\mathrm{conf},y}$ by Assumption 12(iii), and its $\widehat{\Lambda}^{1/2}$-weighted norm is at most $\lambda_{\max}(\widehat{\Lambda})^{1/2} \epsilon_{\mathrm{conf},y}$.

Constraints. The estimator of Eq. 24 minimizes the same quadratic over the convex set defined by column stochasticity of $\widetilde{S}_y$, which contains ϑ; it is therefore the $\widehat{\Lambda}$-weighted projection of $\bar{\vartheta}$ onto that set, and $\|\widehat{\vartheta}_y^{\mathrm{lin}} - \vartheta\|_{\widehat{\Lambda}} \le \|\bar{\vartheta} - \vartheta\|_{\widehat{\Lambda}}$.

Combination. The Euclidean norm is at most the $\widehat{\Lambda}$-weighted norm divided by $\lambda_{\min}(\widehat{\Lambda})^{1/2} \ge (N/2)^{1/2} \sigma_{\mathrm{PE}}$, and Eq. 29 gives $\lambda_{\max}(\widehat{\Lambda})^{1/2} \le \sqrt{2}\, \widehat{B}\, \lambda_{\min}(\widehat{\Lambda})^{1/2} / \sigma_{\mathrm{PE}}$. Hence, with probability at least $1 - \xi$,

$$\begin{aligned} \big\| \widehat{\vartheta}_y^{\mathrm{lin}} - \vartheta_y^{\mathrm{lin}} \big\|_2 \le{}& \frac{2\sqrt{2}\, \sigma_\zeta}{\sigma_{\mathrm{PE}}} \sqrt{\frac{\frac{1}{2} p_y^{\mathrm{lin}} \bar{c}_0 + \log(1/\xi)}{N_{\mathrm{trans}}}} \\ &+ \frac{\sqrt{2}\,(1 + L_o + L_\Xi B_{\vartheta,y})}{\sigma_{\mathrm{PE}}} \epsilon_{\mathrm{state}} + \frac{\sqrt{2}\, \bar{B}}{\sigma_{\mathrm{PE}}} \epsilon_{\mathrm{conf},y}. \end{aligned}$$

Since $\frac{1}{2} p_y^{\mathrm{lin}} \bar{c}_0 + \log(1/\xi) \le \max\{\frac{1}{2}\bar{c}_0, 1\}\big(p_y^{\mathrm{lin}} + \log(1/\xi)\big)$, this yields Eq. 8 with a constant $C > 0$ depending only on the fixed regularity quantities $\bar{c}_0$, L_o, L_Ξ, $B_{\vartheta,y}$, $\bar{B}$, and σ_{PE}. □

The rate does not scale with the number of clients: uploads sharing a semantic time correspond to the same transition pair, so $N_{\mathrm{trans}} \le |\mathcal{I}_p| - 1$ (Proposition 6). Additional clients can nevertheless reduce $\epsilon_{\mathrm{state}}$ through more informative state inference.

Corollary 1 (Sample Complexity of the Linear Blocks)**.** *Under the hypotheses of Theorem 2, the sampling term of Eq. 8 is at most ϵ simultaneously for all charts with probability at least $1 - \xi$ if, for every chart y,*

$$N_{\mathrm{trans},y} \ge C^2 \frac{\sigma_\zeta^2}{\sigma_{\mathrm{PE}}^2} \frac{p_y^{\mathrm{lin}} + \log(|\mathcal{Y}|/\xi)}{\epsilon^2}.$$

Thus the requirement is linear in $p_y^{\mathrm{lin}} = 2d_v^2 + N_y d_v + (|\mathcal{M}| - 1)d_v + N_y(N_y - 1)$, whereas the corresponding dense affine chart map has $D_y(D_y + 1)$ parameters.

Proof. Apply Theorem 2 to chart y with failure probability $\xi/|\mathcal{Y}|$, solve its sampling term for $N_{\mathrm{trans},y}$, and take a union bound over y. □

Proposition 7 (Identifiability of the Dispersion Block). *Under Assumptions 9 and 11, let $A \in GL(d_v)$ satisfy $A\,\Delta W\,A^\top = A_y^V \Delta W (A_y^V)^\top$ for every $\Delta W \in \mathcal{W}_y$. Then $A = \pm A_y^V$, so the congruence map $W \mapsto A_y^V W (A_y^V)^\top$ is uniquely determined. Moreover, for every primitive i and any $j \in \mathcal{T}_y$,*

$$Q_{y,i} = [\overline{F}^\circ(x_j)]_{W_i} - A_y^V W_{i,s_j} (A_y^V)^\top,$$

and hence $Q_{y,i}$ is uniquely determined.

Proof. By Assumption 9, A_y^V is invertible; set $U := (A_y^V)^{-1} A$. Then $U \Delta W U^\top = \Delta W$ for every $\Delta W \in \mathcal{W}_y$, so $U \in \mathcal{G}_y = \{\pm I\}$ by Assumption 11. Both signs induce the same congruence, and the expression for $Q_{y,i}$ follows from the dispersion update of Eq. 5 applied to $\overline{F}^\circ(x_j)$. □

D.5 Proof of Theorem 3

Lemma 7 (Rollout Budget). *For every $u \in \mathcal{K}_p$ and $h \geq 0$, $\|F_\phi^h(u) - F^h(u)\|_2 \leq E_h$.*

Proof. Both maps send $\mathcal{K}_p$ into itself, by Assumption 1 and by construction of $\Pi_{\mathcal{K}_p}$. Let $e_n := \|F_\phi^n(u) - F^n(u)\|_2$, so $e_0 = 0$. Since $F^n(u) \in \mathcal{K}_p$,

$$e_{n+1} \leq \big\|F_\phi(F_\phi^n u) - F_\phi(F^n u)\big\|_2 + \big\|F_\phi(F^n u) - F(F^n u)\big\|_2 \leq L_\phi e_n + \epsilon_F,$$

and induction gives $e_h \leq \epsilon_F \sum_{n<h} L_\phi^n = E_h$. No linearity of F is used. □

Lemma 8 (Sensitivity of the Regularized Consumer). *Let $\varphi_{k,t} : \mathbb{R}^{d_k^{\mathrm{loc}}} \to \mathbb{R} \cup \{+\infty\}$ be proper, closed and σ_L-strongly convex, differentiable at $x_k^\star$, and let $x_k^+(p) := \arg\min_x \varphi_{k,t}(x) + \frac{1}{2}\nu_{k,t}^{\mathrm{eff}} \|x - p\|_2^2$ with $\nu_{k,t}^{\mathrm{eff}} > 0$. With $\kappa_{k,t} := \nu_{k,t}^{\mathrm{eff}}/(\sigma_L + \nu_{k,t}^{\mathrm{eff}})$ and $c_k := \|x_k^+(x_k^\star) - x_k^\star\|_2$, for all p, p',*

$$\|x_k^+(p) - x_k^+(p')\|_2 \leq \kappa_{k,t}\|p - p'\|_2, \qquad \|x_k^+(p) - x_k^\star\|_2 \leq \kappa_{k,t}\|p - x_k^\star\|_2 + c_k, \qquad c_k \leq \frac{\|\nabla \varphi_{k,t}(x_k^\star)\|_2}{\sigma_L + \nu_{k,t}^{\mathrm{eff}}}.$$

Proof. Write $\nu = \nu_{k,t}^{\mathrm{eff}}$ and $x_a = x_k^+(p_a)$. Optimality gives $\nu(p_a - x_a) \in \partial \varphi_{k,t}(x_a)$, and strong monotonicity of $\partial \varphi_{k,t}$ gives $\nu\langle (p_1 - x_1) - (p_2 - x_2), x_1 - x_2\rangle \geq \sigma_L \|x_1 - x_2\|_2^2$, that is, $\nu\langle p_1 - p_2, x_1 - x_2\rangle \geq (\sigma_L + \nu)\|x_1 - x_2\|_2^2$; the Cauchy–Schwarz inequality yields the first bound. The second follows from the first with $p' = x_k^\star$ and the triangle inequality. For the third, $h(x) := \varphi_{k,t}(x) + \frac{\nu}{2}\|x - x_k^\star\|_2^2$ is $(\sigma_L + \nu)$-strongly convex with $0 \in \partial h(x_k^+(x_k^\star))$ and $\nabla h(x_k^\star) = \nabla \varphi_{k,t}(x_k^\star)$, and strong monotonicity gives $\|\nabla \varphi_{k,t}(x_k^\star)\|_2 \geq (\sigma_L + \nu) c_k$. □

Proof of Theorem 3. The rollout bound is Lemma 7.

Part (i). Conditioned on the delayed-state estimate $z \in \mathcal{K}_p$ at semantic time $t - \tau$, the propagation error of L1^0 is exactly $\|z - F^\tau(z)\|_2$, and that of L1 is $\|F_\phi^\tau(z) - F^\tau(z)\|_2 \leq E_\tau$ by Lemma 7. The hypothesis $E_\tau < \|F^\tau(z) - z\|_2$ makes the latter strictly smaller.

Part (ii). By Assumption 2, $z_t \in \mathcal{K}_p$, and by Lemma 4, $\|\bar{z}_{t|t} - z_t\|_2 \leq \|\widehat{z}_{t|t} - z_t\|_2 \leq \epsilon_{\mathrm{state}}$. Let $p_0 := \Pi_k^{\mathrm{loc}} \bar{z}_{t|t}$ and $p_2 := \Pi_k^{\mathrm{loc}} F_\phi^\Delta(\bar{z}_{t|t})$. Using $\|\Pi_k^{\mathrm{loc}}\|_2 \leq 1$ and the triangle inequality,

$$\|p_0 - x_k^\star\|_2 \geq \big\|\Pi_k^{\mathrm{loc}}(F^\Delta(z_t) - z_t)\big\|_2 - \big\|\Pi_k^{\mathrm{loc}}(\bar{z}_{t|t} - z_t)\big\|_2 - \big\|\Pi_k^{\mathrm{loc}} F^\Delta(z_t) - x_k^\star\big\|_2 \geq D_\Delta - \epsilon_{\mathrm{state}} - \epsilon_{\mathrm{loc}},$$

$$\|p_2 - x_k^\star\|_2 \leq \big\|F_\phi^\Delta(\bar{z}_{t|t}) - F^\Delta(\bar{z}_{t|t})\big\|_2 + \big\|F^\Delta(\bar{z}_{t|t}) - F^\Delta(z_t)\big\|_2 + \epsilon_{\mathrm{loc}} \leq E_\Delta + \Gamma_\Delta^\star \epsilon_{\mathrm{state}} + \epsilon_{\mathrm{loc}},$$

where the last step uses Lemma 7 with $u = \bar{z}_{t|t}$ and Assumption 8. The hypothesis $E_\Delta < D_\Delta - (1 + \Gamma_\Delta^\star)\epsilon_{\mathrm{state}} - 2\epsilon_{\mathrm{loc}}$ is exactly the condition under which the upper bound for p_2 is strictly below the lower bound for p_0, so $\|p_2 - x_k^\star\|_2 < \|p_0 - x_k^\star\|_2$. By Lemma 8, the post-update error is certified by $p \mapsto \kappa_{k,t}\|p - x_k^\star\|_2 + c_k$, which is strictly increasing in $\|p - x_k^\star\|_2$ since $\kappa_{k,t} > 0$, so the certified bound is strictly smaller at p_2. □

The two conditions reflect the asymmetry between the two temporal uses. Retrospective assimilation compares a model error with the drift, since the evidence fixes the state at semantic time $t - \tau$; prospective prediction must additionally absorb the propagation of the current estimation error and the coordinate mismatch, $(1 + \Gamma_\Delta^\star)\epsilon_{\mathrm{state}} + 2\epsilon_{\mathrm{loc}}$, because no observation of the use-time state exists.

E Experimental Details and Additional Results

E.1 Experimental Setup

Datasets and federations. MELD (Poria et al., 2019) (text, audio, and video), CrisisMMD (Alam et al., 2018) (text and images), and CREMA-D (Cao et al., 2014) (audio and video) follow the preprocessing, feature extraction, and task heads of FedMultimodal (Feng et al., 2023). CooperScene (Wu et al., 2026) retains its original cooperative 3D-detection architecture and is used for qualitative comparison on matched scenes. We preserve persistent client identities, with $K = 20$ speakers on MELD, $K = 5$ source events on CrisisMMD, and $K = 12$ actors on CREMA-D, yielding natural statistical heterogeneity without synthetic repartitioning. Clients further differ in modality capability and encoder capacity. Client partitions, modality traces, participation schedules, and encoder assignments are fixed per seed and reused across methods.

Temporal stress. Each seed defines a 100-window service trace over $\mathcal{M}_{k,s}$, $\tau_{k,t}$, and $\Delta_{k,t}$, replayed identically across methods. **S0** removes temporal mismatch while retaining client and encoder heterogeneity. **S1** combines Markovian modality availability, modality-conditioned delays sampled from $\mathrm{Unif}\{0, \ldots, 8\}$, and $\Delta_{k,t} = 2$. The availability process uses stationary drop probability 0.30 and persistence 0.80, producing temporally correlated missing modalities rather than independent dropout. Delay assignment is coupled to modality payload with seed-specific jitter, inducing the nonzero $\rho_{\tau\mathcal{M}}$ studied in Fig. 4(a). Latent evolution preserves the fixed label space: MELD follows its natural dialogue chronology, while CrisisMMD additionally uses client-, time-, and modality-dependent affine feature drift.

Controlled theory validation. Fig. 4(b–d) and E-G3 use a CREMA-D-calibrated environment with latent ground truth. States evolve under a structured transition in Φ_{str}, and uploads follow the joint emission model of Sec. 4.2 under the same observation-, arrival-, and use-time protocol. Calibration parameters are fixed before comparison, and this environment is used only for theory-facing diagnostics requiring latent ground truth. All methods observe only the generated uploads; latent states and transition parameters are used solely for evaluation.

Implementation and baselines. MELD and CrisisMMD use AdamW with learning rates 5×10^{-4} and 10^{-3}, batch sizes 16 and 64, and two local epochs; CREMA-D follows FedMultimodal optimization conventions. Baselines use official or FedMultimodal implementations when available, with matched backbones, local budgets, and service traces; method-specific hyperparameters are selected using validation data only. Asynchronous methods process uploads at recorded arrival times, while synchronous methods retain their native round barriers. FedLWM uses 16-dimensional normalized representations, $d_v = 4$ canonical coordinates, eight active primitives, $L_{\mathrm{lag}} = 4$, and 5% anchor rounds. The backbone-parameter protocol is fixed across matched comparisons, isolating the latent inference–prediction channel.

E.2 Mechanism and Consistency Validation

Frame residuals measure the relative change of pooled geometry, modality-offset, and cross-modal statistics after independent client-wise frame perturbations followed by each reference construction. Observation geometry is summarized by the principal angle between information subspaces and the corresponding difference in mean staleness. ϵ_{xmod} measures held-out-modality coordinate prediction error, the ACI gap is the coverage error of nominal 90% state intervals, and ϵ_θ, $\epsilon_{\mathcal{L}}$, and ϵ_μ are relative discrepancies between FedLWM world-state inference and centralized EM.

Table 5 isolates the mechanisms underlying FedLWM rather than downstream accuracy alone. HSRF gives the lowest residuals across all three pooled statistic families, with the clearest reductions in geometry and modality offsets. More importantly, changing the observed modality subset increases the information-subspace angle from 10.7° to about 35°, whereas introducing delay–modality coupling leaves this directional separation nearly unchanged but increases $|\Delta\bar{\tau}|$ from 0.10 to 1.87 windows. This separates the two ingredients of modality-conditioned temporal misalignment: modality determines the information direction, while coupling separates the semantic times informing those directions. Joint rather than marginal cross-modal assembly further raises E-G1$'$ from 52.55 to 54.38, reduces ϵ_{xmod} from 1.114 to 0.987, and narrows the ACI gap from 1.93 to 1.61

Table 5: **Mechanism and consistency validation on MELD under S1.** Frame residuals test empirical coordinate consistency across three uploaded statistic families; principal angles and staleness gaps characterize modality-conditioned observation geometry. The lower blocks evaluate joint cross-modal inference and the centralized limit of world-state estimation.

Empirical Frame Consistency				Conditional Observation Geometry			
Reference	**Geom.↓**	**Offset↓**	**X-modal↓**	**Condition**	**Angle (°)↑**	**IQR (°)**	**$\lvert\Delta\bar{\tau}\rvert$ (windows)**
EMA	0.182	0.977	0.990	Same subset	10.7	[7.6, 13.8]	0.00
VICReg	0.935	1.641	1.065	Different, uncoupled	35.4	[34.0, 36.8]	0.10
HSRF	**0.133**	**0.922**	**0.989**	**Different, coupled**	**36.0**	[31.3, 40.8]	**1.87**
Cross-modal E-step				**Centralized-limit Consistency**			
Assembly	**E-G1′↑**	**ϵ_{xmod} ↓**	**ACI gap (pp)↓**	**Comparison**	**ϵ_θ ↓**	**$\epsilon_{\mathcal{L}}$ ↓**	**ϵ_μ ↓**
Marginal product	52.55	1.114	1.93	Distributed assembly vs. batch EM	3.29×10^{-15}	3.66×10^{-16}	2.72×10^{-15}
Joint	**54.38**	**0.987**	**1.61**	–	–	–	–

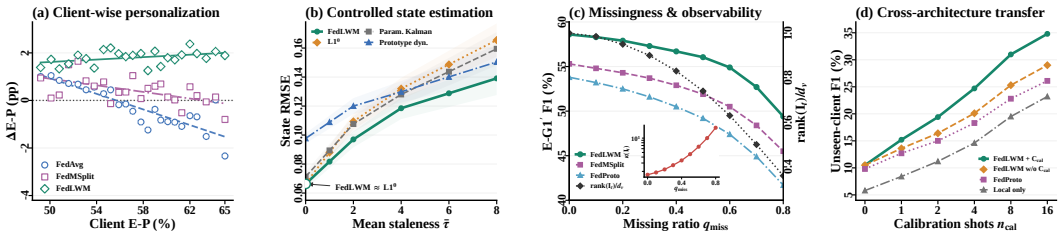

Figure 6: **Extended evaluation.** **(a)** Client-level personalization gain over local training versus standalone client performance. **(b)** Controlled latent-state RMSE under increasing mean staleness; structured and identity transport coincide when no retrospective propagation is required. **(c)** E-G1′ under increasing modality missingness, together with the normalized rank and condition number of the pooled information matrix. **(d)** E-G2 performance of an unseen heterogeneous receiver versus calibration-set size, with and without $\mathcal{C}_{\text{cal}}$.

pp. Finally, distributed sufficient-statistic assembly matches batch EM to machine precision in the contemporaneous full-participation limit.

E.3 Extended Evaluation and Design Diagnostics

Extended robustness and transfer. Fig. 6 probes when the shared world state remains useful under client and observation heterogeneity. FedLWM yields positive client-level gains across standalone performance, whereas parameter-sharing gains diminish for stronger local clients; calibration fiducials likewise improve adaptation to an unseen encoder by enlarging the fixed reference span. Temporally, FedLWM coincides with identity L1 transport at zero staleness but separates as delay grows, outperforming generic parametric filtering and prototype dynamics. Increasing modality missingness simultaneously lowers the normalized rank and worsens the conditioning of pooled information, yet FedLWM degrades more gradually by combining joint partial-observation likelihoods with the predictive state prior. The widening gap under degraded rank and conditioning ties the robustness gain to state observability rather than generic regularization. Likewise, the benefit of $\mathcal{C}_{\text{cal}}$ persists across calibration budgets, indicating that it enlarges the transferable reference span rather than merely improving cold start. Together, these results show that shared latent inference transfers across heterogeneous clients while becoming most useful as temporal misalignment and partial observability intensify.

Cross-dataset generalization. Table 6 extends the recognition results to CREMA-D. Under contemporaneous S0, FedLWM is competitive rather than uniformly dominant, trailing the strongest personalized baseline by 0.32 pp while leading E-G1 by 1.88 pp. Under S1, however, it leads the strongest federated baselines by 2.22 pp in E-P and 3.91 pp in E-G1, with margins of 4.18 pp under modality dropout and 1.89 pp on the unseen receiver. This reproduces the stress-amplified pattern observed in the main benchmarks, supporting that FedLWM's advantage tracks temporal misalignment, partial observability, and receiver heterogeneity rather than a generic static-model advantage.

Design diagnostics. Fig. 7 examines representation capacity and the principal resource choices of FedLWM. HSRF retains broader spectral support than EMA and VICReg, while frame locking reduces residual gauge error without collapsing representation uniformity. In this MELD configu-

Table 6: **Quantitative comparison of audio-visual emotion recognition on CREMA-D.** Entries are means ± sample standard deviations over five matched seeds. **S0** retains cross-client heterogeneity with contemporaneous evidence, whereas **S1** additionally introduces dynamic modality availability, latent drift, and modality-conditioned staleness. E-P and E-G1 evaluate personalized and pooled-global performance; E-G1′ applies 50% test-time modality dropout to the S0 shared model, and E-G2 evaluates an unseen speaker using eight support samples. Centralized trains the same backbone on pooled data without world-model components; under S1 it inherits modality availability and drift but incurs no communication delay, and it is excluded from the federated ranking. † denotes parameter-sharing methods with matched homogeneous encoders.

	CREMA-D – UAR (%)					
	S0		S1		Missing	Unseen
Method	**E-P**	**E-G1**	**E-P**	**E-G1**	**E-G1′**	**E-G2**
Centralized	47.35 ± 0.92	52.46 ± 0.78	44.18 ± 1.41	50.27 ± 1.19	47.08 ± 1.36	34.63 ± 3.27
FedAvg† (McMahan et al., 2017)	42.14 ± 1.14	45.18 ± 0.96	34.87 ± 1.83	38.92 ± 1.57	39.47 ± 1.68	24.34 ± 4.12
FedProx† (Li et al., 2020)	42.57 ± 1.08	46.52 ± 0.91	35.41 ± 1.76	39.58 ± 1.49	39.95 ± 1.61	25.87 ± 3.96
FedAsync† (Xie et al., 2019)	41.93 ± 1.27	46.31 ± 1.05	36.96 ± 1.62	41.24 ± 1.38	39.28 ± 1.74	27.52 ± 4.08
AdaMasFL† (Yan et al., 2025)	42.36 ± 1.19	46.67 ± 0.99	37.42 ± 1.58	42.01 ± 1.31	40.36 ± 1.66	27.19 ± 3.88
HAMFL (Qi & Li, 2024)	43.52 ± 1.02	47.63 ± 0.93	37.14 ± 1.71	42.27 ± 1.44	42.94 ± 1.47	28.46 ± 3.71
FedCola (Sun et al., 2024)	44.08 ± 0.97	48.71 ± 0.87	37.82 ± 1.64	43.35 ± 1.36	42.61 ± 1.52	29.13 ± 3.64
FedAFD (Tan et al., 2026)	$\mathbf{45.93 \pm 0.94}$	48.26 ± 0.95	37.35 ± 1.69	42.98 ± 1.41	42.17 ± 1.58	28.72 ± 3.79
FedMSplit (Chen & Zhang, 2022)	44.36 ± 1.05	49.18 ± 0.84	38.66 ± 1.55	41.02 ± 1.29	45.12 ± 1.34	29.84 ± 3.52
PEPSY (Nguyen et al., 2026)	45.12 ± 0.99	49.54 ± 0.81	$\underline{39.73 \pm 1.48}$	44.86 ± 1.24	44.73 ± 1.39	30.27 ± 3.58
FedTaste (Liang et al., 2026)	44.87 ± 1.03	$\underline{49.86 \pm 0.83}$	39.21 ± 1.52	$\underline{45.47 \pm 1.21}$	$\underline{45.64 \pm 1.31}$	$\underline{31.58 \pm 3.49}$
FedLWM	$\underline{45.61 \pm 0.98}$	$\mathbf{51.74 \pm 0.86}$	$\mathbf{41.95 \pm 1.44}$	$\mathbf{49.38 \pm 1.26}$	$\mathbf{49.82 \pm 1.35}$	$\mathbf{33.47 \pm 3.53}$

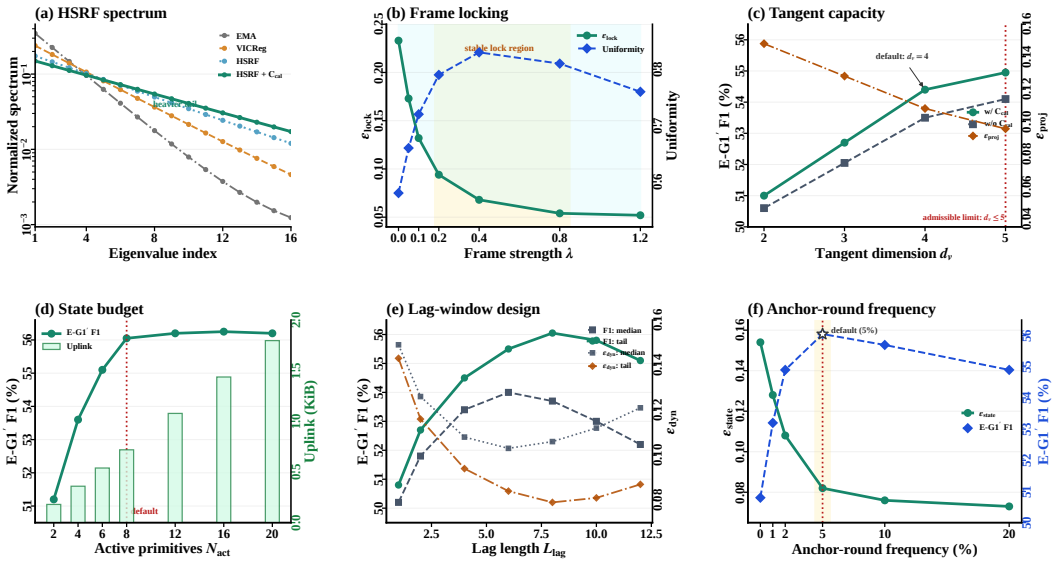

Figure 7: **Mechanism and design diagnostics. (a)** Normalized representation spectra induced by alternative reference-frame constructions. **(b)** Frame-locking residual and representation uniformity as a function of frame strength; shading indicates the stable operating region. **(c)** E-G1′ and reference-span projection residual versus admissible tangent dimension, with default $d_v = 4$. **(d)** E-G1′ and regular-round world-channel uplink versus the number of active primitives. **(e)** E-G1′ and propagation error versus lag-window length under median-delay and modality-conditioned tail-delay criteria. **(f)** State-estimation error and E-G1′ versus anchor-round frequency; the selected operating point is marked.

ration, $d_v \leq 5$, with performance saturating near the default $d_v = 4$. For state inference, accuracy saturates near eight active primitives while uplink continues to grow, the modality-conditioned delay tail gives a better fixed-lag operating point than the population median, and most benefits of anchor calibration appear by the selected 5% rate. Beyond these operating points, additional representation or inference budget mainly increases cost with limited downstream benefit. These trends justify the reference span, state budget, lag window, and anchor frequency used elsewhere.

Transfer to cooperative perception. Fig. 8 extends the evidence beyond multimodal recognition: across matched CooperScene frames, FedLWM recovers ground-truth vehicles missed by FedAvg, with BEV detections confirming localization. Repeated recoveries across scenes and sensing condi-



Figure 8: Qualitative comparison of cooperative 3D detection on CooperScene. Within each two-row group, the top and bottom rows show FedAvg (blue) and FedLWM (red), respectively. Green dashed boxes denote ground truth. The yellow-bordered close-ups highlight vehicles missed by FedAvg and recovered by FedLWM, while the lower-right insets show the corresponding bird's-eye-view detections.

tions indicate that the benefit extends beyond recognition. Because both methods are evaluated on matched scenes under the same detection architecture, these differences isolate the contribution of the shared latent context rather than model capacity. These cases suggest that temporally aligned shared context preserves scene information absent from an individual current observation, demonstrating practical viability and generalizability beyond classification benchmarks.