EARLY CLASSIFICATION OF TIME-DOMAIN SOURCES USING PHYSICALLY PLAUSIBLE FUTURE LIGHT-CURVE HYPOTHESES

Anonymous authors
Paper under double-blind review

ABSTRACT

The early classification of time-domain astronomical events represents a frontier challenge in AI4Science, where models must infer the nature of complex physical phenomena from highly sparse, irregularly sampled multivariate time series restricted to a brief causal prefix. To overcome this highly degenerate inverse problem, we propose a generalized probabilistic framework that reformulates early sequential classification into a causal "generate-then-evaluate" inference paradigm. Rather than forcing a premature, deterministic prediction, we design a physics-informed conditional variational autoencoder (CVAE) that preserves epistemic uncertainty by sampling a multi-hypothesis manifold of physically viable future trajectories. Crucially, this generative space is regularized via domain-specific penalties, acting as a powerful domain-aware denoiser that filters spurious instrumental artifacts and strictly enforces astrophysical boundary conditions. Finally, a Continuous Thought Machine (CTM) dynamically evaluates these physically bounded simulations against newly arriving photometry, accumulating predictive Bayesian evidence via a gated sequential update mechanism. Evaluated on a rigorous, class-balanced PLAsTiCC benchmark of 10,000 transient events across ten classes, our framework improves Top-1 accuracy from 78.19% to 78.81% and significantly reduces balanced log loss from 0.592 to 0.575 (paired gain 95% CI: [0.32, 0.92]). Ultimately, our work demonstrates that maintaining a physically regularized trajectory mixture strictly outperforms deterministic forecasting, delivering superior probability calibration, filtering noisy observations, and establishing an interpretable, scientifically grounded representation learning paradigm for large-scale astronomical surveys.

1 INTRODUCTION

In modern astronomy, discovering time-domain events is fundamentally a real-time sequential classification problem. Wide-field telescopes continuously stream sparse, fragmented multiband snapshots of an object's evolving lifecycle Ivezić et al. (2019). Triggering scarce, specialized follow-up instruments requires categorizing these physical time-series rapidly from a brief initial prefix, well before distinguishing characteristics fully emerge Muthukrishna et al. (2019). For the AI community, this presents a formidable representation learning challenge: accurately categorizing a physical sequence under extreme data sparsity.

Conventional photometric classifiers predominantly operate under a retrospective paradigm, requiring the complete light curve to execute a prediction Lochner et al. (2016), which precludes real-time intervention. Early foundational work addressed this limitation by developing specialized early-time architectures (e.g., RAPID, SCONE) Muthukrishna et al. (2019); Qu & Sako (2022) and continuous-time sequence models for irregular sampling Kidger et al. (2020). However, these classical approaches exhibit critical limitations: they are typically engineered for specific transient classes, struggle to seamlessly fuse multimodal observations, and fail to natively incorporate prior physical constraints of different celestial objects. Furthermore, while early adaptations of latent-variable generative models—such as VAEs Kingma & Welling (2022) and Conditional VAEs Sohn et al. (2015)—have provided probabilistic representations for simulating alternative futures, they lack a rigorous mechanism to systematically generate multimodal, physically bounded trajectories

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.

and evaluate them as class-conditional hypotheses. Although physics-informed machine learning rigorously embeds domain constraints into data-driven models Karniadakis et al. (2021), leveraging these priors for sequential, multimodal astronomical inference remains an open challenge.

The fundamental objective of this paper is therefore not merely predicting an early point-estimate label, but determining whether physically bounded future trajectories can serve as testable hypotheses for sequential Bayesian inference. To address this, we propose a novel physics-inspired multimodal framework that reformulates early classification as a causal generate-and-verify process:

$$\mathcal{P}_K \longrightarrow \{\tilde{\mathcal{F}}_K^{(m,y)}\}_{m=1}^{M} \longrightarrow \text{future verification} \longrightarrow y. \tag{1}$$

Specifically, we deploy a Continuous Thought Machine (CTM) Darlow et al. (2025) as the core light-curve generation engine. Coupled with a physics-informed conditional variational autoencoder (CVAE), the CTM synthesizes multimodal, class-conditional future hypotheses from the observed prefix. By enforcing astrophysical inductive biases during generation, this module actively suppresses simulated trajectories that violate physical domain constraints. As subsequent observations arrive, their predictive log-likelihoods under these CTM-generated candidate sets are quantified as relative Bayesian evidence and dynamically fused with the prefix-conditioned prior for final classification. Evaluated on a ten-class PLAsTiCC benchmark, our framework demonstrates that physically regularized future simulations extract highly discriminative evidence beyond the initial prefix, establishing a robust early classification paradigm while strictly preserving causal deployment protocols.

2 Data

2.1 PLAsTiCC Dataset and Target Classes

We evaluate our framework on the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC) dataset Kessler et al. (2019), a comprehensive benchmark of simulated transient astronomical events designed to replicate the observational cadence of the Legacy Survey of Space and Time (LSST) Ivezić et al. (2019). Each celestial object is formulated as an irregularly sampled multivariate time series across six photometric passbands (u, g, r, i, z, y). The input representation to our model concatenates sequence-level features (Modified Julian Date [MJD], passband, flux, flux uncertainty, and a detection flag) with object-level metadata (redshift, Galactic reddening $E(B-V)$, and validity indicators).

To rigorously test early classification performance under different observation conditions, we select ten classes Kessler et al. (2019): *explosive transients*: SN Ia (90), SN Ia-91bg (67), SN Iax (52), SN Ib/c (62), SN II (42), and SLSN-I (95); *periodic variables*: RR Lyrae (92) and eclipsing binaries (EB, 16); and *stochastic long-timescale sources*: tidal disruption events (TDE, 15) and active galactic nuclei (AGN, 88). This taxonomic diversity ensures that the model is forced to generalize across highly heterogeneous evolutionary trajectories.

Because the raw PLAsTiCC dataset exhibits extreme class imbalance, we apply the eligibility criteria detailed in Appendix A to curate a more balanced experimental cohort. Specifically, we subsample nine of the classes to $15{,}000$ instances each while retaining all available TDE objects. This curation yields a final dataset of $147{,}071$ objects, $98{,}450$ for training, $19{,}811$ for validation, and $28{,}810$ for test (see Appendix Table 3 for the complete dataset composition).

2.2 Data Preprocessing For Early Classification Of Transients

We aggregate the raw photometry into irregularly sampled six-band observation data points, applying a strictly causal normalization scheme parameterized solely by prefix statistics to prevent look-ahead bias. Full preprocessing pipelines are provided in Appendix A. To categorically preclude temporal data leakage, we partition the dataset into object-disjoint training, validation, and locked test sets with the method discussed in Appendix A. This guarantees that all K temporal states of a specific astronomical object remain strictly isolated within a single split.

Because our framework classifies transient events by forecasting and validating future photometric trajectories, we formalize early classification as a sequential inference task indexed over discrete observing nights. The initial detection establishes the first causal state, $K = 1$, occurring at epoch t_1. Each subsequent night yielding a valid observation increments the state index K. To construct the

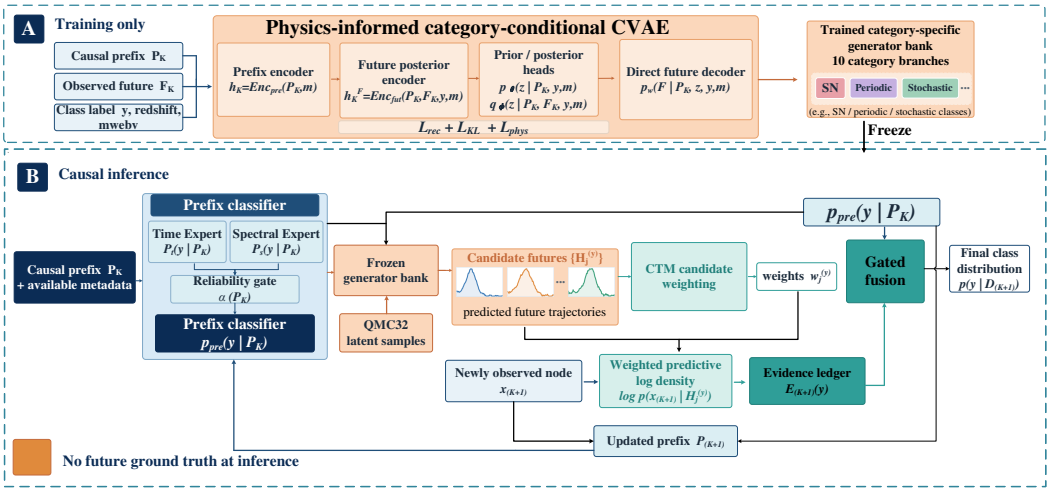

Figure 1: Overview of the proposed generate-and-evaluate framework. During training, category-conditional CVAEs learn physically constrained future hypotheses from observed prefixes and unobserved future trajectories. During inference, future observations are replaced by generated candidate trajectories, which are evaluated by the CTM and subsequently accumulated as predictive Bayesian evidence for gated classification.

temporal dataset, a given celestial source contributes all sequential states within a 50-day window following its initial detection ($t_K - t_1 \leq 50\,\mathrm{d}$), conditioned on the availability of a valid future observation window (the complete early-observation protocol is detailed in Appendix Figure 4).

Formally, at state K, the observed causal prefix $\mathcal{P}_K$ and the corresponding future trajectory $\mathcal{F}_K$ are defined as:

$$\mathcal{P}_K = \big\{(t_i, b_i, f_i, \sigma_i) \mid t_1 - 20\,\mathrm{d} \leq t_i \leq t_K\big\}, \tag{2}$$

$$\mathcal{F}_K = \big\{(t_i, b_i, f_i, \sigma_i) \mid t_K < t_i \leq t_K + 120\,\mathrm{d}\big\}. \tag{3}$$

To enforce temporal shift invariance, all time steps t_i are normalized relative to the initial detection t_1. During inference, the model strictly conditions on the causal prefix $\mathcal{P}_K$ alongside any available metadata, explicitly precluding look-ahead bias. Conversely, the unobserved future trajectory $\mathcal{F}_K$ is reserved exclusively as a target for the generative module. Crucially, we treat $\mathcal{F}_K$ not as an absolute ground truth, but as a stochastic realization of the underlying physical process, intrinsically accounting for the variance and measurement uncertainties characteristic of observational astronomy.

3 Method

3.1 Problem Formulation and Causal Inference Framework

The inference input at state K is defined as $\mathcal{I}_K = (\mathcal{P}_K, \mathbf{m})$, where $\mathcal{P}_K$ is the sequentially observed causal prefix and $\mathbf{m}$ contains all available metadata (e.g., redshift, Galactic reddening). The target label y belongs to a ten-class set $\mathcal{Y}$. Standard prefix-only classifiers model the posterior directly as $\hat{y}_K^{\mathrm{pre}} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid \mathcal{I}_K)$. However, because sparse early observations provide an inherently ambiguous representation of the underlying physical process, robust classification fundamentally requires understanding both the current morphology and the anticipated future trajectory.

Figure 1 illustrates our proposed "generate-and-evaluate" paradigm. Rather than forcing a direct prediction from sparse data, our framework generates diverse, class-conditional future hypotheses, evaluates their predictive likelihood against sequentially arriving observations, and dynamically updates the class posteriors. For each candidate class $y \in \mathcal{Y}$, a generative module models a distribution of possible future trajectories:

$$\big\{\tilde{\mathcal{F}}_K^{(m,y)}\big\}_{m=1}^{M} \sim p_\omega(\mathcal{F} \mid \mathcal{I}_K, y), \qquad M = 32. \tag{4}$$

Crucially, y here indexes a counterfactual hypothesis ("If this object were class y, what would its future look like?") rather than the classification label. As new photometry arrives, it is evaluated against these simulated futures to form a cumulative predictive evidence state, $e_{\leq K}$. The final classification decision is then fused as:

$$\hat{y}_K = \arg\max_{y \in \mathcal{Y}} p\big(y \mid \mathcal{I}_K, \{\tilde{\mathcal{F}}_K^{(m,y)}\}_{m=1}^{M}, e_{\leq K}\big). \tag{5}$$

Consequently, a single observed prefix supports two complementary mechanisms: a direct class prior based on the prefix alone, and a manifold of simulated futures that serve as testable hypotheses for incoming measurements.

3.2 Physics-Informed Future Light-Curve Generation

To capture the stochastic nature of astronomical evolution, we model the future as a continuous predictive distribution $p_\omega(\mathcal{F} \mid \mathcal{I}_K, y)$. We instantiate the generator as a conditional variational autoencoder (CVAE) Kingma & Welling (2022); Sohn et al. (2015). The prefix encoder projects $\mathcal{P}_K$ into a sequential memory state $\mathbf{H}_K$ and a global summary vector $\mathbf{h}_K$. The condition vector $\mathbf{c}$ encodes the hypothesized class identity and relevant metadata, while $\mathbf{s}_K$ summarizes prefix statistics (band-wise mean, scatter, and detection fractions). The prefix-conditioned prior and the target-informed training posterior are formulated as:

$$p_\psi(z \mid \mathcal{I}_K, y) = \mathcal{N}\big(z; \mu_\psi(\mathbf{h}_K, \mathbf{c}, \mathbf{s}_K), \sigma_\psi^2(\mathbf{h}_K, \mathbf{c}, \mathbf{s}_K)\big), \tag{6}$$

$$q_\phi(z \mid \mathcal{I}_K, \mathcal{F}_K) = \mathcal{N}(z; \mu_\phi, \sigma_\phi^2). \tag{7}$$

The decoder emits a six-band Gaussian distribution directly at specified query times on a fixed 22-point temporal grid spanning 1–120 d after t_K. By evaluating the decoder in a single forward pass, we circumvent the compounding distribution shifts that typically degrade autoregressive forecasting models (decoder parameterization is detailed in Appendix D).

During Stage 1 training, the CVAE minimizes a data-driven reconstruction objective:

$$\mathcal{L}_{\mathrm{S1}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{free}}\mathcal{L}_{\mathrm{Huber}}. \tag{8}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ is the reconstruction loss, $\mathcal{L}_{\mathrm{KL}}$ regularizes the divergence between the posterior as defined in 7 and prior as defined in 6, and $\mathcal{L}_{\mathrm{Huber}}$ provides auxiliary robust error penalization. Because purely data-driven objectives can produce visually plausible but physically impossible trajectories, Stage 2 optimization introduces class-specific physical constraints. This stage structurally enforces domain-aware inductive biases, ensuring the learned representations strictly adhere to astrophysical boundaries Karniadakis et al. (2021) (detailed in Appendix D).

3.3 Multiple Future Hypotheses and Candidate Trajectory Modeling

In the early-time regime, mapping a sparse observational prefix to a complete future trajectory constitutes a highly degenerate, ill-posed inverse problem; a single early prefix often admits multiple distinct evolutionary paths that are all physically viable. Consequently, drawing a single latent vector z from the prior, as defined in 6, would prematurely collapse this rich hypothesis space into a single deterministic mode, fatally discarding the epistemic and aleatoric uncertainties inherent to transient phenomena. To mathematically preserve this uncertainty, each class generator must instead emit a frozen, multimodal ensemble of $M{=}32$ trajectory candidates for a given $\mathcal{I}_K$. To ensure low-discrepancy, stratified coverage of this physically permissible latent space, latents are sampled using scrambled Sobol sequences with antithetic pairing Sobol' (1967). These samples are decoded and interpolated onto a frozen 64-point observer-time grid, yielding a deterministic, information-dense candidate library.

To robustly evaluate this ensemble of trajectories, we deploy a shared, class-agnostic CTM Darlow et al. (2025). We specifically select the CTM for its capacity to perform continuous, iterative latent reasoning; rather than relying on a rigid, single-pass feed-forward aggregation, it dynamically reorganizes temporal features over continuous internal "thinking" steps. This allows the network to comprehensively evaluate the relative plausibility of each simulated candidate against the complex morphological cues within the sparse prefix. The CTM processes the prefix tokens alongside the

M candidates, outputting normalized aggregation weights that represent the relative plausibility of each candidate:

$$w_K^{(m,y)} = \mathrm{CTM}\big(\mathcal{I}_K, \{\tilde{\mathcal{F}}_K^{(\ell,y)}\}_{\ell=1}^M, \mathcal{M}_K^{(y)}\big), \qquad \sum_{m=1}^{M} w_K^{(m,y)} = 1, \quad w_K^{(m,y)} \geq 0. \tag{9}$$

The CTM is supervised by targets derived from the true observed future: candidates that exhibit lower prediction error and higher physical validity receive larger soft target weights. These weights parameterize the candidate mixture in equation 10, and the highest-weighted candidate serves as the representative trajectory for physical-consistency reporting. Target construction and ablation weightings are provided in Appendix E and Table 9.

3.4 Prior-Guided Predictive Evidence Accumulation

The standalone prefix classifier establishes an initial class prior $\pi_K(y) = p_\theta(y \mid \mathcal{I}_K)$. At the very first observation state ($K{=}1$), before any sequential future data is available to test the hypotheses, the final decision naturally reduces to this prior: $\arg\max_y \pi_1(y)$. As the observation state advances from $K{-}1$ to K, newly acquired photometry (denoted as Δ_K) transitions from being an unobserved future state into the observed causal prefix. This newly arrived data acts as the empirical signal required to rigorously test the hypotheses generated at the previous state ($\mathcal{I}_{K-1}$). Each candidate trajectory is scored against Δ_K via a Student-t log-density, denoted as $\ell_K^{(m,y)}$ (Appendix E). By applying the CTM aggregation weights from the previous state, we obtain the unified log-predictive score for the class:

$$s_K(y) = \frac{1}{N_K} \log \sum_{m=1}^{M} w_{K-1}^{(m,y)} \exp\big(\ell_K^{(m,y)}\big), \tag{10}$$

where N_K is the number of valid observation cells in Δ_K. Conceptually, a larger $s_K(y)$ indicates that the simulated hypotheses for class y successfully anticipated the new physical measurements.

We track this predictive success over time. Initializing a cumulative evidence state at $e_1(y) = 0$, the evidence dynamically accumulates for $K \geq 2$:

$$e_K(y) = e_{K-1}(y) + \big(s_K(y) - \bar{s}_K\big), \qquad \bar{s}_K = \frac{1}{|\mathcal{Y}|} \sum_{y' \in \mathcal{Y}} s_K(y'). \tag{11}$$

By centering the predictive score s_K across all classes in $\mathcal{Y}$, the state effectively isolates the relative Bayesian evidence supporting competing physical hypotheses. Finally, we fuse the prefix prior with this accumulated evidence via a learned gating mechanism:

$$\lambda_K(y) = \big(\log \pi_K(y) - \overline{\log \pi_K}\big) + g_K\big(e_K(y) - \overline{e_K}\big), \tag{12}$$

where the gate $g_K \in [0, 1]$ is calibrated on the validation set. The final predicted probability is obtained via a standard softmax operator over $\lambda_K(y)$.

3.5 Training and Inference Procedure

To rigorously enforce causal information boundaries and preclude temporal data leakage, the generative CVAE, the CTM, and the prefix classifier are optimized as fully decoupled modules. The generative CVAE is trained first, progressing from Stage 1 (unconstrained data-driven reconstruction) to Stage 2 (imposition of astrophysical constraints). Upon convergence, the synthesized candidate trajectory library is permanently frozen. Subsequently, the CTM is trained on this static library, utilizing the true observed futures as supervision to learn optimal, physically grounded trajectory rankings. Concurrently, the prefix classifier is independently optimized on the causal prefixes and their corresponding ground-truth taxonomic labels.

During the validation phase, we calibrate the sequential fusion gate g_K in tandem with standard temperature scaling to ensure robust confidence calibration Guo et al. (2017). Once calibrated, all network weights and fusion hyperparameters are strictly locked. Crucially, during the inference stage, the framework strictly emulates the causal deployment setting of real-time survey alerts: the true future trajectory $\mathcal{F}_K$ remains entirely unobserved, and the sequential evidence increment Δ_K is exclusively populated by photometry that has causally materialized by epoch t_K. Comprehensive training schedules and hyperparameter configurations are detailed in Appendix F.

Table 1: Object-level final-K performance on the class-balanced locked test cohort. BLL denotes balanced log loss; arrows indicate the preferred direction, and bold marks the better value.

Method	Top-1 ↑	Top-2 ↑	MRR ↑	BLL ↓
Prefix-only	78.19	91.35	0.8717	0.5915
Prefix + predictive evidence	**78.81**	**91.52**	**0.8753**	**0.5750**

4 Experiments and Results

4.1 Experimental Setup

We evaluate our framework on the class-balanced PLAsTiCC protocol defined in Section 2. To ensure strict causal evaluation, all observation states (K) from a given celestial object are locked to a single partition. The primary evaluation utilizes a locked test cohort of 10,000 objects (1,000 per class, spanning 43,642 object–K states). For each inference step, the generative module produces a frozen library of $M{=}32$ scrambled-Sobol trajectory hypotheses. Sequential evidence parameters and the K-stratified fusion gate were calibrated using five object-level validation folds alongside standard temperature scaling Guo et al. (2017). Once calibrated, these parameters were strictly frozen for a single, definitive replay of the locked test set.

We compare our full generate-and-evaluate framework against a standard Prefix-only classification baseline. Additionally, to isolate the effect of the CTM, we perform ablation studies over the aggregation policy using the same frozen candidate bank (detailed in Table 9). Primary metrics encompass Top-1/Top-2 accuracy, Mean Reciprocal Rank (MRR), and balanced log loss, evaluated at both the object-level (final available K) and stratified across sequential evidence states ($K \in \{1, 2, 3, 4+\}$).

4.2 Overall Early-classification Performance

Integrating predictive generative evidence significantly improved early classification performance across the locked test set (Table 1). Compared to the Prefix-only baseline, our gated evidence fusion increased Top-1 accuracy from 78.19% to 78.81% (a paired gain of 0.62 percentage points; 95% CI: 0.32–0.92). Top-2 accuracy and MRR saw commensurate improvements, indicating that the generative evidence refines both the primary decision boundary and the broader taxonomic ranking.

Crucially, stratifying the performance by state K confirms the causal mechanism of our framework (Figure 2). At $K{=}1$, before any new photometric evidence has materialized to test the generated hypotheses, Top-1 accuracy matches the baseline exactly. However, once causal evidence accumulates, Top-1 accuracy cleanly diverges: improving from 67.48% to 67.81% at $K{=}2$, and from 75.84% to 76.42% at $K{=}3$.

Furthermore, predictive probability quality improved substantially, with overall balanced log loss dropping from 0.5915 to 0.5750 (paired 95% CI: 0.0134–0.0198). In the context of time-domain astronomy, this significant reduction in log loss is highly consequential. Early-time classification is primarily used to trigger scarce, expensive spectroscopic follow-up resources based on confidence thresholds. The improved BLL confirms that evidence fusion does not merely flip marginal hard predictions, but systematically sharpens the posterior density assigned to the true underlying physical class, directly optimizing resource allocation.

4.3 Ablation Experiments and Candidate Trajectory Analysis

To investigate how the network resolves the epistemic uncertainty of the generated manifold, we ablated the candidate aggregation policy while holding the 32-candidate library constant (Table 2).

The CTM-weighted posterior mixture achieved the highest Top-1 accuracy (78.81%), highest MRR (0.8753), and lowest balanced log loss (0.5750) among all evaluated strategies. By contrast, forcing the network to commit to a single deterministic future via hard selection (argmax) significantly degraded Top-1 accuracy to 78.47%, decisively proving that maintaining a broad, multi-hypothesis manifold (Mixture) is strictly superior to prematurely collapsing the uncertainty space.

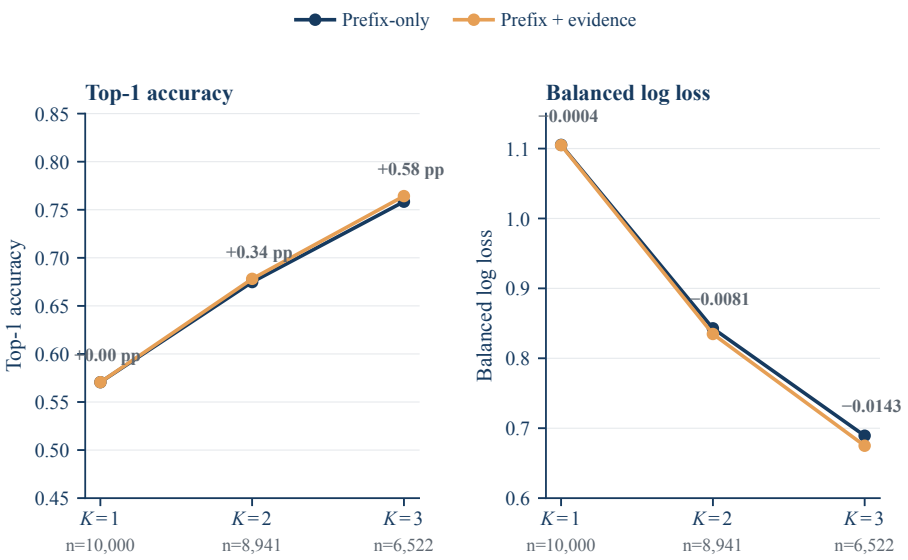

Figure 2: Stage-wise early-classification performance. Top-1 accuracy is unchanged at $K=1$, when no new observation can test a generated future, and increases after the evidence-bearing updates at $K=2$ and $K=3$. Balanced log loss follows the same stage-wise pattern.

Table 2: Candidate aggregation ablation on the same class-balanced locked test cohort. All policies use the same 32-candidate bank and calibrated fusion; BLL denotes balanced log loss.

Strategy	Top-1 ↑	Top-2 ↑	MRR ↑	BLL ↓
CTM posterior mixture	**78.81**	91.52	**0.8753**	**0.5750**
Fixed uniform	78.70	**91.60**	0.8749	0.5755
Hard argmax	78.47	91.59	0.8738	0.5757
Posterior medoid	78.80	91.46	0.8752	0.5753
Physics minimum	78.34	91.46	0.8728	0.5820

Notably, while a fixed-uniform mixture performed competitively in hard classification metrics (78.70% Top-1, with a marginal 0.11 percentage point difference from CTM), the CTM demonstrated its indispensable architectural value in two distinct, AI4Science-critical dimensions. First, the CTM maintained the lowest balanced log loss across all policies, proving that dynamic, observation-conditioned candidate ranking is necessary for optimal probabilistic calibration (a prerequisite for triggering expensive telescope follow-ups). Second, unlike uniform weighting—which blindly treats all hypotheses as black boxes—the CTM's dynamic internal reasoning successfully identifies and isolates the single most plausible representative trajectory. This explicit selection mechanism bridges predictive accuracy with physical interpretability, providing the foundation for the physical consistency verification analyzed in Section 4.5.

4.4 Class-wise Performance and Evidence Evolution

Taxonomic evaluation reveals that the framework easily separates distinct astrophysical mechanisms but appropriately reflects the physical degeneracies inherent to early-time observations (Appendix Table 5). Recall was exceptionally high for distinct phenotypes like SLSN-I (93.6%) and RR Lyrae (93.3%). Conversely, confusion was tightly localized to highly degenerate subclasses: e.g., 153 SN Iax objects were misclassified as SN Ib/c, and 151 SN Ib/c objects as SN Iax (see Appendix Figure 7).

Importantly, generative evidence fusion actively resolved several of these ambiguities. Relative to the Prefix-only baseline, recall improved for SN Iax (56.5% → 61.8%), SN Ia-91bg (78.2% → 80.2%), and SN II (62.0% → 63.9%). This indicates that while early-time morphological ambiguities cannot be entirely eliminated, sequential trajectory testing successfully redistributes probability

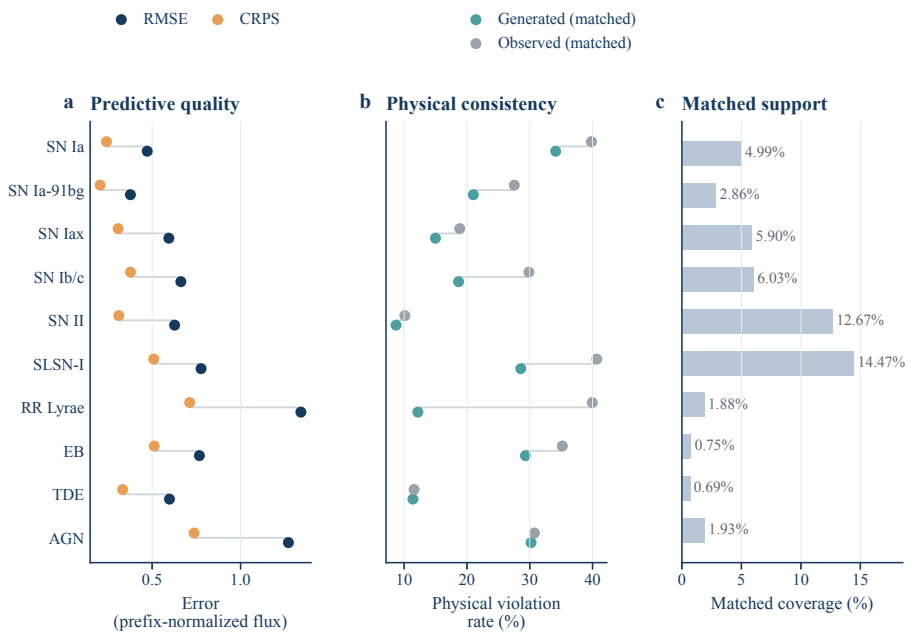

Figure 3: Predictive quality and physical consistency of generated futures. The three panels report (a) predictive trajectory quality, (b) physical-consistency violations, and (c) matched support for the CTM-selected candidates. Physical consistency is evaluated on valid matched checks, and the matched-support fraction is reported with the corresponding comparison.

mass toward the correct subclass within a given morphological family. Appendix Figure 6 visualizes this dynamic evidence accumulation at the case level.

4.5 Quality and Physical Consistency of Generated Futures

Finally, we evaluate the physical realism of the generative CVAE on the class-specific test partitions (125,478 object–K states). Using the maximum-CTM-weight candidate to represent a realizable trajectory, Appendix Table 6 reports prefix-normalized RMSE and Continuous Ranked Probability Score (CRPS) Gneiting & Raftery (2007).

Beyond simple reconstruction error, we quantified the framework's adherence to strict astrophysical boundaries (Appendix Table 7). Astoundingly, when evaluating physical violations at matched observer times, the generated trajectories frequently exhibited *lower* violation rates than the raw, noisy observed photometry itself. Across all matched checks, the generative model yielded a 21.66% violation rate compared to 36.55% for the true observational data—an excess reduction of -14.89 percentage points. Crucially, this reduction is not an artifact of naive low-pass smoothing; rather, it confirms that our Stage-2 physical regularization structurally acts as a powerful domain-aware denoiser. By actively filtering out spurious instrumental artifacts and atmospheric noise (the "dirty" observational data), the generative module ensures that the simulated hypotheses remain strictly bounded by astrophysical laws, thereby supplying the Bayesian classifier with a cleansed, physically rigorous evidence signal.

5 Discussion

This study establishes a fundamental shift in temporal forecasting for AI4Science by reformulating early time-domain classification from a deterministic prediction task into a dynamic, "generate-and-evaluate" Bayesian inference paradigm. Rather than forcing a single, uniquely correct forecast from an impoverished causal prefix, we construct a manifold of class-conditional, physically viable future trajectories. These simulated futures are not treated as absolute predictions; instead, they encode alternative evolutionary hypotheses. As newly arriving photometry materializes, it provides

the empirical signal to rigorously test these hypotheses. The complete absence of an accuracy update at $K=1$, juxtaposed with the significant predictive gains observed at $K=2$ and $K=3$, validates this mechanism: physically regularized futures contribute critical discriminative evidence exclusively when subsequent observations arrive to test them (as illustrated by the transition from unresolved trajectory alternatives to updated class posteriors in Appendix Figures 5 and 6).

The preservation of a diverse, physically grounded candidate distribution lies at the heart of this forecasting method. By utilizing scrambled-Sobol sampling, we maintain 32 low-discrepancy evolutionary alternatives per class, successfully mitigating the premature collapse of epistemic uncertainty caused by sparse data. Crucially, our ablations decisively prove that retaining this multimodal trajectory mixture (Mixture) yields substantially lower balanced log loss than any deterministic, single-trajectory policy (Argmax). Furthermore, while the learned CTM weights provided only marginal gains in hard Top-1 accuracy over uniform weighting, the CTM proved indispensable for optimizing predictive posterior density (BLL) and extracting interpretable, physically consistent representative trajectories from the hypothesis manifold. By structurally filtering out spurious instrumental noise while strictly adhering to astrophysical laws, this model transcends traditional classification. Ultimately, the decisive takeaway is the foundational superiority of our physically regularized, multi-hypothesis probabilistic evaluation over classical deterministic forecasting, establishing a highly robust paradigm for AI4Science.

Analyzing the remaining classification errors highlights both the advantages and the frontiers of embedding physical properties into temporal machine learning models. Generative evidence fusion successfully improved recall for several challenging supernova subclasses specifically because the Stage-2 regularization enforced strict astrophysical boundaries. However, the dominant residual confusions remained tightly localized within distinct morphological families (e.g., SN Iax vs. SN Ib/c, RR Lyrae vs. EB, and TDE vs. AGN). Predictive evidence effectively redistributed probability mass within these families rather than entirely eliminating their intrinsic ambiguities. These edge cases emerge when sparse early observations legitimately admit multiple continuations with overlapping physical properties, or when the generated trajectories diverge from the realized physical pattern (indicated by the positive violation excesses for AGN and SN Ia). Disentangling these causal error contributions will require dedicated future analyses explicitly contrasting our physics-informed generator against a purely data-driven, "no-physics" baseline.

Finally, while this framework establishes a new benchmark on the rigorously controlled PLAsTiCC dataset Kessler et al. (2019), transitioning this AI4Science methodology to live survey operations presents a formidable sim-to-real challenge. Although motivated by the upcoming LSST program Ivezić et al. (2019), direct deployment on live LSST or ZTF alert streams Bellm et al. (2019); Patterson et al. (2019) must contend with severe domain shifts in observational cadence, sensor calibration, class prevalence, data missingness, and atmospheric noise. Consequently, prospective evaluation on real-time alert streams, alongside matched-candidate ablations of the physical priors, form the decisive next steps for validating this causal temporal forecasting paradigm in observational astronomy.

AI USE STATEMENT

We used generative AI tools to assist with language polishing of portions of the manuscript and searches for background information and relevant literature.

REFERENCES

Joseph P. Anderson, Santiago González-Gaitán, Mario Hamuy, Claudia P. Gutiérrez, et al. Characterizing the v-band light-curves of hydrogen-rich type ii supernovae. *The Astrophysical Journal*, 786(1):67, 2014. doi: 10.1088/0004-637X/786/1/67. URL https://doi.org/10.1088/0004-637X/786/1/67.

Eric C. Bellm, Shrinivas R. Kulkarni, Matthew J. Graham, et al. The zwicky transient facility: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, 2019. doi: 10.1088/1538-3873/aaecbe. URL https://doi.org/10.1088/1538-3873/aaecbe.

Luke Darlow, Ciaran Regan, Sebastian Risi, Jeffrey Seely, and Llion Jones. Continuous thought machines, 2025. URL https://arxiv.org/abs/2505.05522v4.

Ryan J. Foley, P. J. Challis, R. Chornock, Mohan Ganeshalingam, et al. Type iax supernovae: A new class of stellar explosion. *The Astrophysical Journal*, 767(1):57, 2013. doi: 10.1088/0004-637X/767/1/57. URL https://doi.org/10.1088/0004-637X/767/1/57.

Avishay Gal-Yam. Luminous supernovae. *Science*, 337(6097):927–932, 2012. doi: 10.1126/science.1203601. URL https://doi.org/10.1126/science.1203601.

Avishay Gal-Yam. Observational and physical classification of supernovae. In *Handbook of Supernovae*, pp. 195–237. Springer, 2017. doi: 10.1007/978-3-319-21846-5_35. URL https://doi.org/10.1007/978-3-319-21846-5_35.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.04599.

Stephan Hachinger, Paolo A. Mazzali, Stefan Taubenberger, Rüdiger Pakmor, and Wolfgang Hillebrandt. Spectral analysis of the 91bg-like type ia sn 2005bl: low luminosity, low velocities, incomplete burning. *Monthly Notices of the Royal Astronomical Society*, 399:1238–1254, 2009. doi: 10.1111/j.1365-2966.2009.15403.x. URL https://doi.org/10.1111/j.1365-2966.2009.15403.x.

Nina Hernitschek and Keivan G. Stassun. The impact of observing strategy on the reliable classification of standard candle stars: Detection of amplitude, period, and phase modulation (blazhko effect) of rr lyrae stars with lsst. *The Astrophysical Journal Supplement Series*, 258(1):4, 2021. doi: 10.3847/1538-4365/ac3baf. URL https://doi.org/10.3847/1538-4365/ac3baf.

Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, et al. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, mar 2019. doi: 10.3847/1538-4357/ab042c. URL https://doi.org/10.3847/1538-4357/ab042c.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5. URL https://doi.org/10.1038/s42254-021-00314-5.

R. Kessler, G. Narayan, A. Avelino, E. Bachelet, R. Biswas, P. J. Brown, D. F. Chernoff, A. J. Connolly, M. Dai, S. Daniel, R. Di Stefano, M. R. Drout, L. Galbany, S. González-Gaitán, M. L. Graham, R. Hložek, E. E. O. Ishida, J. Guillochon, S. W. Jha, D. O. Jones, K. S. Mandel, D. Muthukrishna, A. O'Grady, C. M. Peters, J. R. Pierel, K. A. Ponder, A. Prša, S. Rodney, V. A. Villar, (The LSST Dark Energy Science Collaboration, the Transient, and Variable Stars Science Collaboration). Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501, jul 2019. doi: 10.1088/1538-3873/ab26f1. URL https://doi.org/10.1088/1538-3873/ab26f1.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4a5876b450b45371f6cfe5047ac8cd45-Paper.pdf.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Michelle Lochner, Jason D. McEwen, Hiranya V. Peiris, Ofer Lahav, and Max K. Winter. Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series*, 225(2):31, aug 2016. doi: 10.3847/0067-0049/225/2/31. URL https://doi.org/10.3847/0067-0049/225/2/31.

C. L. MacLeod, Ž. Ivezić, C. S. Kochanek, et al. Modeling the time variability of sdss stripe 82 quasars as a damped random walk. *The Astrophysical Journal*, 721(2):1014–1033, 2010. doi: 10.1088/0004-637X/721/2/1014. URL https://doi.org/10.1088/0004-637X/721/2/1014.

Daniel Muthukrishna, Gautham Narayan, Kaisey S. Mandel, Rahul Biswas, and Renée Hložek. Rapid: Early classification of explosive transients using deep learning. *Publications of the Astronomical Society of the Pacific*, 131(1005):118002, sep 2019. doi: 10.1088/1538-3873/ab1609. URL https://doi.org/10.1088/1538-3873/ab1609.

Maria T. Patterson, Eric C. Bellm, Ben Rusholme, et al. The zwicky transient facility alert distribution system. *Publications of the Astronomical Society of the Pacific*, 131(995):018001, 2019. doi: 10.1088/1538-3873/aae904. URL https://doi.org/10.1088/1538-3873/aae904.

M. M. Phillips. The absolute magnitudes of type ia supernovae. *The Astrophysical Journal*, 413: L105, 1993. doi: 10.1086/186970. URL https://doi.org/10.1086/186970.

Andrej Prša, Natalie Batalha, Robert W. Slawson, et al. Kepler eclipsing binary stars. i. catalog and principal characterization of 1879 eclipsing binaries in the first data release. *The Astronomical Journal*, 141(3):83, 2011. doi: 10.1088/0004-6256/141/3/83. URL https://doi.org/10.1088/0004-6256/141/3/83.

Helen Qu and Masao Sako. Photometric classification of early-time supernova light curves with scone. *The Astronomical Journal*, 163(2):57, jan 2022. doi: 10.3847/1538-3881/ac39a1. URL https://doi.org/10.3847/1538-3881/ac39a1.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9. URL https://doi.org/10.1016/0041-5553(67)90144-9.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

Stefan Taubenberger. The extremes of thermonuclear supernovae. In *Handbook of Supernovae*, pp. 317–373. Springer, 2017. doi: 10.1007/978-3-319-21846-5_37. URL https://doi.org/10.1007/978-3-319-21846-5_37.

Sjoert van Velzen, Thomas W.-S. Holoien, Francesca Onori, Tiara Hung, et al. Optical-ultraviolet tidal disruption events. *Space Science Reviews*, 216(8), 2020. doi: 10.1007/s11214-020-00753-z. URL https://doi.org/10.1007/s11214-020-00753-z.

Table 3: PLAsTiCC target classes and selected object counts. All eligible TDE objects are retained; the other nine classes are sampled for class balance.

Class	Variability type	Selected	Train	Val.	Test
SN Ia	Explosive transient	15,000	10,000	2,000	3,000
SN Ia-91bg	Explosive transient	15,000	10,000	2,000	3,000
SN Iax	Explosive transient	15,000	10,000	2,000	3,000
SN Ib/c	Explosive transient	15,000	10,000	2,000	3,000
SN II	Explosive transient	15,000	10,000	2,000	3,000
SLSN-I	Explosive transient	15,000	10,000	2,000	3,000
RR Lyrae	Periodic variable	15,000	10,000	2,000	3,000
EB	Periodic variable	15,000	10,000	2,000	3,000
TDE	Stochastic / long-timescale	12,071	8,450	1,811	1,810
AGN	Stochastic / long-timescale	15,000	10,000	2,000	3,000
Total	–	147,071	98,450	19,811	28,810

A Dataset Preprocessing and Splitting Protocols

A sequential photometric measurement is retained strictly if the timestamp, flux, and uncertainty are finite, the uncertainty is positive, and the passband index falls within $\{0, \ldots, 5\}$. To construct robust input representations, raw photometry is grouped into irregular observation nodes. Repeated measurements within a temporal node are combined via optimal inverse-variance weighting, mitigating localized instrumental noise:

$$\sigma_f = \Big(\sum_i w_i\Big)^{-1/2}, \qquad w_i = \sigma_i^{-2}. \tag{13}$$

To ensure rigorous early-time evaluation, a celestial object is deemed eligible only if it satisfies the following criteria: at least one valid detection, baseline photometry extending $20\,\mathrm{d}$ prior to the first detection, a strictly positive prefix-derived scale, and the existence of subsequent valid measurements to supervise the $120\,\mathrm{d}$ future generation grid. To categorically preclude any temporal or object-level data leakage, data splits are deterministically enforced at the object level using a SHA-256 hash function evaluated on the tuple $(\mathrm{seed}, \mathrm{class\ ID}, \mathrm{object\ ID})$ with $\mathrm{seed} = 42$. This cryptographic hashing guarantees that all sequential K states of a given object remain strictly confined to a single immutable partition. The final selected class distributions are reported in Table 3.

B Experimental Configuration and Extended Results

To ensure full reproducibility and to provide a granular diagnostic of our generate-and-evaluate framework, this section details the comprehensive experimental configurations and reports the extended, class-wise classification metrics.

B.1 Configuration and Ablation Protocol

Table 4 summarizes the exact experimental pipeline and evaluation protocols utilized for the ten-class early-classification task. To rigorously isolate the contribution of the Continuous Thought Machine (CTM) in resolving epistemic uncertainty, we designed the evaluation such that the 32-candidate generative bank, the prefix classifier, and the calibration data remain strictly identical across all ablations. Consequently, the compared baseline policies (Uniform mixture, Hard selection, Posterior medoid, and Physics-minimum) solely perturb the downstream candidate aggregation strategy. Furthermore, we explicitly emphasize the decoupled nature of our validation phase: the five object-level folds used for evidence fitting are completely disjoint from the validation-calibration split utilized to lock the sequential fusion gates (g_K) and temperature scales. This ensures that the final replay on the locked test cohort represents a strictly unbiased out-of-sample evaluation.

Table 4: Experimental configuration for the ten-class early-classification evaluation.

Component	Setting
Evaluation unit	Object-level final-K prediction; row-level object–K results are auxiliary
Locked test cohort	10 classes, 1,000 objects per class, 10,000 objects and 43,642 object–K states
Candidate bank	32 scrambled-Sobol hypotheses per class on a frozen 64-point grid through $+120\,\mathrm{d}$
Primary policy	Current-prefix logits plus increment-only evidence from the CTM posterior mixture
Comparisons	Prefix-only, uniform mixture, hard selection, posterior medoid, and physics-minimum diagnostic
Validation	Five object-level folds for evidence fitting; disjoint validation-calibration split for gates, temperatures, and class residuals
Metrics	Top-1, Top-2, MRR, balanced log loss, per-class recall and precision, and paired bootstrap 95% intervals

B.2 Taxonomic Resolution and Physical Degeneracies

While the object-level Top-1 accuracy provides a macroscopic view of model performance, Table 5 provides a micro-level, class-wise diagnostic of the framework's taxonomic resolution on the locked test cohort.

As anticipated by astrophysical domain knowledge, the predictive performance strongly correlates with the inherent morphological distinctiveness of the transient phenomena. The framework achieves exceptionally high F1 scores for classes characterized by unambiguous temporal signatures, such as strictly periodic variables (RR Lyrae: 93.1%, EB: 93.1%) and highly energetic Superluminous Supernovae (SLSN-I: 88.1%).

Conversely, the model faces the highest epistemic uncertainty when distinguishing within the structurally similar Type I supernova family (e.g., SN Iax with an F1 of 61.1%, and SN Ib/c with an F1 of 62.8%). This localized confusion is not an algorithmic failure, but rather a faithful reflection of physical reality: in their early epochs, these explosive transients exhibit highly degenerate photometric trajectories driven by overlapping thermodynamic and radiative processes. Despite this inherent physical ambiguity, the generative predictive evidence actively redistributes probability mass within these morphological families to optimize the final balanced log loss, demonstrating the robust domain-awareness of the classification framework.

Table 5: Per-class performance of the Prefix + predictive evidence model on the class-balanced locked test cohort. F1 is the harmonic mean of the reported recall and precision and is shown as a derived summary.

Class	Recall ↑	Precision ↑	F1 ↑
SN Ia	81.6	71.6	76.3
SN Ia-91bg	80.2	87.1	83.5
SN Iax	61.8	60.4	61.1
SN Ib/c	63.3	62.3	62.8
SN II	63.9	75.6	69.3
SLSN-I	93.6	83.3	88.1
RR Lyrae	93.3	92.9	93.1
EB	92.9	93.3	93.1
TDE	85.5	77.7	81.4
AGN	72.0	86.6	78.6
Macro average	78.8	79.1	78.7

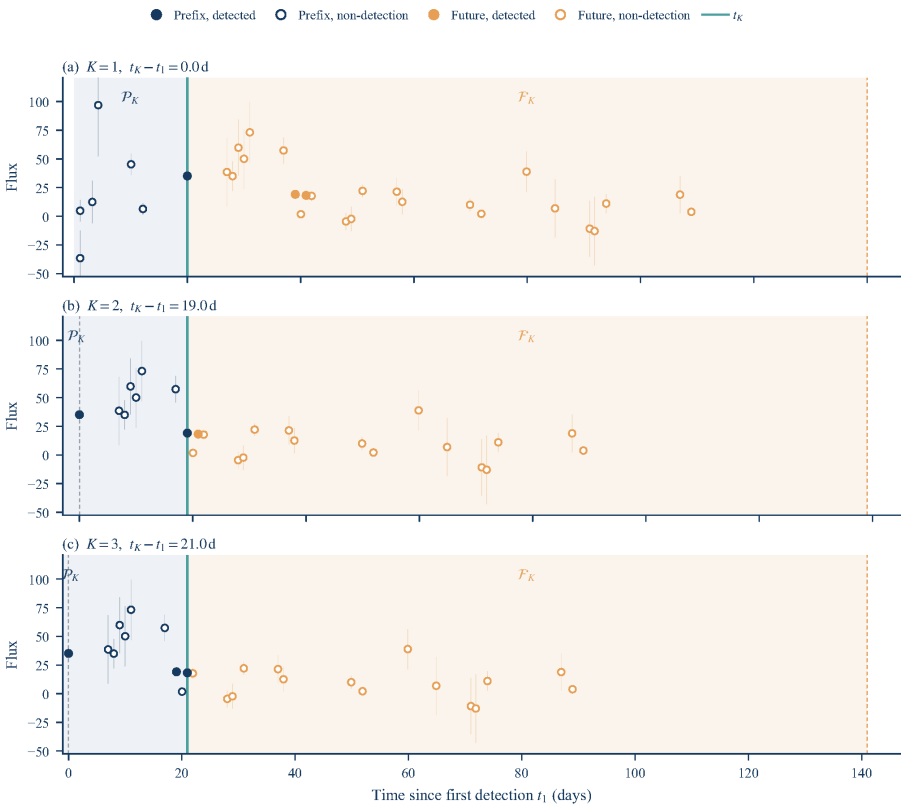

Figure 4: Early-observation protocol. Each state is indexed by the detection night K. The prefix contains observations available at t_K, whereas the future window is used only as generator supervision.

C Supplementary Figures and Diagnostics

To provide a deeper, transparent view into the internal mechanics of our generate-and-evaluate framework, this section visualizes the causal evaluation protocol, the dynamic evolution of predictive evidence, and the physical consistency of the generated trajectories.

C.1 Causal Protocol and Temporal Masking

Figure 4 illustrates the strict causal masking protocol employed during sequential inference. By indexing observations strictly by the detection night K, we enforce a rigid boundary between the known causal prefix ($\mathcal{P}_K$) and the unobserved future ($\mathcal{F}_K$). This visualization confirms that our evaluation strictly emulates realistic survey deployment, ensuring that no future photometric information inadvertently leaks into the prefix classifier or the candidate hypothesis evaluation at state K.

C.2 Visualizing the Hypothesis Manifold and Evidence Accumulation

The core theoretical claim of this paper—that preserving a multi-hypothesis manifold is superior to deterministic forecasting—is visually validated in Figure 5 and Figure 6. Figure 5 showcases the M=32 generated trajectory candidates for representative SN Ia, TDE, and AGN cases. Rather than collapsing to a single mean line, the generative CVAE explicitly models the epistemic uncertainty of the early-time prefix, fanning out into diverse yet physically plausible evolutionary paths.

As time advances, Figure 6 demonstrates how this generated manifold drives the sequential Bayesian update. At early states ($K = 1$), the model relies heavily on the prefix-only prior. However, as new photometry incrementally arrives, the evidence ledger dynamically accumulates positive or negative

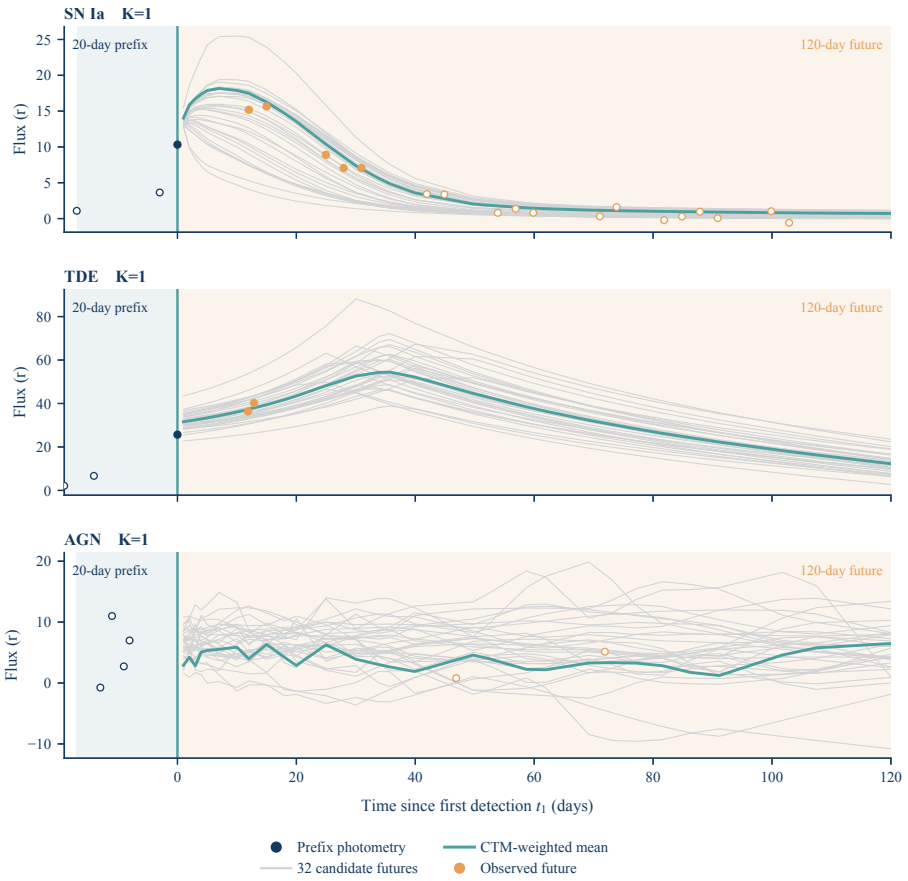

Figure 5: Generated future trajectory hypotheses. The candidate set, CTM-weighted mean, and later photometry are shown for representative SN Ia, TDE and AGN cases.

log-predictive scores. The gated fusion smoothly transitions the probability mass, illustrating how raw photometric data is mathematically converted into discriminative Bayesian evidence to refine the final class posteriors.

C.3 Taxonomic Confusion and Physical Degeneracy

Figure 7 presents the detailed 10×10 confusion matrix on the locked test cohort. From an AI4Science perspective, the residual errors are highly informative: they are not randomly distributed, but are strictly bounded within distinct astrophysical families. The most prominent block-diagonal confusions occur between SN Iax and SN Ib/c (thermonuclear vs. stripped-envelope core-collapse supernovae with similar early rise times), RR Lyrae and EB (periodic variables), and TDE and AGN (stochastic nuclear transients). This indicates that while the framework successfully distinguishes fundamentally different physical mechanisms, it correctly preserves the inherent observational degeneracies of early-time astrophysics where distinct subclasses share highly overlapping morphological prefixes.

C.4 Trajectory Quality and Domain-Aware Denoising

Tables 6 and 7 evaluate the physical realism of the CTM-selected representative trajectories. Table 6 establishes the baseline generative accuracy via standard RMSE and CRPS metrics across all ten classes.

More importantly, Table 7 provides critical proof of the framework's capacity for domain-aware denoising. When comparing the physical constraint violations of the generated trajectories against

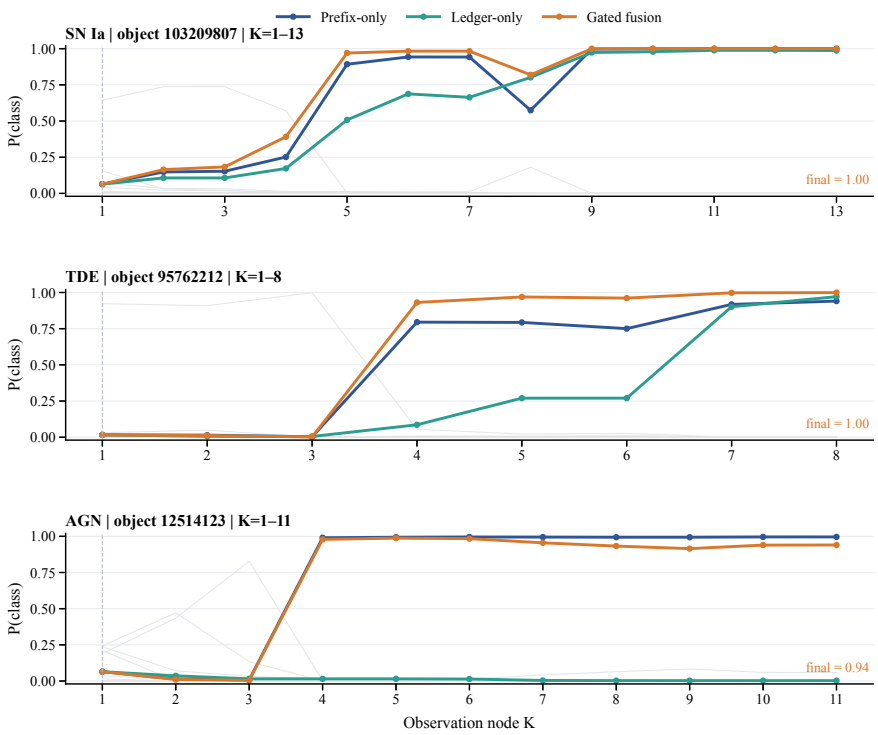

Figure 6: Case-level predictive evidence accumulation. Prefix probabilities, the evidence ledger, and gated fused probabilities are shown as additional observations arrive.

Table 6: Predictive trajectory quality of the maximum-CTM-weight candidate on the generator test partitions. RMSE and CRPS use prefix-normalized flux. Generated violation and generated valid coverage are computed on the fixed future query grid.

Class	RMSE ↓	CRPS ↓	Generated violation (%)	Generated valid cov. (%)
SN Ia	0.4352	0.2658	24.57	56.25
SN Ia-91bg	0.3868	0.2438	10.19	80.57
SN Iax	0.5409	0.3246	12.78	91.15
SN Ib/c	0.6429	0.4225	13.39	90.65
SN II	0.6064	0.3710	7.28	97.24
SLSN-I	0.7720	0.5877	32.02	32.63
RR Lyrae	1.4113	0.9321	0.55	99.92
EB	0.6609	0.5018	0.11	99.90
TDE	0.6061	0.3722	42.62	64.82
AGN	1.2129	0.7657	4.23	57.19
Overall (fixed grid)	–	–	15.16	75.88

the real observed curves at matched timestamps, the generative model frequently exhibits a *negative* excess (e.g., RR Lyrae: -24.91%, SLSN-I: -7.18%). Overall, the generated futures yield a 21.66% violation rate compared to 36.55% for the raw observations. This quantitatively proves that our Stage-2 physical regularization does not merely fit the data; it actively filters out non-physical instrumental artifacts and atmospheric noise, ensuring the downstream Bayesian classifier evaluates hypotheses that are strictly bound by astrophysical reality.

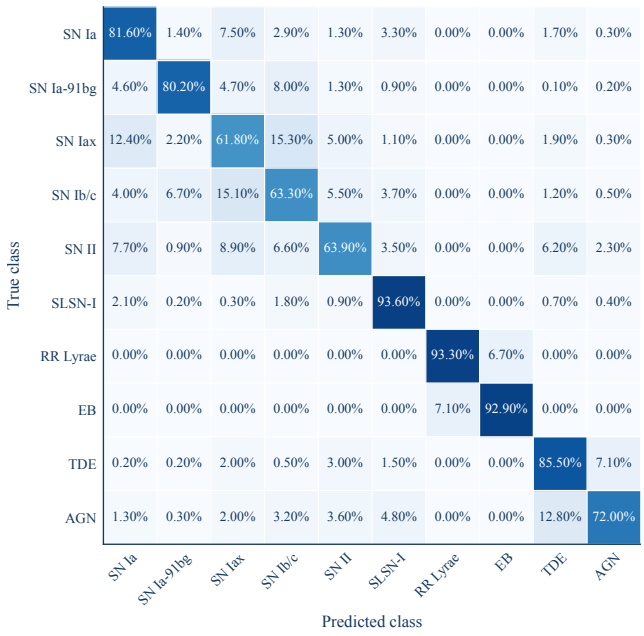

Figure 7: Confusion matrix on the class-balanced locked test cohort. The largest exchanges occur between morphologically related classes, including SN Iax and SN Ib/c, SN Ia and SN Iax, RR Lyrae and EB, and TDE and AGN.

Table 7: Matched physical consistency between the generated trajectory and the observed curve at deployment observation times. Excess is generated minus observed violation rate in percentage points; matched coverage is the fraction of checks valid for both curves.

Class	Observed violation (%)	Generated violation (%)	Excess (pp)	Matched cov. (%)
SN Ia	40.65	44.02	+3.37	4.78
SN Ia-91bg	28.15	24.19	-3.96	2.81
SN Iax	19.02	14.34	-4.68	5.63
SN Ib/c	30.15	19.00	-11.15	5.83
SN II	10.43	10.22	-0.21	11.85
SLSN-I	41.54	34.36	-7.18	13.89
RR Lyrae	39.69	14.78	-24.91	1.72
EB	34.42	33.99	-0.43	0.69
TDE	11.13	10.94	-0.19	0.68
AGN	30.80	34.77	+3.97	2.01
Overall (matched)	36.55	21.66	-14.89	1.85

D GENERATOR DETAILS

At any queried deployment time t, the generative decoder models the joint multiband flux distribution as a factorized Gaussian, conditioned on the sampled latent state z and the causal prefix $\mathcal{I}_K$:

$$p_\omega(\mathbf{f}_t \mid z, \mathcal{I}_K) = \prod_{b=1}^{6} \mathcal{N}(f_{t,b}; \mu_{t,b}, \sigma_{t,b}^2). \tag{14}$$

To faithfully capture distinct astrophysical mechanisms, the latent representations and network heads are explicitly tailored to specific transient families. For strictly periodic variables (RR Lyrae and EB), the latent variable z is constrained to encode canonical period, amplitude, and phase dynamics. Conversely, long-timescale stochastic sources (TDE and AGN) are driven by turbulent accretion processes; consequently, their generators employ an auxiliary stochastic-process head to model inherent red-noise variance rather than deterministic morphological evolution.

Table 8: Class-dependent physical regularization. Penalties are evaluated only at valid query times.

Class	Reference frame	Physical constraints
SN Ia	Peak-relative	Peak magnitude, Δm_{15}, late-time slope, and calibrated SN Ia ranges Phillips (1993); Gal-Yam (2017).
SN Ia-91bg	Peak-relative	Peak magnitude, Δm_{15}, secondary-peak suppression, and faster decline Taubenberger (2017); Hachinger et al. (2009).
SN Iax	Peak-relative	Peak magnitude, Δm_{15}, late decline; broader peak and slower range Foley et al. (2013).
SN Ib/c	Peak-relative	Peak magnitude, Δm_{15}, late decline, red-band secondary-peak suppression Gal-Yam (2017).
SN II	Peak-relative	Peak range and IIP/IIL/IIN morphology Gal-Yam (2017); Anderson et al. (2014).
SLSN-I	Peak-relative	Peak luminosity, width, long-duration decline Gal-Yam (2012).
RR Lyrae	Phase-relative	Recurrence, cross-band phase, amplitude Hernitschek & Stassun (2021).
EB	Phase-relative	Recurrence, eclipse timescale, eclipse depth Prša et al. (2011).
TDE	Peak-relative	Post-peak decline, rebrightening suppression, power-law tail, color van Velzen et al. (2020).
AGN	Observer timescale	Structure-function envelope, stochastic amplitude, cross-band coherence MacLeod et al. (2010); Kessler et al. (2019);the empirical bluer-when-brighter term is an implementation-specific color constraint.

Each class-conditional generator is optimized independently. Stage 1 focuses purely on data-driven representation learning via direct-grid reconstruction, utilizing both prior and posterior latent samples. To mitigate premature latent posterior collapse, the KL divergence term in equation 8 incorporates a free-bits hinge loss coupled with a 30-epoch monotonic warmup schedule.

Because purely data-driven objectives can synthesize trajectories that violate fundamental thermodynamic or radiative limits, Stage 2 introduces differentiable physical regularization and causal temporal anchoring:

$$\mathcal{L}_{\mathrm{S2}} = \mathcal{L}_{\mathrm{S1}} + \lambda_{\mathrm{phy}}\mathcal{L}_{\mathrm{phy}} + \lambda_{\mathrm{repeak}}\mathcal{L}_{\mathrm{repeak}}. \tag{15}$$

$\mathcal{L}_{\mathrm{phy}}$ is summed only on valid query times. $\mathcal{L}_{\mathrm{repeak}}$ ties the predicted peak to a causal anchor estimated from $\mathcal{P}_K$ at deployment. Table 8 lists the primary constraints.

E Candidate Trajectory Modeling and Robust Evidence Formulation

To robustly approximate the multimodal predictive distribution for a hypothesized class y, we draw an ensemble of $M{=}32$ candidates:

$$z^{(m,y)} = \mu_\psi^{(y)} + \sigma_\psi^{(y)} \odot \varepsilon^{(m)}, \qquad m = 1, \ldots, M, \tag{16}$$

Crucially, naive Monte Carlo sampling in high-dimensional latent spaces suffers from high variance and sparse coverage. To optimize the Bayesian evidence estimation, we formulate this as a Quasi-Monte Carlo (QMC) process. The noise vectors $\varepsilon^{(m)}$ are drawn using $M/2$ scrambled Sobol points paired with their antithetic complements $1 - u$, mathematically mapped via the Gaussian quantile function. This ensures a low-discrepancy, variance-reduced stratification of the physically permissible latent space. Decoding follows Section 3.2, with interpolation anchoring the curves onto the frozen 64-point CTM grid.

When new photometry Δ_K arrives, candidate fluxes on the grid are interpolated to the newly observed timestamps. To ensure Bayesian robustness against heteroscedastic, non-Gaussian instrumental outliers (e.g., cosmic ray hits or transient atmospheric anomalies), each valid observational cell is scored using a heavy-tailed Student-t density with $\nu{=}3$ degrees of freedom. The candidate score $\ell_K^{(m,y)}$ is the sum of these robust cell log-densities. The accumulated mixture in equation 10 is explicitly normalized by N_K to render asynchronous observing nights with varying numbers of measurements mathematically comparable.

To train the CTM, supervision is derived exclusively from the unobserved future. To align the latent representations, we construct a soft oracle target that balances prediction accuracy with physical va-

lidity. For candidate m, let d_m^{pred} denote the predictive divergence (integrating the residual quadratic form and empirical CRPS Gneiting & Raftery (2007)), and d_m^{phy} denote the class-specific physical violation cost when the validity mask is present. These are normalized by positive robust scale factors r_{pred} and r_{phy}. The joint quality cost and the resulting soft oracle target are:

$$c_m = \frac{d_m^{\text{pred}}}{r_{\text{pred}}} + \lambda_{\text{phy}} \frac{d_m^{\text{phy}}}{r_{\text{phy}}}, \qquad q_m = \frac{\exp(-c_m/\tau_{\text{rank}})}{\sum_{j \in \mathcal{V}} \exp(-c_j/\tau_{\text{rank}})}, \tag{17}$$

where $\mathcal{V}$ restricts the softmax strictly to candidates with finite prediction costs and valid physical supervision; if no physical feature is valid, the second penalty is omitted. This soft target explicitly forces the CTM to seek a Pareto equilibrium between data-fidelity and astrophysical laws. The CTM's final aggregated weights are supervised via a Kullback-Leibler ranking loss:

$$\mathcal{L}_{\text{rank}} = -\sum_{m \in \mathcal{V}} q_m \log w^{(m)}. \tag{18}$$

To stabilize this complex alignment across multiple "thought ticks," the Stage 3 optimization employs a balanced multi-objective scheme incorporating posterior prediction, expected joint risk, entropy matching, and auxiliary tick-level penalties. The deployment representative trajectory used for diagnostic reporting is isolated deterministically via:

$$m^* = \arg\max_{m \in \mathcal{V}} w^{(m)}. \tag{19}$$

The formal configuration utilizes $M = 32$ candidates, 10 CTM thought ticks, $\tau_{\text{rank}} = 0.2$, a tick-target temperature of 0.50, and a five-epoch physical penalty warmup. The multi-objective coefficients are strictly balanced as: $\lambda_{\text{prediction}} = 1$, $\lambda_{\text{CRPS}} = 0.5$, $\lambda_{\text{rank}} = 0.25$, $\lambda_{\text{physics-rank}} = 0.80$, $\lambda_{\text{expected-risk}} = 0.25$, $\lambda_{\text{tick-prediction}} = 0.25$, $\lambda_{\text{tick-physics}} = 0.25$, $\lambda_{\text{tick-aggregation}} = 0.25$, and $\lambda_{\text{entropy}} = \lambda_{\text{regret}} = 0.05$. The remaining auxiliary coefficients follow the fixed configuration deployed for the locked replay. Table 9 details the ablation weightings applied to the identical 32-candidate library. The primary proposed policy corresponds to the CTM posterior mixture as defined in 9.

Table 9: Weighting strategies on the same 32-candidate library. Main results use CTM posterior mixture.

Strategy	Weight	Role
CTM posterior mixture	equation 9	Main policy: relative weights on the set
Uniform mixture	$w^{(m)} = 1/M$	Multiple hypotheses without CTM
Hard selection	$w^{(m^*)} = 1$, $m^* = \arg\max_m w^{(m)}$	Collapse to one hypothesis
Posterior medoid	Trajectory nearest the weighted mean	Single representative curve
Physical minimum cost	Lowest Stage-2 physical penalty	Physics ranking instead of CTM

F Calibration and Decoupled Training Schedule

To rigorously preserve causal information boundaries and preclude temporal data leakage, the training pipeline is strictly decoupled. Stage 1 and Stage 2 of the generative CVAE are optimized sequentially and subsequently frozen. The CTM is then trained independently on the resulting static candidate library. Crucially, the unobserved true future is utilized exclusively as a training oracle to compute the supervision targets. As formalized in equation 17, this soft target constructs a Pareto-optimal signal proportional to the exponential of the negative joint cost, explicitly balancing normalized predictive divergence with class-specific astrophysical boundary constraints.

Through this supervision, the CTM jointly optimizes its internal, step-wise reasoning mechanism (tick-level weights) and the final aggregated posterior. For operational deployment, our policy bifurcates the utility of these CTM outputs: the full aggregated posterior mixture drives the Bayesian predictive evidence (preserving multi-hypothesis uncertainty), whereas the single maximum-weight candidate is isolated strictly for deterministic physical consistency diagnostics. Concurrently, the

prefix classifier is optimized purely on the causal prefixes, deployment-available metadata, and ground-truth taxonomic labels, ensuring it remains entirely agnostic to the generative trajectory hypotheses during training.

Finally, a distinct validation split is employed for post-hoc confidence calibration. We fit the sequence-aware fusion gate g_K, temperature scaling factors for both the prefix and evidence logits, and class-specific residuals, dynamically stratified by the causal observation state K. Post-calibration, all model parameters and hyperparameters are immutably locked. The locked test set is then subjected to a single, rigorous forward pass, strictly evaluating the fused probabilities under emulated real-time streaming constraints, ensuring a robust, out-of-sample assessment of the framework.