RESOLVING LATENT REPRESENTATION COLLAPSE IN SCHRÖDINGER BRIDGES FOR MULTI-TASK INFERENCE AND CALIBRATED FORECASTING OF PANEL TIME SERIES: AN APPLICATION TO ASTROPHYSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generative modeling via paired Schrödinger Bridges faces severe limitations in sequential forecasting of sparse, irregularly sampled panel time series. In streaming contexts – such as real-time multi-band supernova light curve monitoring – continuous, noisy non-detections during the critical pre-peak phase creates an asymmetric many-to-one mapping deficiency. This sparse conditioning triggers latent representation collapse, leading to miscalibrated uncertainty intervals and severe classification prior saturation toward dominant classes. In this work, we present a framework that maps panel time series into a continuous hidden space, advancing a hidden-layer dynamic variance Image-to-Image Schrödinger Bridge (I2SB) solver to model extension intervals without data-level noise corruption. We solve the uncertainty quantification task by generating parametric prediction intervals derived directly from the hidden-space I2SB trajectories paired with a post-hoc fine-tuned UQNet uncertainty network, cleanly decoupling trajectory synthesis from downstream tasks. Crucially, categorical classification is resolved via a novel pre-softmax logit fusion technique at the terminal boundary. This architecture ensembles an abstract latent-space energy-based model prior with a standalone temporal neural network trained directly on uncompressed temporal flux trajectories using class-explicit inverse-frequency weights. Concurrently, light-curve-driven redshift estimation operates as a completely independent, parallel channel. Evaluated on highly imbalanced astrophysical benchmarks, our joint pipeline significantly improves early-stage calibration – substantially improving empirical coverage in the pre-peak regime – while maximizing macro classification F1-score and reducing prior saturation.

1 INTRODUCTION

Conditional generative modeling has emerged as a cornerstone paradigm for learning complex mappings between paired, non-stationary data distributions across a wealth of domains, ranging from computer vision to structural biology. Recently, frameworks such as Image-to-Image Schrödinger Bridges (I2SB) (Liu et al., 2023) have demonstrated exceptional capability in modeling continuous-time transport dynamics between disparate distributions by casting generative synthesis as a diffusion-based bridge problem. While highly effective for stationary settings or dense, regularly sampled images, adapting these architectures to the sequential forecasting of irregularly sampled, sparse panel time series exposes an inherent topological bottleneck. In real-time streaming contexts, such systems face a severe *many-to-one non-injective challenge*: a highly asymmetric, nested sequence of increasingly long historical prefixes all map forward to a single, full-trajectory target that itself remains fundamentally incomplete and heteroscedastic in its sampling.

Prior attempts to resolve this representation degeneracy typically use strict multi-task regularizers or joint latent-space classifiers (e.g. Assran et al., 2023) that evaluate the compressed features directly at the generative finish line. However, because the high-entropy pre-peak latent space is inherently saturated, these models succumb to strong historical class imbalances. When predicting within a collapsed latent manifold, standard activation layers lose geometric resolution, causing the model to

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

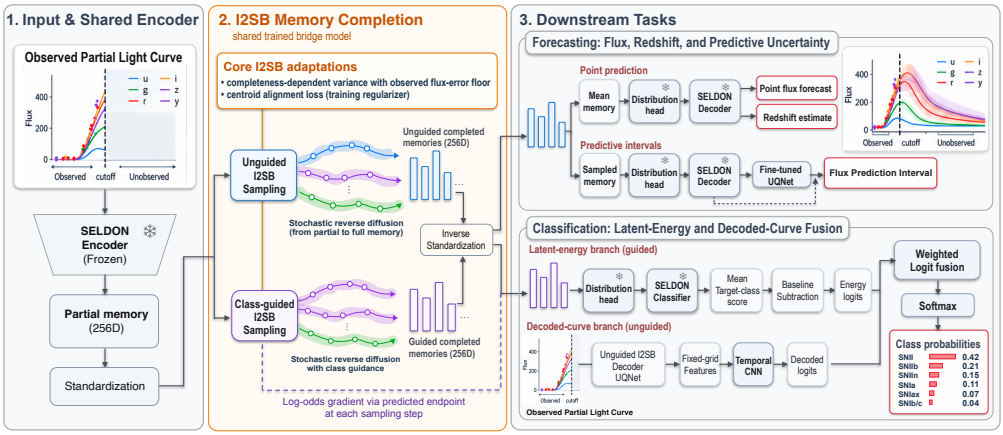

Figure 1: Overview of the proposed framework.

output heavily skewed, overconfident classification bubbles or default completely to uninformative background priors.

To address these limitations, we introduce a modular framework that decouples generative trajectory synthesis from downstream tasks in continuous hidden space; see Figure 1. To resolve the architectural many-to-one collapse under high data sparsity, we enforce a geometric centroid alignment regularizer within the latent space, paired with a heteroscedastic dynamic variance schedule for the hidden-space I2SB solver. This generative track is bounded by a scalar floor directly derived from the mean observed telescope flux error, protecting the core transport dynamics from being corrupted by physical noise floors. We evaluate the continuous hidden trajectories using a post-hoc fine-tuned UQNet to output high-fidelity future mean curves and calibrated parametric prediction intervals.

Crucially, we break the pre-peak classification deadlock by abandoning pure latent-space prediction. Instead, we propose a novel pre-softmax logit fusion approach executed at the terminal bridge boundary. This decision-layer framework functions as a linear opinion pool of two complementary experts: an abstract latent-space energy-based model (EBM Xie et al., 2016) that acts as a structural prior filter ($S_y - \bar{S}_y$), and a standalone temporal convolutional neural network (CNN) trained directly on the uncompressed temporal flux trajectories output by the decoder. By optimizing the CNN on raw, uncompressed physical curves scaled by class-explicit inverse-frequency weights, we allow explicit future shape hypotheses – such as the rapid thermodynamic rise of a Type Ia supernova – to directly out-vote the biased latent prior. Light-curve-driven photometric redshift estimation is an independent parallel processing channel to ensure a clean separation of independent physical tasks.

Evaluated on large-scale astrophysical transient benchmarks, our joint pipeline demonstrates improvements across forecasting, uncertainty quantification, and classification. Out I2SB+UQNet successfully delivers empirical coverage probabilities in the challenging pre-peak regime. Our pre-softmax logit fusion framework further improves pre-peak macro F1 and calibration relative to the latent-energy branch, while improving class-wise recall for several supernova classes.

Our core contributions can be summarized as follows: (1) We present a unified modular framework in Figure 1 that maps sparse panel time series into a continuous hidden dimension space, advancing a dynamic variance I2SB solver to model extension intervals without data-level noise corruption. (2) We decouple generative uncertainty quantification (UQ) from downstream identification by pairing hidden-space trajectories with a fine-tuned UQNet to output calibrated parametric prediction intervals. (3) We introduce a pre-softmax logit fusion framework that blends abstract EBM priors with standalone temporal classifiers operating directly on uncompressed trajectories, resolving latent representation collapse and eliminating class prior saturation. (4) We validate our model on highly imbalanced astrophysical benchmarks, demonstrating concurrent breakthroughs in early-stage empirical UQ coverage and pre-peak macro classification F1-score.

We use the astrophysical dataset primarily for illustrative purposes; however, the framework itself is domain-agnostic and directly transferable to other panel time series contexts.

2 RELATED WORKS

Continuous-Time Neural Representations for Irregular Time Series Modeling irregularly sampled, sparse, and heteroscedastic panel time series remains a foundational challenge in deep learning. Traditional recurrent networks (RNNs) struggle with non-uniform continuous-time steps, leading to the development of Neural ODEs (Chen et al., 2018) and continuous-time architectures, e.g., GRU-ODE-Bayes (De Brouwer et al., 2019). These architectures treat state trajectories as continuous latent flows driven by parameterized vector fields, inherently accounting for missing intervals. In observational astrophysics, this framework was adapted by a pipeline called Supernova Explosions Learned by Deep ODE Networks (SELDON; Wu et al., 2026) to learn fixed representation backbones optimized for multi-band light curve interpolation and continuous-time reconstruction. While continuous encoders excel at summarizing historical data, forcing them to compress sparse, high-entropy prefixes – such as pre-peak transient timelines dominated by non-detections – triggers severe latent representation collapse. We present a new framework that builds upon these continuous-time backbones but shifts the task of trajectory extension into an expressive, hidden dimension space to protect downstream dynamics from early-stage data-level degradation.

Generative Bridge Models and Image-to-Image Diffusion Generative diffusion models (Ho et al., 2020; Song et al., 2020) have redefined conditional synthesis by framing distribution mapping as a stochastic drift process. While standard diffusion setups rely on Gaussian noise limits, Schrödinger Bridge formulations and I2SB (Liu et al., 2023) simulate direct transport dynamics between two non-trivial paired data endpoints. This framework is highly effective for image translation or fixed-grid infilling, but faces strict non-injective mapping challenges when applied to time-series forecasting, where a nested sequence of varying historical prefixes maps onto a single full-trajectory target. Related sequential diffusion approaches often suffer from error accumulation or are heavily disrupted by physical noise floors when operating directly in data-level coordinates (Tashiro et al., 2021). We resolve this limitation by implementing a heteroscedastic dynamic variance schedule for the hidden-space I2SB solver, explicitly bounded by an observational flux noise floor to decouple generation from raw physical noise.

UQ in Sequence Generation Robust UQ is critical for streaming time-series deployment, where downstream tasks depend on the reliability of forecasting and their prediction intervals. While vanilla diffusion models offer non-parametric sampling, they lack direct calibration guarantees under highly sparse conditioning. Post-hoc conformal prediction and other approaches (e.g., UQNet architectures) have been utilized to enforce explicit coverage bounds in deep regression systems (Shafer & Vovk, 2007; Nix & Weigend, 1994; Pearce et al., 2018). We decouple the generative trajectory synthesis of the Schrödinger Bridge from parametric UQ estimation by building a post-hoc fine-tuned UQNet. This joint integration produces parametric prediction intervals that are explicitly optimized to match nominal empirical coverage probabilities, even during high-uncertainty early-phase non-detections.

Decision Fusion and Imbalanced Categorical Classification Multi-class transient classification at early observational phases is severely hindered by class imbalances and prior saturation (Charnock & Moss, 2017; Fernandes et al., 2025). Standard architectures that execute classification via joint latent heads or post-generation re-encoding frequently smooth over distinct morphological variations, collapsing predictions into dominant cosmic abundance baselines. Ensembling techniques typically utilize Product of Experts (PoE) formulations or late decision fusion in probability space to mitigate individual classifier biases (Hinton, 2002). However, standard probability averaging suffers from gradient saturation when models are heavily overconfident. Our proposed pre-softmax logit fusion framework acts as a linear opinion pool directly in logit space. By ensembling a calibrated latent-space EBM prior with a standalone temporal classifier operating on uncompressed decoded curves, we allow explicit geometric future shape hypotheses to directly out-vote biased latent priors without losing linear separability.

3 METHODS

As illustrated in Figure 1, a partial light curve is mapped into a continuous hidden representation using the pretrained SELDON encoder (Wu et al., 2026). Its masked GRU-ODE architecture processes the sparse, heteroscedastic observations in continuous time and produces a fixed-length mem-

ory representation that serves as the initial boundary of the bridge. Our newly developed framework completes this partial representation using I2SB and leverages the generated representations for downstream forecasting, UQ, and classification. In the following sections, we first describe the frozen SELDON backbone, followed by the hidden-space partial-to-full bridge and its downstream evaluation components.

3.1 Frozen SELDON Representation Backbone

We use the pretrained SELDON model as a fixed representation and reconstruction backbone throughout our framework. SELDON is designed for sparse, irregularly sampled, and heteroscedastic multiband light curves, applying a GRU-ODE encoder and a parametric decoder to continuous-time light-curve reconstruction. We refer to the original SELDON work by Wu et al. (2026) for the full architectural details.

For each supernova, we encode both a partial light curve and its corresponding full light curve using the frozen SELDON encoder. We denote the resulting 256-dimensional encoder-memory representations as $m_{\text{partial}}, m_{\text{full}} \in \mathbb{R}^{256}$ which serve as the two endpoints. The proposed I2SB model operates directly in this 256-dimensional encoder-memory space, learning a stochastic bridge from m_{partial} toward the corresponding m_{full}. Within the SELDON backbone, the frozen variational distribution head maps a memory representation $m \in \mathbb{R}^{256}$ to the parameters of a six-dimensional diagonal Gaussian latent distribution, $z \in \mathbb{R}^6$ where $q_\phi(z \mid m) = \mathcal{N}\left(\mu_\phi(m), \operatorname{diag}\left(\sigma_\phi^2(m)\right)\right)$.

After bridge sampling, each generated memory representation is mapped through the frozen SELDON distribution head to obtain z_{mean}. The latent mean is then passed through the frozen SELDON decoder to reconstruct the multiband light curve for forecasting, and is also used as the representation for downstream classification. The SELDON backbone remains frozen during training of the proposed bridge framework, so that the proposed framework learns to complete partial representations without modifying the pretrained SELDON representation space.

3.2 Partial-to-Full Latent Schrödinger Bridge

Given a partial light curve and its corresponding full light curve, we construct a stochastic bridge between their frozen SELDON encoder-memory representations. We adapt the paired I2SB (Liu et al., 2023) formulation to this representation space and introduce two modifications tailored to partially observed light curves: a completeness-aware variance rule and an object-level alignment objective for the many-to-one partial-to-full mapping.

The 256-dimensional SELDON memories are standardized using dimension-wise statistics computed from the training partial and full representations. For each partial observation i, we define the standardized bridge endpoints as $x_{\text{partial},i} = \tilde{m}_{\text{partial},i}, \quad x_{\text{full},i} = \tilde{m}_{\text{full},i}$. Following the I2SB convention, the reverse bridge is initialized from the observed partial representation $x_{\text{partial},i}$ and evolves toward the corresponding full representation $x_{\text{full},i}$. Let $\{\beta_n\}_{n=0}^{N-1}$ denote the base symmetric bridge schedule, and define the forward and backward cumulative variances $A_n = \sum_{k=0}^{n} \beta_k, B_n = \sum_{k=n}^{N-1} \beta_k$. Let $s_i > 0$ denote the observation-dependent variance multiplier defined in the next section, which remains fixed across bridge timesteps for a given partial observation i. The intermediate training state is sampled as $X_{i,n} \sim \mathcal{N}\left(\frac{B_n x_{\text{full},i} + A_n x_{\text{partial},i}}{A_n + B_n}, s_i \frac{A_n B_n}{A_n + B_n} I_{256}\right)$. The base unscaled bridge is recovered when $s_i = 1$. Equivalently, scaling both cumulative variances by s_i preserves the endpoint-conditioned interpolation mean while multiplying the bridge covariance by s_i. We train the bridge network to predict the normalized residual target $v_{i,n}^\star = \frac{X_{i,n} - x_{\text{full},i}}{\sqrt{s_i A_n}}$. Given the intermediate state, normalized bridge timestep, and partial endpoint, the network prediction is $\widehat{v}_{i,n} = v_\theta\left(X_{i,n}, t_n, x_{\text{partial},i}\right), t_n = \frac{n+1}{N}$. The corresponding full-memory endpoint estimate is $\widehat{x}_{\text{full},i}^{(n)} = X_{i,n} - \sqrt{s_i A_n}\, \widehat{v}_{i,n}$. The prediction network is an MLP conditioned on the intermediate bridge state, partial-memory endpoint, and a 64-dimensional sinusoidal timestep embedding, with two 512-unit SiLU hidden layers and a 256-dimensional output. We use $N = 1000$ discrete bridge indices with a symmetric base schedule whose maximum discrete variance increment is $0.3/N = 3 \times 10^{-4}$.

Completeness-Aware Bridge Variance The uncertainty associated with completing a partial light curve should depend on how much of its temporal evolution has already been observed. A very early light curve generally admits a wider range of plausible future trajectories, whereas a more complete observation provides stronger constraints on the full representation. We therefore assign each partial observation a scalar variance multiplier s_i that remains fixed along its bridge trajectory.

We first quantify temporal completeness as $r_i = \frac{t_{\text{last},i} - t_{\min}}{t_{\max} - t_{\min}}, r_i \in [0,1]$, where $t_{\min}$ and $t_{\max}$ are the minimum and maximum last-observed times among the cached partial views of the same object. A completeness-dependent variance scale is then defined as $s_{\text{dynamic},i} = \exp(-2r_i)$.

Temporal completeness alone may underestimate uncertainty when the observed photometry remains noisy. Let $S = \sum_{k=0}^{N-1} \beta_k$ denote the total variance of the base schedule, and let $\bar{\sigma}_{\text{obs},i}$ denote the mean normalized flux error over valid observed points. We define $u_i = 5\bar{\sigma}_{\text{obs},i}, u_{i,\text{cap}} = \min\left(u_i,\, 0.75\sqrt{S}\right)$, and use the final per-observation scale $s_i = \max\left(\exp(-2r_i), \frac{u_{i,\text{cap}}^2}{S}\right)$.

The same multiplier is applied to both cumulative bridge variances, $a_{i,n} = s_i A_n, \quad b_{i,n} = s_i B_n$. Consequently, the endpoint-conditioned interpolation mean is unchanged, while the conditional bridge covariance is scaled by s_i: $X_{i,n} = \mu_{i,n} + \sqrt{V_{i,n}}\epsilon, \epsilon \sim \mathcal{N}(0, I)$, with $\mu_{i,n} = \frac{B_n x_{\text{full},i} + A_n x_{\text{partial},i}}{A_n + B_n}, V_{i,n} = s_i \frac{A_n B_n}{A_n + B_n}$. This scaling is used consistently in forward bridge sampling, target normalization, endpoint reconstruction, and reverse sampling. In particular, for a reverse transition from bridge index n to an earlier selected index $\ell < n$, the posterior update is $X_{i,\ell} = \mu^{\text{rev}}_{i,n\to\ell} + \sqrt{V^{\text{rev}}_{i,n\to\ell}}\xi, \xi \sim \mathcal{N}(0, I)$, where $\mu^{\text{rev}}_{i,n\to\ell} = \frac{A_n - A_\ell}{A_n}\widehat{x}^{(n)}_{\text{full},i} + \frac{A_\ell}{A_n} X_{i,n}, V^{\text{rev}}_{i,n\to\ell} = s_i \frac{(A_n - A_\ell)A_\ell}{A_n}$. Thus, the same observation-specific uncertainty scale governs both training and generation. This mechanism acts in SELDON memory space and is distinct from the flux-space predictive uncertainty model introduced later.

Many-to-One Alignment Objective Our paired training data contain multiple partial observations of each underlying light curve, all associated with the same full-memory target. To account for this many-to-one structure, we group partial views according to their light-curve identifier and formulate the bridge objective at the object level.

Let $g = 1, \ldots, G$ index objects and $j = 1, \ldots, J_g$ index the partial views sampled for object g. The base I2SB regression loss is

$$\mathcal{L}_{\text{I2SB}} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{J_g}\sum_{j=1}^{J_g}\frac{1}{256}\left\|v_\theta^{(g,j)} - v_{\text{target}}^{(g,j)}\right\|_2^2.$$

This assigns equal weight to each object regardless of the number of partial views sampled for that object. When all objects contribute the same number of views, it reduces to the ordinary mean over sampled views. We additionally introduce an object-level centroid-alignment term. For each object, we compare the centroid of its sampled partial-memory endpoints with the centroid of the corresponding predicted full-memory endpoints:

$$\mathcal{L}_{\text{centroid}} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{256}\left\|\frac{1}{J_g}\sum_{j=1}^{J_g} x_{\text{partial}}^{(g,j)} - \frac{1}{J_g}\sum_{j=1}^{J_g}\widehat{x}_{\text{full}}^{(g,j)}\right\|_2^2.$$

Here, $\widehat{x}_{\text{full}}^{(g,j)}$ is the endpoint estimate obtained at the bridge index sampled for that training example, rather than the intermediate noisy state. Both centroids are computed in the standardized 256-dimensional SELDON memory space.

The final bridge-training objective is $\mathcal{L} = \mathcal{L}_{\text{I2SB}} + \lambda_{\text{centroid}}\mathcal{L}_{\text{centroid}}, \lambda_{\text{centroid}} = 0.15$. The centroid term constrains the discrepancy between the group-level mean partial and predicted representations; it does not force each predicted endpoint to match its corresponding partial memory individually. The paired full-memory representation remains the target of the I2SB regression loss, while the centroid term provides an additional object-level constraint across multiple partial views.

Bridge Sampling and Light-curve Reconstruction At inference time, we initialize the reverse process at the standardized partial-memory endpoint, $X_{i,N-1} = x_{\text{partial},i}$, while retaining $x_{\text{partial},i}$

as a fixed conditioning input throughout the trajectory. The observation-dependent variance multiplier s_i is also held fixed. We use 50 reverse transitions selected from the original $N = 1000$ bridge indices rather than constructing a separate inference-time schedule.

For a transition from index n to an earlier index $\ell < n$, we first estimate the full-memory endpoint,

$$\widehat{x}^{(n)}_{\mathrm{full},i} = X_{i,n} - \sqrt{s_i A_n}\, v_\theta\left(X_{i,n}, t_n, x_{\mathrm{partial},i}\right),$$

and sample the reverse posterior as

$$X_{i,\ell} = \frac{A_n - A_\ell}{A_n}\widehat{x}^{(n)}_{\mathrm{full},i} + \frac{A_\ell}{A_n}X_{i,n} + \sqrt{s_i \frac{(A_n - A_\ell)A_\ell}{A_n}}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{256}).$$

The final transition is deterministic and returns the predicted full-memory endpoint.

We repeat the stochastic reverse process to obtain $M = 256$ completed-memory samples. Each terminal sample is transformed back to the original SELDON memory coordinates by reversing the training-set standardization,

$$\widehat{m}_i^{[m]} = \mu_{\mathrm{train}} + \sigma_{\mathrm{train}} \odot \widehat{x}^{[m]}_{\mathrm{full},i}.$$

The frozen SELDON variational head then maps each generated memory to its six-dimensional latent mean,

$$z_i^{[m]} = \mu_\phi\left(\widehat{m}_i^{[m]}\right),$$

which is decoded into a light curve using the frozen SELDON decoder,

$$\widehat{f}_i^{[m]}(\tau, b) = \mathcal{T}^{-1}\left[D_\psi\left(z_i^{[m]}, \tau, b\right)\right],$$

where $\mathcal{T}^{-1}$ converts the decoder output back to physical flux units. The same frozen SELDON decoder also returns physical parameter outputs, from which we extract the corresponding redshift estimate; no additional redshift prediction model is trained.

For point forecasting, we average the generated memories before decoding, $\overline{m}_i = \frac{1}{M}\sum_{m=1}^{M}\widehat{m}_i^{[m]}$, and compute the flux forecast and redshift estimate from $\mu_\phi(\overline{m}_i)$. This differs from averaging the individually decoded flux trajectories because the mapping from memory to flux is nonlinear. Individual decoded samples are retained for predictive uncertainty estimation in the following section.

3.3 Predictive Uncertainty Quantification

The stochastic bridge generates variability over completed SELDON memory representations, but this variability alone does not necessarily provide calibrated predictive intervals in flux space. We therefore augment the bridge samples with SELDON's pretrained uncertainty network (UQNet), which models pointwise residual uncertainty after decoding. For a fair comparison, we train separate UQNet instances for the partial SELDON baseline and the I2SB-completed forecasts, using the same uncertainty formulation and loss.

For the I2SB branch, UQNet fine-tuning uses a completed-memory prediction $\widehat{m}_i$ obtained by deterministic I2SB sampling, with the reverse posterior noise omitted. The frozen SELDON variational head produces $z_i = \mu_\phi\left(\widehat{m}_i\right) \in \mathbb{R}^6$, and the frozen decoder produces a reconstruction at query time τ and band b. For the partial baseline, the corresponding UQNet is trained separately using the original partial SELDON representation in place of the I2SB-completed representation.

In both cases, UQNet receives a 520-dimensional input consisting of the scalar decoder-space reconstruction, a 512-dimensional photometric-band embedding, the six-dimensional latent mean z_i, and the scalar query time. Its output is transformed using a softplus function to obtain a positive decoder-space uncertainty scale:

$$\sigma_{\mathrm{dec},i}(\tau, b) = \mathrm{softplus}\left(U_\omega\left(\left[\widehat{f}_{\mathrm{dec},i}(\tau, b), e_b, z_i, \tau\right]\right)\right).$$

We convert this scale to physical flux units using

$$\sigma_{\mathrm{flux},i}(\tau, b) = \sigma_{\mathrm{dec},i}(\tau, b)\left(\left|\widehat{f}_{\mathrm{pred},i}(\tau, b)\right| + 1\right)\ln 10,$$

where $\widehat{f}_{\mathrm{pred},i}$ denotes the corresponding predicted flux in physical units.

Each UQNet is trained using a Gaussian negative log-likelihood over valid observed points only. Let $\mathcal{O}_i$ denote the positions satisfying both the observation and padding masks for partial view i, and let $Q = \sum_i |\mathcal{O}_i|$ be the number of such points in the minibatch. The loss is

$$\mathcal{L}_{\mathrm{UQ}} = \frac{1}{Q} \sum_i \sum_{q \in \mathcal{O}_i} \frac{1}{2} \left[\frac{\left(f_{i,q} - \widehat{f}_{i,q} \right)^2}{\sigma^2_{\mathrm{flux},i,q}} + \log \left(\sigma^2_{\mathrm{flux},i,q} \right) \right].$$

At inference time, for the I2SB + UQNet model, we apply the I2SB-trained UQNet separately to each stochastically generated and decoded bridge sample. For memory sample $m = 1, \ldots, M$ and query point (τ, b), we construct a flux sample as

$$f_i^{[m]}(\tau, b) = \widehat{f}_i^{[m]}(\tau, b) + \gamma \, \sigma^{[m]}_{\mathrm{flux},i}(\tau, b) \, \epsilon^{[m]}_{i,\tau,b}, \qquad \epsilon^{[m]}_{i,\tau,b} \sim \mathcal{N}(0, 1).$$

For the Partial + UQNet baseline, the separately trained partial UQNet provides the corresponding flux-space uncertainty around the fixed partial SELDON point forecast, without stochastic bridge completion. The Gaussian perturbations are drawn independently across samples and query points, conditional on the corresponding decoded forecast. UQNet therefore adds pointwise residual uncertainty without introducing a separate model for residual correlations across time or band. We use $\gamma = 1$ in the reported experiments. Pointwise predictive intervals are obtained from empirical quantiles of the resulting flux samples; a nominal 80% interval uses the 10th and 90th percentiles. For I2SB + UQNet, this construction combines stochastic memory-completion variability with learned flux-domain residual uncertainty, whereas the Partial + UQNet baseline contains only the learned residual-uncertainty component.

3.4 Downstream Classification

We evaluate whether I2SB-completed representations recover class-discriminative information that is missing from partial observations. Our final classifier combines two complementary sources of information: (i) latent-space class compatibility from the pretrained SELDON energy model, and (ii) decoded light-curve morphology and predictive uncertainty from a temporal convolutional classifier. The two branches are fused in logit space before the final softmax.

Class-guided latent-energy branch. The pretrained SELDON energy model assigns the compatibility score $s(z, c) = -E(z, c)$ to the six-dimensional latent representation z and candidate class c. For each candidate class, the bridge maximizes the target-versus-rest log-odds $J_c(z) = s(z, c) - \log \sum_{j \neq c} \exp s(z, j)$. At each reverse step t, the predicted completed endpoint is mapped to SELDON's z_{mean}, and the gradient of J_c is propagated back to the current bridge state. We average the terminal target-class scores over stochastic samples to obtain $S_{i,c}$. A class-dependent first-detection baseline $\bar{S}_c$, estimated using training data only, gives the calibrated energy logits $L^E_{i,c} = S_{i,c} - \bar{S}_c$.

Decoded light-curve branch. The second branch uses unguided I2SB samples decoded through the frozen SELDON distribution head and decoder. UQNet provides sample-wise uncertainty, which is combined with bridge variability as $\sigma^2_{\mathrm{total}} = \mathrm{Var}_m \left(f^{(m)}_{\mathrm{I2SB}} \right) + \mathrm{mean}_m \left[\left(\sigma^{(m)}_{\mathrm{UQNet}} \right)^2 \right]$. We train this classifier using class-weighted cross-entropy and denote its output logits by $L^D_{i,c}$.

Pre-softmax fusion. We combine the two branches in logit space: $L^{\mathrm{fused}}_{i,c} = (1 - \alpha) L^E_{i,c} + \alpha L^D_{i,c}$, with final class probabilities $p(y = c \mid i) = \mathrm{softmax}_c \left(L^{\mathrm{fused}}_{i,c} \right)$. We select a single global value $\alpha = 0.825$ using validation data. Additional implementation details for both classification branches are provided in Appendix B.1.

Table 1: Forecasting and predictive-interval performance across three independent seeds. Nominal PI coverage is 80%. "I2SB only" uses I2SB-completed memories with bridge-only intervals; "Partial + UQNet" uses the partial SELDON point forecast with a UQNet optimized for the partial baseline; and "I2SB + UQNet" uses the I2SB-completed point forecast with a UQNet optimized for I2SB forecasts.

(a) All observation stages

Method	NRMSE	RMSE	Coverage	PI width
I2SB only	0.2612 ± 0.0012	46.92 ± 0.29	0.293 ± 0.003	55.83 ± 0.68
Partial + UQNet	0.2629 ± 0.0000	49.96 ± 0.00	0.697 ± 0.020	64.94 ± 5.94
I2SB + UQNet	$\mathbf{0.2612 \pm 0.0012}$	$\mathbf{46.92 \pm 0.29}$	$\mathbf{0.787 \pm 0.005}$	94.36 ± 2.24

(b) Pre-peak

Method	NRMSE	RMSE	Coverage	PI width
I2SB only	0.1946 ± 0.0032	109.74 ± 0.71	0.499 ± 0.011	134.96 ± 3.03
Partial + UQNet	0.2041 ± 0.0000	124.85 ± 0.00	0.504 ± 0.026	84.32 ± 9.61
I2SB + UQNet	$\mathbf{0.1946 \pm 0.0032}$	$\mathbf{109.74 \pm 0.71}$	$\mathbf{0.733 \pm 0.004}$	164.32 ± 4.43

4 EXPERIMENTS AND RESULTS

Dataset We use simulated multiband supernova light curves from the ELAsTiCC (Narayan & ELAsTiCC Team, 2023) dataset, with observations in the u, g, r, i, z, y bands and six supernova classes. The training, validation, and test splits contain 26 190, 3275, and 3273 unique objects, respectively, with all partial views of an object assigned to the same split. We construct five time-based partial views per object at fractions $\{0, 0.25, 0.5, 0.75, 1.0\}$ of its eligible observation span, retaining all observations up to each cutoff and requiring at least six observations and six detections.

Model and inference We compare the original partial SELDON representation with I2SB-completed representations. For uncertainty evaluation, we further compare bridge-only sampling with I2SB combined with the learned flux-space UQ module. The SELDON backbone remains frozen throughout I2SB training, and the UQNet is trained separately with SELDON and I2SB fixed. The I2SB checkpoint is selected by latent validation loss. Forecasting and UQ use 50 reverse transitions and 256 stochastic memory samples per partial view, with 80% prediction intervals from 10th and 90th percentiles. For classification, we combine the class-guided latent-energy branch with the decoded-curve temporal classifier using pre-softmax fusion. Model and fusion settings are selected using validation data.

Evaluation Forecasting is evaluated on observations held out after each partial-view cutoff, across all stages and on a pre-peak subset whose last observation precedes the full light-curve maximum. We report root mean squared error (RMSE) and normalized RMSE (nRMSE) for point forecasts, and empirical pointwise coverage and mean interval width at nominal 80% coverage for uncertainty. nRMSE divides each view's RMSE by its maximum absolute true flux. Classification is evaluated on all partial views and the pre-peak subset using accuracy, macro recall, macro F1, NLL, and ECE. Flux metrics are averaged over eligible views, and all results are reported over three random seeds.

4.1 FORECASTING PERFORMANCE

We first evaluate whether I2SB memory completion improves prediction of future flux relative to directly decoding the partial SELDON representation. Both methods use the same frozen SELDON variational head and decoder. For I2SB, the point prediction is obtained by averaging the generated completed memories and then decoding the resulting mean representation. Table 1 reports forecasting and prediction-interval performance across three independent seeds. Averaged across the three runs, I2SB completion reduces all-stage RMSE from 49.96 to 46.92 ± 0.29 and pre-peak RMSE from 124.85 to 109.74 ± 0.71. The corresponding NRMSE also decreases from 0.2629 to 0.2612 ± 0.0012 across all stages and from 0.2041 to 0.1946 ± 0.0032 in the pre-peak subset. An example light curve forecast from the test set is provided in Figure 2.

Table 2: Held-out classification performance averaged over three seeds.

Subset	Method	Acc. (%)	Macro Rec. (%)	Macro F1 (%)	NLL / ECE (%)
Entire	Energy ($\alpha = 0$)	55.94	61.87	52.41	1.423/23.00
	Fusion ($\alpha = 0.825$)	70.54	**68.80**	64.93	0.727/3.10
	Model B ($\alpha = 1$)	**71.40**	66.71	**65.87**	**0.722**/**2.94**
Pre-peak	Energy ($\alpha = 0$)	26.95	31.29	25.06	2.578/44.35
	Fusion ($\alpha = 0.825$)	43.08	**34.83**	**33.97**	1.424/**3.53**
	Model B ($\alpha = 1$)	**45.65**	33.99	33.46	**1.391**/4.24

The reduction is more pronounced in the pre-peak regime, where substantially less of the light-curve evolution has been observed. At the same time, pre-peak RMSE remains considerably larger than the aggregate error across observation stages, indicating that early forecasting remains challenging. The relatively small variation across I2SB seeds suggests that the observed improvement is not driven by a single initialization. As an auxiliary diagnostic, we also examine redshift estimates extracted from the frozen SELDON decoder. Although redshift is not an explicit objective of the I2SB bridge, the residual distribution becomes more concentrated around zero as additional observations are available. Detailed distributions and quantitative residual metrics are provided in Appendix C. The reliability of prediction intervals is evaluated separately in Section 4.2.

4.2 Predictive Uncertainty

We next evaluate whether stochastic bridge completions provide sufficient predictive uncertainty and how UQNet affects interval coverage. Table 1 compares bridge-only intervals, the partial SELDON baseline with a dedicated UQNet, and I2SB-completed forecasts with a separately fine-tuned UQNet.

Bridge-only intervals exhibit substantial undercoverage. At nominal 80% coverage, I2SB sampling alone achieves only 0.293 ± 0.003 across all stages and 0.499 ± 0.011 in the pre-peak subset. The corresponding intervals are relatively narrow, indicating that variation across completed-memory samples alone does not capture the full predictive uncertainty in future flux. Adding UQNet increases these values to 0.787 ± 0.005 and 0.733 ± 0.004, respectively, exceeding the corresponding partial-baseline coverages of 0.697 ± 0.020 and 0.504 ± 0.026. The combined model therefore approaches nominal coverage across all stages, while residual undercoverage remains in the pre-peak regime. These coverage gains are accompanied by wider intervals, highlighting the need to assess coverage and sharpness jointly. UQNet does not alter the point forecasts.

As an additional calibration check, we evaluate I2SB+UQNet on the pre-peak subset at multiple nominal interval levels. Empirical coverage increases from 0.648 at 68% nominal coverage to 0.754, 0.843, and 0.894 at 80%, 90%, and 95%, respectively, increasing monotonically but remaining below nominal throughout. Full results are provided in Appendix A.3.

4.3 I2SB-Based Downstream Classification

We select the decoded-curve classifier and fusion weight using validation data. Model B, which incorporates both the decoded forecast and predictive uncertainty, is retained for the final classifier. The fusion coefficient is selected as $\alpha = 0.825$. Classifier, class-weighting, and fusion-weight ablations are reported in Appendix B.2. Table 2 compares baseline-corrected, guided-I2SB energy branch ($\alpha = 0$), the selected fusion ($\alpha = 0.825$), and the decoded-curve classifier ($\alpha = 1$) on the entire test set and the pre-peak subset. Relative to the energy-only branch, fusion improves accuracy from 55.94% to 70.54% on the entire test set and from 26.95% to 43.08% before peak, while substantially reducing NLL and ECE. Model B alone achieves slightly higher accuracy, whereas fusion yields higher macro recall on the entire test set and modestly higher macro recall and macro F1 in the pre-peak regime. Figure 4 shows class-wise recall for the selected fusion. Pre-peak classification is substantially harder, with recall of 58.1% for SNII, 48.5% for SNIa, and 44.5% for SNIb/c, but only 9.0% for SNIIb and 15.4% for SNIax, indicating persistent ambiguity from limited early observations. Full confusion matrices, post-peak results, and class-wise metrics are provided in Appendix B.3.

AI USE STATEMENT

Generative AI tools were used for manuscript drafting and language editing, programming and debugging assistance, and discussions related to methodological choices and interpretation of experimental results. All AI-assisted content and code were reviewed and validated by the authors, who made all final decisions regarding the methodology, experiments, analysis, and conclusions.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of the results presented in this work, we have provided comprehensive implementation details including but not limited to our network architectures, methodology, hyperparameter configurations, algorithmic specifications, experimental setups in the main text and the Appendix. Specifically, some detailed results are deferred to the Appendix. The training data used is publicly available through `https://portal.nersc.gov/cfs/lsst/DESC_TD_PUBLIC/ELASTICC/`. While the full source code and implementation artifacts are currently withheld to maintain double-blind reviewing compliance, clean and documented source code to reproduce all experiments, along with pre-trained models, will be publicly released upon publication.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *arXiv e-prints*, art. arXiv:2301.08243, January 2023. doi: 10.48550/arXiv.2301.08243.

Tom Charnock and Adam Moss. Deep recurrent neural networks for supernovae classification. *The Astrophysical Journal Letters*, 837(2):L28, mar 2017. doi: 10.3847/2041-8213/aa603d. URL `https://doi.org/10.3847/2041-8213/aa603d`.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf`.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/455cb2657aaa59e32fad80cb0b65b9dc-Paper.pdf`.

Guilherme Grancho D. Fernandes, Marco A. Barroca, Mateus dos Santos, and Rafael S. Oliveira. Classification of transient astronomical object light curves using lstm neural networks, 2025. URL `https://arxiv.org/abs/2511.17564`.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 08 2002. ISSN 0899-7667. doi: 10.1162/089976602760128018. URL `https://doi.org/10.1162/089976602760128018`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf`.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023.

Gautham Narayan and ELAsTiCC Team. The Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC). In *American Astronomical Society Meeting Abstracts #241*, volume 241 of *American Astronomical Society Meeting Abstracts*, pp. 117.01, January 2023.

D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pp. 55–60 vol.1, 1994. doi: 10.1109/ICNN.1994.374138.

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4075–4084. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/pearce18a.html.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction, 2007. URL https://arxiv.org/abs/0706.3188.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020. URL https://arxiv.org/abs/2011.13456.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235765577.

Jiezhong Wu, Jack O'Brien, Jennifer Li, M. S. Krafczyk, Ved G. Shah, Amanda R. Wasserman, Daniel W. Apley, Gautham Narayan, and Noelle I. Samia. Seldon: Supernova explosions learned by deep ode networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26922–26930, 2026.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. A Theory of Generative ConvNet. *arXiv e-prints*, art. arXiv:1602.03264, February 2016. doi: 10.48550/arXiv.1602.03264.

A Light Curve Forecasting

A.1 Representative Forecasts

Figure 2 demonstrates our new approach on forecasted light curves of a simulated ELAsTiCC Type Ia supernova with forecasts and their prediction intervals from the base SELDON model with corresponding UQNet (left), SELDON with I2SB (middle), and SELDON with UQNet+I2SB (right). The shaded points to the left of the vertical black dashed line are the points fed into the encoder while the unshaded points to the right are unseen.

A.2 Bridge Component Ablation

We evaluate the contribution of the bridge adaptations through a sequential component ablation. Configuration A uses the base bridge without centroid alignment, completeness-dependent variance scaling, or an observational flux-error floor. Configuration B adds centroid alignment with weight 0.15; configuration C additionally introduces completeness-dependent variance scaling; and configuration D further incorporates the flux-error floor. All configurations use the same data caches, grouped sampling protocol, network architecture, and optimizer settings, with 40 training epochs comprising 80,000 updates. Evaluation uses 256 stochastic samples and 50 reverse steps. UQNet is excluded from this experiment, so the reported intervals reflect bridge-induced variability alone.

Table 3 reports mean and sample standard deviation across three seeds. Adding centroid alignment produces small changes in the reported point metrics: all-stage RMSE decreases from 47.28 to 47.17, while pre-peak RMSE decreases from 109.49 to 109.43. These changes are small relative to the reported between-seed variation and do not, by themselves, establish a substantial independent forecasting benefit.

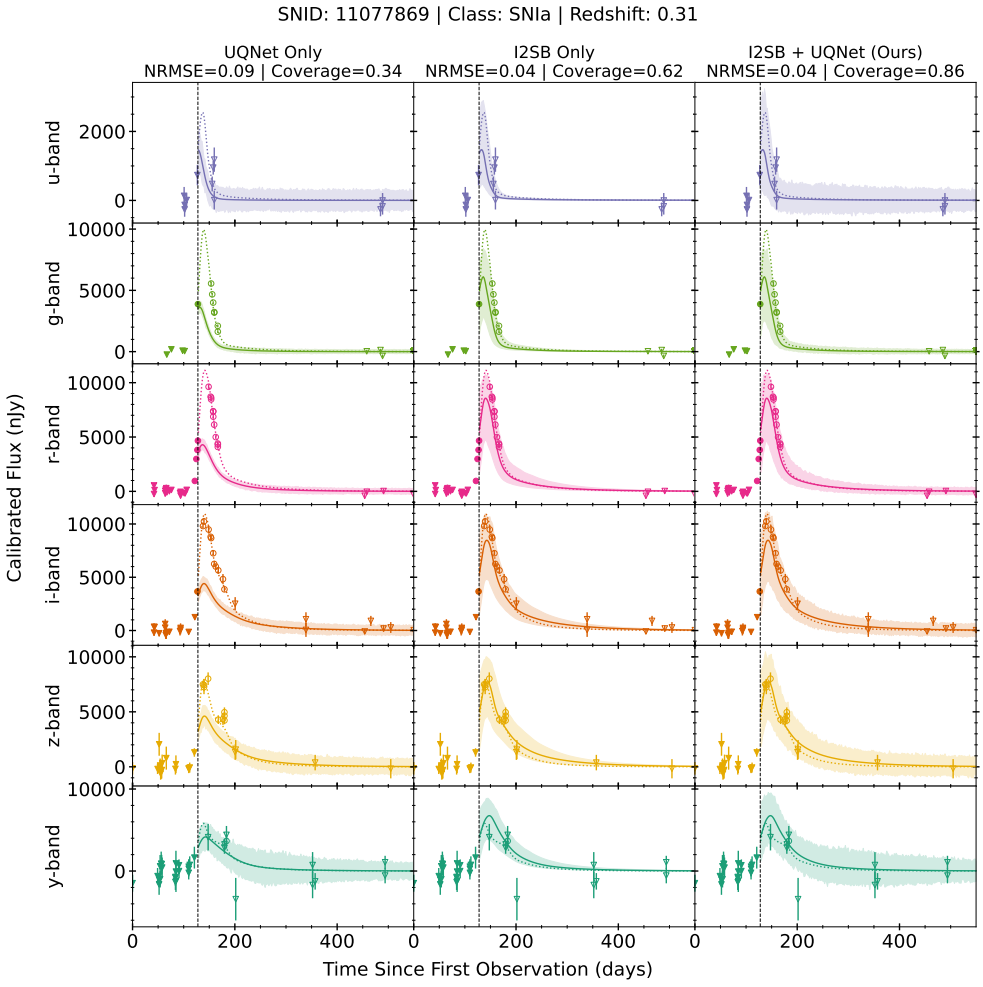

Figure 2: Forecasted light curves of a simulated ELAsTiCC Type Ia supernova with predictions and their prediction intervals from the base SELDON model with corresponding UQNet (left), SELDON with I2SB (middle), and SELDON with UQNet+I2SB (right). The dotted line shows the SELDON forecast when the entire light curve is observed. The solid line is the prediction of each method conditioned strictly on observations to the left of the vertical black dashed lines in each frame. Downward triangles in the light curve show flux measurements that were below the signal-to-noise threshold of the survey with negative values caused by simulated image subtraction artifacts.

Completeness-dependent variance scaling changes the balance between point accuracy and predictive spread. In the pre-peak subset, adding this component increases coverage from 0.469 to 0.520 and mean interval width from 123.02 to 145.86, while RMSE increases from 109.43 to 110.17. Adding the flux-error floor reduces RMSE to 109.74 and width to 134.96, but also reduces coverage to 0.499. Configuration D achieves the lowest mean all-stage RMSE, whereas configuration B has the lowest mean pre-peak RMSE among the evaluated configurations.

These results indicate trade-offs rather than uniform gains from every component. Because the components are added sequentially, each comparison measures an incremental effect conditional on the preceding components. None of the bridge-only configurations approaches nominal 80% coverage, supporting the need for the separate flux-domain uncertainty model evaluated in the main text.

Table 3: Sequential bridge component ablation. A: base bridge; B: A + centroid alignment; C: B + completeness-dependent variance; D: C + observational flux-error floor. Values are mean ± sample standard deviation across three seeds. UQNet is not used. RMSE denotes mean per-view future-flux RMSE. Coverage is evaluated against a nominal level of 0.80; width should be interpreted jointly with coverage.

Evaluation set	Configuration	NRMSE	RMSE	Coverage	Width
All-stage	A	0.2612 ± 0.0013	47.28 ± 0.32	0.289 ± 0.003	54.25 ± 0.63
	B	0.2612 ± 0.0012	47.17 ± 0.37	0.289 ± 0.003	53.98 ± 0.55
	C	0.2620 ± 0.0015	47.11 ± 0.58	0.301 ± 0.002	59.11 ± 1.29
	D	0.2612 ± 0.0012	46.92 ± 0.29	0.293 ± 0.003	55.83 ± 0.68
Pre-peak	A	0.1929 ± 0.0051	109.49 ± 2.05	0.469 ± 0.012	123.24 ± 3.22
	B	0.1928 ± 0.0050	109.43 ± 2.07	0.469 ± 0.011	123.02 ± 2.81
	C	0.1971 ± 0.0037	110.17 ± 0.96	0.520 ± 0.009	145.86 ± 3.32
	D	0.1946 ± 0.0032	109.74 ± 0.70	0.499 ± 0.011	134.96 ± 3.03

Table 4: Nominal-level calibration of I2SB + UQNet in a single reference run, separate from the three-seed experiments. The all-stage and pre-peak evaluations contain 13,151 and 2,757 eligible views, respectively. Coverage is reported as a percentage.

	All-stage		Pre-peak	
Nominal (%)	Coverage	Width	Coverage	Width
68	70.63	76.40	64.82	125.24
80	80.86	99.42	75.37	163.89
90	88.82	129.49	84.30	215.47
95	92.92	156.37	89.37	263.11

A.3 Calibration across Nominal Interval Levels

We additionally assess predictive coverage at nominal levels of 68%, 80%, 90%, and 95%. This diagnostic uses a single reference I2SB + UQNet run with 256 stochastic samples, 50 reverse steps, and unit UQ perturbation scale. It is separate from the three-seed evaluation reported above. For each nominal level, central pointwise intervals are constructed from the corresponding empirical quantiles. Coverage and width are averaged over eligible partial views.

Table 4 shows that empirical coverage increases with the nominal level in both evaluation sets. Across all observation stages, empirical coverage is 70.6%, 80.9%, 88.8%, and 92.9%, respectively. Agreement is closest to nominal at the 80% level, while the 90% and 95% intervals show undercoverage.

Pre-peak intervals remain below nominal coverage at every evaluated level, attaining 64.8%, 75.4%, 84.3%, and 89.4%. Increasing the nominal level widens these intervals from 125.24 at 68% to 263.11 at 95%, but does not eliminate the coverage deficit. The absolute pre-peak coverage gaps range from approximately 3.2 to 5.7 percentage points. These results demonstrate that agreement at a single nominal level does not establish calibration across interval levels or observation regimes.

B Additional Classification Details

B.1 Classification Implementation Details

Energy-guided sampling. We generate one I2SB trajectory for each of the $K = 6$ class hypotheses. At reverse step t, I2SB predicts the completed endpoint memory $\hat{m}_{0,t}$ from the current bridge state x_t. After inverse standardization, $\hat{m}_{0,t}$ is mapped through the frozen SELDON distribution head to obtain $z_{\mathrm{mean},t}$.

For target class c, the guidance gradient is

$$g_{c,t} = \nabla_{x_t} \left[s(z_{\mathrm{mean},t}, c) - \log \sum_{j \neq c} \exp s(z_{\mathrm{mean},t}, j) \right].$$

We first clip its Euclidean norm to one and multiply it by the guidance scale $\eta = 0.006$. The resulting displacement is further constrained by the magnitude of the corresponding unguided I2SB transition:

$$\|\Delta x_{\mathrm{guide},t}\|_2 \le \rho \left\|x_{t-1}^{\mathrm{I2SB}} - x_t\right\|_2, \qquad \rho = 0.01.$$

The guided transition is therefore

$$x_{t-1} = x_{t-1}^{\mathrm{I2SB}} + \Delta x_{\mathrm{guide},t}.$$

We use 50 reverse steps and $M_E = 32$ stochastic samples per class track. Posterior noise is sampled independently across the six tracks. SELDON and I2SB parameters remain frozen throughout this procedure.

At the terminal state, we retain the diagonal target-class score from each hypothesis track and average it over the stochastic samples:

$$S_{i,c} = \frac{1}{M_E} \sum_{m=1}^{M_E} s\left(z_{i,c}^{(m)}, c\right).$$

First-detection score baseline. The class-dependent baseline $\bar{S}_c$ is estimated using training data only. We select 500 pre-peak training objects per class that contain at least six non-detections, giving 3,000 baseline objects in total. For each object, we retain the chronologically first true detection. Every later detection is changed to a non-detection and its flux and error are replaced with a real training non-detection from the same band and the same within-band error decile. Observation times and the total number of measurements are preserved.

The modified light curves are passed through I2SB and SELDON using 50 reverse steps and 16 stochastic samples. The baseline for output class c is the mean terminal score over the modified training sample:

$$\bar{S}_c = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} S_{b,c}.$$

The calibrated energy logit is

$$L_{i,c}^{E} = S_{i,c} - \bar{S}_c.$$

Decoded forecast features. The decoded-curve branch uses unguided I2SB sampling with 50 reverse steps and $M_D = 8$ stochastic samples. Each completed memory is mapped through the frozen SELDON distribution head and decoder on a fixed grid of 64 query points spanning $[0, 32]$ in the model's normalized time coordinate, for each of the six photometric bands.

For each band and grid location, we calculate

$$\mu_f = \frac{1}{M_D} \sum_{m=1}^{M_D} f^{(m)}$$

and

$$\sigma_{\mathrm{total}}^2 = \mathrm{Var}_m\left(f^{(m)}\right) + \frac{1}{M_D} \sum_{m=1}^{M_D} \left(\sigma_{\mathrm{UQNet}}^{(m)}\right)^2.$$

Thus, the first term represents variation across I2SB forecasts and the second represents the mean conditional variance predicted by UQNet.

Irregular observed measurements are assigned to their nearest grid locations. When multiple measurements occupy the same cell, their fluxes are combined using inverse-variance weighting. The resulting feature tensor has shape $6 \times 6 \times 64$ and contains

$$[\mu_f,\ \log \sigma_{\mathrm{total}},\ f_{\mathrm{obs}},\ \log \sigma_{\mathrm{obs}},\ m_{\mathrm{obs}},\ m_{\mathrm{det}}].$$

Here, m_{obs} indicates the presence of an observed measurement and m_{det} indicates whether that measurement is a detection. Each feature channel is standardized using statistics computed from the training set only.

Temporal classifier architecture. The same temporal encoder is applied independently to each band. It contains three one-dimensional convolutional blocks:

$$6 \xrightarrow{\text{Conv1D(5)}} 32 \xrightarrow{\text{Conv1D(5,stride=2)}} 64 \xrightarrow{\text{Conv1D(3,stride=2)}} 64.$$

Each convolution is followed by batch normalization and a GELU activation. Adaptive average pooling produces one 64-dimensional representation per band. The six band representations are concatenated and passed through

$$384 \to 192 \to 6$$

fully connected layers, with GELU activation and dropout of 0.2 before the final classification layer.

Class weighting and optimization. The decoded-curve classifier is trained using weighted cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\frac{\sum_{i=1}^{B} w_{y_i} \log p_D(y_i \mid x_i)}{\sum_{i=1}^{B} w_{y_i}}, \qquad w_c = \left(\frac{N}{KN_c}\right)^{\gamma_c}.$$

Here, B is the minibatch size, $K = 6$ is the number of classes, and N_c is the number of training views belonging to class c, excluding the internal holdout. We define $N = \sum_{c=1}^{K} N_c$.

For the selected Scenario 2 model, the class-dependent exponents are

c	SNII	SNIIb	SNIIn	SNIa	SNIax	SNIb/c
γ_c	0.50	0.75	0.75	0.50	0.45	0.35

We train with AdamW using a learning rate of 3×10^{-4}, weight decay of 10^{-4}, batch size 256, and gradient-norm clipping at one. Training runs for at most 50 epochs with early-stopping patience of eight.

A class-stratified 10% internal holdout is constructed at the object level, ensuring that all partial views of an object remain in the same partition. The selected checkpoint maximizes the average of the entire-set and pre-peak macro recalls on this internal holdout.

Fusion implementation. The saved decoded-classifier outputs are probabilities rather than raw network logits. We therefore use

$$L^D_{i,c} = \log p_D(c \mid i)$$

in the fusion. Since

$$\log p_D(c \mid i) = \ell^D_{i,c} - \log \sum_j \exp \ell^D_{i,j},$$

where $\ell^D_{i,c}$ is the raw network logit, the removed term is constant across classes for a given object and cancels under the final softmax. Consequently, using $\log p_D$ is equivalent to using the raw decoded classifier logits in

$$p(y = c \mid i) = \text{softmax}_c \left[(1-\alpha)L^E_{i,c} + \alpha L^D_{i,c}\right].$$

B.2 Classification Ablations and Additional Results

Decoded-classifier selection. We compared two temporal classifiers; Model A and Model B. Model A uses the mean decoded forecast together with observed photometry, observational errors, and observation/detection masks. Model B additionally includes a predictive-uncertainty channel combining I2SB sample variance and UQNet residual variance. The two models showed similar validation performance, but Model B was retained so that the classifier could explicitly use uncertainty from both I2SB and UQNet. We also examined alternative class-weighting schemes. Strong scalar inverse-frequency weighting increased recall for some minority classes but reduced overall accuracy. A class-dependent weighting scheme (Scenario 2) provided a more balanced recall–precision trade-off than Scenario 1 and was therefore retained for the final decoded-curve classifier.

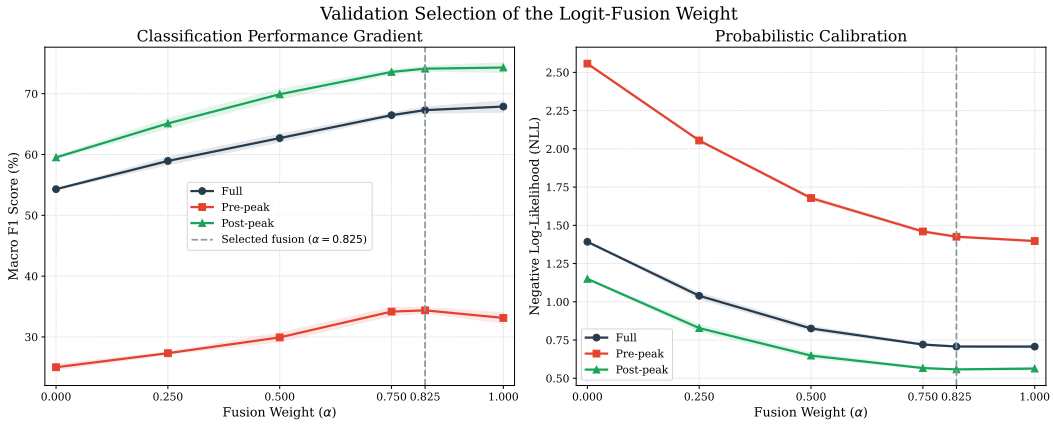

Figure 3: Validation selection of the pre-softmax fusion coefficient α. The left panel shows macro F1 and the right panel shows negative log-likelihood for the entire, pre-peak, and post-peak validation sets. The dashed line indicates the selected global value, $\alpha = 0.825$. Shaded regions show the sample standard deviation across three seeds.

Fusion-weight selection. The fusion coefficient was selected using validation data only. We first evaluated a coarse grid over $\alpha \in [0, 1]$ and then searched $\alpha \in [0.75, 1.0]$ at intervals of 0.025. As shown in Figure 3, pre-peak macro F1 was maximized at $\alpha = 0.825$. We therefore selected $\alpha = 0.825$ as a single global value and fixed it across seeds and observation stages before held-out test evaluation.

Here, $\alpha = 0$ uses only the baseline-corrected guided energy logits, whereas $\alpha = 1$ uses only the decoded-curve classifier. The selected value retains a contribution from both branches; it does not simultaneously optimize every metric or observation subset.

B.3 Class-Wise and Additional Test Results

Table 5 reports the class-wise precision, recall, and F1 score of the selected fusion model. Performance varies substantially with observation stage. On the entire test set, the classifier achieves its strongest F1 scores for SNII, SNIIn, and SNIa. SNIIb and SNIax remain more difficult: SNIIb has relatively low recall and precision, while SNIax retains moderate recall but lower precision.

Before peak, SNIIb and SNIax are the most challenging classes. Their recalls decrease to 8.99% and 15.42%, respectively, with correspondingly low precision. SNIIn exhibits a different pattern: its pre-peak precision remains relatively high at 62.00%, but its recall decreases to 33.47%, indicating conservative predictions that miss many true SNIIn events.

Class separation improves substantially after peak. Post-peak recall exceeds 65% for all classes except SNIIb, while SNII, SNIIn, and SNIa achieve F1 scores above 80%. Nevertheless, SNIIb and SNIax retain comparatively low precision, indicating persistent false-positive predictions for these classes even after peak.

C Redshift Diagnostics

Although redshift is not an explicit objective of the I2SB bridge, we examine estimates returned by the frozen SELDON decoder as additional observations become available. This reference-run diagnostic follows the same 3,273 test objects across seven target observation stages, using 256 stochastic memory samples and 50 reverse steps. Predictions are obtained by decoding the mean completed memory through the frozen distribution head and decoder.

Figure 5 shows the residual

$$r_z = z_{\mathrm{obs}} - z_{\mathrm{pred}}.$$

Table 5: Class-wise performance of the selected fusion ($\alpha = 0.825$). Values are mean $\pm$ sample standard deviation across three seeds.

Subset	Class	Recall (%)	Precision (%)	F1 (%)
Entire	SNII	78.30 ± 3.57	76.95 ± 2.69	77.53 ± 0.45
	SNIIb	48.28 ± 4.28	34.43 ± 2.34	40.17 ± 2.79
	SNIIn	71.94 ± 2.10	79.62 ± 2.54	75.54 ± 0.43
	SNIa	85.58 ± 1.26	74.27 ± 1.08	79.51 ± 0.24
	SNIax	66.45 ± 2.42	37.56 ± 1.61	47.96 ± 1.09
	SNIb/c	62.26 ± 2.60	77.16 ± 1.50	68.87 ± 0.97
Pre-peak	SNII	58.06 ± 1.96	51.02 ± 1.41	54.28 ± 0.31
	SNIIb	8.99 ± 5.06	11.05 ± 1.98	9.61 ± 4.04
	SNIIn	33.47 ± 2.37	62.00 ± 3.02	43.39 ± 1.71
	SNIa	48.51 ± 5.53	32.73 ± 0.73	39.01 ± 1.88
	SNIax	15.42 ± 3.15	6.49 ± 0.76	9.10 ± 1.12
	SNIb/c	44.53 ± 1.61	53.09 ± 0.97	48.43 ± 1.35
Post-peak	SNII	82.42 ± 3.96	83.01 ± 3.18	82.61 ± 0.49
	SNIIb	55.30 ± 5.77	36.49 ± 1.68	43.93 ± 2.88
	SNIIn	80.29 ± 2.04	81.72 ± 2.43	80.96 ± 0.34
	SNIa	93.04 ± 0.62	85.65 ± 0.60	89.19 ± 0.05
	SNIax	76.92 ± 3.21	46.95 ± 3.63	58.18 ± 2.11
	SNIb/c	65.81 ± 3.33	82.24 ± 2.14	73.05 ± 1.16

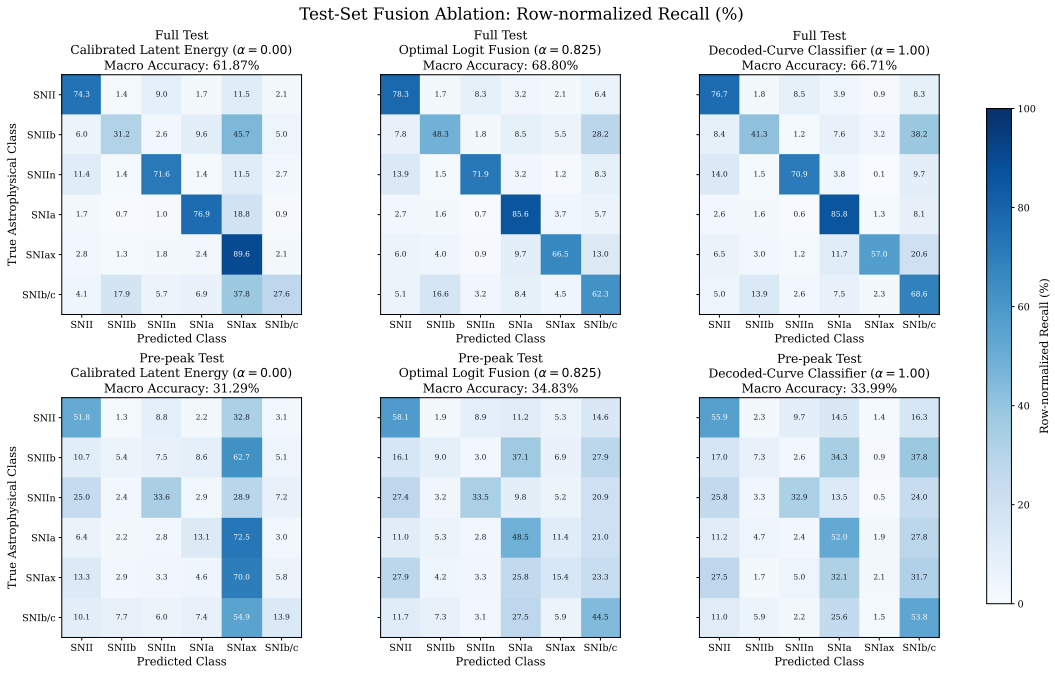

Figure 4: Row-normalized confusion matrices for the calibrated latent-energy classifier ($\alpha = 0$), selected fusion ($\alpha = 0.825$), and decoded-curve classifier ($\alpha = 1$) on the entire test set (top) and pre-peak subset (bottom). Values are percentages averaged over three seeds.

MAE decreases from 0.1510 at the earliest stage to 0.0901 at the most complete stage, while RMSE decreases from 0.2397 to 0.1776. Under this residual convention, the mean residual changes from -0.0105 to 0.0021.

The stage labels are target time-completeness ratios used to select cached partial views, rather than exact observed ratios. For example, the target ratio of 0.50 corresponds to an actual median ratio

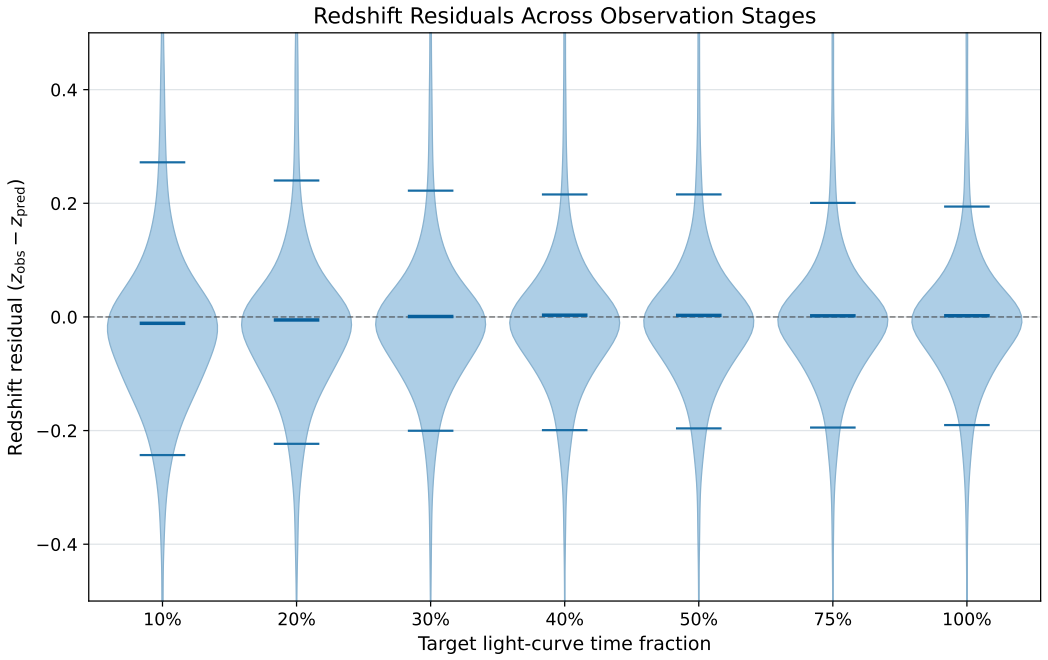

Figure 5: Redshift residuals across increasing light-curve observation stages. The central horizontal markers show the mean residuals, with lower and upper markers indicating the 5th and 95th percentiles, respectively. The dashed line indicates zero residual. Residuals become narrower and remain centered near zero as additional observations become available.

of 0.344. These results describe the evolution of I2SB-derived redshift estimates with observation stage; they do not by themselves establish an improvement over direct decoding of partial memories.