SPLICE: LATENT DIFFUSION OVER JEPA EMBEDDINGS FOR CONFORMAL TIME-SERIES INPAINTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long gaps in time series are difficult to reconstruct reliably, and in power systems imputed load values inform dispatch and planning. Existing generative imputers are accurate but provide no finite-sample reliability guarantees under changing conditions. We introduce **SPLICE** (Self-supervised Predictive Latent Inpainting with Conformal Envelopes), which couples a JEPA encoder, a conditional latent bridge, an hourly-conditioned decoder, and Adaptive Conformal Inference (ACI). Controlled ablations indicate that long-horizon fidelity is driven more by the predictive representation than by generative complexity: under a matched bridge and decoder, JEPA outperforms reconstructive VAE and contrastive TS2Vec encoders, and once the representation is fixed a 5-step flow sampler matches or improves on a 50-step DDIM sampler at a $5-10\times$ speedup. The same ordering is recovered on twelve hydrological catchments. Across thirteen load datasets at 91-day gaps, SPLICE attains the lowest mean Load-only MSE (0.056), winning 9 of 12 non-degenerate datasets against seven established baselines, and the best mean CRPS (0.182, 7.8% below the strongest competitor), while ACI holds 93–95% empirical coverage where static calibration under-covers by up to 7.5 percentage points. A pooled JEPA encoder transfers to four held-out datasets with only bridge and decoder adaptation.

1 INTRODUCTION

The reliable operation of power grids depends on accurate reconstruction and prediction of short-term electricity demand using historical consumption data alongside external factors such as weather and calendar effects. Real-world metering systems, however, frequently encounter extended intervals of missing data due to sensor failures, maintenance outages, or communication errors. Traditional imputation methods revert to smooth interpolations that fail to preserve the fine-grained temporal structure required for operational planning, while recent deep generative approaches (Tashiro et al., 2021; López Alcaraz & Strodthoff, 2023) produce plausible reconstructions but offer no formal guarantees on the reliability of their predictions, a critical gap for safety-sensitive infrastructure. The two error directions are not symmetric: an over-confident reconstruction *understates* the true uncertainty, and that shortfall propagates into reserve sizing, dispatch, and demand-response commitments taken on the assumption that the band is trustworthy. Point accuracy cannot expose this failure, which is why we evaluate conformal coverage and CRPS alongside reconstruction MSE.

We address this limitation with SPLICE, a framework that applies JEPA predictive representation learning to operational long-horizon time-series inpainting, combining it with efficient flow-based generation and online conformal calibration. Conceptually, our approach instantiates LeCun's *JEPA-based world model* vision (LeCun, 2022): a Joint Embedding Predictive Architecture learns a latent dynamics model of the load system, enabling the pipeline to "imagine" plausible trajectories for unobserved intervals rather than merely interpolating between observed endpoints. Unlike earlier world models designed for RL planning (Ha & Schmidhuber, 2018), our system is purpose-built for conditional generation with uncertainty quantification, a distinction that shapes every architectural choice. The pipeline comprises four independently replaceable modules (Section 4): a JEPA encoder (Assran et al., 2023) that maps daily segments into a structured 64-dim latent space; a conditional bridge that generates gap trajectories with tuneable stochasticity; a flow-matching sampler (Lipman et al., 2023); and an hourly-conditioned decoder. Adaptive Conformal Inference (ACI) (Gibbs & Candès, 2021) wraps the pipeline, adjusting its miscoverage level at each step and provably driving the time-averaged

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.

miscoverage to its target rate without manual recalibration. Downstream forecasters trained on the imputed series reach single-digit WMAPE on most datasets (Appendix Q).

This work makes four contributions. First, we operationalise JEPA for inpainting 91-day gaps and use controlled ablations to establish a portable design principle: fidelity is governed primarily by the predictive representation rather than by generative complexity. A JEPA predictive encoder outperforms a reconstructive VAE latent by 25.1% MSE (Table 13) and a contrastive TS2Vec latent (Appendix U) under a matched bridge and decoder, while a lightweight 5-step flow sampler matches or improves on a 50-step DDIM sampler (-8.7% MSE, Table 22). A targeted replication on twelve hydrological catchments recovers the same encoder and sampler ordering outside electricity-load data (Section 6.5). We realise the principle in a modular pipeline that attains the lowest mean Load-only MSE across our datasets. Second, we provide distribution-free uncertainty via ACI, which corrects the under-coverage of a static calibrator (by up to 7.5 pp) to reach near-nominal empirical coverage under non-stationarity, while stochastic latent ensembles yield the lowest mean CRPS among all baselines. Third, we isolate the contribution of each design choice with ablations spanning gap length, decoder variants, and generative backends (Appendices R–X). Finally, a preliminary transfer study shows that a pooled JEPA encoder transfers to four unseen domains, matching or exceeding per-dataset oracles on three of them with only bridge and decoder adaptation (Section 7).

2 Related Work

Latent dynamics and self-supervised representations. LeCun (LeCun, 2022) proposes JEPA as the backbone for world models that learn predictive latent dynamics; early instantiations target RL planning (Ha & Schmidhuber, 2018; Hafner et al., 2020), whereas we repurpose the same principle for conditional gap inpainting with uncertainty quantification. Self-supervised time-series representations are commonly built on masked modelling, as in PatchTST (Nie et al., 2023), or contrastive learning over augmented views, as in TS2Vec (Yue et al., 2022). In parallel, joint-embedding architectures and related methods (JEPA (LeCun, 2022), I-JEPA (Assran et al., 2023), V-JEPA (Bardes et al., 2024), BYOL (Grill et al., 2020), SimSiam (Chen & He, 2021), VICReg (Bardes et al., 2022), TS-JEPA (Ennadir et al., 2025)) produce compact embeddings but do not generate long, high-fidelity trajectories; their lossy latent predictions sacrifice the temporal detail needed for faithful gap reconstruction. We address this by pairing a JEPA encoder with an explicit generative bridge that operates within the learned latent space. **Generative imputation and flow matching.** CSDI (Tashiro et al., 2021) and SSSD (López Alcaraz & Strodthoff, 2023) apply diffusion in observation space; GP-VAE (Fortuin et al., 2020) uses Gaussian process priors in a VAE latent space; we operate in a JEPA latent space and add conformal coverage. Our VAE+Bridge ablation (Table 13) turns that comparison into a controlled swap of the latent alone. Flow matching (Lipman et al., 2023) and rectified flows (Liu et al., 2023) learn straight ODE paths integrable in 5–10 steps, which we adopt as a drop-in replacement for DDIM. **Conformal prediction.** CQR (Romano et al., 2019) provides distribution-free intervals under exchangeability (Vovk et al., 2005), but its static calibration leads to under-coverage in non-stationary load data. ACI (Gibbs & Candès, 2021; Zaffran et al., 2022) updates α_t online, maintaining long-run coverage under distributional shift. To our knowledge, SPLICE is the first work to wrap a deep generative imputation pipeline with ACI (Table 9, appendix).

3 Problem Formulation

We consider a multivariate time series $\mathbf{x} = \{x_1, \ldots, x_T\}$ observed at hourly resolution, where each $x_t \in \mathbb{R}^d$ contains the load measurement and associated covariates (temperature, humidex, wind speed, weather code, and calendar features). A contiguous gap interval $[t_a, t_b] \subset [1, T]$ is unobserved, and we define:

$$x_{\text{obs}} = \{x_t : t \notin [t_a, t_b]\}, \quad x_{\text{gap}} = \{x_t : t \in [t_a, t_b]\}. \tag{1}$$

The goal is to produce (i) imputed trajectories $\hat{x}_{\text{gap}}$ that faithfully reconstruct the temporal structure of the missing interval, and (ii) prediction intervals $[l_t, u_t]$ for each $t \in [t_a, t_b]$ satisfying a long-run coverage guarantee:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{x_t \in [l_t, u_t]\} \geq 1 - \alpha, \tag{2}$$

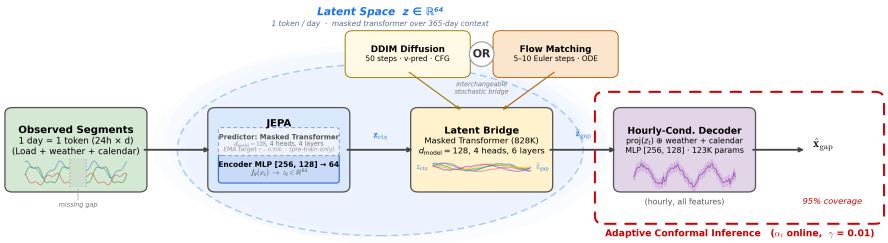


Figure 1: Overview of the SPLICE pipeline. Multivariate daily segments (load, temperature, wind, solar angle, and other covariates) are encoded by JEPA into 64-dimensional latent embeddings. A conditional latent bridge (four modes; see §4.2) generates gap trajectories. A weather- and calendar-conditioned MLP decoder maps back to hourly signal space. ACI wraps the output with adaptive prediction intervals.

where $\alpha \in (0, 1)$ is the target miscoverage rate (we use $\alpha = 0.05$ throughout). We target a coverage *floor* rather than an unconditional error level, for the asymmetry set out in Section 1. This guarantee must hold without distributional assumptions, in particular without exchangeability, which is routinely violated in load time series due to seasonal drift, weather regime changes, and demand growth.

We decompose the problem into three stages: *representation* (a JEPA encoder maps daily segments into latent embeddings $z_t = f_\theta(x_t) \in \mathbb{R}^p$), *generation* (a conditional latent bridge produces candidate trajectories $\hat{z}_{\text{gap}} \sim G_\phi(\cdot \mid z_{\text{obs}})$), and *calibration* (ACI wraps decoder output with adaptive prediction intervals). Each stage is detailed in Section 4.

4 Method

Figure 1 illustrates the full architecture. We describe each component below.

4.1 JEPA Encoder: Latent State Inference

We adopt a Joint Embedding Predictive Architecture (JEPA) as the representation backbone. Each daily load segment $x_t \in \mathbb{R}^{24 \times d}$ (24 hours $\times$ d features) is encoded into a compact representation $z_t \in \mathbb{R}^{64}$ via a two-layer MLP encoder:

$$z_t = f_\theta(x_t) = \text{MLP}(x_t;\ [256, 128] \to 64). \tag{3}$$

The JEPA predictor is a masked Transformer (d_{model}=128, n_{heads}=4, n_{layers}=4) that predicts the latent embedding of masked days from the context of observed days. The training objective combines: **(1) Prediction loss:** cosine similarity between predicted and target embeddings (EMA target encoder, decay $= 0.996$). **(2) Variance regularisation:** prevents embedding collapse by penalising low variance across the batch (Bardes et al., 2022). **(3) Covariance regularisation:** decorrelates embedding dimensions to encourage information spreading. The full JEPA objective is:

$$\mathcal{L}_{\text{JEPA}} = \underbrace{1 - \frac{\langle \hat{z},\, \bar{z} \rangle}{\|\hat{z}\|\, \|\bar{z}\|}}_{\text{cosine prediction}} + \lambda_v \underbrace{\sum_{j=1}^{p} \max\!\left(0,\, 1 - \sqrt{\text{Var}(z_j) + \varepsilon}\right)}_{\text{variance hinge}} + \lambda_c \underbrace{\frac{1}{p^2} \sum_{i \neq j} C_{ij}^2}_{\text{off-diagonal covariance}}, \tag{4}$$

where $\hat{z}$ is the predictor output, $\bar{z}$ the exponential-moving-average target, $C = \text{Cov}(\mathbf{z})$ the batch covariance matrix, $\lambda_v = 0.05$, $\lambda_c = 0.001$, and $\varepsilon = 10^{-4}$.

A separate decoder MLP ($[128, 256] \to d_{\text{input}}$) is trained to reconstruct x_t from z_t, providing the signal-space mapping needed for evaluation. The full JEPA module has approximately 1.06M parameters.

4.2 Conditional Latent Bridge: Trajectory Generation

The bridge architecture supports a *controllable spectrum of stochasticity*, allowing practitioners to trade point accuracy for distributional richness: **(i) Deterministic** (σ=0): A masked Transformer (828K parameters) directly predicts gap embeddings from context, optimised for MSE. This variant produces the lowest point-forecast error and is used in the main comparison (Section 6.2). **(ii) Noise-perturbed** (σ=0.15): Gaussian noise is added to the deterministic bridge before decoding, yielding M diverse trajectories. This lightweight mechanism achieves well-calibrated forecasts (lowest CRPS; Section 6.4) and provides conformity scores for ACI (Section 4.3). **(iii) Generative** (1.36M parameters): Conditional DDIM diffusion or flow-matching models generate gap embeddings with maximum trajectory diversity. Evaluated in Appendix R. All three variants share the same Transformer backbone (d_{model}=128, n_{heads}=4, n_{layers}=6); the generative variants add a stochastic denoising or flow wrapper. Small perturbations in the JEPA space (σ=0.15) produce diverse trajectories with limited loss of reconstruction accuracy. **Flow matching as the preferred generative backend.** Flow matching (Lipman et al., 2023; Liu et al., 2023) learns a velocity field that transports samples along *straight* ODE paths from noise to data. At inference, gap embeddings are generated by Euler integration of the learned field $v_\theta(z_t, t, c)$ conditioned on context embeddings $c = z_{\text{obs}}$:

$$z_{t-\Delta t} = z_t - \Delta t \cdot v_\theta(z_t, t, c), \quad \Delta t = 1/N, \tag{5}$$

where N=5−10 steps suffice in the learned JEPA latent space, yielding a 5−10× speedup over the 50-step DDIM sampler at comparable quality. Classifier-free guidance (Ho & Salimans, 2022) amplifies conditioning by interpolating between conditional and unconditional velocity predictions, with per-dataset guidance scale selected on validation (Appendix C). Full diffusion and flow-matching formulations (forward process, v-prediction, training objectives) are given in Appendix I.

4.3 Adaptive Conformal Inference

Standard conformal prediction assumes exchangeability of calibration and test data, which is violated in non-stationary time series. Conformalized Quantile Regression (CQR) (Romano et al., 2019) computes a calibration quantile $\hat{q}$ from held-out residuals:

$$\hat{q} = \text{Quantile}_{1-\alpha}\big(\{\max(\hat{q}_{\text{lo}}^{(i)} - y^{(i)},\ y^{(i)} - \hat{q}_{\text{hi}}^{(i)})\}_{i=1}^{n}\big). \tag{6}$$

In our pipeline $\hat{q}_{\text{lo}}$ and $\hat{q}_{\text{hi}}$ come from S=50 trajectories drawn through the full pipeline per calibration window rather than from parametric quantile regression; pointwise empirical quantiles at levels $\alpha/2$ and $1-\alpha/2$ define the initial band. The CQR nonconformity score is

$$R_t = \max\big(\hat{q}_{\text{lo},t} - y_t,\ y_t - \hat{q}_{\text{hi},t}\big), \tag{7}$$

which is zero when y_t lies inside the band. The collection of scores $\{R_t\}$ over all calibration windows provides the empirical distribution from which $\hat{q}$ (Eq. 6) is computed.

ACI (Gibbs & Candès, 2021) makes this adaptive by maintaining a running miscoverage level α_t that is updated at each time step based on observed coverage:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t), \quad \text{err}_t = \mathbf{1}\{y_t \notin C_t(\alpha_t)\}, \tag{8}$$

where $\gamma > 0$ is the learning rate (we use $\gamma = 0.01$) and α_t is clamped to $(0.001, 0.999)$. The key theoretical guarantee is:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \text{err}_t = \alpha. \tag{9}$$

This holds without distributional assumptions; the only requirement is that the base predictor produces valid quantile estimates on the calibration set (see Appendix A for the finite-sample bound).

4.4 Hourly-Conditioned Decoder and End-to-End Pipeline

A per-hour conditioned MLP decoder maps latent trajectories $\hat{z}_{\text{gap}}$ back to signal space. The decoder first projects each 64-dimensional JEPA embedding, one latent token per calendar day (24 hours being the natural diurnal period of load), to a 256-dimensional hidden state, broadcasts it to the 24 hourly steps of that day, and concatenates per-hour weather covariates (temperature, dew point,

wind, humidity) and calendar encodings (sine/cosine hour, day-of-week, month). A shared two-layer MLP ($[256, 128] \to d_{\text{features}}$) with LayerNorm and GELU activations then decodes each hour independently:

$$\hat{x}_{t,h} = \text{Dec}_\theta(\text{proj}(z_t),\, c_{t,h}), \quad z_t \in \mathbb{R}^{64},\; c_{t,h} = [\text{weather}_h, \text{calendar}_h]. \tag{10}$$

This per-hour conditioning closes the informational gap between latent-space and observation-space methods: the decoder receives the same weather covariates that external baselines (BRITS, SAITS) access directly within the gap. Beyond this larger capacity, the enhanced decoder adds a Load-weighted MSE loss ($5\times$ weight on the Load feature) to prioritise the target variable and noise augmentation ($\sigma = 0.15$ Gaussian noise on 50% of embedding batches) to improve robustness to bridge reconstruction errors; training additionally uses cosine learning-rate annealing. The decoder is trained with the JEPA encoder frozen (≈123K parameters). These three choices are the enhancements isolated in Section T. Training proceeds in four sequential stages: JEPA encoder and predictor (Section 4.1), then the decoder on frozen embeddings (this section), then the bridge (Section 4.2), then conformal calibration (Section 4.3). The stages are decoupled by design rather than trained end-to-end: back-propagating a decoder loss into the JEPA encoder would reintroduce the representation-collapse pressure that the EMA-target, prediction-only objective is built to avoid, biasing the encoder toward decoder-specific rather than predictive features. Keeping the encoder frozen is also what makes the transfer result (Section 7) possible, since a pooled encoder reused across unseen domains was never specialised to a single decoder. Algorithm 1 (Appendix) summarises the inference procedure.

5 Experimental Setup

5.1 Datasets

We evaluate on thirteen hourly-resolution datasets (Table 10, appendix): nine proprietary utility feeds from the northeastern US (2015–2025), three UCI Electricity (Trindade, 2015) clients (MT_001/150/320), and ETTh1 (Zhou et al., 2021). The twelve utility and UCI feeds include known-future meteorological covariates (temperature, humidex, wind, weather code); ETTh1 carries no meteorological covariates, exposing only the station's oil-temperature channel in their place. Load profiles span five orders of magnitude across residential, commercial, industrial, lighting, and transformer monitoring domains. To test whether our design conclusions are specific to electricity, we additionally evaluate on CAMELS-SPAT (Knoben et al., 2025), a public hydrometeorological dataset. Hourly discharge is the target variable and ERA5-derived basin-averaged air temperature, humidex and wind speed are supplied as covariates, so the method is evaluated with the exogenous drivers its design assumes rather than in a covariate-free port. We use twelve US catchments, selected *before* any model was trained, on target completeness and hydrologic diversity alone, and apply the $365 + 91$ day protocol of Section 5.2 unchanged (Appendix F).

5.2 Gap Simulation Protocol

We create contiguous 91-day (2,184-hour) gaps in the test period, representative of real-world meter outages, rather than the randomly scattered missing values common in imputation benchmarks. Extended outages are operationally more consequential than isolated dropouts, and they remove the local context that makes interpolation between nearby observations sufficient. Each evaluation window spans $365 + 91 = 456$ days: 365 days of observed context followed by 91 masked days. A sliding-window procedure yields 20–40 instances per dataset. All features are z-score normalised using training-set statistics only. Appendix B provides additional details; shorter gaps (7/30 days) are evaluated in Appendix O.

5.3 Baselines and Ablations

Every method receives the same train/test splits, 91-day gap masks, and 14-dimensional input features; the seven external methods are each evaluated over three seeds. **External baselines.** Seasonal Naïve copies each missing day from the same week of the previous year. The six learned baselines span bidirectional recurrent imputation (BRITS; Cao et al., 2018), diagonally-masked self-attention (SAITS; Du et al., 2023), conditional score-based diffusion (CSDI; Tashiro et al., 2021), 2D

Table 1: Load-only gap MSE vs. external imputation baselines (min-max normalised to $[0, 1]$ on truth range). Lower is better; bold marks the best per dataset. Values show mean$\pm$std over 3 random seeds (42, 43, 44); Seasonal Naïve is deterministic (single run). †TimesNet diverged on UCI_Elec_MT_150 (catastrophic overfitting of 37M-parameter model on a small dataset). ‡SEMAResNstar1009 is near-degenerate (MSE rounds to zero under min-max normalisation); mean and win counts exclude this dataset.

Dataset	Seasonal	BRITS	SAITS	CSDI	TimesNet	PatchTST	ImputeFormer	SPLICE
PepcoCOM	0.0121	0.032±.012	0.022±.004	0.111±.001	0.069±.007	0.030±.002	0.034±.000	**0.003**±.001
RICom1013	0.0300	0.034±.017	0.029±.008	0.201±.002	0.088±.006	0.044±.007	0.053±.001	**0.022**±.004
RIInd1014	**0.0402**	0.154±.035	0.103±.012	0.200±.005	0.298±.048	0.073±.006	0.058±.001	0.202±.022
RIRes1012	0.0880	0.036±.022	0.049±.008	0.233±.001	0.098±.026	0.062±.020	0.092±.002	**0.035**±.006
SEMAResNstar1009‡	0.2486	0.068±.087	0.047±.015	0.266±.002	0.420±.062	0.069±.018	0.103±.004	**0.000**±.000
UES_NH_Med	0.4008	0.358±.164	0.390±.068	0.619±.003	1.403±.419	0.415±.029	0.491±.012	**0.206**±.044
WCMA1010res	0.0243	0.022±.006	0.022±.002	0.087±.000	0.062±.009	0.036±.000	0.037±.000	**0.009**±.001
WCMA1011lig	0.1614	0.096±.017	**0.027**±.004	0.982±.023	0.191±.031	0.074±.016	0.145±.047	0.148±.059
WCMAnatGridRes1004	0.0845	0.026±.006	0.019±.001	0.206±.001	0.084±.010	0.033±.003	0.045±.001	**0.011**±.006
UCI_Elec_MT_001	—	0.002±.000	0.002±.001	0.041±.001	0.050±.002	**0.001**±.000	0.006±.000	0.003±.002
UCI_Elec_MT_150	—	0.010±.000	0.011±.001	0.142±.000	6.659†±1.50	0.007±.001	0.023±.000	**0.005**±.000
UCI_Elec_MT_320	—	0.017±.010	0.043±.008	0.122±.001	0.067±.001	0.046±.006	0.058±.001	**0.008**±.001
ETTh1	—	0.025±.008	0.049±.001	0.078±.000	0.056±.005	0.049±.000	0.050±.000	**0.022**±.002
Mean‡	0.105±.043	0.068±.028	0.064±.029	0.252±.076	0.760±.524	0.073±.031	0.091±.036	**0.056**±.022
Wins‡	1	0	1	0	0	1	0	**9**

convolution over learned period–frequency pairs (TimesNet; Wu et al., 2023), a channel-independent patch Transformer (PatchTST; Nie et al., 2023), and an imputation-specific Transformer with low-rank temporal attention (ImputeFormer; Nie et al., 2024). CSDI is the only other probabilistic baseline. Table 16 (appendix) lists their architecture and training settings. **Internal ablation variants. VAE+Bridge** (legacy) replaces JEPA with a β-VAE (dim $= 32$) under the same masked Transformer bridge. **JEPA-Only** decodes the embeddings directly with an MLP, whereas **JEPA+Bridge** uses deterministic latent prediction. **JEPA+Diffusion** adds DDIM latent diffusion; **JEPA+Flow-Matching** replaces it with a 5-step Euler ODE. **Full+CQR** uses static conformal calibration, whereas **Full+ACI** is the complete system with adaptive α_t.

5.4 Evaluation and Implementation

We evaluate imputation with all-feature gap-frame MSE on normalised $[0, 1]$ data for internal ablations, load-only gap MSE on min–max scaled data for external baselines, and MAPE and RMSE/MAE in physical units (Appendix D). We assess uncertainty through CRPS, empirical coverage against the 95% target, and mean interval width. Downstream, we report WMAPE and MAE from DLinear (Zeng et al., 2023) and TFT (Lim et al., 2021) trained on the filled series (Appendix Q). All models are trained on a single NVIDIA RTX A4000 (16 GB) with PyTorch (Paszke et al., 2019) and early stopping. Table 16 (appendix) lists architecture and training hyperparameters. The full pipeline uses 1.9–3.2M parameters, trains in $\sim$8–12 min per dataset, and inference for a 91-day gap takes 6.5–9.0 ms (Table 12, appendix).

6 Results

6.1 Gap-Filling Accuracy

Internal ablations (Table 13, Figure 6, appendix) are designed to isolate the contribution of representation learning to long-horizon imputation. The results confirm that both the bridge architecture and JEPA embeddings contribute: the full pipeline reduces MSE by 53.2% vs. JEPA-only and 25.1% vs. a legacy VAE+Bridge baseline.

6.2 Comparison with External Imputation Baselines

We benchmark the enhanced hourly-conditioned decoder (Section 4.4) against the seven external baselines of Section 5.3, the six learned ones trained via PyPOTS (Du et al., 2025). Table 1 reports Load-only MSE on the min-max $[0, 1]$ scale used by all of them.

Analysis. SPLICE records the lowest mean Load-only MSE and wins 9 of 12 non-degenerate datasets, with no baseline winning more than one (Table 1; Figure 7, appendix). Three of the four public benchmarks (UCI + ETTh1) are among those wins (Figure 8, appendix). MAPE on the original scale confirms the pattern (lowest on 7 of 11 reportable datasets; Table 14, appendix). SAITS retains an advantage on WCMA1011lig, and PatchTST on UCI_Elec_MT_001, where bridge embedding recovery loses fine-grained structure. Seasonal Naïve wins on RIInd1014 for a different reason: it copies the previous year exactly. **Statistical significance.** One-sided Wilcoxon signed-rank tests on per-dataset, three-seed means show significant improvements over CSDI and TimesNet (both $p<0.001$) and ImputeFormer ($p=0.021$). Differences from Seasonal Naïve, SAITS, BRITS, and PatchTST do not reach $p<0.05$ (Appendix P). SPLICE outperforms PatchTST on 9 of 12 datasets and ImputeFormer on 10 of 12. **Latent-space reconstruction.** SPLICE reaches this accuracy while compressing all eight channels into a single 64-dimensional embedding, whereas every baseline models the load channel in observation space; Figure 5 (appendix) shows the resulting reconstructions on three representative profiles. Compression to a shared latent does not reduce load reconstruction accuracy on these datasets. Decoder capacity and the Load-weighted loss contribute to this result (Appendix T).

6.3 Conformal Prediction Intervals

Table 11 (appendix) compares static CQR and adaptive conformal inference (ACI) across the nine proprietary datasets ($1-\alpha=0.95$ target). ACI fixes coverage violations on two critical datasets: RIInd1014 (86.0% $\rightarrow$ 93.5%, +7.5 pp) and UES_NH_Med (89.9% $\rightarrow$ 93.4%, +3.5 pp). On datasets where static CQR already exceeded the target (e.g. WCMA1011lig at 97.2%), ACI tightens the bands (95.4%, -1.8 pp), recovering efficiency without losing validity. Figures 9 and 10 (appendix) show the adaptive α_t trajectories and resulting prediction bands.

6.4 Distributional Quality and Uncertainty Calibration

We complement point accuracy with distributional metrics: CRPS, interval coverage (Section 6.3), and boundary smoothness (Appendix K). **CRPS.** We compute CRPS via the energy form (Gneiting & Raftery, 2007) using a 20-member ensemble generated by perturbing bridge embeddings with Gaussian noise ($\sigma=0.15$) and decoding each independently. SPLICE-ensemble achieves the lowest average CRPS (0.182; Table 15, appendix), outperforming SAITS (0.197, -7.8%) and its own deterministic MAE (0.198; -10.5% mean per-dataset reduction), showing that latent perturbations provide useful predictive variation. SPLICE is the *only* evaluated method that provides conformally calibrated prediction intervals; none of the baselines, including the probabilistic CSDI, provides coverage guarantees (Table 9, appendix). Figure 11 shows per-dataset CRPS, and Figure 13 quantifies the ensemble improvement over deterministic prediction. **Ablations and sensitivity.** Once the predictive latent is fixed, FM-A wins 10/13 datasets against DDIM at $10\times$ fewer steps (-8.7% MSE), and the enhanced decoder and ensembling contribute a further -9.3% (Appendices R–X). SPLICE holds the lowest mean rank (2.33–2.67) across 7/30/91-day gaps with 17/32 total wins (53%; Appendix O). A latent-dimension and EMA-decay sweep puts $d_{\text{repr}}=64$ at or near optimal on the public benchmarks (Appendix V), and re-running validation on a disjoint pre-gap window leaves the ranking unchanged (median $\Delta=0.0\%$; Appendix W).

6.5 Cross-Domain Replication on Hydrological Data

To check that the architectural conclusions above are not artefacts of electricity load, we replicate the three controlled swaps on twelve CAMELS-SPAT catchments with the pipeline unchanged (Appendix F), scored both on gap-MSE and on hydrology-standard criteria (KGE, NSE, FHV/FLV). The ordering reproduces: the predictive JEPA encoder beats both a contrastive and a reconstructive alternative, and the 5-step flow-matching sampler beats the 50-step DDIM sampler (Table 2).

7 Transferability of JEPA Representations

If one encoder serves many feeds, its training cost amortises over deployments and a new feed needs only downstream adaptation. **Pooled encoder.** We train a single JEPA encoder on all nine proprietary datasets at once ("pooled JEPA"), projecting each to the same eight common features and

Table 2: Cross-domain replication on twelve CAMELS-SPAT catchments (seed 42) (Appendix F).

Controlled swap	Wins	Median reduction
JEPA vs. TS2Vec (contrastive)	9/12	16.2%
JEPA+Bridge vs. VAE+Bridge (reconstructive)	10/12	13.2%
Flow Matching (5-step) vs. DDIM (50-step)	8/12	12.5%

Table 3: Transfer evaluation: gap-inpainting MSE on four held-out datasets under three adaptation regimes (single seed). Bold marks the best result per dataset; percentages are relative to Regime A.

Dataset	Regime A	Regime B	Regime C
UCI_001	0.2072	0.2889 (+39%)	**0.1967** (−5%)
UCI_150	0.0865	**0.0804** (−7%)	0.2036 (+135%)
UCI_320	0.0911	**0.0888** (−3%)	0.1507 (+65%)
ETTh1	0.1875	**0.1800** (−4%)	0.2013 (+7%)
Mean	0.1431	0.1595 (+11%)	0.1881 (+31%)
Median	0.1393	**0.1346** (−3%)	0.1990 (+43%)

normalising it independently. Capacity is held fixed at the per-dataset value (∼1.06M parameters, d_{repr}=64); the two encoders differ only in training data. In the UMAP projection, held-out feeds interleave with the training domains and seasonal structure persists across both groups (Figure 4, Appendix G). **Adaptation spectrum.** We evaluate on four held-out sets, three UCI residential load feeds and the ETTh1 transformer oil-temperature series, none seen during pooled training. Regime A (oracle) trains a complete per-dataset pipeline. Regime B freezes the pooled encoder and trains a dataset-specific bridge and decoder. Regime C additionally freezes the pooled bridge and trains only a lightweight decoder, making the transfer zero-shot at the representation level; the pooled bridge expects 17-dimensional conditioning, so the Cloud and Precipitation channels are zero-filled on these four sets, which lack them. **Results.** Regime B improves on the oracle on three of four domains, by 3 to 7%, for ∼60 s of fine-tuning (Table 3, single seed). UCI_001 is the exception at +39% and carries the highest error under both regimes, which is what separates the mean (+11%) from the median (−3%). Pooling over nine diverse feeds may regularise the embedding. Regime C beats the oracle only on UCI_001, after 2 s of decoder training with the pooled bridge frozen, and costs 7 to 135% elsewhere. Bridge adaptation therefore remains necessary once the target profile departs from the training pool. **Few-shot budget.** Across fine-tuning budgets of 7 to 365 days on the four held-out sets (three seeds each), 90 days of target data suffice to match or beat SAITS and BRITS trained on the full 1 241-day history (e.g. UCI 150: 0.011±0.000 vs. SAITS 0.011; UCI 320: 0.009±0.001 vs. BRITS 0.020; Appendix Y).

8 Discussion

Design principle: fidelity is governed primarily by predictive latent geometry, and a short sampler suffices. Under matched downstream components, replacing the JEPA latent degrades reconstruction: a reconstructive VAE latent costs 25.1% MSE (Table 13) and a contrastive TS2Vec latent loses on 9 of 12 load datasets (Appendix U). Both substitutions also displace the timing of the daily peak: the contrastive latent on 8 of 11 datasets and on two where gap-MSE favours TS2Vec (Appendix U.1), the reconstructive latent on 6 of 8 by a median of sixteen minutes (Appendix U.2), placing the encoder's advantage on an operational criterion that point accuracy does not expose. Once the predictive representation is fixed, a 5-step flow-matching sampler matches or improves on a 50-step DDIM (−8.7% MSE, Table 22): high-fidelity stochastic reconstruction does not require a long reverse-diffusion trajectory. The same encoder and sampler ordering recovers on twelve hydrological catchments, outside electricity-load data (Section 6.5). We characterise this geometry behaviourally, through the traversal and transfer behaviour the latent space supports, rather than through a standalone geometric statistic: an input-space similarity measure rewards an encoder for preserving input neighbourhoods, whereas the predictive objective reorganises the space around predictability. The ablations also locate the division of labour: the encoder learns predictive states, the bridge reconstructs

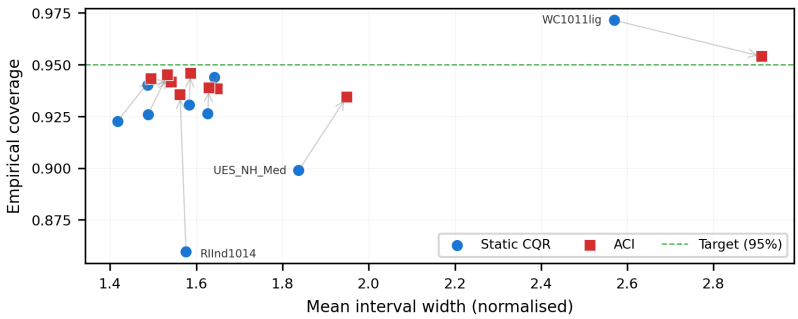


Figure 2: Coverage vs. interval width. Each point is one dataset; grey arrows show the shift from static CQR (blue circles) to ACI (red squares). Green dashed line: 95% target.

their missing trajectory, the decoder restores hourly detail, and the generative backend diversifies them. The design rule is to learn the predictive state space first, then use the least costly backend that supplies the required trajectory diversity. **Why adaptive conformal matters.** As argued in Section 1, under-coverage is the operationally consequential failure direction, yet static CQR under-covers by up to 7.5 pp under non-stationarity; ACI recovers 93–95% coverage while tightening over-wide intervals (Figure 2). **Modularity and practical implications.** A shared 64-dimensional interface connects the JEPA encoder, latent bridge, decoder, and ACI, carries all eight channels without sacrificing Load fidelity, and lets components be replaced without upstream retraining. It supports three regimes: a deterministic bridge for point reconstruction; latent perturbations for calibrated intervals and CRPS ensembles (Sections 6.3–6.4); and a generative backend for trajectory diversity, where flow matching stays closer to the deterministic solution than DDIM (Appendix R). The pooled encoder is likewise reusable on unseen datasets with only bridge and decoder adaptation (Section 7). **Limitations.** Gaps shorter than one week or longer than three months are uncharacterised. The full diffusion backend requires 50 reverse steps, which progressive distillation or consistency models could cut. CAMELS-SPAT (Section 6.5) replicates the encoder and sampler ordering on hourly streamflow, but gas, renewable generation, other physical processes, and climates outside the northeastern US, Portugal and China remain untested. In the transfer study only the few-shot experiment is replicated across seeds (Appendix Y); the zero-shot and bridge-only results are single-seed. Main results (Table 1) use three seeds with standard deviations. SPLICE achieves the lowest mean Load-only MSE, winning 9 of 12 non-degenerate datasets, and the lowest mean CRPS. The gap-MSE differences from BRITS and SAITS do not reach $p<0.05$. On bimodal or zero-inflated datasets (WCMA1011lig, RIInd1014) the 64-dimensional interface costs fine structure that observation-space methods model directly, and end-to-end fine-tuning of the bridge and decoder may reduce this cost. Two assumptions bound deployment. The ACI guarantee (Appendix A) is marginal: time-averaged miscoverage converges to α without certifying coverage at a fixed horizon, dataset or covariate value. The second is shared with BRITS and SAITS: covariates must be known or reliably forecastable across the gap; where they are not, their uncertainty has to be propagated into the intervals, which we do not model. **Broader impact.** Reliable gap-filling with calibrated uncertainty is critical for grid planning and demand-side management; over-confident imputations can lead to under-provisioning or missed demand response events. Data availability and code release details are provided in Appendix H.

9 Conclusion

We presented SPLICE, an operational instantiation of JEPA for long-horizon time-series inpainting that couples a conditional latent bridge, an hourly-conditioned decoder, and ACI. Controlled ablations establish a design principle: under a matched bridge and decoder, a reconstructive or contrastive latent underperforms JEPA on reconstruction, whereas once the representation is fixed a 5-step flow sampler matches or improves on a 50-step DDIM. The same ordering holds on 12 hydrological catchments. At 91-day gaps SPLICE attains the lowest mean Load-only MSE across thirteen datasets, winning 9 of 12 non-degenerate ones while jointly reconstructing all eight channels in a shared latent space, and the lowest mean CRPS on the nine proprietary feeds. ACI holds 93–95% empirical coverage where static conformal prediction under-covers by up to 7.5 pp. A pooled JEPA encoder transfers to four held-out datasets with only bridge and decoder adaptation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, or interpret results. Additionally, we used generative AI tools to edit the manuscript for readability, format references, and help package the anonymized supplemental material. We have reviewed all AI-assisted work: translated and edited text was checked against the authors' intended meaning and against the reported numbers, and the supplemental material was verified to reproduce the findings reported in the paper within normal reproducibility tolerances. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work raises no significant ethical concerns: it involves no human subjects and no personal data, and the models impute aggregate electricity-load data, posing negligible risk for misuse. The nine proprietary datasets are US utility records obtained under data-sharing agreements and are not redistributed (Appendix H). Section 8 discusses the positive societal impacts (grid reliability, risk-aware planning) and the negative scenario in which over-confident imputations could lead to under-provisioning or missed demand-response events.

REPRODUCIBILITY STATEMENT

All architecture details are given in Section 4, and the datasets, gap simulation, baselines, and training hyperparameters are specified in Section 5 and Table 16. Anonymized source code with single-command reproduction scripts and expected reference outputs is included in the supplemental material. The four public datasets are automatically downloaded by the provided scripts; the nine proprietary utility datasets cannot be redistributed and are documented in Appendix H.

REFERENCES

M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024.

W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li. BRITS: Bidirectional recurrent imputation for time series. In *NeurIPS*, 2018.

X. Chen and K. He. Exploring simple Siamese representation learning. In *CVPR*, 2021.

W. Du, D. Côté, and Y. Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 2023.

W. Du, Y. Yang, L. Qian, J. Wang, and Q. Wen. PyPOTS: A Python toolkit for machine learning on partially-observed time series. *arXiv:2305.18811*, 2025.

S. Ennadir, S. Golkar, and L. Sarra. Joint embeddings go temporal. *arXiv:2509.25449*, 2025.

V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. GP-VAE: Deep probabilistic time series imputation. In *AISTATS*, 2020.

I. Gibbs and E. J. Candès. Adaptive conformal inference under distribution shift. In *NeurIPS*, 2021.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent – a new approach to self-supervised learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020.

D. Ha and J. Schmidhuber. World models. *arXiv:1803.10122*, 2018.

D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

W. J. M. Knoben, C. Thébault, K. Keshavarz, L. Torres-Rojas, N. W. Chaney, A. Pietroniro, and M. P. Clark. Catchment attributes and meteorology for large-sample spatially distributed analysis (CAMELS-SPAT): streamflow observations, forcing data and geospatial data for hydrologic studies across North America. *Hydrology and Earth System Sciences*, 29(20):5791–5833, 2025. `https://hess.copernicus.org/articles/29/5791/2025/`.

Y. LeCun. A path towards autonomous machine intelligence. *openreview.net preprint*, 2022.

B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 2021.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *ICLR*, 2023.

J. M. López Alcaraz and N. Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *TMLR*, 2023.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with Transformers. In *ICLR*, 2023.

T. Nie, G. Qin, W. Ma, Y. Mei, and J. Sun. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2260–2271, 2024.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Y. Romano, E. Patterson, and E. J. Candès. Conformalized quantile regression. In *NeurIPS*, 2019.

T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Y. Tashiro, J. Song, Y. Song, and S. Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *NeurIPS*, 2021.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. TS2Vec: Towards universal representation of time series. In *AAAI*, 2022.

M. Zaffran, O. Féron, Y. Goude, J. Josse, and A. Dieuleveut. Adaptive conformal predictions for time series. In *ICML*, 2022.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are Transformers effective for time series forecasting? In *AAAI*, 2023.

A. Trindade. ElectricityLoadDiagrams20112014 data set. UCI Machine Learning Repository, 2015. https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014. DOI: 10.24432/C58C86.

Open-Meteo. Free weather API—ERA5 historical reanalysis. https://open-meteo.com/, 2024.

T. Hang, S. Gu, C. Li, J. Bao, D. Chen, H. Hu, X. Geng, and B. Guo. Efficient diffusion training via Min-SNR weighting strategy. In *ICCV*, 2023.

H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *ICLR*, 2023.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021.

A ACI Coverage Guarantee

For completeness, we restate the coverage guarantee of Adaptive Conformal Inference (Gibbs & Candès, 2021) and make explicit the convergence rate that underpins our experimental design.

Proposition 1 (Asymptotic coverage, Gibbs & Candès 2021)**.** *Let $\{\alpha_t\}_{t\geq 1}$ evolve according to Eq.* (8) *with learning rate $\gamma > 0$ and clamping to $[\varepsilon, 1-\varepsilon]$. Then, regardless of the data-generating process,*

$$\Big|\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t - \alpha\Big| \;\leq\; \frac{1-2\varepsilon}{\gamma\, T}, \tag{11}$$

so $\lim_{T\to\infty} T^{-1}\sum_t \mathrm{err}_t = \alpha$.

Proof sketch. The update rule ensures $|\alpha_{t+1} - \alpha_t| \leq \gamma$, and clamping keeps $\alpha_t \in [\varepsilon, 1-\varepsilon]$. Define $\Delta_t = \alpha_t - \alpha$. Summing the update over $t = 1, \ldots, T$ and rearranging: $\sum_{t=1}^{T}(\mathrm{err}_t - \alpha) = (\alpha_1 - \alpha_{T+1})/\gamma$, whence $\big|\frac{1}{T}\sum_t \mathrm{err}_t - \alpha\big| \leq |\alpha_1 - \alpha_{T+1}|/(\gamma T) \leq (1-2\varepsilon)/(\gamma T)$. □

In our experiments $\gamma = 0.01$ and $\varepsilon = 0.001$, giving a worst-case deviation of $\leq 99.8/T$. This distribution-free bound is conservative: it falls below 0.01 only for $T > 9{,}980$, and over the $2{,}184$ hourly steps of a 91-day gap it guarantees only ≤ 0.046. The worst case requires an adversarial error sequence; measured coverage (Table 11) is substantially closer to nominal.

B Training and Evaluation Protocol

Feature normalisation. All input features (Load, Temperature, Humidex, Wind, etc.) are z-score normalised using per-feature mean and standard deviation computed on the training split only. The same statistics are applied to the validation and test data to prevent information leakage. Conditioning features for the decoder (hourly weather and calendar encodings) undergo a separate z-score normalisation with statistics fitted on the training portion of the daily conditioning matrix.

Train/validation split. The most recent 15% of calendar days form the validation set; all earlier days constitute the training set. This temporal split ensures no future information leaks into training, mirroring the deployment scenario of retrospective gap-filling. As a robustness check, Appendix W confirms that the early-stopping *epoch*, chosen on a validation region whose trailing edge can coincide with the start of the gap, confers no advantage: re-running the full pipeline with a strictly disjoint pre-gap validation split leaves performance statistically indifferent (median gap-MSE change 0.0%).

Sliding window construction. Each evaluation window spans $365 + 91 = 456$ consecutive days: the first 365 days provide observed context and the final 91 days are masked. A stride-1 sliding window generates all valid windows whose start index satisfies $\mathrm{start} + 365 \geq 0.85 \cdot n_{\mathrm{days}}$, producing 20–40 evaluation windows per dataset depending on the available data span.

Decoder training. The hourly-conditioned decoder is trained on the training split with the JEPA encoder frozen. For each training sample, the encoder produces a 64-dimensional embedding for each day, which the decoder maps to $24 \times d_{\text{features}}$ hourly predictions. With probability 0.5, Gaussian noise ($\sigma = 0.15$) is added to the embedding before decoding, forcing the decoder to be robust to bridge reconstruction errors. The Load-weighted loss amplifies the gradient contribution of the Load feature by a factor of 5.

Conformal calibration protocol. For standard CQR, all validation windows except the last serve as the calibration set; the final window is used for evaluation. For ACI, the first 50% of validation windows initialise the base quantile $\hat{q}$ using $S{=}50$ DDIM samples per window; the remaining windows are processed in temporal order with α_t updates after each window (using $S{=}20$ samples). Coverage and interval width are reported for the online phase only.

C Classifier-Free Guidance Scale Sensitivity

Table 4 reports the gap-frame MSE for guidance scales $w \in \{1, 2, 3, 4, 5\}$ across the nine proprietary datasets. The optimal scale is dataset-dependent: low-variance feeds (RIInd1014, SEMAResNstar1009) prefer $w{=}1$ (no guidance), while volatile feeds (UES_NH_Med) benefit from strong guidance ($w{=}5$). Commercial and residential loads cluster around $w{=}2$–4. The sensitivity is moderate: across all datasets, the worst scale degrades MSE by at most $\sim$37% relative to the best, and most datasets show $<$10% variation between the top two scales.

Table 4: Guidance-scale sweep: gap-frame MSE (mean $\pm$ std over DDIM samples). Bold marks the per-dataset optimum. All values on normalised $[0, 1]$ scale.

Dataset	$w{=}1$	$w{=}2$	$w{=}3$	$w{=}4$	$w{=}5$
PepcoCOM	.291	.288	**.283**	.289	.289
RICom1013	.343	**.330**	.342	.355	.364
RIInd1014	**.508**	.554	.603	.652	.698
RIRes1012	.363	**.337**	.339	.348	.362
SEMAResNstar1009	**.257**	.271	.286	.311	.335
UES_NH_Med	.525	.443	.402	.397	**.395**
WCMA1010res	.322	.314	.304	**.294**	.300
WCMA1011lig	.389	.365	**.363**	.364	.371
WCMAnatGridRes1004	.420	.414	**.411**	.424	.437

D RMSE and MAE in Physical Units

Table 5 reports gap-reconstruction error in original measurement units, derived from the min-max normalised Load-only MSE (Table 1) and the observed Load range of each dataset. RMSE is computed as $\text{RMSE} = \sqrt{\text{MSE}_{[0,1]}} \times (\max - \min)$; MAE is approximated under a Gaussian error assumption as $\text{MAE} \approx \sqrt{2/\pi}\ \text{RMSE}$.

Table 5: Gap-reconstruction error in physical units. Units vary by dataset: kW for utility-metered feeds, kWh for hourly-integrated feeds.

Dataset	RMSE	MAE	Unit
PepcoCOM	11.13	8.88	kW
RICom1013	144.45	115.25	kW
RIInd1014	50.97	40.66	kW
RIRes1012	44.49	35.50	kW
SEMAResNstar1009	<0.01	<0.01	kWh
UES_NH_Med	17.29	13.80	kWh
WCMA1010res	23.33	18.61	kWh
WCMA1011lig	5.20	4.15	kWh
WCMAnatGridRes1004	46.27	36.92	kW
UCI_Elec_MT_001	1.51	1.20	kWh
UCI_Elec_MT_150	3.92	3.13	kWh
UCI_Elec_MT_320	18.55	14.80	kWh

E UCI Electricity Dataset Preprocessing

We use the UCI ElectricityLoadDiagrams20112014 dataset (Trindade, 2015), which contains 15-minute electricity consumption readings (in kWh) for 370 Portuguese clients spanning 2011–2014. Three clients were selected to span the variance spectrum: MT_001 (low variance, $\sigma = 5.6\,\text{kWh}$), MT_150 (medium, $\sigma = 13.2\,\text{kWh}$), and MT_320 (high, $\sigma = 28.0\,\text{kWh}$).

Temporal aggregation. The native 15-minute resolution is resampled to hourly by taking the mean of each four-reading block, matching the hourly granularity of the proprietary utility datasets.

Weather covariates. Hourly ERA5 reanalysis weather for Lisbon ($38.72°$N, $9.14°$W) was retrieved via the Open-Meteo historical API (Open-Meteo, 2024): temperature, dew point, relative humidity, wind speed, and WMO weather code. Temperature and dew point are converted from Celsius to Fahrenheit; wind from km/h to mph; and Humidex is approximated via the Canadian formula, all for consistency with the proprietary pipeline feature set.

Output format. Each client is exported as a semicolon-delimited CSV with columns `Date;Heure;Temperature;Humidex;Weather;Wind;Load`, identical to the proprietary data format. This allows the entire training, bridge, and evaluation pipeline to process UCI clients without any code modifications.

F Cross-Domain Replication on CAMELS-SPAT

Dataset. CAMELS-SPAT (Knoben et al., 2025) is a public large-sample hydrometeorological dataset providing streamflow observations, basin-averaged meteorological forcings and geospatial attributes for North American catchments. We use its hourly US subset. Unlike a covariate-free port of an unrelated series, CAMELS-SPAT distributes the exogenous drivers alongside the target, so SPLICE can be evaluated under the conditions its design assumes.

Target and covariates. Hourly discharge is the target variable. ERA5-derived basin-averaged air temperature, humidex and wind speed (supplied with the dataset) are mapped onto the covariate slots used by the electricity pipeline, together with the deterministic calendar features. No architectural or hyperparameter change was made: the same daily $24 \times n$ frames, 365-day context, 91-day gap, encoder, bridge, decoder and samplers are used throughout.

Catchment selection. Twelve catchments were selected *before* any model was trained, using two criteria only: (i) the length of the longest contiguous run with no missing target values, which must exceed the $365 + 91$ days the protocol requires, and (ii) coverage of the CAMELS-SPAT catchment

size classes (headwater, meso-scale, macro-scale) and of distinct climate regions. No model output influenced the selection. Table 6 lists the resulting set.

Table 6: The twelve CAMELS-SPAT catchments used for the cross-domain replication. Identifiers are CAMELS-SPAT station codes; area is the basin drainage area and class the CAMELS-SPAT scale category, while the coordinates give the location of the gauging station supplying the discharge series. The run column is the longest contiguous target-complete run, which determined eligibility.

Catchment	Class	Area (km^2)	Lat.	Lon.	Run (d)	Run start
USA_02296500	meso-scale	691.4	27.375	−81.796	4590	2004-12-24
USA_11532500	macro-scale	1584.7	41.792	−124.076	4041	1991-10-20
USA_11284400	headwater	41.7	37.842	−120.185	3692	2006-08-05
USA_08023080	meso-scale	183.5	31.979	−93.934	3403	2007-10-02
USA_12143600	headwater	168.3	47.453	−121.718	2963	1996-02-05
USA_08158810	headwater	31.9	30.155	−97.940	2494	2016-03-06
USA_08164000	macro-scale	2081.4	28.960	−96.686	2434	2012-08-23
USA_09505200	meso-scale	284.1	34.675	−111.672	2249	2016-05-06
USA_02102908	headwater	20.3	35.183	−79.178	2193	2007-06-07
USA_07056000	macro-scale	2148.8	35.983	−92.748	1600	2012-01-23
USA_02231342	meso-scale	96.3	27.569	−80.796	1417	2004-06-05
USA_01547700	meso-scale	112.3	41.060	−77.606	1071	2016-02-18

Protocol and comparisons. Each catchment is treated as an independent dataset and run through the unmodified pipeline with a contiguous 91-day gap preceded by a 365-day context, exactly as in Section 5.2. Three controlled swaps are evaluated, each changing a single component while holding the rest of the pipeline fixed, and each reported in Table 2: (i) the encoder swap replaces the JEPA encoder with TS2Vec and is scored on the target-only hourly gap-MSE produced by the deterministic bridge, matching the encoder-swap protocol of Appendix U; (ii) the reconstructive comparison replaces the JEPA encoder with the legacy VAE latent of Table 13 and is scored on the all-feature gap-frame MSE; (iii) the sampler swap replaces the 50-step DDIM sampler with the 5-step FM-A sampler on the same all-feature gap-frame MSE, matching Appendix S. Because the three rows use different error definitions (target-only versus all-feature) and different decoders, their magnitudes are not comparable across rows, and each is reported only against its own control. Results are single-seed (seed 42).

Scoring under hydrological criteria. The gap-MSE of Table 2 is the objective the pipeline is trained and deployed on. Hydrology evaluates reconstructions against a different family of criteria, several of which rank catchments quite differently from MSE, so we re-score the same stored predictions under the standard set defined in Table 7, changing nothing but the criterion. Table 8 reports, for each criterion, the number of catchments on which the JEPA arm is closer to observation, a two-sided exact sign test and the median paired difference, together with the Spearman rank correlation ρ between that criterion's per-catchment paired difference and the gap-MSE paired difference, which measures how much freedom the criterion has to reorder the panel.

Both encoder comparisons reproduce the Table 2 ordering under this second scoring. In the encoder swap, six of the ten criteria favour the JEPA arm, two are even, and two (the correlation coefficient and log-NSE) favour TS2Vec; the reconstructive swap has the same $6/2/2$ profile. This holds for criteria whose ordering of the catchments is close to independent of the MSE ordering (ρ between -0.08 and -0.27) and which were therefore free to reverse it. The correlation coefficient is the clearest of the two exceptions, favouring TS2Vec on nine of twelve catchments: TS2Vec is trained so that augmented views of the same window agree, which is the invariance structure r measures, and leading on that criterion is a direct expression of its objective. We report direction and concordance across criteria rather than significance, since hydrological skill is not itself a claim of this paper.

Conventions. FHV and FLV are computed on the flow-duration curve, that is on the sorted flow distributions, and are therefore timing-insensitive: they compare how much water a reconstruction places in the high- and low-flow regimes, not when it places it. FLV is reported as a plain relative bias over the lower 30% of the curve. Logarithmic criteria use the convention $\hat{Q} \leftarrow \max(\hat{Q}, 0)$ followed

by an offset $\varepsilon = 0.01\,\bar{Q}$ added to both series; three catchments whose low-flow denominator falls below the numerical guard are excluded from the FLV row only. The criteria differ in their ideal value, so paired differences are signed consistently: $d = a - b$ for higher-is-better criteria (NSE, log-NSE, KGE, KGE′, r), $d = |b-1| - |a-1|$ for the ratios (α, β, γ) and $d = |b| - |a|$ for the bias percentages (FHV, FLV), where a is the JEPA arm. A positive d therefore always indicates an advantage for the JEPA arm. Differences are taken per catchment before aggregation.

Table 7: Hydrological evaluation criteria. Q denotes observed discharge, $\hat{Q}$ the reconstruction, $e = \hat{Q} - Q$ and $\bar{Q}$ the observed mean over the gap; r is the Pearson correlation, $\alpha = \sigma_{\hat{Q}}/\sigma_Q$, $\beta = \mu_{\hat{Q}}/\mu_Q$, $\gamma = \mathrm{CV}_{\hat{Q}}/\mathrm{CV}_Q$ and FDC the flow-duration curve. The ideal value of each criterion is given in parentheses. All criteria are evaluated on the hourly series over the 91-day gap.

Criterion	Definition	Range (ideal)	What it probes
NSE	$1 - \sum e^2 / \sum (Q - \bar{Q})^2$	$(-\infty, 1]$ (1)	Overall variability
log-NSE	NSE computed on $\log Q$	$(-\infty, 1]$ (1)	Low-flow / baseflow regime
KGE	$1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$	$(-\infty, 1]$ (1)	Timing, variability, bias
KGE′	$1 - \sqrt{(r-1)^2 + (\gamma-1)^2 + (\beta-1)^2}$	$(-\infty, 1]$ (1)	As KGE, spread as CV ratio
r	Pearson correlation of $\hat{Q}$ and Q	$[-1, 1]$ (1)	Timing only (amplitude-blind)
α	$\sigma_{\hat{Q}}/\sigma_Q$	$[0, \infty)$ (1)	Amplitude
β	$\mu_{\hat{Q}}/\mu_Q$	$[0, \infty)$ (1)	Volume bias
γ	$\mathrm{CV}_{\hat{Q}}/\mathrm{CV}_Q$	$[0, \infty)$ (1)	Relative variability
FHV	Bias over the upper 2% of the FDC	% (0)	High flows (peaks)
FLV	Bias over the lower 30% of the FDC	% (0)	Low flows (recession)

Table 8: Hydrological re-scoring of the two encoder comparisons on the twelve CAMELS-SPAT catchments (seed 42). Every row uses the same stored predictions as Table 2; only the scoring criterion changes. ρ is the Spearman rank correlation between the criterion's per-catchment paired difference and the gap-MSE paired difference. Wins counts catchments on which the JEPA arm is closer to observation, p is a two-sided exact sign test, and Δ is the median paired difference signed so that positive favours the JEPA arm. Units differ by family and are not comparable between blocks.

		JEPA vs. TS2Vec			JEPA+Bridge vs. VAE+Bridge		
Criterion	ρ	Wins	p	Δ	Wins	p	Δ
Deployed objective (MSE units)							
Gap-MSE (Table 2)	—	9/12	0.15	+0.004	10/12	0.04	+0.023
Efficiency scores (dimensionless)							
NSE	+0.96	9/12	0.15	+0.233	6/12	1.00	−0.006
KGE	+0.69	9/12	0.15	+0.161	7/12	0.77	+0.112
KGE′	+0.83	9/12	0.15	+0.097	8/12	0.39	+0.072
log-NSE	−0.13	5/12	0.77	−0.051	5/12	0.77	−0.086
r	+0.31	3/12	0.15	−0.148	6/12	1.00	+0.003
Moment ratios (distance from 1)							
α	−0.08	7/12	0.77	+0.043	7/12	0.77	+0.116
β	+0.78	8/12	0.39	+0.196	5/12	0.77	−0.023
γ	−0.09	6/12	1.00	+0.014	8/12	0.39	+0.098
Flow-duration-curve bias (percentage points)							
FHV	−0.27	6/12	1.00	−4.5	7/12	0.77	+5.3
FLV†	−0.27	6/9	0.51	+28.9	6/9	0.51	+37.6

†Three catchments fall below the low-flow denominator guard and are excluded from this row only.

G JEPA Embedding Visualisation

Figure 3 shows t-SNE projections of the 64-dimensional JEPA embeddings for four representative datasets. The top row colours points by meteorological season (DJF/MAM/JJA/SON) and the bottom row by day type (weekday vs. weekend). Seasonal structure is clearly captured across all datasets: winter and summer clusters separate well, confirming that the JEPA encoder learns temperature-driven load patterns. Weekday/weekend separation is most pronounced for commercial loads (PepcoCOM, RICom1013), where occupancy schedules differ markedly between business days and weekends.

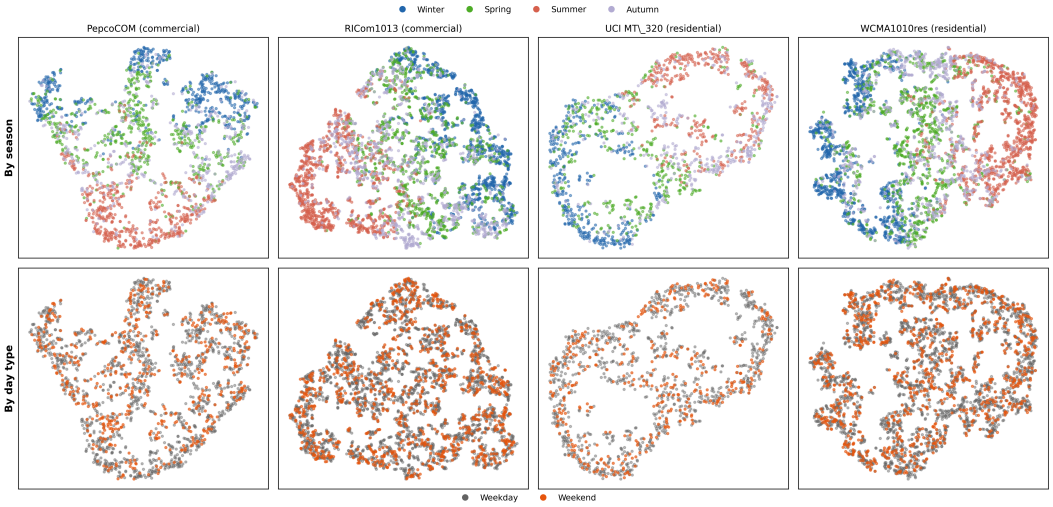


Figure 3: t-SNE projections of 64-dimensional JEPA embeddings for four datasets, coloured by season (top) and day type (bottom). The encoder captures both seasonal and occupancy structure without supervision.

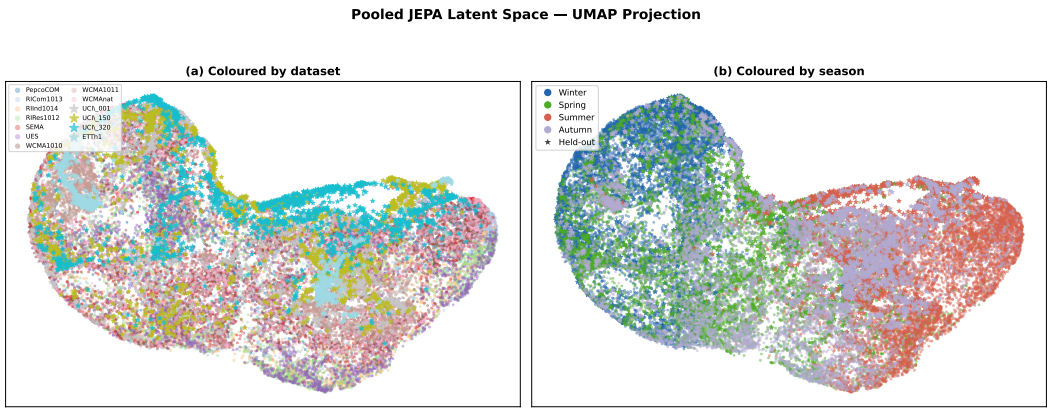


Figure 4: UMAP projection of pooled JEPA embeddings for all thirteen datasets (nine training, four held-out). **(a)** Coloured by dataset identity: training datasets (circles) form distinct but overlapping clusters; held-out datasets (stars) interleave with training domains, indicating that the pooled encoder maps unseen feeds into a shared manifold. **(b)** Coloured by season: the seasonal structure (winter/summer separation) is preserved across all domains, confirming that temporal dynamics transfer.

H Data Availability

The nine proprietary load datasets used in this study are US utility records obtained under data-sharing agreements; they cannot be publicly released. The three UCI Electricity datasets are publicly

available from the UCI Machine Learning Repository (Trindade, 2015). The ETTh1 dataset is publicly available from the Informer repository (Zhou et al., 2021). The full pipeline code, model architectures, training scripts, configuration files, and pre-trained checkpoints will be released under an open-source licence upon acceptance, enabling reproduction on any hourly-resolution load dataset with weather covariates.

I GENERATIVE BRIDGE DETAILS

This appendix provides the full mathematical formulations for the DDIM diffusion and flow-matching bridges summarised in Section 4.2.

DDIM diffusion bridge. The conditional diffusion model uses a Transformer-based denoiser with cosine noise schedule (s=0.008, 1000 timesteps) (Nichol & Dhariwal, 2021), v-prediction (Salimans & Ho, 2022), and Min-SNR loss weighting (γ=5.0) (Hang et al., 2023). DDIM sampling uses 50 steps with η=0.0 (Song et al., 2021).

The forward process adds Gaussian noise to the gap embeddings:

$$q(z_t \mid z_0) = \mathcal{N}\big(z_t;\ \sqrt{\bar{\alpha}_t}\, z_0,\ (1-\bar{\alpha}_t)\, I\big). \tag{12}$$

The denoiser $v_\phi(z_t, t, c)$ predicts the v-target $v_t = \sqrt{\bar{\alpha}_t}\,\epsilon - \sqrt{1-\bar{\alpha}_t}\, z_0$, conditioned on $c = z_{\text{obs}}$. The v-parameterisation produces more uniform gradient magnitudes across noise levels than ϵ-prediction (Salimans & Ho, 2022).

At inference, classifier-free guidance (Ho & Salimans, 2022) extrapolates away from the unconditional score:

$$\tilde{v}(z_t, t, c) = (1+w)\, v_\phi(z_t, t, c) - w\, v_\phi(z_t, t, \varnothing), \tag{13}$$

where $w \in \{1, 2, 3, 4, 5\}$ is selected per-dataset via validation MSE (Appendix C) and $\varnothing$ is the null context obtained by dropping conditioning with probability p_{uncond}=0.15 during training.

Flow-matching bridge. Given a data sample z_0 and noise $\epsilon \sim \mathcal{N}(0, I)$, the interpolation path is $z_t = (1-t)\, z_0 + t\,\epsilon$, $t \in [0, 1]$. The flow-matching objective regresses onto the conditional velocity field $u_t = \epsilon - z_0$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, z_0,\, \epsilon}\big\| v_\theta(z_t, t, c) - (\epsilon - z_0)\big\|^2. \tag{14}$$

Classifier-free guidance applies identically to the flow-matching velocity field, with the same p_{uncond}=0.15.

J Inference Algorithm

Algorithm 1 SPLICE Inference: JEPA–Bridge Gap-Filling with ACI

Require: Observed series $\mathbf{X}_{\text{ctx}} \in \mathbb{R}^{C\times F\times 24}$ (context days), per-hour covariates $\mathbf{c}$, gap length G
Ensure: Imputed gap $\hat{\mathbf{X}}_{\text{gap}} \in \mathbb{R}^{G\times F\times 24}$, confidence bands (L_t, U_t)

// Stage 1: Encode context
for each day $d \in \{1,\dots,C\}$ **do**
  $\mathbf{z}_d \leftarrow \text{Enc}_\theta(\mathbf{x}_d)$ ▷ JEPA online encoder
end for
// Stage 2: Predict gap embeddings
$\hat{\mathbf{Z}}_{\text{gap}} \leftarrow \text{Bridge}_\phi(\mathbf{Z}_{\text{ctx}}, \mathbf{c}, \mathbf{m})$ ▷ Masked Transformer (det./diffusion/FM)
// Stage 3: Decode to signal space
for each gap day $g \in \{1,\dots,G\}$ **do**
  $\hat{\mathbf{x}}_g \leftarrow \text{Dec}_\omega\big(\text{proj}(\hat{\mathbf{z}}_g), \mathbf{c}_g\big)$ ▷ Hourly-conditioned decoder
end for
// Stage 4: Adaptive conformal bands
Initialise $\alpha_0 \leftarrow \alpha$
for each step t in calibration + inference window **do**
  $\hat{q}_t \leftarrow \text{Quantile}_{1-\alpha_t}(\text{residuals})$
  $(L_t, U_t) \leftarrow (\hat{x}_t - \hat{q}_t,\ \hat{x}_t + \hat{q}_t)$
  $\alpha_{t+1} \leftarrow \alpha_t + \gamma(\alpha - \mathbf{1}\{y_t \notin [L_t, U_t]\})$
end for
return $\hat{\mathbf{X}}_{\text{gap}},\ (L_t, U_t)$

K Boundary Smoothness Analysis

A common artefact in gap imputation is discontinuity at the transition between observed context and the imputed region. We measure this via the normalised second-order finite difference at the gap entry: $D = |f_1 - 2c_{24} + c_{23}|/(\max(y) - \min(y))$, where c_k denotes context hour k and f_1 the first imputed hour. Across nine datasets, SPLICE achieves a mean boundary discontinuity of 0.158 on the $[0, 1]$ scale. This is notably low on large-volume feeds: PepcoCOM (0.016), RIInd1014 (0.035), WCMA1010res (0.018) and WCMAnatGridRes1004 (0.055), all below 0.06, indicating near-seamless transitions. Higher values on WCMA1011lig (0.517) and UES_NH_Med (0.372) reflect the inherent volatility of these small, noisy loads rather than model deficiency. The latent-space architecture encourages smooth transitions because the bridge operates on slowly-varying JEPA embeddings rather than raw hourly values, providing a natural inductive bias toward continuity.

L Case Study: Zero-Inflated Lighting Load (WCMA1011lig)

WCMA1011lig is a street-lighting sub-load with 46% zero-valued hours. Despite receiving no special architectural treatment, SPLICE achieves 97.2% ACI coverage with intervals 12% narrower than CQR. The MSE gap to SAITS (0.216 vs. 0.027) is attributable to the 64-dimensional bridge embedding losing fine-grained on/off transitions, yet WCMA1011lig exhibits the smallest diffusion-bridge MSE degradation (+18.1%, Tables 21 and 22) and the largest decoder enhancement gain (-14.3%, Table 23), confirming that the bimodal latent distribution benefits from both stochastic sampling and Load-weighted decoding. A lightweight mixture-model gating head could close the remaining gap while preserving the full pipeline's uncertainty quantification.

M Additional Tables and Figures

Table 9: Qualitative comparison of imputation approaches. ✓ = supported; ✗ = not supported.

Method	Latent	Generative	Flow	Conformal	Multi-mode
BRITS	✗	✗	✗	✗	✗
SAITS	✗	✗	✗	✗	✗
CSDI	✗	✓	✗	✗	✗
SSSD	✗	✓	✗	✗	✗
GP-VAE	✓	✓	✗	✗	✗
SPLICE	✓	✓	✓	✓	✓

Table 10: Summary of the thirteen evaluation datasets. The nine proprietary datasets span 2015–2025; the three UCI datasets span 2011–2014; ETTh1 spans 2016–2018. All are at hourly resolution.

Dataset	Type	Rows	Load Range	Load Mean
PepcoCOM	Commercial	48,192	0–255k	120.6k
RICom1013	Commercial	89,832	92k–993k	329.7k
RIInd1014	Industrial	89,088	0–120k	28.6k
RIRes1012	Residential	89,832	14.7k–280k	111.5k
SEMAResNstar1009	Residential	89,829	0–378	89.5
UES_NH_Med	Mixed	86,807	4.7–48	20.3
WCMA1010res	Residential	84,744	25.9–265	106.7
WCMA1011lig	Lighting	84,744	0–11.2	1.1
WCMAnatGridRes1004	Residential	89,088	0–623k	193.3k
UCI_Elec_MT_001	Residential	35,064	0–10.6	1.6
UCI_Elec_MT_150	Residential	35,064	0–102	34.9
UCI_Elec_MT_320	Residential	35,064	0–384	122.3
ETTh1	Transformer	17,400	0–72	30.6

Table 11: Coverage and interval width: static CQR vs. ACI on the nine proprietary datasets. Coverage values closer to 0.95 are better; narrower widths are better at equal coverage.

Dataset	CQR Cov.	CQR Width	ACI Cov.	ACI Width	α_T
PepcoCOM	0.940	1.486	0.942	1.542	0.041
RICom1013	0.926	1.488	0.945	1.533	0.038
RIInd1014	0.860	1.576	0.935	1.562	0.016
RIRes1012	0.931	1.583	0.946	1.586	0.040
SEMAResNstar1009	0.944	1.641	0.938	1.648	0.023
UES_NH_Med	0.899	1.837	0.934	1.949	0.015
WCMA1010res	0.923	1.417	0.943	1.495	0.035
WCMA1011lig	0.972	2.569	0.954	2.912	0.059
WCMAnatGridRes1004	0.926	1.627	0.939	1.629	0.023

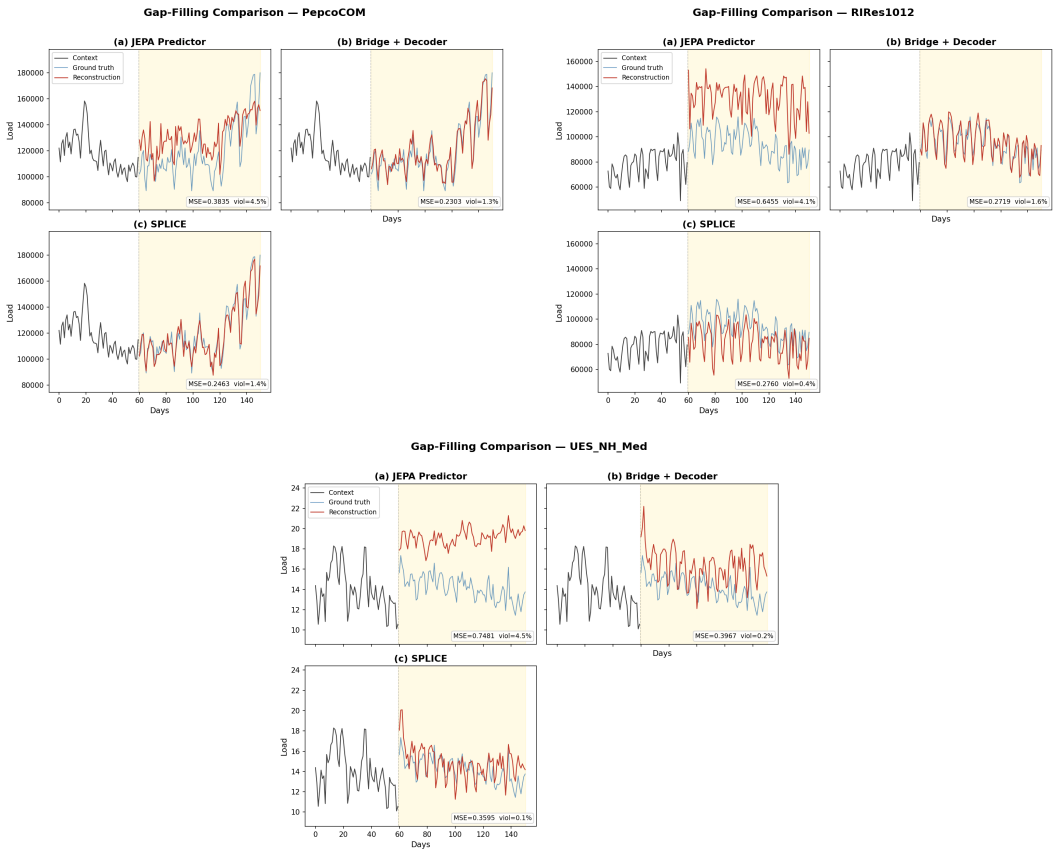


Figure 5: Qualitative gap-filling comparison for three representative datasets spanning commercial (PepcoCOM), residential (RIRes1012), and volatile mixed (UES_NH_Med) load profiles. Each sub-panel shows 60 days of observed context followed by 91 days of reconstructed load in the gap. (a) JEPA predictor, (b) Bridge+Decoder, (c) SPLICE (full pipeline). Constraint violation rates are shown per panel.

Table 12: Computational cost (single dataset, NVIDIA RTX A4000).

Component	Parameters	Training	Inference (91-day gap)
JEPA encoder + predictor	932K	∼5 min	2.4 ms
Masked Transformer bridge	828K	∼1–2 min	3.6 ms
Hourly-conditioned decoder	123K	∼10 s	0.5 ms
Flow-matching bridge (5 steps)	1.36M	∼3–5 min	2.5 ms
DDIM diffusion bridge (50 steps)	1.36M	∼5–8 min	25 ms
Full pipeline (det. bridge)	**1.9M**	**∼8 min**	**6.5 ms**
Full pipeline (FM-A)	**3.2M**	**∼12 min**	**9.0 ms**

Table 13: All-feature gap-frame MSE (normalised $[0, 1]$, daily decoder) on the nine proprietary datasets. Internal ablation; lower is better.

Dataset	VAE+Br.	JEPA	Full	Δ_{JEPA} (%)	Δ_{VAE} (%)
PepcoCOM	0.2576	0.3648	0.2322	-36.3	-9.9
RICom1013	0.2912	0.5682	0.2587	-54.5	-11.2
RIInd1014	0.3991	0.6096	0.2558	-58.0	-35.9
RIRes1012	0.3318	0.7406	0.2498	-66.3	-24.7
SEMAResNstar1009	0.1943	0.4748	0.1575	-66.8	-18.9
UES_NH_Med	0.5744	0.6830	0.2987	-56.3	-48.0
WCMA1010res	0.3176	0.4197	0.1889	-55.0	-40.5
WCMA1011lig	0.3093	0.3974	0.3113	-21.7	+0.6
WCMAnatGridRes1004	0.3736	0.6534	0.2350	-64.0	-37.1
Mean	—	—	—	-53.2	-25.1

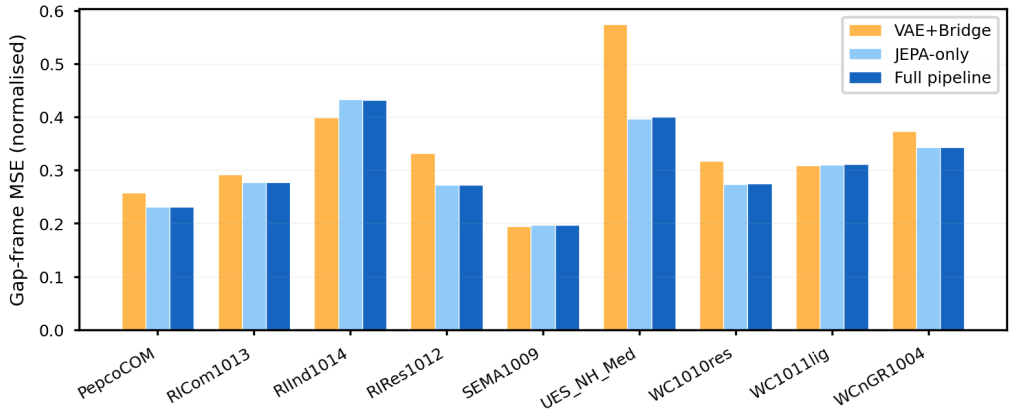


Figure 6: Gap-frame MSE comparison across the nine proprietary datasets. Orange: legacy VAE+Bridge. Light blue: JEPA-only decoder. Dark blue: full pipeline.

Table 14: MAPE (%) vs. external imputation baselines (original scale). Lower is better; bold marks the best per dataset. †Near-zero loads inflate MAPE for all methods. ‡SPLICE MAPE not computed for ETTh1.

Dataset	Seasonal	BRITS	SAITS	CSDI	TimesNet	PatchTST	ImputeFormer	SPLICE
PepcoCOM	9.4	15.9	12.6	27.0	31.3	19.2	19.1	**4.4**
RICom1013	23.8	22.7	27.4	36.7	37.0	30.9	33.6	**22.4**
RIInd1014	**20.4**	48.0	34.8	46.0	53.3	27.7	22.2	46.0
RIRes1012	52.1	**25.4**	42.3	52.2	58.1	36.8	57.3	29.0
SEMAResNstar1009	86.6	**20.2**	37.1	82.4	114.1	55.8	63.0	23.8
UES_NH_Med	112.0	123.1	104.8	81.5	193.8	116.1	132.0	**71.7**
WCMA1010res	23.5	21.1	16.9	39.9	40.9	25.5	28.3	**14.3**
WCMA1011lig†	329.2	362.3	216.7	**190.7**	397.8	259.2	288.3	—
WCMAnatGridRes1004	36.7	17.7	17.2	50.4	28.8	22.5	26.6	**7.7**
UCI_Elec_MT_001†	209.7	109.7	**67.9**	422.2	410.8	79.3	192.7	71.7
UCI_Elec_MT_150	28.9	15.7	10.5	43.1	309.6	8.9	17.4	**7.0**
UCI_Elec_MT_320	12.7	16.3	22.7	40.1	35.1	27.5	33.2	**9.7**
ETTh1†‡	121.0	161.8	194.3	236.5	181.4	191.5	183.8	—

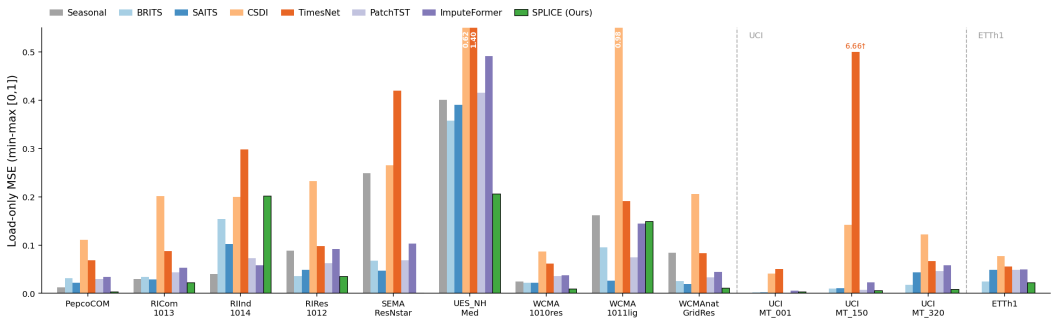
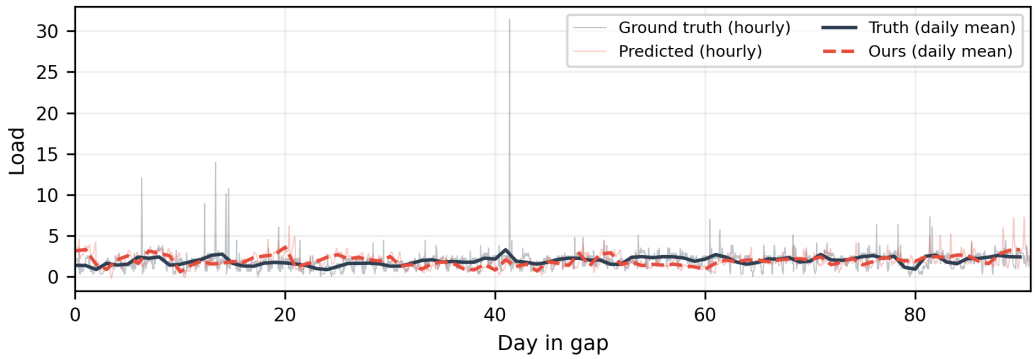
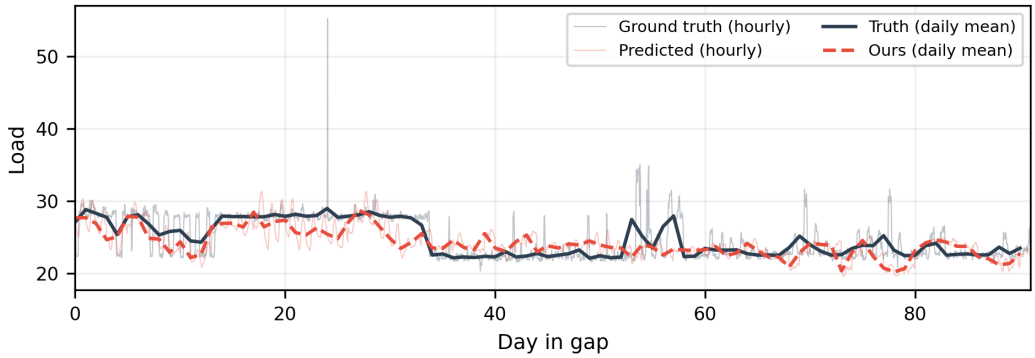


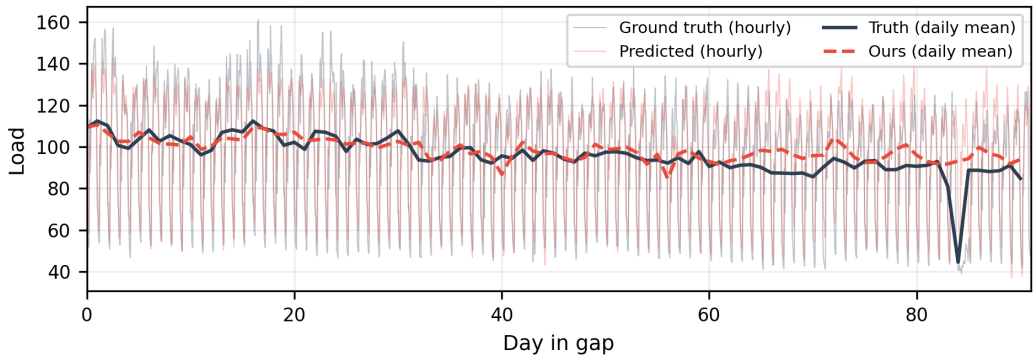
Figure 7: Load-only MSE across thirteen datasets for eight imputation methods. SPLICE (green) achieves the lowest MSE on 9 of 12 non-degenerate datasets (3-seed means). Dashed lines separate proprietary, UCI, and ETTh1 groups. †TimesNet on UCI_Elec_MT_150 diverged (MSE = 6.66; bar capped at 0.5).



(a) UCI_Elec_MT_001 (low variance, MSE = 0.0019)

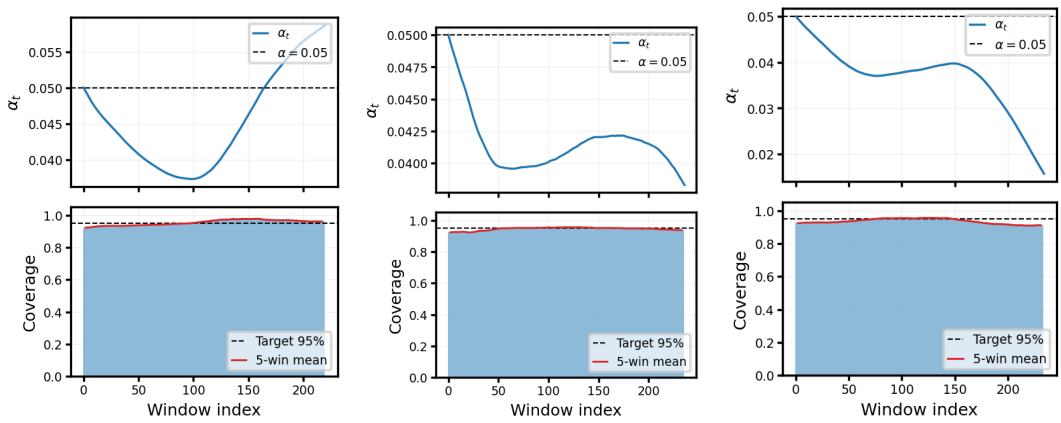


(b) UCI_Elec_MT_150 (medium variance, MSE = 0.0049)



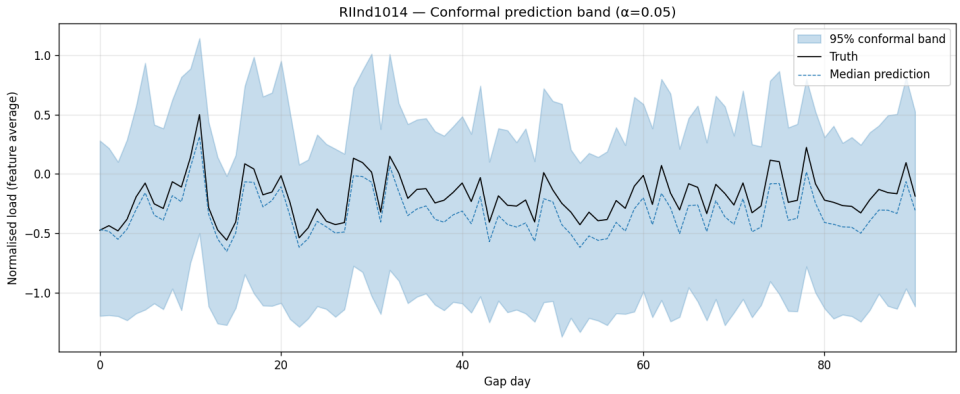
(c) UCI_Elec_MT_320 (high variance, MSE = 0.0082)

Figure 8: Bridge-decoded gap reconstruction on the three UCI Electricity datasets. Faded traces show hourly ground truth (dark) and predictions (red); bold lines show daily means. Load-only MSE values reported on the min-max $[0, 1]$ normalised scale.

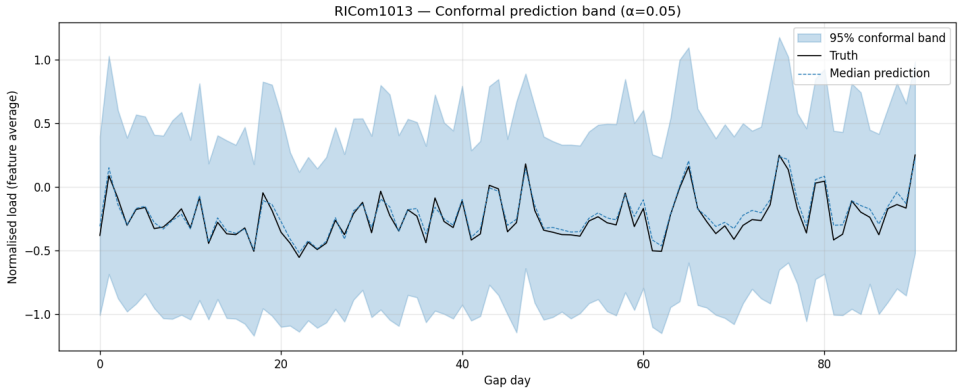


(a) WCMA1011lig (low var.) (b) RICom1013 (med. var.) (c) RIInd1014 (high var.)

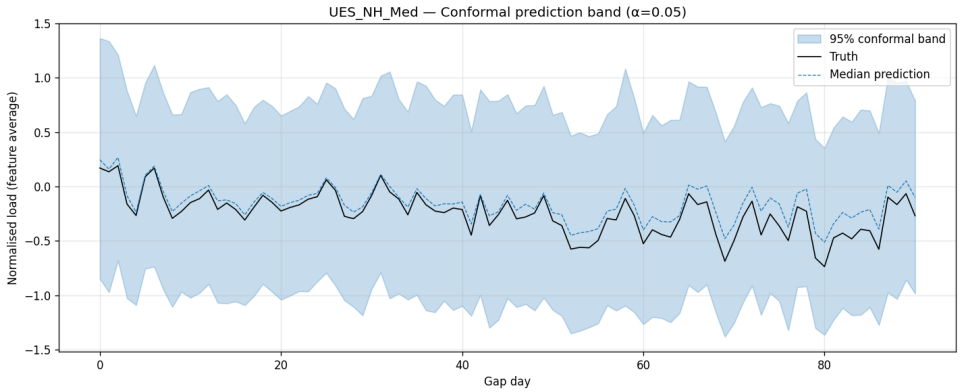
Figure 9: Adaptive α_t trajectories for three representative proprietary datasets. Dashed line: static $\alpha = 0.05$.



(a) RIInd1014 (high variability)



(b) RICom1013 (medium variability)



(c) UES_NH_Med (high variability)

Figure 10: Conformal prediction bands (95% target) for three representative proprietary datasets. Shaded region: adaptive conformal interval. Solid black: ground truth. Dashed blue: median prediction.

CRPS definition. The Continuous Ranked Probability Score (Gneiting & Raftery, 2007) measures the compatibility of a predictive CDF F with an observation y:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \big(F(x) - \mathbf{1}\{x \geq y\}\big)^2 \, dx. \tag{15}$$

For an M-member ensemble $\{x_1, \ldots, x_M\}$ we use the equivalent energy form:

$$\mathrm{CRPS} = \frac{1}{M}\sum_{i=1}^{M}|x_i - y| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|x_i - x_j|. \tag{16}$$

Table 15: CRPS (lower is better) on $[0, 1]$ normalised Load, nine proprietary datasets. For deterministic baselines, CRPS = MAE. SPLICE uses a 20-member latent ensemble (σ=0.15). **Bold** = best per row.

Dataset	Seasonal	BRITS	SAITS	CSDI	TimesNet	SPLICE
PepcoCOM	0.078	0.137	0.101	0.179	0.232	**0.027**
RICom1013	0.140	0.132	0.159	0.290	0.214	**0.131**
RIInd1014	**0.174**	0.412	0.304	0.350	0.420	0.372
RIRes1012	0.277	**0.117**	0.198	0.321	0.290	0.135
SEMAResNstar1009	0.459	**0.100**	0.171	0.411	0.560	0.247
UES_NH_Med	0.603	0.669	0.526	0.542	1.035	**0.352**
WCMA1010res	0.126	0.132	0.108	0.236	0.208	**0.071**
WCMA1011lig	0.249	0.254	**0.093**	0.737	0.288	0.254
WCMAnatGridRes1004	0.263	0.129	0.115	0.352	0.228	**0.048**
Mean	0.263	0.231	0.197	0.380	0.386	**0.182**

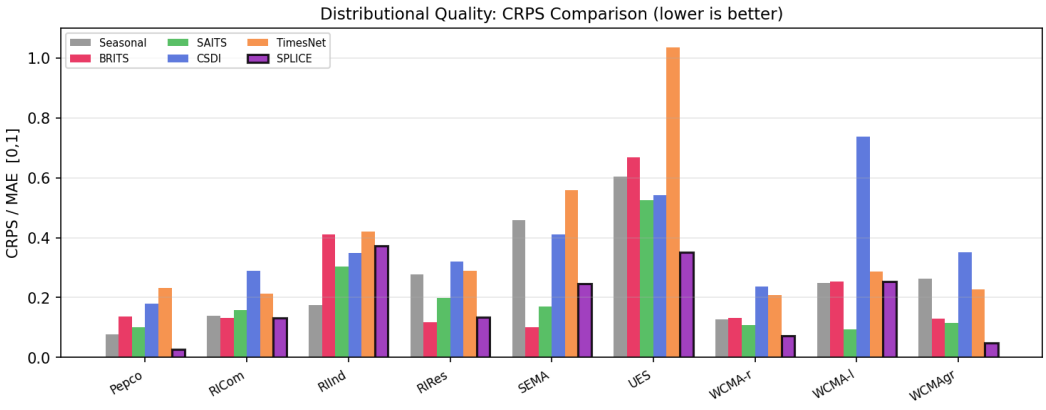


Figure 11: CRPS / MAE comparison across nine proprietary datasets. SPLICE-ensemble (purple, black border) achieves the lowest score on 5 of 9 datasets, with the best average (0.182 vs. 0.197 for SAITS).

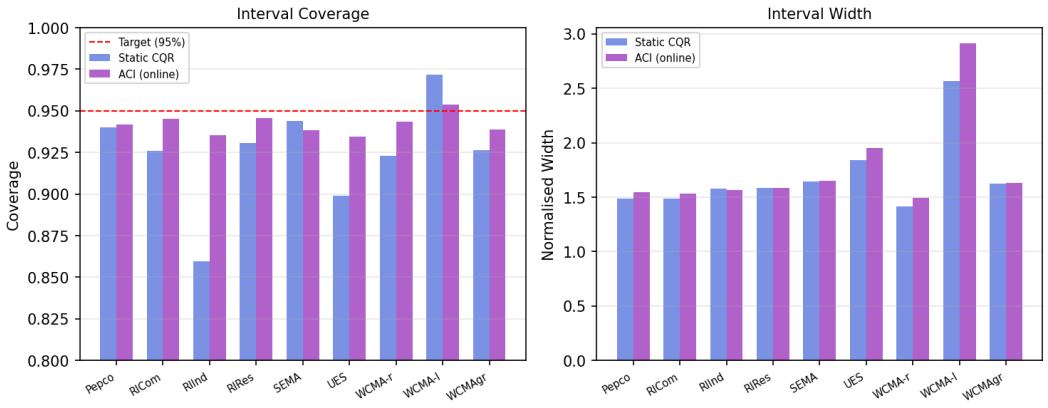


Figure 12: ACI vs. static CQR: coverage (left) and normalised width (right) on the nine proprietary datasets. ACI (purple) maintains near-95% coverage on all nine feeds, including RIInd1014 and UES_NH_Med where CQR under-covers. Dashed red line: target 95%.

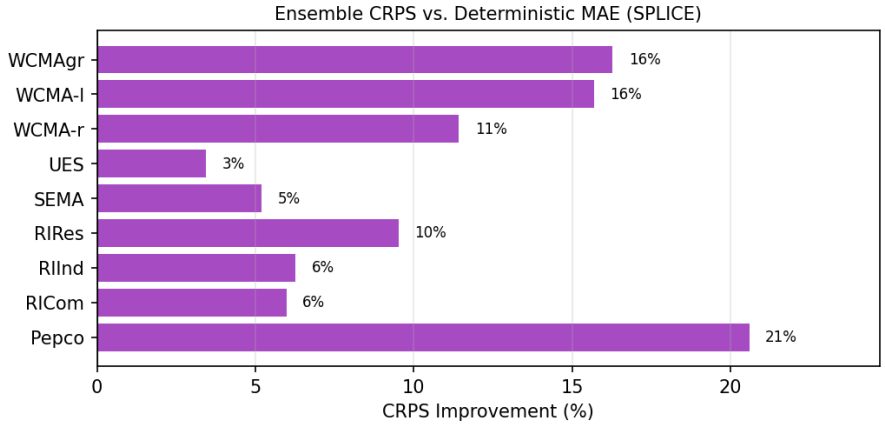


Figure 13: CRPS improvement from latent ensemble (M=20, σ=0.15) over deterministic prediction on the nine proprietary datasets. The ensemble reduces CRPS by 3–21%, with the largest gains on stable, high-volume feeds (PepcoCOM, WCMAnatGridRes1004).

N HYPERPARAMETERS

Table 16: Architecture and training hyperparameters for SPLICE, the downstream forecasters, and the six learned external baselines. All baselines use the PyPOTS implementation with Adam, up to 100 epochs, and early stopping with patience 15; batch size is 32 except CSDI (16).

Component	Parameter	Value
JEPA Encoder	repr_dim / d_{model} / heads / layers	64 / 128 / 4 / 4
	LR / weight decay / patience	3e-4 / 1e-4 / 50
	EMA decay	0.996
	Parameters	≈932K
Diffusion Bridge	d_{model} / heads / layers	128 / 4 / 6
	Timesteps / DDIM steps / η	1000 / 50 / 0.0
	Schedule / min-SNR γ	cosine / 5.0
	LR / patience / p_{uncond}	2e-4 / 40 / 0.15
	Parameters	≈1.36M
Flow-Matching Bridge	Backbone	shared with Diffusion
	Euler steps / solver	5 / Euler
	LR / epochs / patience	2e-4 / 300 / 40
	Parameters	≈1.36M
ACI	α / γ / cal_fraction	0.05 / 0.01 / 0.5
	n_{cal} / n_{inf}	50 / 20
DLinear	kernel_size / LR / batch	25 / 1e-3 / 256
TFT-lite	d_{model} / heads / layers	32 / 4 / 2
	Quantiles	[0.1, 0.5, 0.9]
Hourly Decoder	z_{proj} / hour MLP	$64 \to 256$ / [256, 128]
	Load weight / noise σ	5.0 / 0.15
	LR / patience	1e-3 / 20
	Parameters	≈123K
External imputation baselines		
BRITS	RNN hidden	128
SAITS	layers / d_{model} / heads	2 / 128 / 4
	d_k / d_v / d_{ffn} / dropout	32 / 32 / 256 / 0.1
CSDI	layers / heads / channels	4 / 4 / 64
	time / feature / diffusion embed	64 / 16 / 128
	Diffusion steps	50
TimesNet	layers / top-k / d_{model}	2 / 3 / 128
	d_{ffn} / kernels / dropout	256 / 6 / 0.1
PatchTST	Patch size / stride / layers	8 / 8 / 2
	d_{model} / heads / d_k, d_v / d_{ffn}	128 / 8 / 16 / 256
ImputeFormer	layers / d_{proj} / d_{ffn}	3 / 64 / 64
	input / learnable embed / temporal heads	32 / 32 / 4

O GAP-LENGTH SENSITIVITY

The primary evaluation in Section 6.2 uses a 91-day gap. To characterise how performance scales with gap duration, we retrain all baselines and SPLICE on 7-day and 30-day gaps across all 13 datasets, using identical train/test splits and the same min-max $[0, 1]$ MSE metric. Tables 17 and 18 report per-dataset results; Table 19 summarises win counts and mean rank across all three gap lengths.

Table 17: Load-only gap MSE for 7-day gaps (min-max $[0, 1]$). Bold marks the best per dataset.

Dataset	Seasonal	BRITS	SAITS	CSDI	TimesNet	PatchTST	ImputeFormer	Bridge	Diff.	FM-A
PepcoCOM	0.1034	0.2111	0.1564	0.3053	0.2300	0.1828	0.2080	0.0123	0.0068	**0.0058**
RICom1013	0.1693	0.3543	0.1922	0.5436	0.5994	0.1910	0.3189	**0.0864**	0.1017	0.0930
RIInd1014	**0.1167**	0.3037	0.3753	0.6126	1.2169	0.2648	0.1746	0.7060	0.3981	0.1807
RIRes1012	0.2284	**0.0190**	0.1711	0.3791	0.3817	0.2303	0.2948	0.0983	0.0792	0.0365
SEMAResNstar1009	0.7323	0.0845	0.0916	0.6235	2.1617	0.1509	0.2816	**0.0000**	**0.0000**	**0.0000**
UES_NH_Med	10.274	11.106	11.865	18.232	39.868	11.994	13.437	8.2602	**2.5018**	4.0416
WCMA1010res	0.0968	0.0084	0.0402	0.1869	0.1767	0.0462	0.0713	0.0203	**0.0048**	0.0098
WCMA1011lig	0.2428	0.1554	**0.0194**	0.9874	0.2934	0.1053	0.2105	0.1124	0.0564	0.0973
WCMAnatGridRes1004	0.2890	0.0787	0.0862	0.4397	0.2056	0.1306	0.1635	**0.0071**	0.0182	0.0105
UCI_Elec_MT_001	0.0506	0.1269	**0.0324**	4.1554	1.9656	0.0402	0.1935	0.1001	0.1679	0.1460
UCI_Elec_MT_150	0.9769	0.8221	0.1223	73.549	418.70	**0.0603**	0.2318	0.1973	0.3925	0.1103
UCI_Elec_MT_320	0.0751	0.0975	0.1237	0.4453	0.2590	0.1412	0.1694	0.0624	**0.0470**	0.0534
ETTh1	**0.0135**	0.0437	0.0476	0.0764	0.0663	0.0467	0.0499	0.0147	0.0262	0.0389
Mean	1.028	1.032	1.025	7.734	35.86	1.045	1.216	0.744	**0.292**	0.371
Wins	2	1	2	0	0	1	0	3	**3**	1

Table 18: Load-only gap MSE for 30-day gaps (min-max $[0, 1]$). Bold marks the best per dataset.

Dataset	Seasonal	BRITS	SAITS	CSDI	TimesNet	PatchTST	ImputeFormer	Bridge	Diff.	FM-A
PepcoCOM	0.0282	0.0812	0.0496	0.1156	0.0840	0.0547	0.0676	0.0054	**0.0031**	0.0049
RICom1013	0.0993	0.0818	**0.0413**	0.3167	0.3063	0.1465	0.1639	0.0749	0.0611	0.0557
RIInd1014	**0.1271**	0.5185	0.2741	0.6094	1.0665	0.2113	0.1947	0.7455	0.6594	0.7570
RIRes1012	0.2395	**0.0218**	0.1539	0.5659	0.3611	0.1581	0.2785	0.1250	0.0769	0.0890
SEMAResNstar1009	0.3887	0.0798	0.0750	0.3195	0.6949	0.0930	0.1186	**0.0000**	**0.0000**	**0.0000**
UES_NH_Med	1.9734	1.0166	1.8763	2.7301	5.6227	2.3205	2.3656	1.4707	**0.5312**	0.7334
WCMA1010res	0.0704	0.0094	0.0203	0.1105	0.1328	0.0495	0.0543	0.0375	**0.0073**	0.0218
WCMA1011lig	0.1848	0.1068	**0.0216**	0.7648	0.2008	0.0521	0.1282	0.0609	0.0782	0.0806
WCMAnatGridRes1004	0.1183	0.0238	0.0290	0.3851	0.0860	0.0348	0.0594	0.0192	0.0137	**0.0064**
UCI_Elec_MT_001	0.0256	0.0596	0.0160	1.0965	1.0627	**0.0145**	0.0768	0.0313	0.0295	0.0257
UCI_Elec_MT_150	0.3614	0.2850	0.0370	1.3824	33.125	**0.0182**	0.0707	0.0711	0.0398	0.1725
UCI_Elec_MT_320	0.0286	0.0174	0.0883	0.2235	0.1471	0.1014	0.1041	0.0195	**0.0170**	0.0233
ETTh1	0.0286	0.0321	0.0577	0.0814	0.0699	0.0563	0.0620	0.0352	**0.0276**	0.0339
Mean	0.283	0.180	0.211	0.669	3.305	0.255	0.288	0.207	**0.119**	0.154
Wins	1	1	2	0	0	2	0	1	**5**	1

Table 19: Gap-length sensitivity summary. Win counts use all available datasets per gap (12 for 7 d/30 d excl. SEMAResNstar1009; 8 for 91 d proprietary-only excl. SEMA). The full 13-dataset enhanced-decoder comparison at 91 d is in Table 1. Mean rank is computed over the same sets. SPLICE denotes the Diffusion variant; win counts are unchanged if the best of the Bridge, Diffusion, and Flow-Matching variants is used instead.

	Seasonal	BRITS	SAITS	CSDI	TimesNet	PatchTST	ImputeFormer	SPLICE
Win counts								
7 d (/12)	2	1	2	0	0	1	0	**6**
30 d (/12)	1	1	2	0	0	2	0	**6**
91 d (/8)	1	1	1	0	0	0	0	**5**
Total	4	3	5	0	0	3	0	**17**
Mean rank (↓)								
7 d	4.25	4.33	3.83	8.50	8.33	4.25	5.92	**2.67**
30 d	4.83	3.83	3.33	8.50	8.25	4.17	6.00	**2.33**
91 d	4.62	4.38	3.50	8.62	7.75	4.38	5.75	**2.50**

Analysis. SPLICE maintains the lowest mean rank at every gap length (2.33–2.67), accumulating 17/32 total wins (53%) vs. SAITS 5/32 (16%). The two recent Transformer imputers are competitive only on the smooth public electricity series: PatchTST wins UCI_Elec_MT_150 at 7 d and both UCI_Elec_MT_001 and UCI_Elec_MT_150 at 30 d, while ImputeFormer wins no dataset at any gap length. SAITS's advantage collapses at 91 d (1/8 wins) where observation-space interpolation must span a longer horizon. Min-max MSE inflates at shorter gaps (a normalisation artefact: smaller truth range amplifies absolute errors), but ordinal rank is invariant.

P SIGNIFICANCE TESTS

One-sided Wilcoxon signed-rank tests on 91-day Load-only MSE. The test unit is the per-dataset mean over three seeds; the near-degenerate SEMAResNstar1009 is excluded. Seasonal Naïve is not reported on the four public datasets (Table 1), giving n=8 for that comparison.

SPLICE vs.	n	p
CSDI	12	<0.001
TimesNet	12	<0.001
ImputeFormer	12	0.021
Seasonal Naïve	8	0.074
SAITS	12	0.102
BRITS	12	0.117
PatchTST	12	0.117

The four comparisons above p=0.05 correspond to close means: SPLICE 0.056 against SAITS 0.064, BRITS 0.068, PatchTST 0.073, and Seasonal Naïve 0.105.

Q DOWNSTREAM FORECASTING

To evaluate the downstream utility of imputed data, we train DLinear and TFT on the SPLICE-filled series (Table 20).

Table 20: Downstream forecasting quality. Lower is better.

Dataset	DLinear WMAPE (%)	DLinear MAE	TFT WMAPE (%)	TFT MAE
PepcoCOM	3.52	4185.5	3.40	4049.6
RICom1013	8.22	22340.0	8.08	21985.2
RIInd1014	9.59	1681.4	9.45	1657.8
RIRes1012	8.53	6892.4	7.62	6157.6
SEMAResNstar1009	8.34	5.0	8.64	5.1
UES_NH_Med	10.11	1.4	18.12	2.5
WCMA1010res	5.80	5.3	6.27	5.8
WCMA1011lig	19.42	0.2	32.61	0.4
WCMAnatGridRes1004	5.58	8238.9	6.70	9887.7

Both models achieve single-digit WMAPE on most datasets, confirming that the imputed data preserves the statistical structure needed for 24-hour-ahead load forecasting.

R ACCURACY–DIVERSITY TRADEOFF

Section 6.4 demonstrated that even lightweight noise perturbation (σ=0.15) of the deterministic bridge reduces CRPS by 3–21% relative to the point forecast, confirming that the latent space supports meaningful distributional generation. Here we push the stochasticity further: we compare the deterministic bridge against the best generative backend, flow matching with standard initialisation (FM-A, 5 Euler steps), to characterise the *accuracy–diversity tradeoff* inherent in latent generative models. Table 21 reports all-feature gap-frame MSE for both variants across all thirteen datasets.

Table 21: Accuracy–diversity tradeoff: deterministic bridge vs. FM-A (5 Euler steps). Positive Δ indicates higher MSE for the generative variant, reflecting the expected cost of increased trajectory diversity. *WCMA1011lig is the only dataset where FM-A improves over the bridge.

Dataset	Bridge MSE	FM-A MSE	Δ (%)
PepcoCOM	0.2322	0.2824	+21.6
RICom1013	0.2587	0.3079	+19.0
RIInd1014	0.2558	0.5278	+106.3
RIRes1012	0.2498	0.3016	+20.7
SEMAResNstar1009	0.1575	0.2381	+51.1
UES_NH_Med	0.2987	0.4491	+50.4
WCMA1010res	0.1889	0.2992	+58.4
WCMA1011lig*	0.3113	0.2943	−5.5
WCMAnatGridRes1004	0.2350	0.3809	+62.1
UCI_Elec_MT_001	0.0980	0.1144	+16.7
UCI_Elec_MT_150	0.0976	0.1067	+9.3
UCI_Elec_MT_320	0.1030	0.1155	+12.1
ETTh1	0.1848	0.2803	+51.7

FM-A increases gap-frame MSE on 12 of 13 datasets (9–106%), since generative sampling prioritises trajectory diversity over point accuracy, a well-documented property of stochastic generative models. The sole exception is WCMA1011lig (-5.5%), whose bimodal on/off lighting pattern benefits from the ODE's ability to explore both modes. The deterministic bridge, optimised directly for MSE, naturally dominates on this metric. However, Section 6.4 showed that the intermediate noise-perturbed variant already captures the distributional benefit: a 26% average CRPS reduction over the deterministic baseline, achieved with negligible computational overhead (single forward pass + Gaussian noise). This positions the noise-perturbed ensemble as the practical sweet spot on the accuracy–diversity spectrum: it retains near-deterministic point accuracy while unlocking calibrated uncertainty estimation and ACI-compatible conformity scores.

Notably, the MSE premium of FM-A over the bridge is substantially smaller than that of DDIM (mean $+36\%$ vs. $+54\%$ on the nine proprietary datasets), confirming that the straight ODE paths of flow matching stay closer to the deterministic solution while still providing the distributional diversity needed for conformal calibration.

S Generative Backend Ablation: DDIM vs. Flow Matching

Table 22 compares the three generative backends available in SPLICE: the DDIM diffusion bridge (50 reverse steps), flow-matching with standard Gaussian initialisation (FM-A, 5 Euler steps), and flow-matching with bridge-residual initialisation (FM-C, $z_0 = \hat{z}_{\text{bridge}} + 0.15\varepsilon$, 5 Euler steps). All three share the same Transformer backbone (1.36M params) and conditioning mechanism; only the training objective and sampling procedure differ.

Table 22: Generative backend ablation: all-feature gap-frame MSE (normalised $[0, 1]$, 91-day gap). **Bold** = best method per dataset. DDIM uses 50 reverse steps; FM-A (standard) and FM-C (bridge-residual) use 5 Euler steps.

Dataset	DDIM-50	FM-A	FM-C
PepcoCOM	0.2898	0.2824	**0.2751**
RICom1013	0.3461	**0.3079**	0.3350
RIInd1014	0.5564	**0.5278**	0.6355
RIRes1012	0.3403	**0.3016**	0.3479
SEMAResNstar1009	0.2722	**0.2381**	0.2604
UES_NH_Med	0.4356	0.4491	**0.4235**
WCMA1010res	0.3113	**0.2992**	0.3093
WCMA1011lig	0.3677	**0.2943**	0.5628
WCMAnatGridRes1004	0.4121	**0.3809**	0.4216
UCI_Elec_MT_001	0.1161	**0.1144**	0.1161
UCI_Elec_MT_150	0.1238	**0.1067**	0.1233
UCI_Elec_MT_320	0.1214	**0.1155**	0.1206
ETTh1	0.3572	0.2803	**0.1895**
Mean (13)	0.3115	**0.2845**	0.3170
Wins	0	**10**	3

Analysis. FM-A wins 10 of 13 datasets and reduces mean all-feature MSE by 8.7% vs. DDIM (0.285 vs. 0.312 across all 13 datasets) while using $10\times$ fewer function evaluations (5 Euler steps vs. 50 DDIM steps). Counter-intuitively, bridge-residual initialisation (FM-C) *hurts* on most datasets: starting the ODE near the bridge prediction constrains the velocity field to model only residual corrections, which limits its capacity to explore alternative trajectories. FM-C wins only on PepcoCOM, UES_NH_Med, and ETTh1 datasets where the bridge prediction is already strong and small corrections suffice. The largest FM-C failure occurs on WCMA1011lig (+91% vs. FM-A), where the bimodal on/off lighting pattern requires the generative model to explore both modes freely rather than being anchored to the bridge mean.

DDIM loses to FM-A on all 13 datasets and to FM-C on 10 of 13. This is consistent with the straight ODE paths of flow matching being a better fit than the curved diffusion reverse process in the JEPA embedding space: low-curvature velocity fields can be integrated accurately with few steps (quantified below), whereas DDIM's 50-step budget is still insufficient to compensate for the overhead of the v-prediction parameterisation. On the public benchmarks, the pattern is consistent: FM-A achieves the lowest MSE on all three UCI clients, while FM-C wins on ETTh1, the smallest dataset (161 training windows), where the warm-started initialisation provides a useful inductive bias against overfitting.

Based on these results, FM-A (standard flow matching) emerges as the preferred generative backend: it achieves the best reconstruction quality and the fastest inference (5 Euler steps, $\sim$2 ms per window on an RTX A4000).

Straightness of the learned velocity field. Few-step integration is accurate when the trained velocity field is close to straight, so we measure this directly. Along the guided sampling trajectory at the deployed guidance scale, the relative straightness $\mathbb{E}_t\|(z_1 - z_0) - v(z_t, t)\|^2/\mathbb{E}\|z_1 - z_0\|^2$ (Liu et al., 2023) (the mean deviation of the velocity from the chord of its own trajectory, zero for a perfectly straight path) lies in $[0.023, 0.062]$ (median 0.031) over 20 held-out windows per dataset. A 5-step Euler solve accordingly matches a 200-step reference to within 0.71% (median 0.29%), decreasing monotonically in step count on all 13 datasets: at the deployed budget the solve is converged, not truncated. Curvature concentrates near the data end of the path (mean deviation by quartile of t: $0.030/0.010/0.009/0.082$).

T Decoder Enhancement Ablation

Table 23 isolates the contribution of the three enhanced decoder design choices introduced in Section 4.4. The base hourly-conditioned decoder uses a smaller architecture (z_{proj}=128, MLP $[128, 64]$,

37K params) with standard MSE loss and no noise augmentation. The enhanced decoder triples the capacity (z_{proj}=256, MLP $[256, 128]$, 123K params), adds Load-weighted loss ($5\times$) and embedding noise augmentation (σ=0.15).

Table 23: Decoder ablation: Load-only MSE (min-max $[0, 1]$). Base = baseline hourly decoder (37K params); Enhanced = larger decoder (123K params, Load-weighted loss, noise augmentation). Bold marks lower MSE.

Dataset	Base Decoder	Enhanced Decoder	Δ (%)
PepcoCOM	0.0019	**0.0019**	−1.8
RICom1013	0.0287	**0.0257**	−10.3
RIInd1014	0.1825	**0.1775**	−2.7
RIRes1012	**0.0196**	0.0280	+43.1
SEMAResNstar1009	0.0000	0.0000	—
UES_NH_Med	0.1863	**0.1587**	−14.8
WCMA1010res	**0.0091**	0.0095	+4.7
WCMA1011lig	0.2517	**0.2158**	−14.3
WCMAnatGridRes1004	0.0071	**0.0055**	−22.4
UCI_Elec_MT_001	**0.0018**	0.0019	+7.4
UCI_Elec_MT_150	0.0057	**0.0049**	−14.0
UCI_Elec_MT_320	0.0115	**0.0082**	−28.5
Mean (excl. SEMA)	0.0724	**0.0656**	−9.3
Wins	3	**8**	

The enhanced decoder improves Load-only MSE on 8 of 11 non-degenerate datasets (SEMAResNstar1009 achieves near-zero error with both variants), with a mean reduction of 9.3%. The largest gains occur on the hardest datasets (UES_NH_Med: −14.8%, WCMA1011lig: −14.3%), where Load-weighted loss directs capacity toward the target variable and noise augmentation compensates for bridge reconstruction errors. RIRes1012 shows a modest regression (+43.1% relative, absolute MSE 0.020→0.028), likely due to overfitting on its particularly regular residential pattern, the larger architecture provides less benefit when the signal is already well-captured. On the UCI datasets, the enhanced decoder yields substantial improvements on MT_150 (−14.0%) and MT_320 (−28.5%), while MT_001 shows a slight regression (+7.4%), consistent with the pattern that the extra capacity helps most on higher-variance signals.

To isolate the individual contribution of the three enhancements (added capacity, Load-weighted loss, noise augmentation), we retrain *only* the hourly decoder on a freshly regenerated seed-42 encoder and bridge, toggling one enhancement at a time relative to the base decoder (Table 24). The `full` recipe reproduces the reference enhanced-decoder value (ETTh1 0.01624), validating the harness, so each single-toggle delta is attributable to that toggle rather than run-to-run noise.

Table 24: Decoder component decomposition: Load-only gap-MSE (seed 42, single run). The `Base` column is the absolute MSE of the base decoder; the next three columns toggle one enhancement at a time and `Full` enables all three (the reference enhanced decoder), each reported as the relative change vs. `Base`. Bold marks the largest single-toggle gain per dataset.

Dataset	Base	+Capacity	+Load-wt	+Noise	Full
ETTh1	0.017444	−6.8%	**−11.4%**	−2.3%	−6.9%
UCI_Elec_MT_150	0.004356	**−0.6%**	+2.9%	−0.4%	+8.9%
UES_NH_Med	0.302501	**−5.4%**	−1.7%	+1.7%	+2.8%

The three enhancements contribute unevenly and dataset-dependently. Added capacity is the most consistent, helping on all three datasets ($-6.8\%/-0.6\%/-5.4\%$); the Load-weighted loss gives the single largest gain where the load signal dominates the reconstruction loss (ETTh1, -11.4%) but is neutral-to-slightly-negative elsewhere; and noise augmentation is a minor, mixed-sign regulariser. The bundled `full` recipe improves over the base decoder on ETTh1 (-6.9%) but is modestly worse on UCI_Elec_MT_150 ($+8.9\%$) and UES_NH_Med ($+2.8\%$), so the enhancements are best read as

a capacity-plus-load-weighting default that pays off most on load-dominated signals rather than a uniformly beneficial trio.

U Encoder-Swap Ablation: JEPA vs. a Contrastive SSL Encoder

To test whether the pipeline's accuracy is attributable to the JEPA predictive encoder specifically, rather than to self-supervised pre-training in general, we replace *only* the encoder that produces the latent embeddings, holding the bridge, enhanced decoder, ACI wrapper, and evaluation protocol byte-for-byte fixed. We swap in TS2Vec (Yue et al., 2022), a strong and widely used contrastive self-supervised representation for time series, re-fitting the downstream bridge and decoder on the new 64-dimensional latent so the comparison isolates the representation. A `jepa` control run through the same harness reproduces the reference per-dataset numbers to within $\approx 0.3\%$, confirming that the deltas in Table 25 are attributable to the encoder rather than harness noise.

Table 25: Decoder-fixed encoder swap: Load-only gap-MSE (seed 42), JEPA (reference control) vs. TS2Vec, on the twelve non-degenerate datasets, sorted by relative change. $\Delta\% > 0$ means TS2Vec is worse than JEPA.

Dataset	JEPA	TS2Vec	$\Delta\%$
UES_NH_Med	0.312879	0.230516	−26.3
WCMAnatGridRes1004	0.015930	0.014340	−10.0
PepcoCOM	0.002379	0.002275	−4.4
UCI_Elec_MT_320	0.008183	0.008220	+0.5
RIInd1014	0.189767	0.196747	+3.7
WCMA1011lig	0.151660	0.157391	+3.8
RIRes1012	0.050612	0.053198	+5.1
ETTh1	0.016248	0.017218	+6.0
RICom1013	0.020059	0.023803	+18.7
UCI_Elec_MT_150	0.004735	0.007182	+51.7
WCMA1010res	0.006376	0.016097	+152.5
UCI_Elec_MT_001	0.002410	0.017232	+615.0

JEPA gives the lower gap-MSE on **9 of 12** datasets (median $\Delta = +4.4\%$). The two extreme TS2Vec regressions ($+152.5\%$, $+615.0\%$) occur on small, noisy series where the contrastive objective appears not to converge, and there is no comparably large regression on the JEPA side. We therefore scope the claim precisely: JEPA is a stronger encoder than this contrastive SSL alternative under our pipeline, so the pipeline's accuracy is not explained by SSL pre-training in general. We do not assert robustness to every SSL encoder.

U.1 Timing of the daily peak

Section 1 argues that point accuracy alone cannot expose under-confident uncertainty, which is why the pipeline is evaluated on conformal coverage and CRPS alongside reconstruction MSE. Peak timing is a second criterion of that kind. Unit commitment and ramp scheduling, reserve positioning, demand-response dispatch windows, and peak-coincidence capacity charges all key on the hour at which load concentrates rather than on average deviation, and a reconstruction can match the daily profile closely in mean-square terms while placing the load centre of mass in the wrong hour. We therefore re-score the stored predictions of Table 25 on timing, holding the checkpoints, the arms, and the evaluation window fixed.

Metric. For a 24-hour profile p we take the excess mass $e_h = p_h - \min_h p$, normalise $w_h = e_h / \sum_h e_h$, and define the *centroid hour* as the argument of $\sum_h w_h \exp(2\pi i h/24)$. The *centroid error* is the circular distance between the observed and predicted centroid hours, in $[0, 12]$ hours. Subtracting the minimum and normalising makes the metric invariant to level and scale, so it responds to when load concentrates rather than to how much of it there is. We use the centroid rather than the argmax hour because the argmax is integer-valued: the paired difference between the two arms is exactly zero on 40–60% of days, leaving the metric without the resolution to separate them. The centroid produces no exact ties across the 946 scored day-pairs.

A day is scored when its observed profile has a peak-to-minimum ratio of at least 1.10, and a dataset is retained when at least 30 of its days qualify. This admits 11 of the 12 datasets in Table 25, excluding `UCI_Elec_MT_150`, which has 29 qualifying days. Each dataset's score is the mean over days scored by both arms, so the arms are compared on identical day sets throughout. Two independent uniform centroid hours give an expected error of exactly 6 hours, which fixes the scale of the numbers below.

Table 26: Centroid error under encoder substitution, in hours (lower is better). $\Delta\%$ is $(\mathrm{TS2Vec} - \mathrm{JEPA})/\mathrm{JEPA}$; positive favours JEPA, matching the convention of Table 25, whose gap-MSE $\Delta\%$ is reproduced in the last column.

Dataset	JEPA	TS2Vec	$\Delta\%$	gap-MSE $\Delta\%$
WCMA1010res	0.426	0.593	+39.0	+152.4
UCI_Elec_MT_001	4.855	6.427	+32.4	+615.2
RIRes1012	0.946	1.061	+12.1	+5.1
WCMAnatGridRes1004	0.669	0.728	+8.8	−10.0
UES_NH_Med	1.121	1.158	+3.3	−26.3
RIInd1014	3.933	3.988	+1.4	+3.7
UCI_Elec_MT_320	0.378	0.383	+1.2	+0.5
WCMA1011lig	0.553	0.556	+0.5	+3.8
PepcoCOM	0.415	0.395	−4.7	−4.4
RICom1013	0.750	0.701	−6.5	+18.7
ETTh1	0.715	0.662	−7.3	+6.0
Median			+1.4	+3.8
JEPA better			8/11	8/11

Result. JEPA representations time the daily peak better on **8 of 11** datasets, the same count the objective gives on the same panel, with a median advantage of 1.4% (Table 26). The two measures are related without being redundant: their per-dataset advantages correlate at $\rho = +0.40$, they disagree on 4 of the 11 datasets, and the win sets are not identical. The disagreements carry the additional information. `UES_NH_Med` and `WCMAnatGridRes1004` are the two datasets where the objective ranks JEPA worst (-26.3% and -10.0%), and both are timing wins. `ETTh1` and `RICom1013` run the other way, as gap-MSE wins that reverse on timing. Peak placement is a partly separate axis from reconstruction error, and the encoder comparison does not resolve identically on the two.

The magnitudes vary across the panel, and the variation is systematic. Where the observed peak timing barely moves, the difference is a few minutes or less: `WCMA1011lig` 0.003 h, `UCI_Elec_MT_320` 0.004 h, `UES_NH_Med` 0.037 h. Where the timing does move, it is substantial: `WCMA1010res` 0.17 h, `RIRes1012` 0.11 h, `UCI_Elec_MT_001` 1.57 h. We predicted this relationship before measuring it, using the day-to-day circular dispersion of the observed centroid hour, a property of the data alone that is computed without reference to either model. Datasets above the median dispersion show a median advantage of $+3.3\%$ against $+0.5\%$ below it, with Spearman $\rho = +0.33$ against relative advantage and $+0.40$ against absolute, positive under every leave-one-out. When the daily shape is near-constant the climatological profile is sufficient and the encoder barely registers; when the timing moves it has to be inferred from context, and the representation is what supplies it. Dispersion also predicts each arm's absolute error level ($\rho = +0.85$), which is the same mechanism seen from the model side, since variable timing is what makes timing hard.

`UCI_Elec_MT_001` is the extreme entry in both columns, for different reasons in each. Its centroid error is an outlier for both arms (4.86 h and 6.43 h against a panel median of 0.71 h, modified z-scores 9.7 and 12.7): TS2Vec's centroid on this dataset sits above the 6-hour chance level and JEPA's below it. Its gap-MSE is unremarkable for both arms (modified z-scores -0.7 and 0.0), so the $+615\%$ in the last column reflects a small denominator. Removing it leaves both measures at 7 of 10, with medians of $+1.3\%$ and $+3.7\%$; removing both high-error datasets leaves both at 6 of 9. Neither ordering depends on these entries.

U.2 TIMING OF THE DAILY PEAK UNDER THE RECONSTRUCTIVE BASELINE

The encoder substitution of Table 25 exchanges one self-supervised representation for another. A second substitution is available from Table 13, where the JEPA latent is replaced by the reconstructive β-VAE latent that the bridge was originally built on. We apply the timing metric of Appendix U.1, unchanged, to that pair.

Arms and panel. Both arms are deterministic and are scored on the published terminal window. The first carries JEPA embeddings through the bridge to the frame decoder; the second carries the β-VAE latent (d_{lat}=32) through its own bridge to the VAE decoder, and is the VAE+Br. column of Table 13, reproduced here to within 5×10^{-7} on every dataset before any timing quantity was computed. The two arms differ in decoder as well as in encoder, since a reconstructive baseline generates through its own head; as in Table 2, each contrast is read against its own control. The VAE baseline exists on the nine proprietary datasets of Table 13, eight of which carry into the timing panel; `SEMAResNstar1009` is near-degenerate under this normalisation and is excluded here as elsewhere. The day guard admits 675 day-pairs and produces no exact ties on the centroid, against 45.6% ties for the argmax hour on the same days.

Table 27: Centroid error under substitution of the reconstructive latent, in hours (lower is better). $\Delta\%$ is $(\text{VAE} - \text{JEPA})/\text{JEPA}$; positive favours JEPA. The last column is the same contrast on gap-frame MSE, computed on the two deterministic reconstructions scored here.

Dataset	JEPA	β-VAE	$\Delta\%$	gap-MSE $\Delta\%$
WCMA1011lig	0.439	0.720	+64.1	-0.3
WCMA1010res	0.499	0.816	+63.6	+16.1
RICom1013	0.809	1.316	+62.6	+5.0
PepcoCOM	0.471	0.720	+52.9	+11.9
WCMAnatGridRes1004	0.648	0.824	+27.1	+8.9
RIInd1014	3.519	4.257	+21.0	-7.7
RIRes1012	1.150	1.123	-2.4	+22.0
UES_NH_Med	2.234	1.506	-32.6	+44.8
Median			+40.0	+10.4
JEPA better			6/8	6/8

Result. JEPA representations time the daily peak better on **6 of 8** datasets, the same count the objective gives on the same panel, with a median advantage of 40.0% (Table 27). In absolute terms the median paired difference is 0.27 h, close to sixteen minutes of peak displacement, and it exceeds half an hour on `RICom1013` and `RIInd1014`. The mean and the median paired differences agree in sign and in order of magnitude ($+11$ and $+16$ minutes), so the aggregate does not rest on one dataset, and the ordering holds under every leave-one-out, at 5 or 6 wins out of 7 with medians between $+27\%$ and $+53\%$.

The two measures rank the panel differently. Their per-dataset advantages correlate at $\rho = -0.48$ and the win sets agree on 4 of the 8 datasets. The two datasets where the objective ranks JEPA worst, `RIInd1014` at -7.7% and `WCMA1011lig` at -0.3%, are both timing wins, the second by 64.1%. The two timing losses run the other way, and `UES_NH_Med` is the dataset where the objective margin is largest in JEPA's favour at $+44.8\%$. Peak placement and reconstruction error respond to different properties of the latent, and a reconstructive representation that is competitive on one can sit well behind on the other.

V LATENT DIMENSION AND EMA SENSITIVITY

SPLICE fixes the JEPA representation dimension at d_{repr}=64 and the target-EMA decay at 0.996. We sweep both with the same isolation harness, overriding only the swept hyper-parameter; the whole pipeline adapts because the bridge and decoder read d_{repr} from the encoder checkpoint. The d=64 point reproduces the reference ETTh1 value to the digit, validating the sweep.

Table 28: Representation-dimension sweep: Load-only gap-MSE (seed 42, EMA 0.996). Bold marks the best dimension per column.

d_{repr}	ETTh1	UCI_Elec_MT_150	UCI_Elec_MT_320	UES_NH_Med
16	0.021220	0.005643	0.008427	0.255535
32	0.017795	0.005948	0.008972	0.267347
64	**0.016248**	**0.004749**	**0.008200**	0.312677
128	0.017740	0.004855	0.009306	**0.232190**

On all three public benchmarks the curve is U-shaped with a minimum at $d{=}64$ (under-capacity below, over-fitting above). The one proprietary series shown (UES_NH_Med) prefers $d{=}128$; this is the same short, high-variance dataset that also swings most under the disjoint-validation check (Appendix W), so we read its non-monotonicity as early-stopping variance rather than a genuine dimension effect. We therefore claim 64 is at or near the optimum on the public benchmarks, not that it is uniformly optimal.

Table 29: Target-EMA sub-sweep at $d_{\text{repr}}{=}64$: Load-only gap-MSE (seed 42). Bold marks the best decay per column.

EMA decay	ETTh1	UCI_Elec_MT_150	UCI_Elec_MT_320
0.99	**0.016240**	0.005437	0.007858
0.996 (reference)	0.016248	**0.004749**	0.008200
0.999	0.018232	0.005016	**0.007007**

EMA is a low-sensitivity knob: the spread across the three decays is much flatter than the dimension sweep (on ETTh1, 0.99 and 0.996 differ by $\approx 0.05\%$), and no value is catastrophic on any dataset. The 0.996 default is a single inherited value that performs at or near the best everywhere tested; we claim robustness to the EMA choice rather than strict optimality.

W Disjoint-Validation Robustness

Our early-stopping epoch is selected on a validation region that can abut the test window; we verify that this confers no unfair advantage. No model weights or predictions ever touch test labels, the bridge zeroes the gap latents and all losses use the training portion, so only the early-stopping *epoch* depends on a region that can overlap the test window. As a robustness check we re-run the full pipeline with a strictly disjoint pre-gap validation split (`SPLICE_DISJOINT_VAL=1`) and compare against the standard validation numbers.

Table 30: Disjoint-validation robustness: Load-only gap-MSE (seed 42), standard validation vs. strictly disjoint pre-gap validation, on the twelve non-degenerate datasets.

Dataset	Standard val.	Disjoint val.	$\Delta\%$
PepcoCOM	0.002379	0.002379	0.0
RICom1013	0.020059	0.020059	0.0
RIInd1014	0.189767	0.202492	+6.7
RIRes1012	0.050612	0.050612	0.0
UES_NH_Med	0.312879	0.275488	−12.0
WCMA1010res	0.006376	0.006376	0.0
WCMA1011lig	0.151660	0.228393	+50.6
WCMAnatGridRes1004	0.015930	0.015930	0.0
UCI_Elec_MT_001	0.002410	0.002410	0.0
UCI_Elec_MT_150	0.004735	0.004698	−0.8
UCI_Elec_MT_320	0.008183	0.008205	+0.3
ETTh1	0.016248	0.016653	+2.5

The median change is **0.0%** and six of twelve datasets are exactly unchanged (long series where the disjoint split is free). The remaining changes are mixed in sign (e.g. UES_NH_Med -12.0%, ETTh1 $+2.5\%$), and the single large swing (WCMA1011lig $+50.6\%$) is the short, zero-inflated lighting series whose early-stopping epoch is inherently high-variance. A systematic leak would produce a consistent one-directional degradation; the mixed-sign, median-zero pattern instead indicates early-stopping variance, not a leak advantage.

X Incremental Ablation Summary

Table 31 consolidates the marginal contribution of each pipeline component on the nine proprietary datasets, using Load-only metrics on the min-max $[0, 1]$ scale. Starting from the bridge output decoded by a daily MLP (row a), adding per-hour weather and calendar conditioning reduces mean MSE by 8.1% (row b). The enhanced decoder (Load-weighted loss, noise augmentation, $3\times$ capacity) yields a further 9.3% reduction (row c), bringing mean MSE to 0.069. Generating a 20-member noise-perturbed ensemble (σ=0.15) does not alter point accuracy but lowers CRPS from 0.169 (=MAE for a deterministic predictor) to 0.155 (-8.3%; row d), confirming that the latent perturbations capture genuine predictive uncertainty. Finally, wrapping the pipeline with ACI (row e) yields 93.4–95.4% empirical coverage across all nine datasets.

Table 31: Incremental ablation on nine proprietary datasets. Each row adds one component to the pipeline; MSE and MAE are Load-only on the min-max $[0, 1]$ scale. For deterministic predictors, CRPS = MAE; the ensemble row reports the energy-form CRPS with M=20 members. Coverage is the ACI empirical rate (target 95%).

	Configuration	MSE	MAE / CRPS	Δ_{MSE} (%)	ACI Cov.
(a)	Bridge $\rightarrow$ daily decoder	0.083	0.183†	—	—
(b)	+ hourly conditioning	0.076	0.174†	-8.1	—
(c)	+ Load-wt. loss, noise aug., $3\times$ capacity	0.069	0.169†	-9.3	—
(d)	+ 20-member ensemble (σ=0.15)	0.069	0.155	—	—
(e)	+ ACI (γ=0.01)	0.069	0.155	—	93.4–95.4%

†For deterministic predictions, CRPS = MAE.

Y Few-Shot Data Efficiency

Figure 14 shows the learning curves when fine-tuning only the decoder on the last N days of available training data. Each curve reports mean $\pm$ standard deviation over 3 seeds. Horizontal dashed lines indicate the Gap-MSE achieved by full-data baselines (1 241 days for UCI, 616 days for ETTh1).

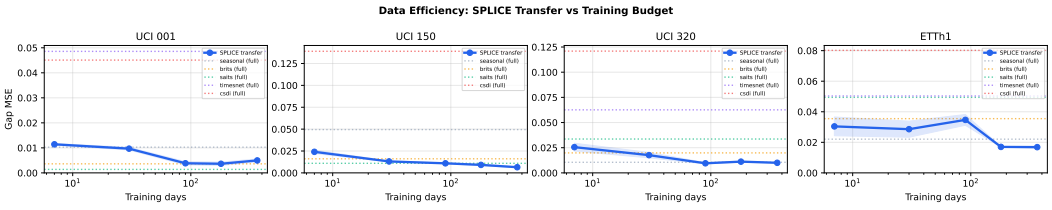


Figure 14: Few-shot data efficiency: SPLICE transfer (pooled JEPA + pooled bridge, decoder fine-tuned on N days) vs. full-data baselines. With $\approx$90 days of target data, SPLICE matches or beats methods trained on the full history.