# Schema-Adaptive Action-Conditioned JEPA for Cross-Machine CNC Transfer under Partial Sensor Overlap

Ayoub Louaye Bouaziz[1] Matthieu Ostertag[2] Anton Demasles[2]

[1]Université de Bretagne Occidentale, Brest, France [2]Mines Nancy, Université de Lorraine, Nancy, France

abouaziz@univ-brest.fr matthieu.ostertag7@etu.univ-lorraine.fr anton.demasles2@etu.univ-lorraine.fr

## Abstract

*Cross-machine deployment of industrial world models requires transfer across changes in machine dynamics, sensing interfaces, sampling regimes, and control units. This study investigates a schema-adaptive action-conditioned Joint-Embedding Predictive Architecture (SAAC-JEPA) for CNC dynamics when a source machine provides 17 canonical sensor channels and a target machine shares only 10. The evaluation protocol contains group-disjoint source splits, source-only normalization, deterministic held-out self-supervised validation, explicit unit audits, and a sealed target-machine test used only after model locking. A matched five-seed comparison shows no clean-source forecasting gain from JEPA pretraining: scratch and pretrained-body models obtain RMSE* $0.811 \pm 0.022$ *and* $0.813 \pm 0.022$*, respectively. A source-only architecture search over 20 candidates then selects a schema-consistent action-conditioned JEPA variant after stability checks across seven seeds. On the single confirmatory target-machine pass, the locked model reaches zero-shot RMSE 0.546,* $R^2 = 0.012$*, and NLL 0.52, improving over persistence but not over official RevIN-equipped PatchTST and iTransformer zero-shot baselines (0.503 and 0.498). A post-lock, pre-declared paired ablation attributes this gap to normalization: RevIN inside the same architecture yields* $0.495 \pm 0.004$ *over three seeds, level with the baselines, but collapses target calibration (NLL 20.6) on stationary context windows. A pre-lock adaptation sweep further suggests that small target support can reduce RMSE to 0.520. The evidence supports a narrow conclusion: source-domain forecasting accuracy alone is insufficient for evaluating industrial predictive representations, and cross-machine adaptation under partial sensor overlap is a distinct evaluation axis for action-conditioned world models.*

## 1. Introduction

Industrial control systems generate multivariate sensor streams while exposing controllable inputs such as spindle speed and axis feed commands. A deployable world model should represent the evolution of the physical process, predict future states under candidate actions, and remain useful when the sensing interface changes across machines. This requirement differs from conventional in-domain time-series forecasting because the target machine can exhibit both *domain shift* and *schema shift*: dynamics, controller conventions, units, sampling rates, and sensor availability can all change simultaneously.

Joint-Embedding Predictive Architectures (JEPAs) learn by predicting representations rather than reconstructing every observed sample [1, 2]. This objective is potentially attractive for industrial transfer because a latent representation can emphasize predictable process dynamics while reducing dependence on machine-specific observation details. However, JEPA training is susceptible to representational collapse, and strong alternatives already exist in time-series forecasting and masked pretraining [3–6]. Action-aware industrial world modeling has also been considered previously [11]. Consequently, the relevant question is not whether JEPA can be applied to industrial data, but whether schema-adaptive, action-conditioned predictive representations remain useful under cross-machine and partial-observation shift.

This study evaluates that question on a CNC source-to-target setting (Figure 1) built from two public datasets: the THWS five-axis CNC milling production dataset with multiple changeovers [9] as the source machine, and the FH JOANNEUM CNC machining data repository with high-frequency energy measurements [10] as the target machine. The source machine contains 17 canonical sensor channels across 62 program sessions, whereas the target exposes a 10-channel transferable subset across 7 independent runs. The protocol was designed to prevent target leakage during model selection and includes source-only normalization, session-grouped splitting, held-out SSL validation, unit audits, collapse diagnostics, matched scratch/pretraining controls, trivial forecasting baselines, and run-level target statistics.

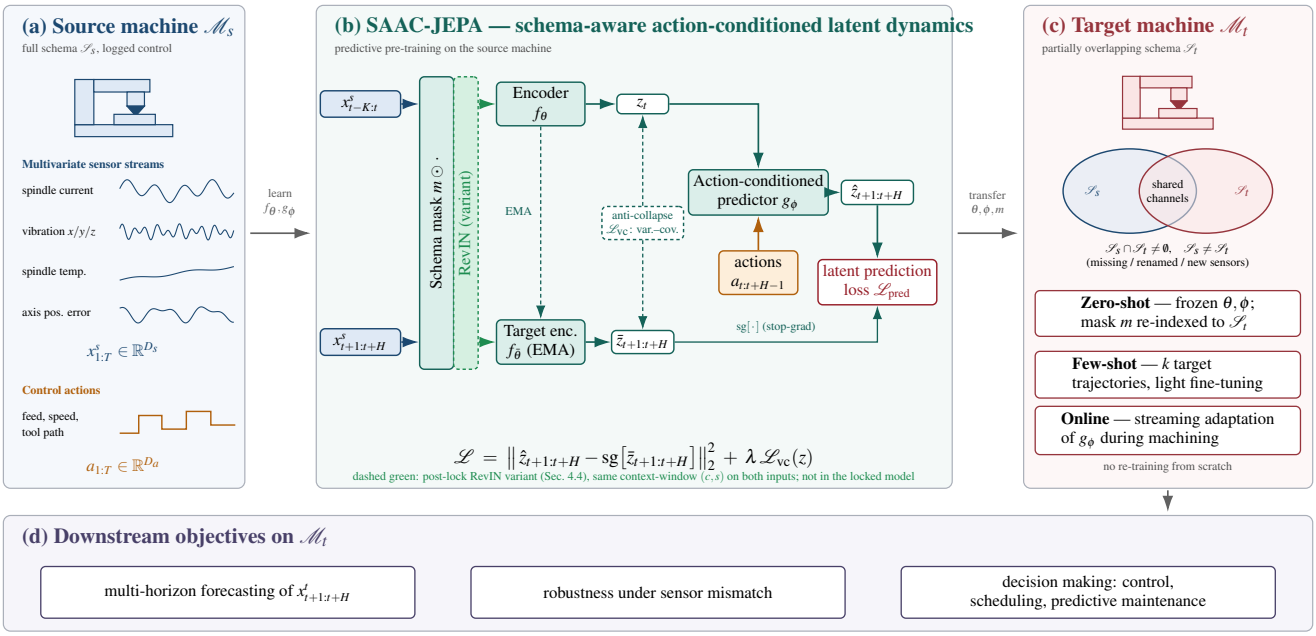



Figure 1. **Overall view of SAAC-JEPA for cross-machine CNC transfer.** A source CNC machine provides multivariate sensor streams and control actions, whereas the target machine exposes only a partially overlapping sensor schema. The model learns action-conditioned latent dynamics using a context encoder, an EMA target encoder, latent prediction, schema masking, and anti-collapse regularization. The learned representation is subsequently evaluated under zero-shot, few-shot, and online adaptation settings. The downstream objective is cross-machine forecasting and decision support under machine and sensor-schema shift. Dashed green blocks mark the post-lock instance-normalized variant (Sec. 4.4); the locked model omits them.

**Contributions.** The contributions are as follows:

- a leakage-audited cross-machine CNC protocol combining machine shift with partial sensor overlap;
- a schema-adaptive action-conditioned JEPA formulation with latent schema consistency, EMA targets, and explicit anti-collapse regularization;
- an empirical diagnosis showing that JEPA pretraining does not improve clean source RMSE over scratch, while representation stability is sensitive to latent-loss weighting, variance–covariance regularization, and EMA dynamics;
- a pre-lock diagnostic adaptation sweep suggesting that limited target support improves cross-machine forecasting relative to zero-shot transfer; it is retained as diagnostic rather than confirmatory evidence;
- a post-lock paired ablation showing that presence-aware instance normalization (RevIN) inside SAAC-JEPA closes the zero-shot RMSE gap to the official RevIN-equipped forecasters ($0.495 \pm 0.004$ against 0.503 and 0.498) while collapsing target calibration, which isolates input normalization, not the latent objective, as the dominant zero-shot factor.

## 2. Related Work

**Joint-embedding predictive learning.** I-JEPA introduced representation prediction as a non-generative self-supervised objective for images [1], later extended to video [2]; VICReg regularizes representation variance and covariance to avoid collapse [3]. The CNC experiments below show that anti-collapse control is operationally necessary rather than auxiliary.

**Time-series forecasting and masked pretraining.** SimMTM reconstructs masked series from complementary masked neighbors [4]; PatchTST uses channel-independent patch tokens [5]; iTransformer inverts the tokenization axis to model multivariate dependencies [6]; RevIN counters distribution shift by instance normalization [7]; Sec. 4.4 tests it inside SAAC-JEPA under sensor presence masks. These methods motivate a protocol in which JEPA is not assumed to dominate source-domain forecasting and must justify its value under transfer and schema mismatch.

**World models and industrial adaptation.** Latent dynamics models support multi-step prediction and planning [8], and actionable world models have been studied for industrial process control [11]. The present setting adds cross-machine transfer, partial sensor overlap, explicit action alignment, source-only normalization, and post-training adaptation on the target machine.

## 3. Problem Formulation

Let $\mathcal{M}_s$ and $\mathcal{M}_t$ denote the source and target CNC machines, with sensor schemas $\mathcal{S}_s$ and $\mathcal{S}_t$. The source observation at time $t$ is $x_t^s \in \mathbb{R}^{D_s}$, and the target exposes only a partially overlapping set of channels, $\mathcal{S}_s \cap \mathcal{S}_t \neq \emptyset$ and $\mathcal{S}_s \neq \mathcal{S}_t$. A binary mask $m_t \in \{0,1\}^{D_s}$ indicates the channels available to the encoder after re-indexing to the canonical source schema. The control vector $a_t$ contains the commanded spindle and axis-feed variables applied over $(t, t+1]$, so $x_{t+1}$ depends on $a_t$ in the usual world-model convention.

Given a context of length $K$ and a future action sequence of length $H$, the predictive objective is

$$p_\Theta(x_{t+1:t+H} \mid x_{t-K:t}, m_{t-K:t}, a_{t:t+H-1}). \tag{1}$$

The mask enters concretely at the data level. Every machine is re-indexed to the canonical source schema before batching, so a channel that a machine does not expose carries the value 0 together with $m = 0$ rather than an unmarked entry. The encoder then builds one token per channel and time step and excludes the absent ones both from the sensor-axis attention and from the pooling that reduces the channel tokens to one representation per time step (falling back to a single channel when a step has none present), so an absent channel is treated differently from a channel that genuinely reads zero.

The forecasting metric measures numerical prediction error. The world-model requirement is stronger: predictions should also respond meaningfully to changes in future actions. This distinction is essential in highly autocorrelated industrial streams, where a model can obtain low short-horizon error while effectively ignoring control inputs. Figure 2 illustrates this operational interpretation. The figure is read as three lanes sharing one time axis that is split at *now*: a physical lane carrying the measured and the predicted process variables together with an illustrative alarm threshold, a latent lane carrying the encoder, the latent dynamics and the decoder, and a control lane carrying the already applied and the candidate action sequences.

## 4. Method

### 4.1. Schema-adaptive context encoding

The context encoder $f_\theta$ receives observed sensor values, the sensor-presence mask, and past actions:

$$z_t^c = f_\theta(x_{t-K:t}, m_{t-K:t}, a_{t-K:t}). \tag{2}$$

Schema masking generates multiple partial views of the same source trajectory. Missing channels are represented structurally rather than replaced before normalization, so the model receives an explicit indication of which measurements are available.

In words, $f_\theta$ first runs a transformer over the tokens of the sensors present at a given time step, reduces them to a single time token by a mean restricted to those present channels, adds a linear projection of the action applied at that step to the time token, and then runs a temporal transformer over the $K$ time tokens; $z_t^c$ is the representation at the last position. The locked configuration uses $d = 256$, eight attention heads, six layers in the temporal transformer and pre-norm blocks (App. G).

### 4.2. Action-conditioned latent prediction

A predictor $g_\phi$ receives the context representation and candidate future actions:

$$\hat{z}_{t+1:t+H} = g_\phi(z_t^c, a_{t:t+H-1}). \tag{3}$$

In practice the context spans $K = 32$ steps and prediction is made at a fixed set of horizons $\mathcal{H} = \{1, 2, 4, 8, 16\}$, $H = 16$, by direct multi-horizon prediction rather than autoregressive rollout. Windows are formed on the 1 Hz grid (Table 1), so the context covers 32 s and the longest horizon 16 s. For each $h \in \mathcal{H}$ the predictor receives the mean of $a_{t:t+h-1}$, i.e. the mean command applied over $(t, t+h]$, so the action conditioning captures the level of the future command rather than its temporal profile. All horizon slots are filled in a single causal pass from the same $z_t^c$: every slot receives the context representation, a horizon embedding and its own mean command, and no predicted latent is fed back as an input to another slot. The slots are the horizons of the evaluation grid, so training and evaluation share the same prediction structure. A target encoder $f_{\bar{\theta}}$ embeds the future observation window into $\bar{z}_{t+1:t+H}$. The target parameters are updated by exponential moving average,

$$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta, \tag{4}$$

and gradients are stopped through the target branch.

The SSL objective is

$$\mathcal{L}_{\text{SSL}} = \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{vic}}\mathcal{L}_{\text{VICReg}} + \lambda_{\text{sch}}\mathcal{L}_{\text{schema}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}}, \tag{5}$$

where $\mathcal{L}_{\text{lat}}$ aligns predicted and target latent trajectories, $\mathcal{L}_{\text{VICReg}}$ prevents low-variance or redundant embeddings, and $\mathcal{L}_{\text{schema}}$ encourages two partial sensor views of the same trajectory to predict compatible future latents. The variance–covariance term combines a variance hinge with a covariance penalty and carries no invariance component, because the latent prediction loss already supplies the alignment between the two branches; in the locked configuration it is applied per horizon to the target latents and to the pooled context latent, and only the context term contributes gradient because the target branch is detached.

Schema consistency is imposed on latent dynamics rather than on an untrained physical output head. During pure self-supervised pretraining the physical head has no supervised target and receives no gradient, so a constraint on its output would be empty; the constraint therefore acts

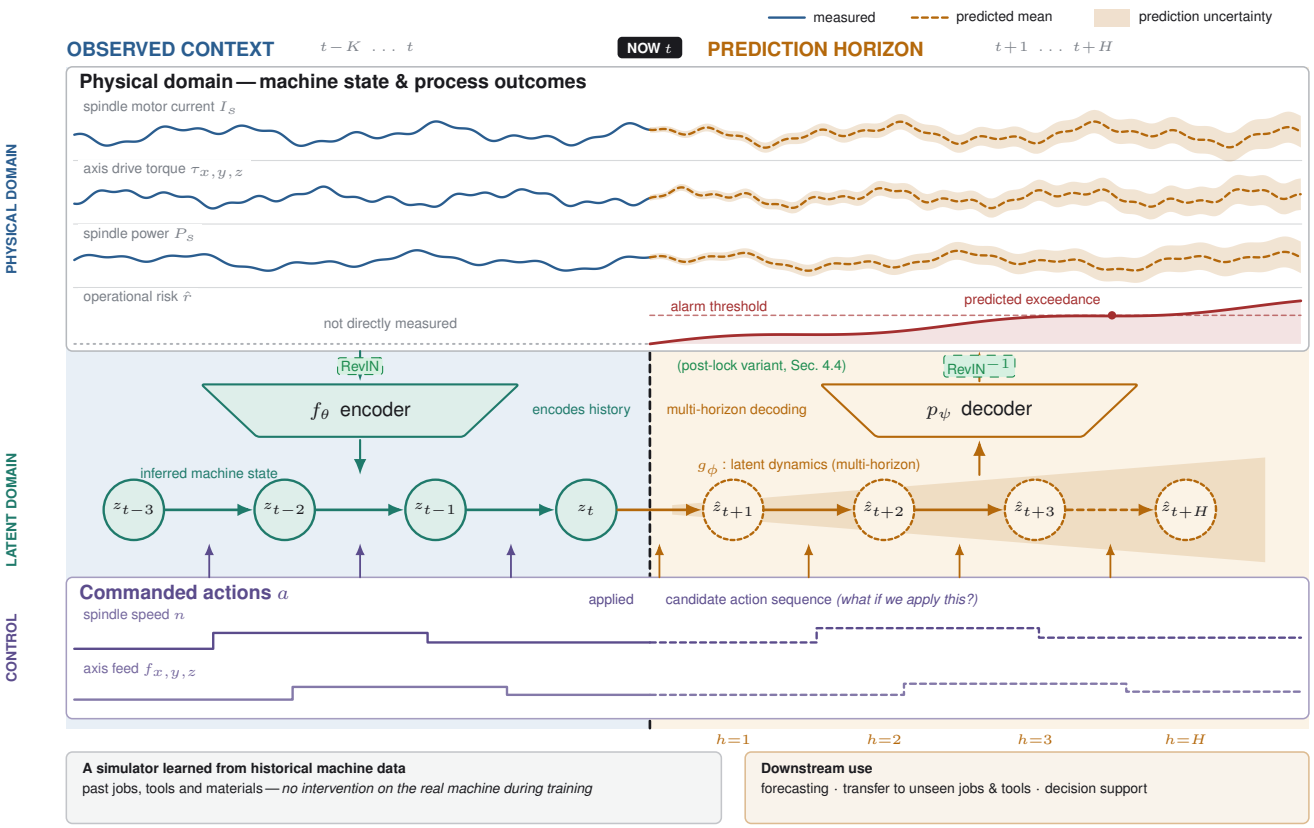



Figure 2. **A CNC world model in practice.** The schematic illustrates the operational interpretation of the learned dynamics model. Recent sensor measurements are encoded into a latent state, future control actions condition multi-horizon latent predictions, and a physical decoder maps predicted latent states to multi-horizon process variables and uncertainty. The figure is conceptual and does not constitute a closed-loop deployment result. Its purpose is to distinguish passive forecasting from action-conditioned simulation of candidate futures. The chain of latent states and the operational-risk lane are illustrative: the implementation predicts all horizons in one pass from the context latent (Sec. 4.2), and the risk variable is a downstream use, not an output of the model in this paper. Dashed green blocks mark the post-lock instance-normalized variant (Sec. 4.4); the locked model omits them.

on the predicted future latents. The ablated view keeps each channel with probability $\kappa = 0.65$ and always retains at least one channel, and the constraint is one-sided: the full-view prediction is detached and used as the reference. $\mathcal{L}_{\text{act}}$ is an action-recovery loss: a small head $r_\omega$ predicts $\bar{a}_h = \text{mean}(a_{t:t+h-1})$ from the pair $(z_t^c, \bar{z}_{t+h})$, so that the latent transition must retain information about the applied action. The future latent in that pair comes from the detached target branch, so the only trainable path runs through $z_t^c$: the objective pressures $z_t^c$ to retain what identifies the applied command; Sec. 6.3 reports the resulting source-validation sensitivity. The selection score of Sec. 5.2 carries the matching diagnostic: a candidate whose source-validation RMSE under shuffled future actions is less than 1.02 times its RMSE under true actions is flagged as making weak use of the action; the locked candidate reaches 1.058. Figure 3 shows both mechanisms.

**Algorithm 1: SAAC-JEPA self-supervised pretraining step (one mini-batch)**

**Input:** windows $(x_{t-K:t}, m_{t-K:t}, a_{t-K:t}, \{x_{t+h}, a_{t:t+h-1}\}_{h\in\mathcal{H}})$; context encoder $f_\theta$; predictor $g_\phi$; action-recovery head $r_\omega$; EMA target encoder $f_{\bar{\theta}}$; mask ratio $\rho$; view keep-probability $\kappa$; weights $\lambda_{\text{lat}}, \lambda_{\text{vic}}, \lambda_{\text{sch}}, \lambda_{\text{act}}$; momentum $\tau$

$\bar{a}_h \leftarrow \text{mean}(a_{t:t+h-1})$ for $h \in \mathcal{H}$
▷ level of the future command

$(\tilde{x}, \tilde{m}) \leftarrow \textsc{Corrupt}(x_{t-K:t}, m_{t-K:t}; \rho)$
▷ channel masking; masked entries set to 0 and marked absent

*RevIN variant only:* $(c, s) \leftarrow \textsc{Stats}(\tilde{x}, \tilde{m})$; $\tilde{x} \leftarrow (\tilde{x} - c)/s$; $x_{t+\mathcal{H}} \leftarrow (x_{t+\mathcal{H}} - c)/s$
▷ per-channel context-window statistics over present entries; identity for unseen channels (Sec. 4.4)

$z_t^c \leftarrow f_\theta(\tilde{x}, \tilde{m}, a_{t-K:t})$
▷ context latent

$\hat{z}_{t+h} \leftarrow g_\phi(z_t^c, \bar{a}_h)$ for all $h \in \mathcal{H}$
▷ one causal pass over the horizon slots

$\bar{z}_{t+\mathcal{H}} \leftarrow \text{sg}\big[f_{\bar{\theta}}(x_{t+\mathcal{H}}, m_{t+\mathcal{H}}, \mathbf{0})\big]$
▷ the target rows are encoded jointly, with zero actions; stop-

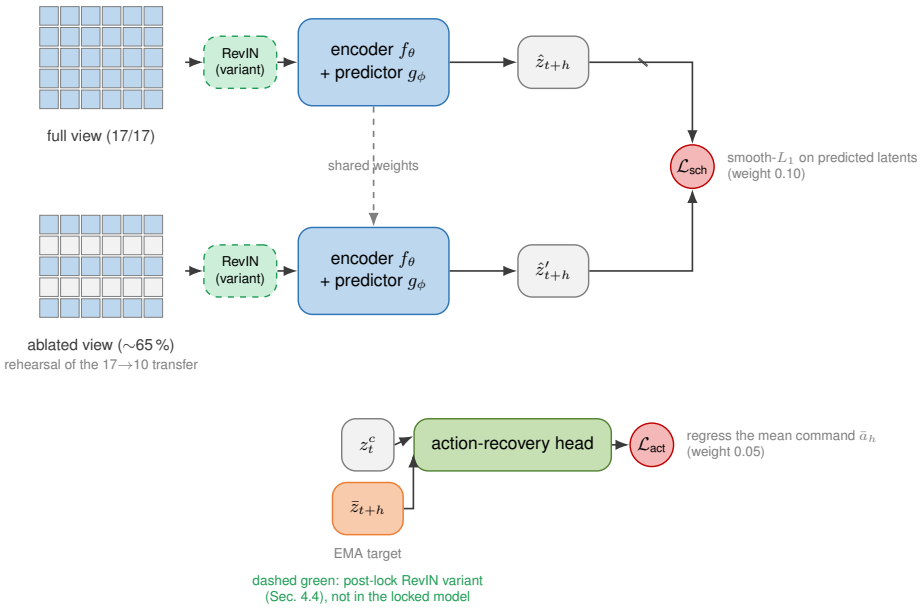



Figure 3. **The two auxiliary objectives of the locked candidate.** *Schema consistency*: the same window is processed twice through the shared encoder and predictor, once with the channel-masked context view and once with a further random channel subset of it (each channel kept with probability $\kappa = 0.65$, at least one channel kept); the predicted latents of the subset view are pulled toward the detached predicted latents of the full view with a smooth-$L_1$ penalty ($\lambda_{\text{sch}} = 0.10$), which exposes the model to reduced-schema views during source training. *Action recovery*: a small head reads the context latent $z_t^c$ and the EMA-target latent $\bar{z}_{t+h}$ and regresses the mean command $\bar{a}_h$ applied over $(t, t+h]$ ($\lambda_{\text{act}} = 0.05$), so the latent transition must retain information about the action. Dashed green blocks mark the post-lock instance-normalized variant (Sec. 4.4); the locked model omits them.

gradient
$\mathcal{L}_{\text{lat}} \leftarrow \frac{1}{|\mathcal{H}|} \sum_h \ell\big(\hat{z}_{t+h}, \bar{z}_{t+h}\big)$
▷ $\ell$: smooth-$L_1$
$\mathcal{L}_{\text{VICReg}} \leftarrow \frac{1}{2}\big[V(z^c_{\text{pool}}) + \frac{1}{|\mathcal{H}|}\sum_h V(\bar{z}_{t+h})\big]$
▷ $z^c_{\text{pool}}$: time-pooled context; $V$: variance hinge and covariance penalty, no invariance term; the target half is constant
$\mathcal{L}_{\text{act}} \leftarrow \frac{1}{|\mathcal{H}|}\sum_h \big\| r_\omega(z_t^c, \bar{z}_{t+h}) - \bar{a}_h \big\|^2$
▷ action recovery
$(\tilde{x}', \tilde{m}') \leftarrow \textsc{SchemaView}(\tilde{x}, \tilde{m}; \kappa)$
▷ keep each channel with probability $\kappa$, at least one
$\hat{z}'_{t+h} \leftarrow g_\phi\big(f_\theta(\tilde{x}', \tilde{m}', a_{t-K:t}), \bar{a}_h\big)$ for $h \in \mathcal{H}$
$\mathcal{L}_{\text{schema}} \leftarrow \frac{1}{|\mathcal{H}|}\sum_h \ell\big(\hat{z}'_{t+h}, \text{sg}[\hat{z}_{t+h}]\big)$
▷ the ablated view must predict the same future latents
$\mathcal{L}_{\text{SSL}} \leftarrow \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{vic}}\mathcal{L}_{\text{VICReg}} + \lambda_{\text{sch}}\mathcal{L}_{\text{schema}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}}$
update $(\theta, \phi, \omega)$ by AdamW on $\nabla\mathcal{L}_{\text{SSL}}$, gradient norm clipped to 1
$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$
▷ after the optimiser step

The locked candidate sets $\rho = 0.35$, $\kappa = 0.65$, $\tau = 0.996$ and $(\lambda_{\text{lat}}, \lambda_{\text{vic}}, \lambda_{\text{sch}}, \lambda_{\text{act}}) = (1.0, 0.05, 0.10, 0.05)$; the complete configuration is listed in App. G. Checkpoints of this stage are selected on a deterministic held-out SSL validation pass that scores the latent and VICReg terms only, never on the training loss. The second stage attaches a fresh probabilistic head and fine-tunes the whole network jointly, adding the Gaussian negative log-likelihood of the physical output to the unchanged self-supervised terms, with schema consistency then measured on the physical output rather than on the latents; checkpoints of this stage are selected on source-validation RMSE (Sec. 4.3).

### 4.3. Physical forecasting head and adaptation

Training therefore proceeds in two stages: self-supervised pretraining with the physical weight set to zero, so that no gradient reaches the output head, followed by joint fine-tuning in which the Gaussian negative log-likelihood is added to the self-supervised terms. After pretraining, a fresh probabilistic head maps latent forecasts to channel-wise means and log variances. The primary fine-tuning protocol uses a fresh head because an output head exposed during pure SSL has no supervised physical target. Three matched controls are retained: full training from scratch, pretrained body with a fresh head, and complete checkpoint transfer.

Target adaptation updates a restricted subset of parameters using target support windows. Zero-shot transfer uses the source-trained model without target updates. Few-shot transfer takes the leading fraction of each target run as support, updates the predictor and the physical head for 50 steps at learning rate $10^{-4}$, and evaluates on the remaining, later windows of the same runs, so support and query windows

are disjoint. Online adaptation is reserved for causal prequential evaluation and is not used in the results reported in this study.

### 4.4. Instance-normalized variant

The locked model normalizes every channel with a fixed source-train z-score. The official forecasting baselines instead apply RevIN [7], which centers and scales each window by its own statistics and inverts the transform on the output. To test whether the zero-shot gap of Sec. 6.6 is a normalization effect rather than a property of the latent objective, a variant of SAAC-JEPA applies a presence-aware RevIN inside the network with the settings of the official baselines (no affine parameters, no last-value subtraction, $\epsilon = 10^{-5}$), so that the two arms differ by this transform only and by no parameter. For each window and channel $k$, the center $c_k$ and scale $s_k$ are computed on the $K$ context rows only and on present entries only, $c_k = \sum_t m_{t,k} x_{t,k} / \sum_t m_{t,k}$ and $s_k^2 = \sum_t m_{t,k}(x_{t,k} - c_k)^2 / \sum_t m_{t,k} + \epsilon$, and are detached from the graph; masked and absent entries are stored as zeros and would bias a plain mean, a choice shared with the official SimMTM pretraining code. A channel never observed in a window keeps the identity transform, so the seven channels absent on the target machine behave exactly as in the locked model. The target rows $x_{t+\mathcal{H}}$ are normalized with the context statistics before entering $f_{\bar{\theta}}$, so no future statistic reaches the model and every latent loss is computed in instance space. The physical head predicts in instance space, and its mean and log-variance are mapped back to the global z-space, $\hat{\mu}_k \leftarrow s_k \hat{\mu}_k + c_k$ and $\log \hat{\sigma}_k^2 \leftarrow \log \hat{\sigma}_k^2 + 2 \log s_k$, before any loss or metric, so the evaluation space, the trivial-predictor anchor and the adaptation procedure are unchanged. With the transform disabled the code path is bitwise that of the locked model, which regression tests enforce.

## 5. Experimental Protocol

### 5.1. Data and audits

The source machine is the Spinner U5-620 five-axis machining center of the THWS production dataset with multiple changeovers [9]; the target machine is provided by the FH JOANNEUM CNC machining data repository [10]. Both datasets are public under CC BY 4.0. The merged table (Table 1) contains 1,533,700 rows and 25 columns. THWS contributes 29,785 samples at 1 Hz, segmented into 62 NC-program sessions after excluding jog periods and splitting at temporal gaps longer than 5 s. JOANNEUM contains 1,503,915 raw controller-cycle records across 7 runs and 13 sessions. The source provides 17 canonical sensor channels, whereas 10 current/torque/power channels are shared with the target. Four mapped action variables represent spindle speed and the commanded X/Y/Z feeds.

Source sessions are divided into 42 training, 11 validation, and 9 test sessions with group-disjoint splitting. The resulting source window counts are 18,267 / 3,423 / 5,189, with zero window overlap across partitions. Windows are indexed inside each (machine, session, run) group and never cross a group boundary, so a group must contain at least $K + \max \mathcal{H} = 48$ rows before it yields a single window, whose context is the $K = 32$ rows preceding $t$ and whose targets are the five rows $t + h$, $h \in \mathcal{H}$. The same requirement explains why the 5% support subset of Fig. 6b contains no complete window. All normalizers are fitted on source training data only. A unit audit detected three mismatches that were corrected deterministically before modeling: JOANNEUM power channels were stored in watts while THWS used kilowatts (factor $10^{-3}$); commanded axis feeds were recorded in mm/s on one machine and mm/min on the other (factor 60); and spindle speed was logged in degrees per second on one machine and rpm on the other (factor 6, verified from position derivatives). The last two matter because the action channels condition the predictor. Target statistics are not used to fit source normalizers. Validation masking is deterministic, DataLoader workers are seeded, checkpoint loading reports missing and unexpected keys, and regression tests cover split integrity, masking determinism, EMA and stop-gradient behavior, and metric aggregation. All experiments ran on one NVIDIA DGX Spark (one GB10 GPU), about six GPU-days in total.

Table 1. Audited cross-machine data protocol.

| | THWS source | JOANNEUM target |
|---|---|---|
| Raw rows | 29,785 | 1,503,915 |
| Independent units | 62 sessions | 7 runs / 13 sessions |
| Working rate | 1 Hz | 1 Hz (from 2 ms cycles) |
| Canonical/shared sensors | 17 | 10 shared |
| Actions | spindle + commanded X/Y/Z feeds | |
| Normalization | source-train only | source normalizer |

### 5.2. Model selection and baselines

The source-only architecture search defines 20 candidates prior to execution. The search spans masking strategies, action-injection mechanisms, VICReg placement, latent-loss weight, horizon aggregation, schema consistency, action recovery, EMA dynamics, model capacity, normalization, and head initialization. Each candidate is first run with four seeds. Ranking uses a composite source-validation score: each component is z-scored within the candidate pool and combined as $0.30$ clean RMSE $+ 0.25$ schema-drop RMSE $+ 0.15$ masked RMSE $+ 0.10$ long-horizon RMSE $+ 0.10$ calibration penalty $+ 0.10$ SSL validation loss, plus penalties of $3.0$ per collapsed seed and $1.0$ for weak action usage (lower is better). The top-5 candidates receive a fifth seed. If the provisional winner changes under confirmation, model locking is refused and additional seeds are added to

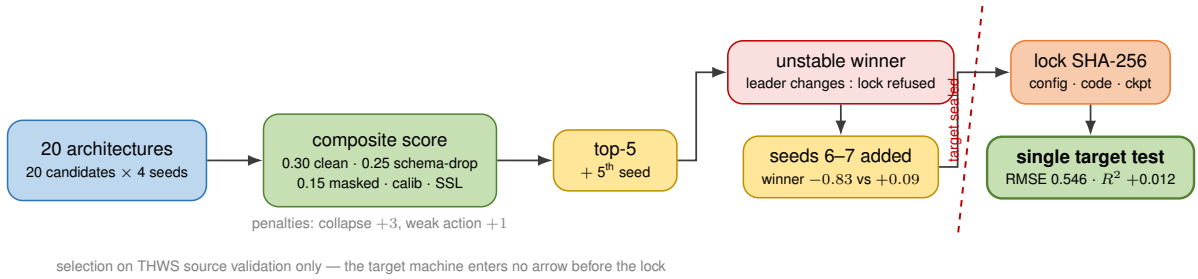



Figure 4. **Leakage-free model-selection pipeline.** Twenty pre-defined SAAC-JEPA candidates are ranked on THWS source validation only, first with four seeds and then with a fifth seed for the top-5. When the provisional winner changes after confirmation, locking is refused and extra seeds are added to the leading candidates. In the present study, the per-horizon VICReg variant led the discovery stage, but the schema-consistency variant with action recovery remained superior after the full stability check and was frozen before a single confirmatory evaluation on the JOANNEUM target machine.

the tied leaders. This occurred here: the per-horizon VICReg variant led the discovery stage, but the schema-consistency variant with action recovery overtook it after confirmation and remained superior after seven valid seeds, after which the configuration, code, normalizer, and checkpoint were frozen before a single confirmatory pass on the JOANNEUM target machine.

The six components are measured on source validation after fine-tuning: clean RMSE on the unmasked source-validation windows; schema-drop RMSE, obtained with the schema restricted to the 10 channels shared with the target; masked RMSE under deterministic corruption masks; long-horizon RMSE, taken as the worst of the five per-horizon RMSE values; a calibration penalty, the absolute difference between the empirical coverage at the nominal 90% level and 0.90; and the best held-out SSL validation loss reached during pretraining. Each component is z-scored within the pool of usable runs, a component with zero variance across that pool contributing 0, and the resulting score is averaged over the seeds of a candidate. Algorithm 2 states the procedure; the locked values are listed in App. G.

**Algorithm 2: Source-only model selection and lock**

**Input:** candidate set $\mathcal{C}$, $|\mathcal{C}| = 20$, written down before any run; source train/validation splits; component weights $w$; penalties $\pi_{\text{col}} = 3$, $\pi_{\text{act}} = 1$; the target machine stays sealed

**for** each $c \in \mathcal{C}$ and each of four seeds **do**
  train $c$ (Algorithm 1, then head fine-tuning); measure the six components on source validation
  drop the run from the pool if the pretraining latents collapse (latent std $< 0.05$ or effective rank $< 10\%$ of $d$); flag post-fine-tuning collapse and weak action use ($\text{RMSE}(\text{shuffled } \bar{a})/\text{RMSE}(\text{true } \bar{a}) < 1.02$)
**end for**
z-score each component within the pool of valid runs; per seed $S \leftarrow \sum_k w_k z_k + \pi_{\text{col}}\,\mathbf{1}[\text{collapsed}] + \pi_{\text{act}}\,\mathbf{1}[\text{weak}]$; $S(c) \leftarrow$ mean over the valid seeds of $c$
$\mathcal{T} \leftarrow$ the five lowest $S$; add a fifth seed to each $c \in \mathcal{T}$; recompute $S$ within $\mathcal{T}$
**while** the leader of $\mathcal{T}$ changed since the previous seed count **do**
  refuse to lock; add seeds to the leading candidates; recompute $S$
  ▷ seven seeds settled it here
**end while**
$c^\star \leftarrow \arg\min_{c \in \mathcal{T}} S(c)$; lock only if $c^\star$ has no collapse flag, no weak-action flag, at least five valid seeds, a stable rank-1 across the confirmation stage, and the confirmation stage was run; freeze configuration, code, normalizer and the best-composite checkpoint with SHA-256 hashes
verify the hashes and evaluate $c^\star$ **once** on the sealed target windows
  ▷ the only pass that touches the target before the lock; the post-lock ablation below is declared separately

**Post-lock declared ablation.** The lock consumed the single declared target pass. One paired ablation was run afterwards, in response to a reviewer remark that SAAC-JEPA lacked the instance normalization used by the target baselines. Its protocol was fixed before any run: the control arm re-uses the locked M03 configuration verbatim (its three re-runs reproduce the V1 validation metrics exactly), the variant arm differs only by the RevIN flag (Sec. 4.4), both arms are trained with the same three seeds through the same two-stage pipeline, and a second, declared opening of the target windows was conditioned on the variant winning all four RMSE components of the composite score on source validation. That verdict was recorded before the target stage ran, and the target stage then evaluated all six runs with no selection on the target. The target windows have therefore been read twice: once by the locked model, which remains the only confirmatory result, and once by this ablation, whose numbers are reported as diagnostic evidence with their seed

count. The lock itself was not revisited.

Baselines include supervised MLP, GRU, LSTM, TCN, DLinear, Transformer, TSMixer, and RSSM models, together with trivial persistence and linear-drift references. For the target comparison, the official PatchTST and iTransformer repositories are trained once on the same source-train windows (single seed, repository defaults, RevIN enabled) and evaluated on the same sealed target windows; they receive the 17 sensors plus the 4 action channels as input, and channels absent on the target are zero-imputed in normalized space (Appendix E). Six in-house reconstruction-pretraining baselines (MAE-, SimMTM-, PatchFormer-, EMIT-, MMR- and LoMaR-style) were implemented but are not reported here; the official SimMTM repository did not converge on this data regime.

### 5.3. Metrics

Forecasting is evaluated using RMSE, MAE, $R^2$, and per-horizon/per-channel errors. The probabilistic head is evaluated with Gaussian NLL and empirical predictive coverage. Where target inferential statistics are reported, the 7 JOANNEUM runs are the independent units; windows are not used as independent replicates. $R^2$ is computed against the pooled mean of the evaluated channels and horizons.

## 6. Results

### 6.1. Corrected JEPA pretraining is stable but does not improve clean-source RMSE

A protocol audit identified three issues in the initial pretraining implementation: checkpoint selection based on training loss, latent schema consistency applied to an untrained physical head, and representational collapse in the EMA target encoder. The corrected pipeline selects checkpoints on deterministic held-out SSL validation, moves schema consistency to latent space, audits checkpoint loading, and applies VICReg to context and target representations. Target-encoder effective rank increased from approximately 5% to 58% of the latent dimensionality after correction. Effective rank is defined here as the exponential of the entropy of the singular values of the centred latent matrix after normalizing them to sum to one, and is reported as a fraction of the latent dimensionality $d$; a run counts as collapsed when the smallest per-dimension standard deviation falls below 0.05 or the effective rank falls below 10% of $d$.

A matched five-seed comparison isolates the effect of pretraining (Table 2a). Training from scratch yields $0.811 \pm 0.022$ RMSE, pretrained body with a fresh head yields $0.813 \pm 0.022$, and complete checkpoint transfer yields $0.812 \pm 0.012$. The scratch-versus-pretraining difference is not significant ($p = 0.76$). Persistence and linear drift remain substantially weaker. Thus, no claim of improved clean-source forecasting from JEPA pretraining is supported by the current evidence.

### 6.2. Representation hyperparameters materially affect stability and forecasting

The latent-loss weight has a pronounced effect on downstream forecasting (Table 2b). The best tested value is $\lambda_{\text{lat}} = 0.25$ with RMSE 0.774; larger values degrade performance to 0.821 at $\lambda_{\text{lat}} = 1$ and 0.845 at $\lambda_{\text{lat}} = 4$. VICReg improves clean-source RMSE from 0.815 without regularization to 0.798 with per-horizon weighting and is additionally required to prevent collapse. Pooled horizon aggregation reaches 0.787 compared with 0.829 for per-horizon aggregation. Slow EMA targets are also preferable, with 0.9995 reaching 0.799 versus 0.808 at 0.999 and approximately 0.83 for faster updates. Two further observations temper the picture: removing the latent loss entirely ($\lambda_{\text{lat}} = 0$) yields 0.787, so the latent prediction objective at its default weight costs clean-source accuracy, and the best clean RMSE among mask ratios (0.804) is obtained with a very light mask (0.03), so the 0.35 mask of the search is retained for schema-drop robustness, not for clean accuracy. All ablations are single-seed runs on the pre-search base configuration. The locked candidate nevertheless keeps $\lambda_{\text{lat}} = 1$, per-horizon aggregation, mask ratio 0.35 and EMA 0.996, because the composite selection score (Sec. 5.2) weights schema-drop and masked robustness, calibration and action usage in addition to clean RMSE, and the discovery grid did not combine the ablation optima with schema consistency and action recovery. These ablations indicate that JEPA stability on CNC time series depends strongly on representation-objective balance rather than on a single default configuration.

The search also spans the corruption applied to the context window and the mechanism by which the action enters the predictor (Figure 5). Five masking modes are available: random masking removes individual (time, channel) entries; block masking removes contiguous runs of 2 to 12 steps inside a channel; channel masking removes whole sensors from the window; event masking removes a random subset, sized by the mask ratio, of the samples whose step-to-step change exceeds a per-channel quantile of the window; and the mixed mode takes the union of the four. The locked candidate uses channel masking, which removes entire sensors from the context and therefore mimics a channel that is absent on the target machine. Three action-injection mechanisms are compared: token injection adds a projection of the mean future command to each horizon slot, FiLM applies a feature-wise affine modulation of the context representation derived from that command, and cross-attention conditions the horizon slots on the action; these are single-seed runs on the pre-search base configuration, and the token entry is the one reported in Table 2b. Source-validation RMSE is 0.821 for token, 0.830 for FiLM and 0.838 for cross-attention, and the locked candidate uses token injection.

Table 2. **Source-validation results.** (a) Matched five-seed comparison of initializations (mean ± SD over seeds; lower is better; trivial baselines here and in Fig. 6a come from two separate evaluation scripts and differ slightly). (b) Selected ablations, single seed each on the pre-search base configuration; best tested setting per factor.

(a) Initialization (five seeds).

| Model initialization | RMSE |
|---|---|
| Scratch | $0.811 \pm 0.022$ |
| Pretrained body + fresh head | $0.813 \pm 0.022$ |
| Complete pretrained checkpoint | $0.812 \pm 0.012$ |
| Linear drift | 0.928 |
| Persistence | 1.135 |

(b) Ablations (single seed).

| Factor | Best tested setting | RMSE |
|---|---|---|
| Action injection | token | 0.821 |
| Mask ratio | 0.03 | 0.804 |
| Latent weight | $\lambda_{\text{lat}} = 0.25$ | 0.774 |
| VICReg | per-horizon, 0.1 | 0.798 |
| Transformer norm | pre-norm | 0.829 |
| Horizon aggregation | pooled | 0.787 |
| EMA | 0.9995 | 0.799 |

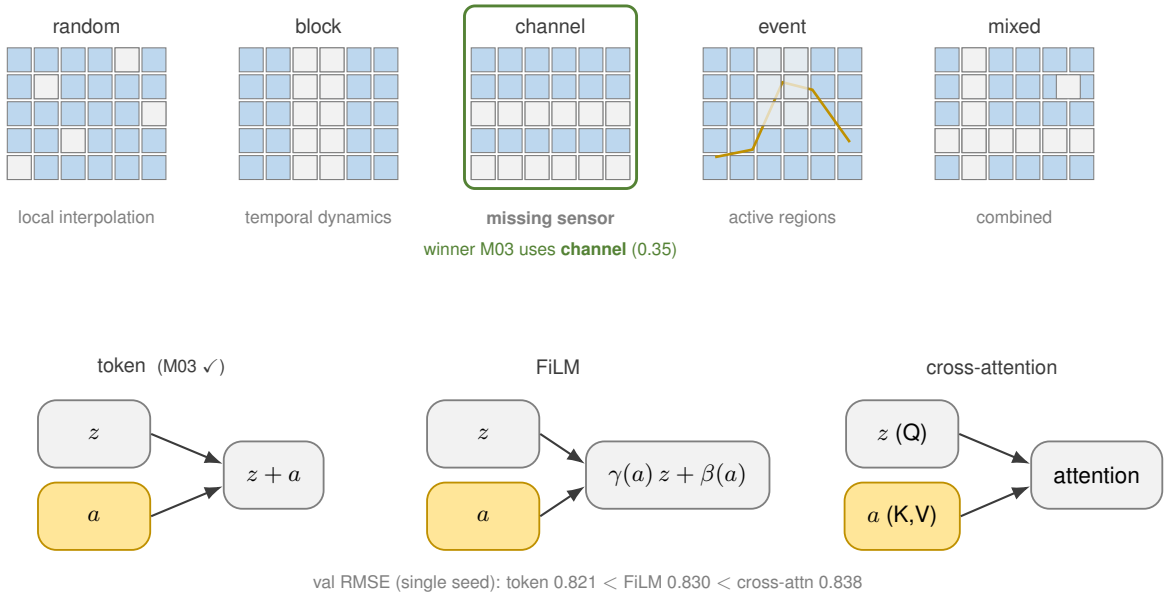
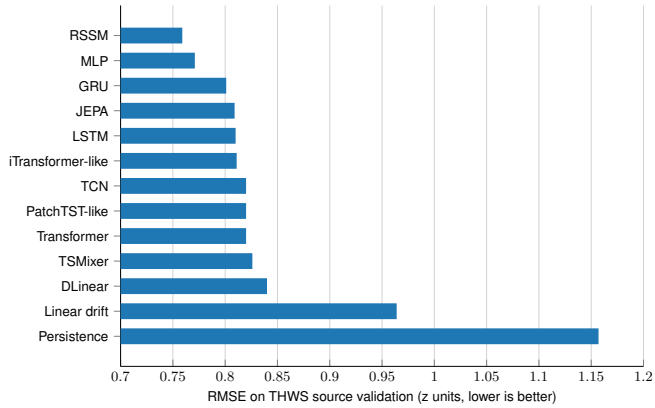



Figure 5. **Masking regimes and action-injection variants spanned by the search.** Top: the five training-time corruption modes (random points, temporal blocks, whole channels, high-activity events, and their union); the locked candidate uses channel masking, which removes entire sensors from the context and mimics a channel absent on the target machine. Bottom: the three action-injection mechanisms of the predictor. Single-seed runs on the pre-search base configuration; the token entry is the one reported in Table 2b: token 0.821, FiLM 0.830, cross-attention 0.838.

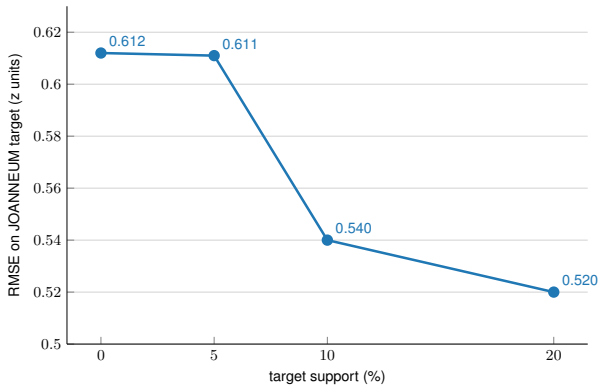
### 6.3. Source-only model selection identifies a schema-consistent action-recovery variant as the locked configuration

The architecture search separates useful design choices from failure modes. The plain JEPA control, which removes anti-collapse regularization, collapses on 4/4 seeds and is excluded from ranking. Large-model scaling is also unstable in this data regime: the largest candidate ($d$=512, eight layers) collapses on 3/4 seeds. In contrast, the selected variant combines latent schema consistency and action recovery while preserving clean-source performance. Its source-validation score is $0.8216 \pm 0.0089$ RMSE with a schema-drop score among the strongest of the valid candidates (0.918) and clear action sensitivity (1.058 under action shuffling); its selection reflects the composite score rather than any single metric.

The stability rule is consequential. The per-horizon VICReg variant ranks first at the four-seed discovery stage, but the schema-consistency variant with action recovery overtakes it after five and then seven seeds. Locking on the initial discovery result would therefore have selected the wrong configuration. Figure 4 summarizes the sealed selection pipeline that keeps the JOANNEUM target data completely outside the architecture search.




(a) Clean-source comparison on THWS validation. "-like" bars are in-house adaptations; the JEPA bar is the pre-search reference model on the source test split, all other bars are source-validation values.




(b) THWS→JOANNEUM few-shot adaptation of the pre-lock model. The 5% point is effectively zero-shot because the support subset holds no complete 48-step window.

Figure 6. **Source forecasting benchmark (a) and pre-lock few-shot transfer (b).**

### 6.4. Strong forecasting baselines remain competitive on the source machine

The clean-source forecasting benchmark is summarized in Fig. 6a. RSSM obtains RMSE 0.759 and MLP 0.771, outperforming the pre-search JEPA reference model at 0.809 (source test split) and the locked candidate at 0.822 (source validation). GRU, LSTM, and iTransformer-like models are also competitive. All learned models outperform linear drift and persistence. The result rules out a generic source-forecasting superiority claim and motivates evaluation on cross-machine adaptation and partial sensor overlap instead.

### 6.5. The locked model transfers zero-shot, but official zero-shot forecasters remain stronger in RMSE

The confirmatory JOANNEUM evaluation is executed once, after locking the schema + action-recovery configuration on source data only. The locked model reaches zero-shot RMSE 0.546, $R^2 = 0.012$, and NLL 0.52 on the target machine, improving over persistence at 0.654 (Table 3). This establishes that the transferred model beats a trivial baseline on an unseen machine despite the reduction from 17 to 10 available channels. The evaluated checkpoint is the single best-composite seed among the seven source-validation seeds, so the target number is a single-seed result. Per horizon, target $R^2$ is 0.035, 0.055, 0.052, $-0.023$ and $-0.059$ at $h = 1, 2, 4, 8, 16$: the transferred model explains a small share of variance at short horizons and falls below the pooled target-mean predictor at 8 and 16 s.

However, official PatchTST and iTransformer baselines with RevIN [7] achieve lower zero-shot RMSE, 0.503 and 0.498, respectively. The claim must therefore remain specific: the present advantage of SAAC-JEPA lies in action-conditioned latent dynamics, probabilistic outputs, and adaptation readiness rather than in state-of-the-art zero-shot RMSE. Section 6.6 shows that this margin disappears once the same instance normalization is applied inside SAAC-JEPA, at a calibration cost.

Only the target columns of Table 3 are a like-for-like comparison: the same 2,457 target windows, the same 10 shared channels, and no target-side selection for any method; the source column mixes the official baselines' source test split with SAAC-JEPA's source validation, as the caption states, and the RevIN row is diagnostic rather than confirmatory.

### 6.6. Instance normalization closes the zero-shot RMSE gap but breaks target calibration

The post-lock paired ablation of Sec. 5.2 tests whether the gap to the official baselines is a normalization effect. On source validation the RevIN variant wins on all three seeds for every RMSE component: clean RMSE $0.819 \pm 0.010 \rightarrow 0.766 \pm 0.001$, schema-drop $0.918 \pm 0.005 \rightarrow 0.887 \pm 0.003$, masked $0.827 \pm 0.010 \rightarrow 0.786 \pm 0.002$ and long-horizon $0.868 \pm 0.007 \rightarrow 0.849 \pm 0.004$, with bootstrap 95% intervals of the paired difference excluding zero (App. H). The calibration penalty falls from 0.124 to 0.035, action sensitivity rises from 1.06 to 1.18, and the seed-to-seed spread shrinks by a factor of five to ten. With three paired seeds the exact sign-flip test cannot go below $p = 0.125$, so these are read as consistent paired wins, not as significance claims.

On the target machine the three control seeds give RMSE $0.555 \pm 0.015$ (the locked seed is 0.546) and the three RevIN seeds $0.495 \pm 0.004$; the paired differences are $-0.046$, $-0.079$ and $-0.054$, and $R^2$ moves from $-0.02$ to $+0.19$ (Table 3). The variant is therefore level with official PatchTST (0.503) and iTransformer (0.498), which were run once each without hyperparameter search, so no state-of-the-art claim follows. The gain concentrates on the

Table 3. Confirmatory JOANNEUM zero-shot transfer after source-only model locking, and the post-lock RevIN ablation. Source column: official baselines are scored on the source test split (5,189 windows, 17 sensors); the SAAC-JEPA rows are the seven-seed (locked) and three-seed (RevIN) mean ± SD on source validation. Target columns: 10 shared channels, 2,457 windows. The locked row is the single declared target pass; the RevIN row is a second, declared read of all three seeds with no selection on the target (Sec. 5.2). Lowest target RMSE in bold.

| Method | Source RMSE | Target zero-shot RMSE | Target NLL | Notes |
|---|---|---|---|---|
| Persistence | — | 0.654 | — | trivial floor |
| PatchTST (official, RevIN) | 0.804 | 0.503 | — | deterministic, single seed |
| iTransformer (official, RevIN) | 0.822 | 0.498 | — | deterministic, single seed |
| SAAC-JEPA (locked) | $0.822 \pm 0.009$ | 0.546 | 0.52 | probabilistic, action-cond.; single declared pass |
| SAAC-JEPA + RevIN (post-lock) | $0.766 \pm 0.001$ | $\mathbf{0.495 \pm 0.004}$ | 20.6 | probabilistic, action-cond.; 3 seeds, diagnostic |

channels whose level shifts most across machines (Z-axis current RMSE 0.84 to 0.20), while spindle current and torque degrade slightly (0.37 to 0.43 and 0.60 to 0.67).

The price is calibration. Target NLL rises from 0.89 to 20.6 and empirical coverage at the nominal 90% level falls from 0.953 to 0.874. On seed 0, 11.5% of the RevIN predictions sit at the log-variance floor of the head ($-8$, against 0% for the control), and the per-channel NLL is dominated by spindle torque (89 nats), Z power (51), spindle current (40) and spindle power (28). The mechanism is the degenerate case of instance normalization on stationary windows: when the spindle is stopped or a power channel is constant over the 32-step context, the window scale is $s_k = \sqrt{\epsilon} \approx 3 \times 10^{-3}$, the term $2 \log s_k$ then pushes the de-normalized log-variance far below any physical value, and a start-up inside the 16-step horizon costs hundreds of nats. RMSE is protected because the mean is de-normalized with the same scale. A latent-health gate also flags the three RevIN runs on its raw-scale criterion (minimum predicted-latent standard deviation 0.040–0.041 against a threshold of 0.05; control 0.059–0.061), whereas the scale-free criterion, effective rank, is twice as high with RevIN (0.45–0.49 against 0.22–0.23) and latent health after fine-tuning is identical; the gate was left unchanged and must become scale-free before RevIN enters any future selection.

The reading for world models is direct. The zero-shot RMSE advantage of the official forecasters was an input-normalization effect, not evidence against the action-conditioned latent objective, since the same objective with the same normalization matches them while keeping actions, uncertainty and adaptation. But pure RevIN, as the baselines use it, is incompatible with a Gaussian head under stationary regimes, which is precisely where an industrial world model must remain uncertainty-aware. The locked model is therefore kept as the reference result and the variant is reported with its cost.

### 6.7. Few-shot adaptation remains promising but is still diagnostic

A pre-lock adaptation sweep suggests that limited target support can further improve transfer (Figure 6b). On the evaluated pre-lock model, zero-shot transfer yields RMSE 0.612, MAE 0.401, and $R^2 = -0.244$. With 10% target support, RMSE decreases to 0.540 and $R^2$ becomes positive (0.046). With 20% support, RMSE reaches 0.520, MAE 0.299, and $R^2 = 0.167$. The 5% support setting does not contain a valid 48-step support window and therefore reduces to the zero-shot model rather than constituting an adaptation failure.

## 7. Discussion

The evidence supports a narrow interpretation. First, JEPA pretraining does not improve clean-source forecasting over matched scratch training. Second, anti-collapse control and source-only selection stability are not minor implementation details: the plain JEPA control collapses completely, large scaling is unstable, and the provisional winner changes during confirmation. Third, cross-machine evaluation provides a distinct axis from in-domain forecasting. The locked schema + action-recovery model transfers zero-shot and outperforms persistence on an unseen machine, even though stronger zero-shot RMSE is currently achieved by official RevIN-equipped forecasting baselines. Fourth, that RMSE margin is an input-normalization effect: the same architecture with RevIN matches the baselines but loses its calibration on stationary windows, so normalization and uncertainty must be designed together rather than borrowed from point forecasters.

Industrial predictive representations should therefore be judged on whether a learned state is controllable, uncertainty-aware, robust to schema shift, and adaptable under limited target data, not on source interpolation accuracy alone.

**Future work.** V2 should (i) train on more than one source machine so that invariance to machine-specific dynamics is measured rather than assumed; (ii) make instance normal-

ization compatible with the probabilistic head, by flooring the window scale in z-units, blending to the identity below a context-variance threshold, or keeping pure RevIN for the mean while predicting the variance in global space, each of which departs from the baselines' pure RevIN and must be reported as such, and make the latent-health gate scale-free before the twenty-candidate search is repeated with normalization enabled; (iii) connect the latent state to planning or model-based RL so it is judged by decision quality; and (iv) evaluate policies off-line before any real-machine test. Causal representation learning stays outside the claim set until the data provide enough intervention diversity (Fig. 7, App. F).

## 8. Limitations

One source and one target machine limit claims of general cross-machine generalization, and the 7 independent JOANNEUM runs restrict statistical power. The target stream is interpreted under a 2 ms controller-cycle assumption, and sensitivity to resampling remains a separate robustness analysis; seven source-only channels have no target counterpart. The few-shot curve was obtained before the model lock and must be re-measured on the locked configuration before being treated as confirmatory. The confirmatory target result is a single checkpoint evaluated once, and the target windows were then read a second time by the declared RevIN ablation, so no further target evaluation of this pipeline can be called sealed; that ablation has three paired seeds, its runs carry a raw-scale latent-health flag, and its calibration collapse is diagnosed but not fixed. Run-level confidence intervals and action-sensitivity tests on the target machine have not yet been computed for the locked model. On the pre-lock model, zeroing or shuffling future actions changed target RMSE by less than 0.005, so action use under transfer is unverified and is the first item to test; that model was also under-calibrated (67% empirical coverage at nominal 90%). The ablations in Table 2b are single-seed, and the official baselines are single-seed runs without hyperparameter search. Finally, the present paper does not yet evaluate downstream policy learning or off-policy control, so the world-model interpretation remains predictive rather than decision-theoretic.

## 9. Conclusion

This study presents a schema-adaptive, action-conditioned JEPA with an audited cross-machine CNC protocol under partial sensor overlap. JEPA pretraining does not improve clean-source RMSE over matched scratch training; source-only search still identifies a stable schema-consistent, action-recovery configuration that transfers zero-shot to an unseen machine above persistence, while official RevIN-based forecasters remain stronger in raw zero-shot RMSE; a post-lock paired ablation attributes that margin to instance normalization, which the same architecture recovers at the cost of target calibration. The contribution is a controlled benchmark and methodology for industrial predictive representations under joint machine shift, schema shift, and action conditioning, not a state-of-the-art claim.

## Code and data availability

Code, configuration files, the twenty candidate specifications, and the audit scripts are available at https://github.com/ostertagmatthieu-dev/saac-jepa; animated versions of the schematics are on the project page, https://ostertagmatthieu-dev.github.io/saac-jepa/. The source and target datasets are public under CC BY 4.0: THWS five-axis CNC milling (https://doi.org/10.5281/zenodo.14094887) and the FH JOANNEUM CNC machining repository (https://doi.org/10.17632/gtvvwmz7r7.2).

## References


[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629, 2023. 1, 2

[2] Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. arXiv:2404.08471, 2024. 1, 2

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. 1, 2

[4] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. SimMTM: A simple pre-training framework for masked time-series modeling. In *NeurIPS*, 2023. 2

[5] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023. 2

[6] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024. 1, 2

[7] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *ICLR*, 2022. 2, 6, 10

[8] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, pages 2555–2565, 2019. 2

[9] Mario Martinez, Anna-Maria Schmitt, Andreas Schiffler, and Bastian Engelmann. Production data set for five-axis CNC milling with multiple changeovers. *Scientific Data*, 12:1067, 2025. doi:10.1038/s41597-025-05294-0. Dataset: https://doi.org/10.5281/zenodo.14094887. 1, 6

[10] Markus Brillinger, Muaaz Abdul Hadi, Stefan Trabesinger, Johannes Schmid, and Florian Lackner. CNC machining data repository: Geometry, NC code & high-frequency energy consumption data for aluminum and plastic machining. *Data in Brief*, 61:111814, 2025. Dataset: https://doi.org/10.17632/gtvvwmz7r7.2. 1, 6

[11] Peng Yan, Ahmed Abdulkadir, Gerrit A. Schatte, Giulia Aguzzi, Joonsu Gha, Nikola Pascher, Matthias Rosenthal, Yunlong Gao, Benjamin F. Grewe, and Thilo Stadelmann. Learning actionable world models for industrial process control. arXiv:2503.01411, 2025. 1, 2

# Schema-Adaptive Action-Conditioned JEPA for Cross-Machine CNC Transfer under Partial Sensor Overlap

Supplementary Material

## A. Architecture Search: Twenty Candidates

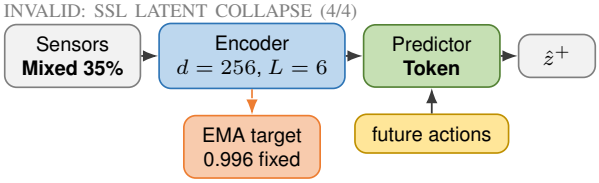
The architecture search contains 20 pre-defined SAAC-JEPA variants. Candidate ranking used **THWS source validation only**; the JOANNEUM target data were excluded from candidate ranking. The cards below show the common computational skeleton and highlight the mechanism changed by each candidate. Metrics are means over valid seeds from the architecture-search log. The final stability check reported in the closure log selected **the schema-consistency variant with action recovery** after seven valid seeds for the two leading candidates, after which code, configuration, normalizers, and checkpoint were frozen before the confirmatory target pass.

**Reading the cards.** Each diagram follows *masked sensor context* → *context encoder* → *action-conditioned predictor* → *future latent*, with an EMA target branch used for latent self-supervision. Tags report the masking regime, action-injection mechanism, latent objective, schema/action auxiliary losses, EMA schedule, capacity, normalization, and fine-tuning head initialization. Action sensitivity is RMSE(shuffled actions)/RMSE(true actions); values above one indicate measurable action usage.

### Plain JEPA control

COLLAPSE 4/4

INVALID: SSL LATENT COLLAPSE (4/4)




mask: Mixed 0.35 | action: Token | VICReg: None | place: -

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.0 | $\lambda_{act}$=0.0 | target: Per-horizon | latent LN: off

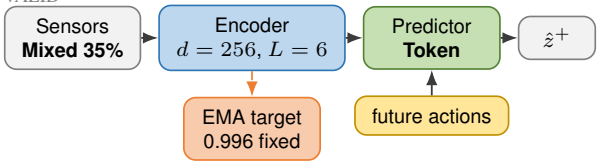
EMA: 0.996 fixed | norm: z-score | head: Fresh **Source validation:** No valid source metrics; candidate excluded before ranking.

**Selection:** Discovery exclu; confirmation exclu; invalid: ssl_latent_collapse.

### JEPA + per-horizon VICReg

7/7 seeds

VALID




mask: Mixed 0.35 | action: Token | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.0 | $\lambda_{act}$=0.0 | target: Per-horizon | latent LN: off

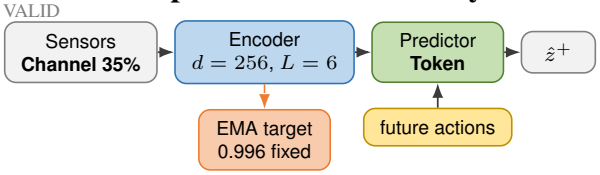
EMA: 0.996 fixed | norm: z-score | head: Fresh **Source validation:** Clean 0.8336 ± 0.0033 | schema 0.9494 | action 1.0350 | eff.rank 0.3189

**Selection:** Discovery #1 (-0.346); confirmation #2 (0.093).

### Schema-adaptive latent consistency

4/4 seeds

VALID




mask: Channel 0.35 | action: Token | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.0 | target: Per-horizon | latent LN: off

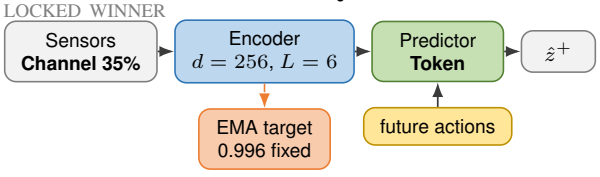
EMA: 0.996 fixed | norm: z-score | head: Fresh **Source validation:** Clean 0.827 ± 0.0058 | schema 0.9252 | action 1.0437 | eff.rank 0.3792

**Selection:** Discovery #11 (-0.272).

### Schema + action recovery

LOCKED

LOCKED WINNER




mask: Channel 0.35 | action: Token | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off

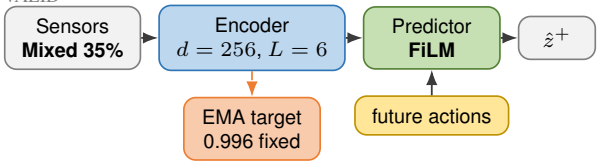
EMA: 0.996 fixed | norm: z-score | head: Fresh **Source validation:** Clean 0.8216 ± 0.0089 | schema 0.9183 | action 1.0583 | eff.rank 0.4273

**Selection:** Discovery #5 (-0.296); confirmation #1 (-0.830).

### FiLM action injection

5/5 seeds

VALID




mask: Mixed 0.35 | action: FiLM | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off

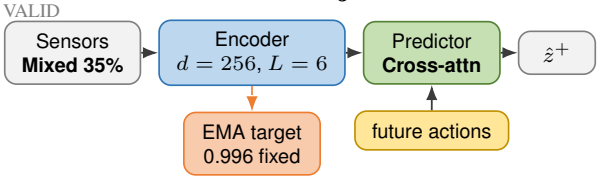
EMA: 0.996 fixed | norm: z-score | head: Fresh **Source validation:** Clean 0.8346 ± 0.0067 | schema 0.952 | action 1.0346 | eff.rank 0.3289

**Selection:** Discovery #3 (-0.318); confirmation #4 (0.149).

## Cross-attention action injection

5/5 seeds

VALID




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

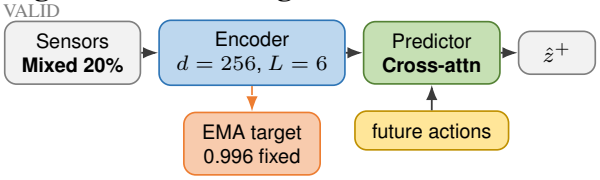
**Source validation:** Clean 0.8333 ± 0.0034 | schema 0.9488 | action 1.0336 | eff.rank 0.3149

**Selection:** Discovery #4 (-0.3); confirmation #3 (0.142).

## Light mixed masking

4/4 seeds

VALID




mask: Mixed 0.2 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

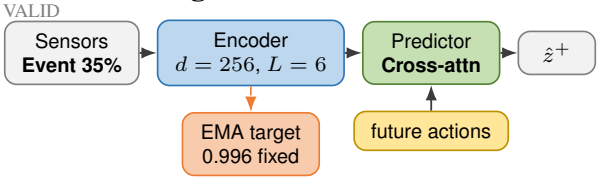
**Source validation:** Clean 0.8269 ± 0.0102 | schema 0.9432 | action 1.047 | eff.rank 0.3687

**Selection:** Discovery #9 (-0.28).

## Event masking

4/4 seeds

VALID




mask: Event 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

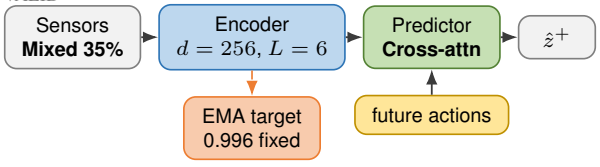
**Source validation:** Clean 0.8505 ± 0.0077 | schema 0.9567 | action 1.0295 | eff.rank 0.3278

**Selection:** Discovery #6 (-0.291).

## Latent LayerNorm ON

4/4 seeds

VALID




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: ON | EMA: 0.996 fixed | norm: z-score | head: Fresh

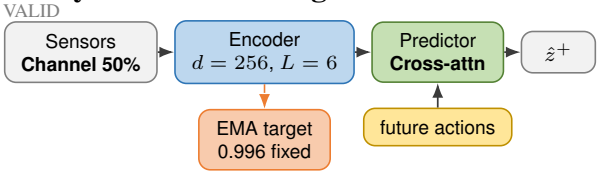
**Source validation:** Clean 0.8143 ± 0.0164 | schema 0.9545 | action 1.074 | eff.rank 0.3616

**Selection:** Discovery #10 (-0.273).

## Heavy channel masking

4/4 seeds

VALID




mask: Channel 0.5 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.2 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

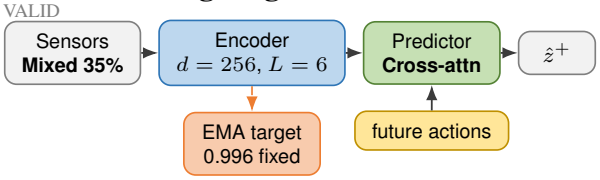
**Source validation:** Clean 0.8213 ± 0.0075 | schema 0.9287 | action 1.0549 | eff.rank 0.416

**Selection:** Discovery #13 (-0.232).

## Pooled VICReg/target

4/4 seeds

VALID




mask: Mixed 0.35 | action: Cross-attn | VICReg: Pooled | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Pooled | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

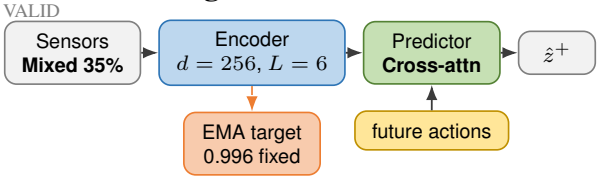
**Source validation:** Clean 0.8139 ± 0.0117 | schema 0.9228 | action 1.079 | eff.rank 0.4005

**Selection:** Discovery #14 (-0.229).

## Low latent weight

4/4 seeds

VALID




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=0.5 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

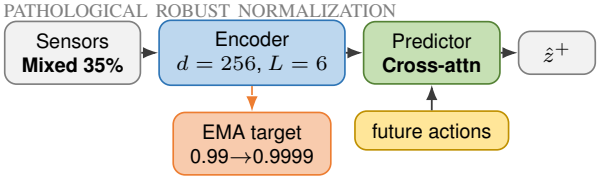
**Source validation:** Clean 0.8019 ± 0.0088 | schema 0.9065 | action 1.0925 | eff.rank 0.4145

**Selection:** Discovery #12 (-0.237).

## Robust source normalization

IQR PATHOLOGY

PATHOLOGICAL ROBUST NORMALIZATION




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.99→0.9999 | norm: Robust | head: Fresh

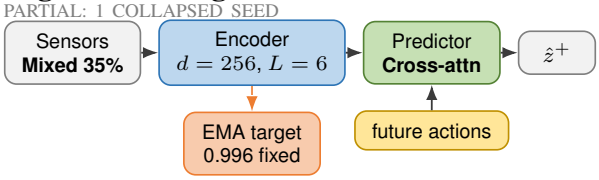
**Source validation:** Clean 0.9813 ± 0.0360 | schema 0.9814 | action 1.0088 | eff.rank 0.0213

**Selection:** Discovery #17 (6.25).

## High latent weight

1 COLLAPSE

PARTIAL: 1 COLLAPSED SEED




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=2.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.996 fixed | norm: z-score | head: Fresh

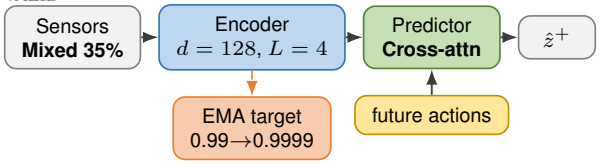
**Source validation:** Clean 0.8438 ± 0.0015 | schema 0.9499 | action 1.0214 | eff.rank 0.2906

**Selection:** Discovery #2 (-0.323); confirmation #5 (1.177); invalid: ssl_latent_collapse.

## Compact model

4/4 seeds

VALID




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.99→0.9999 | norm: z-score | head: Fresh

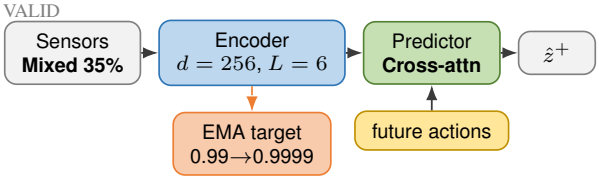
**Source validation:** Clean 0.8485 ± 0.0006 | schema 0.9439 | action 1.0119 | eff.rank 0.3485

**Selection:** Discovery #16 (0.462).

## Scheduled EMA

4/4 seeds

VALID




mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.99→0.9999 | norm: z-score | head: Fresh

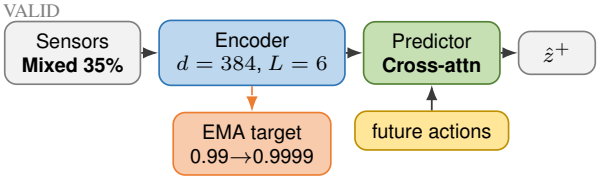
**Source validation:** Clean 0.8345 ± 0.003 | schema 0.9574 | action 1.0311 | eff.rank 0.3002

**Selection:** Discovery #7 (-0.291).

## Medium-wide model

4/4 seeds

VALID

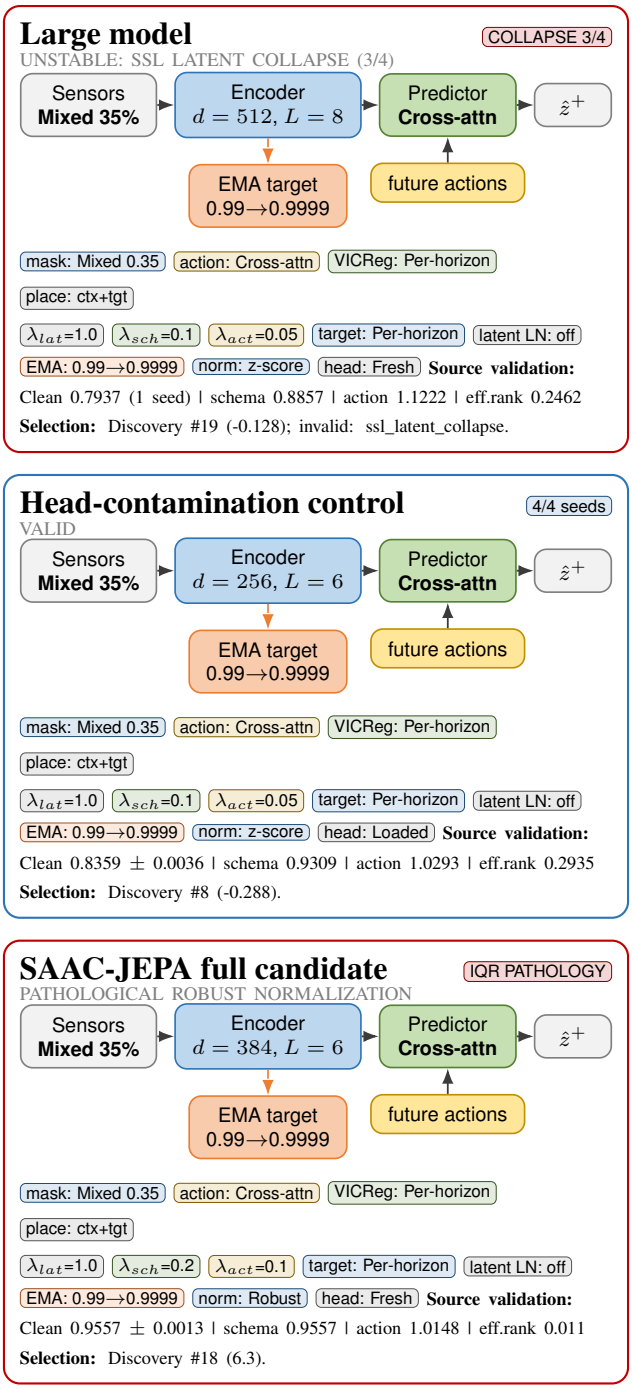



mask: Mixed 0.35 | action: Cross-attn | VICReg: Per-horizon | place: ctx+tgt

$\lambda_{lat}$=1.0 | $\lambda_{sch}$=0.1 | $\lambda_{act}$=0.05 | target: Per-horizon | latent LN: off | EMA: 0.99→0.9999 | norm: z-score | head: Fresh

**Source validation:** Clean 0.8148 ± 0.0016 | schema 0.9325 | action 1.0758 | eff.rank 0.305

**Selection:** Discovery #15 (-0.084).




## B. Complete Candidate Configuration Matrix

Table 4 lists the exact design factors varied across the twenty candidates. "ctx+tgt" denotes VICReg placement on context and EMA-target representations. All candidates use the same source-only data splits and evaluation protocol; only the listed architectural or training factors vary.

## C. Source-Validation Search Outcomes

The table below reports the source-only architecture-search outcomes. Composite scores are z-normalized within their respective selection pools and therefore should not be compared across discovery and confirmation columns. The plain control was excluded because all four seeds collapsed. The large model collapsed in three of four discovery seeds. The two robust-normalization candidates exhibited the pathology associated with near-zero IQR channels. The final closure-stage stability check selected the schema-consistency variant with action recovery after seven valid seeds for the leading pair.

## D. Architecture-Search Interpretation

**Anti-collapse is necessary.** The plain control removes VICReg and collapses for all four discovery seeds. This matches the corrected pretraining audit, in which target-encoder effective rank increased from approximately 5% to 58% after anti-collapse regularization was moved to the appropriate latent representations.

**Capacity scaling is not monotonic.** The large model ($d = 512$, eight layers) collapses in three of four seeds, whereas the locked candidate uses $d = 256$ and six layers. The search therefore does not support a "larger is better" interpretation at the available source-data scale.

**Robust normalization requires channel-specific safeguards.** The two robust-normalization candidates produce pathological errors because near-constant source channels can have an interquartile range close to zero. These candidates are retained in the appendix because they document a real failure mode of the search space rather than an omitted negative result.

**The locked candidate reflects the intended inductive bias.** The locked candidate combines channel masking, latent schema consistency, explicit action recovery, per-horizon VICReg on context and target representations, and a fresh downstream head. Its source-validation advantage is not defined by the lowest clean RMSE alone; the selection score also rewards schema-drop robustness, masked robustness, long-horizon performance, calibration, SSL validation, and action use.

## E. Official Forecasting Baseline Protocol

The official PatchTST (`yuqinie98/PatchTST`, commit `204c21e`, supervised tree) and iTransformer (`thuml/iTransformer`, commit `c2426e6`) repositories were run through a shared dataset adapter so that both consume the same session-bounded windows as SAAC-JEPA: context 32 steps, prediction length 16, scored at horizons $\{1, 2, 4, 8, 16\}$, 1 Hz grid, windows never crossing a session boundary (18,267 / 3,423 / 5,189 source windows and 2,457 target windows, identical to the SAAC-JEPA counts). Inputs are the 17 canonical sensors plus the 4 action channels (21 channels, all forecast jointly); RMSE is reported on the 17 sensors for the source test split and on the 10 channels shared with the target for the sealed target pass. Values are z-scored with the source-train normalizer of the main pipeline; the 7 sensors absent on the target are zero-imputed in normalized space and excluded from scoring. Both baselines were trained once (single seed 20260820, 100 epochs, early-stopping patience 15, batch 128, MSE loss, RevIN enabled without affine parameters, other hyperparameters at repository defaults) with no hyperparameter search; the same RevIN settings are reused for the SAAC-JEPA variant of App. H. The official SimMTM repository (commit `169513b`) diverged to NaN at two learning rates under the same protocol and is therefore not reported. Persistence on

Table 4. Complete architecture-search configuration.

| Candidate | Mask | Act. | VICReg | $\lambda_{lat}$ | $\lambda_{sch}$ | $\lambda_{act}$ | EMA | $d/L$ | Norm | LN | FT head |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plain control | mixed/0.35 | Token | none | 1.0 | 0.0 | 0.0 | 0.996 | 256/6 | zscore | off | fresh |
| Per-horizon VICReg | mixed/0.35 | Token | per-h | 1.0 | 0.0 | 0.0 | 0.996 | 256/6 | zscore | off | fresh |
| Schema consistency | channel/0.35 | Token | per-h | 1.0 | 0.1 | 0.0 | 0.996 | 256/6 | zscore | off | fresh |
| Schema + action rec. | channel/0.35 | Token | per-h | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| FiLM action | mixed/0.35 | FiLM | per-h | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Cross-attn action | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Event masking | event/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Heavy channel mask | channel/0.5 | Cross-attn | per-h | 1.0 | 0.2 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Light mixed mask | mixed/0.2 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Latent LayerNorm | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | on | fresh |
| Pooled VICReg | mixed/0.35 | Cross-attn | pool | 1.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Low latent weight | mixed/0.35 | Cross-attn | per-h | 0.5 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| High latent weight | mixed/0.35 | Cross-attn | per-h | 2.0 | 0.1 | 0.05 | 0.996 | 256/6 | zscore | off | fresh |
| Scheduled EMA | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | sched. | 256/6 | zscore | off | fresh |
| Robust normalization | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | sched. | 256/6 | robust | off | fresh |
| Compact model | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | sched. | 128/4 | zscore | off | fresh |
| Medium-wide model | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | sched. | 384/6 | zscore | off | fresh |
| Large model | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | sched. | 512/8 | zscore | off | fresh |
| Loaded-head control | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.1 | 0.05 | sched. | 256/6 | zscore | off | loaded |
| Full candidate | mixed/0.35 | Cross-attn | per-h | 1.0 | 0.2 | 0.1 | sched. | 384/6 | robust | off | fresh |

Table 5. Source-validation outcomes for the twenty candidates. Lower RMSE/schema-drop is better; action sensitivity greater than one indicates that shuffled future actions degrade prediction.

| Candidate | Seeds | Clean RMSE | Schema | Masked | Long | Calib. | Action sens. | Eff. rank | Discovery / confirm. |
|---|---|---|---|---|---|---|---|---|---|
| Plain control | 0/4 | – | – | – | – | – | – | – | excluded / excluded |
| Per-horizon VICReg | 7/7 | $0.8336 \pm 0.0033$ | 0.9494 | 0.8411 | 0.8797 | 0.1035 | 1.0350 | 0.3189 | #1 (-0.346) / #2 (0.093) |
| Schema consistency | 4/4 | $0.827 \pm 0.0058$ | 0.9252 | 0.8344 | 0.8747 | 0.1245 | 1.0437 | 0.3792 | #11 (-0.272) / – |
| Schema + action rec. | 7/7 | $0.8216 \pm 0.0089$ | 0.9183 | 0.8298 | 0.8709 | 0.1315 | 1.0583 | 0.4273 | #5 (-0.296) / #1 (-0.830) |
| FiLM action | 5/5 | $0.8346 \pm 0.0067$ | 0.952 | 0.8412 | 0.8793 | 0.0957 | 1.0346 | 0.3289 | #3 (-0.318) / #4 (0.149) |
| Cross-attn action | 5/5 | $0.8333 \pm 0.0034$ | 0.9488 | 0.8415 | 0.8786 | 0.1003 | 1.0336 | 0.3149 | #4 (-0.3) / #3 (0.142) |
| Event masking | 4/4 | $0.8505 \pm 0.0077$ | 0.9567 | 0.8492 | 0.8976 | 0.125 | 1.0295 | 0.3278 | #6 (-0.291) / – |
| Heavy channel mask | 4/4 | $0.8213 \pm 0.0075$ | 0.9287 | 0.8297 | 0.8713 | 0.1407 | 1.0549 | 0.416 | #13 (-0.232) / – |
| Light mixed mask | 4/4 | $0.8269 \pm 0.0102$ | 0.9432 | 0.8364 | 0.8752 | 0.1171 | 1.047 | 0.3687 | #9 (-0.28) / – |
| Latent LayerNorm | 4/4 | $0.8143 \pm 0.0164$ | 0.9545 | 0.8216 | 0.866 | 0.1268 | 1.074 | 0.3616 | #10 (-0.273) / – |
| Pooled VICReg | 4/4 | $0.8139 \pm 0.0117$ | 0.9228 | 0.8221 | 0.8657 | 0.1435 | 1.079 | 0.4005 | #14 (-0.229) / – |
| Low latent weight | 4/4 | $0.8019 \pm 0.0088$ | 0.9065 | 0.8127 | 0.8535 | 0.1421 | 1.0925 | 0.4145 | #12 (-0.237) / – |
| High latent weight | 4/5 | $0.8438 \pm 0.0015$ | 0.9499 | 0.8521 | 0.8865 | 0.0976 | 1.0214 | 0.2906 | #2 (-0.323) / #5 (1.177) |
| Scheduled EMA | 4/4 | $0.8345 \pm 0.003$ | 0.9574 | 0.842 | 0.8804 | 0.0807 | 1.0311 | 0.3002 | #7 (-0.291) / – |
| Robust normalization | 4/4 | $0.9813 \pm 0.0360$ | 0.9814 | 0.9816 | 1.0126 | 0.0399 | 1.0088 | 0.0213 | #17 (6.25) / – |
| Compact model | 4/4 | $0.8485 \pm 0.0006$ | 0.9439 | 0.856 | 0.8883 | 0.0666 | 1.0119 | 0.3485 | #16 (0.462) / – |
| Medium-wide model | 4/4 | $0.8148 \pm 0.0016$ | 0.9325 | 0.8237 | 0.8686 | 0.1147 | 1.0758 | 0.305 | #15 (-0.084) / – |
| Large model | 1/4 | 0.7937 (1 seed) | 0.8857 | 0.8018 | 0.8558 | 0.1156 | 1.1222 | 0.2462 | #19 (-0.128) / – |
| Loaded-head control | 4/4 | $0.8359 \pm 0.0036$ | 0.9309 | 0.8416 | 0.8818 | 0.0818 | 1.0293 | 0.2935 | #8 (-0.288) / – |
| Full candidate | 4/4 | $0.9557 \pm 0.0013$ | 0.9557 | 0.9558 | 0.9866 | 0.0075 | 1.0148 | 0.011 | #18 (6.3) / – |

the same target windows gives RMSE 0.654; on the source test split it gives 1.128 (17 sensors). The official baselines forecast the action channels poorly (target RMSE 2.01 and 2.27), which is expected because commanded actions are exogenous inputs rather

than process outcomes.

## F. Concept of Study and V2 Roadmap

The overall concept of the study is a single pipeline from historical machine data to a deployable control decision:

1. **Source machine**: real production data (sensors and commanded actions) from the source CNC machine.
2. **SAAC-JEPA**: schema-adaptive, action-conditioned JEPA pretraining on the source data.
3. **Compact representation of the dynamics**: a latent state that summarizes the machine's physical evolution.
4. **Transfer to the target machine**: zero-shot use of the source-trained representation under partial sensor overlap.
5. **Adaptation with a few target samples**: few-shot or online updates using limited data from the new machine.
6. **World model**: the adapted latent dynamics used as a simulator of the target machine.
7. **Simulation of candidate actions**: rolling out alternative command sequences in the world model.
8. **Reinforcement learning**: policy optimization inside the learned simulator.
9. **New control policy**: the resulting candidate policy for the target machine.
10. **Doubly robust evaluation on historical data**: off-policy evaluation of the candidate policy with historical logs before any physical test.
11. **Decision policy**: the evaluated policy that is retained for deployment or decision support.

The present V1 paper covers steps 1–5 (steps 4 and 5 as zero-shot and pre-lock few-shot evidence). Figure 7 shows how V2 extends this backbone.

## G. Locked Configuration

Table 6 lists the hyperparameters of the locked candidate (M03, "schema + action recovery"), the configuration frozen with SHA-256 hashes before the single confirmatory pass on the target machine. Values from `configs/search_v2/M03.yaml` in the code repository; the RevIN variant is `configs/revin/M03_revin.yaml`, identical except for the flag.

## H. Paired RevIN Ablation

This appendix reports the post-lock ablation of Sec. 6.6 in full. Both arms were trained with `scripts/68_revin_ablation.py`, which runs the two-stage pipeline of `scripts/41` (self-supervised pretraining, fresh-head fine-tuning, common-space validation bundle) once per seed and arm: the control arm is `configs/search_v2/M03.yaml` verbatim, the variant arm is the same file with `model.revin.enabled` set (App. G). The three control re-runs reproduce the V1 validation metrics of seeds 0–2 exactly. Six runs of about 72 minutes each were executed sequentially on the GB10. Eighteen regression tests cover the transform (equivariance of the de-normalized mean under affine changes of the input, insensitivity of the statistics to masked entries, round-trip identity, identity on unseen channels, bitwise equality of the disabled path with the V1 code, and loading of V1 checkpoints).

**Source validation.** Table 7 lists the nine validation components of the composite score in the common evaluation space, with the paired difference, its bootstrap 95% interval over the three paired seeds, and the per-seed win count. The anchor row is the trivial-predictor RMSE of the evaluation space and is identical on both arms, which confirms that de-normalization returns to the same metric space. With $n = 3$ pairs the exact sign-flip test cannot go below $p = 0.125$; the intervals and win counts are the evidence. The latent-health gate flags the three RevIN runs as `ssl_latent_collapse` on its raw-scale sub-criterion only (Table 8); the scale-free sub-criterion is better with RevIN, latent health after fine-tuning is identical, and the gate was not modified.

**Declared second target read.** The V1 lock consumed the single declared target pass. The second opening of the JOANNEUM windows was decided before the ablation ran and conditioned on the variant winning the four RMSE components of Table 7 on source validation; that verdict was written to `summary.json` before the target stage started, and the target stage then scored all six runs with the protocol of `scripts/45` (model rebuilt from the run configuration, source-train normalizers, 2,457 windows, 10 shared channels) and no selection on the target. Table 9 lists every run. Aggregates: control RMSE $0.555 \pm 0.015$, MAE 0.357, $R^2 = -0.02$, NLL 0.89; RevIN RMSE $0.495 \pm 0.004$, MAE 0.276, $R^2 = +0.19$, NLL 20.6; paired RMSE differences $-0.046$, $-0.079$, $-0.054$. Coverage at the nominal 90% level is 0.953 for the control and 0.874 for RevIN. The per-channel diagnosis of the NLL increase (seed 0: 11.5% of predictions at the log-variance floor; spindle torque 89 nats, Z power 51, spindle current 40, spindle power 28) and its mechanism are given in Sec. 6.6.

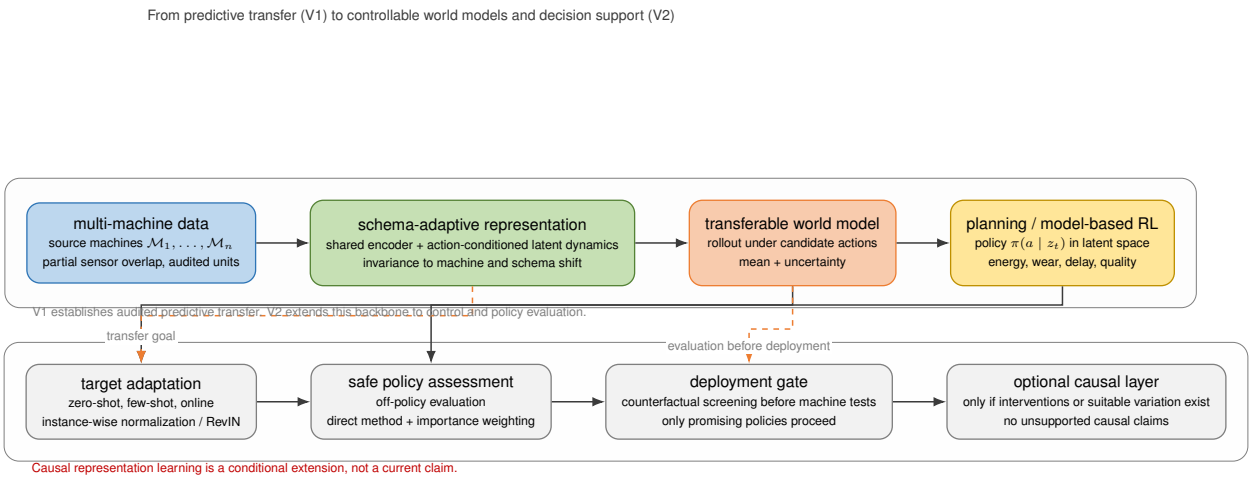



Figure 7. V1 establishes audited cross-machine predictive transfer with schema-adaptive latent dynamics. V2 extends this backbone to multi-machine representation learning, target adaptation with stronger normalization, action-conditioned planning or model-based RL, and policy assessment through off-policy evaluation before any physical deployment. Optional causal representation learning is shown as a conditional branch rather than as a current contribution.

Table 6. Locked configuration of SAAC-JEPA (candidate M03).

| Group | Hyperparameter | Value |
|---|---|---|
| Architecture | Latent dimension $d$ | 256 |
| | Attention heads | 8 |
| | Encoder layers (temporal transformer) | 6 |
| | Predictor layers | 4 |
| | Feed-forward width | 1024 |
| | Dropout | 0.1 |
| | Action injection | token |
| | Transformer normalization | pre-norm |
| Windows | Context length $K$ | 32 |
| | Horizons $\mathcal{H}$ | $\{1, 2, 4, 8, 16\}$ |
| | Normalization | source z-score (train split only) |
| Masking | Mode | channel |
| | Mask ratio `masking.ratio` $\rho$ | 0.35 |
| | Channel drop probability `channel_drop_prob` | 0.15 |
| Objectives | Latent loss | smooth $L_1$ |
| | Latent weight $\lambda_{\text{lat}}$ | 1.0 |
| | Latent LayerNorm | off |
| | VICReg mode / placement | per-horizon / context + target |
| | VICReg weight $\lambda_{\text{vic}}$ | 0.05 |
| | VICReg coefficients (sim/var/cov) | 25 / 25 / 1 (invariance coefficient unused) |
| | Target aggregation | per-horizon |
| | Schema-consistency weight $\lambda_{\text{sch}}$ | 0.10 |
| | Schema-view keep probability $\kappa$ | 0.65 (set in the trainer, not in the YAML) |
| | Action-recovery weight $\lambda_{\text{act}}$ | 0.05 |
| | Physical weight (fine-tuning stage) | 1.0 |
| Optimization | EMA momentum $\tau$ | 0.996 |
| | Epochs (per stage) | 100 |
| | Batch size | 128 |
| | Learning rate | $3 \times 10^{-4}$ |
| | Weight decay | 0.01 |
| | Gradient-norm clip | 1.0 |
| | Mixed precision | enabled |
| | Early-stopping patience | 15 |
| Protocol | Group split key | session |
| | Fine-tuning head | fresh |
| | Probabilistic head | Gaussian mean and log-variance |
| Instance norm. | RevIN `model.revin.enabled` | off (locked); on in the variant of App. H |
| | Affine parameters | none |
| | Last-value subtraction | none |
| | $\epsilon$ | $10^{-5}$ |

Table 7. Paired RevIN ablation on source validation (three seeds, mean ± SD; common evaluation space; lower is better except action sensitivity and effective rank). Δ is RevIN minus control; the interval is a bootstrap 95% interval of the paired difference.

| Component | M03 (control) | M03 + RevIN | Δ | 95% interval | RevIN wins |
|---|---|---|---|---|---|
| Clean RMSE | $0.8190 \pm 0.0103$ | $0.7663 \pm 0.0012$ | $-0.0527$ | $[-0.0640, -0.0457]$ | 3/3 |
| Schema-drop RMSE | $0.9180 \pm 0.0047$ | $0.8867 \pm 0.0034$ | $-0.0313$ | $[-0.0357, -0.0248]$ | 3/3 |
| Masked RMSE | $0.8272 \pm 0.0096$ | $0.7862 \pm 0.0023$ | $-0.0409$ | $[-0.0508, -0.0335]$ | 3/3 |
| Long-horizon RMSE | $0.8683 \pm 0.0073$ | $0.8487 \pm 0.0039$ | $-0.0196$ | $[-0.0234, -0.0167]$ | 3/3 |
| Calibration penalty $\lvert\mathrm{cov}_{90} - 0.9\rvert$ | $0.1242 \pm 0.0270$ | $0.0349 \pm 0.0022$ | $-0.0892$ | $[-0.1201, -0.0691]$ | 3/3 |
| SSL validation loss | $1.1874 \pm 0.0189$ | $1.0640 \pm 0.0214$ | $-0.1234$ | $[-0.1660, -0.0952]$ | 3/3 |
| Action sensitivity (shuffled / true) | $1.0639 \pm 0.0243$ | $1.1790 \pm 0.0013$ | $+0.1151$ | $[+0.1004, +0.1426]$ | 3/3 |
| Effective-rank fraction | $0.4266 \pm 0.0717$ | $0.4626 \pm 0.0062$ | $+0.0360$ | $[-0.0028, +0.1118]$ | 1/3 |
| Evaluation-space anchor RMSE | $0.9575 \pm 0.0000$ | $0.9575 \pm 0.0000$ | 0 | $[0, 0]$ | — |
| NLL (model space) | 3.02 | 1.49 | | | |
| Coverage at nominal 90% | 0.776 | 0.865 | | | |

Table 8. Latent-health gate on the predicted latents (`trainers.latent_health`). Thresholds: minimum standard deviation $< 0.05$ (raw latent units, input-scale dependent) or effective-rank fraction $< 0.10$ (scale-free). SSL: after pretraining; FT: after fine-tuning.

| Arm | Seed | $\mathrm{Std}_{\min}$ SSL | Rank SSL | $\mathrm{Std}_{\min}$ FT | Rank FT |
|---|---|---|---|---|---|
| M03 | 0 | 0.061 | 0.223 | 0.391 | 0.468 |
| M03 | 1 | 0.060 | 0.227 | 0.327 | 0.344 |
| M03 | 2 | 0.059 | 0.231 | 0.365 | 0.468 |
| M03 + RevIN | 0 | 0.040 | 0.488 | 0.280 | 0.467 |
| M03 + RevIN | 1 | 0.041 | 0.470 | 0.291 | 0.456 |
| M03 + RevIN | 2 | 0.041 | 0.446 | 0.294 | 0.465 |

Table 9. Second, declared target read: every run of the paired ablation on the 2,457 JOANNEUM windows (10 shared channels). Persistence and the V1 locked seed are given for reference.

| Arm | Seed | RMSE | MAE | $R^2$ | NLL |
|---|---|---|---|---|---|
| Persistence | — | 0.654 | 0.316 | $-0.420$ | — |
| M03 (V1 locked pass) | 0 | 0.546 | — | $+0.012$ | 0.52 |
| M03 | 0 | 0.5456 | 0.3524 | $+0.0118$ | 0.516 |
| M03 | 1 | 0.5716 | 0.3682 | $-0.0847$ | $-0.002$ |
| M03 | 2 | 0.5470 | 0.3496 | $+0.0067$ | 2.147 |
| M03 + RevIN | 0 | 0.4994 | 0.2796 | $+0.1720$ | 22.48 |
| M03 + RevIN | 1 | 0.4924 | 0.2722 | $+0.1950$ | 15.07 |
| M03 + RevIN | 2 | 0.4928 | 0.2753 | $+0.1937$ | 24.35 |