

A Staged GRU-Based Framework for Robust Deep Modeling of MIMO Hammerstein Systems Under Correlated Measurement Noise

Essia Ben Alaia

Jouf University

Slim Dhahri

Jouf University

Mourad Elloumi

University of Gafsa

Afrahan Alanazi

Jouf University

Sahar Almenwer

`smalmenwer@ju.edu.sa`

Jouf University

Omar Naifar

University of Sfax

Article

Keywords: MIMO Hammerstein systems, GRU, robust deep modeling, correlated measurement noise, staged learning, residual-noise modeling

Posted Date: April 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9351439/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Staged GRU-Based Framework for Robust Deep Modeling of MIMO Hammerstein Systems Under Correlated Measurement Noise

Essia Ben Alaia¹, Slim Dhahri¹, Mourad Elloumi^{2,3}, Afrah Alanazi⁴, Sahar Almenwer^{5,*},
and Omar Naifar^{6,7}

¹Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Al-Jouf, Saudi Arabia; slim.dhahri@isimsf.rnu.tn

²Faculty of Sciences of Gafsa, University of Gafsa, Gafsa, Tunisia

³Laboratory of Sciences and Technology of Automatic Control and Computer Engineering, National School of Engineering of Sfax, University of Sfax, P.O. Box 1173, Sfax 3038, Tunisia; mourad.elloumi@enis.tn

⁴Department of Information System, College of Computer and Information Sciences, Jouf University, Saudi Arabia; aoalenzzy@ju.edu.sa

⁵Department of Computer Science, College of Computer Science and Information, Jouf University, Saudi Arabia; smalmenwer@ju.edu.sa

⁶Control and Energy Management Laboratory, National School of Engineering of Sfax, University of Sfax, P.O. Box 1173, Sfax 3038, Tunisia; omar.naifar@enis.tn

⁷Higher Institute of Applied Sciences and Technology of Kairouan, University of Kairouan, Kairouan, Tunisia

Abstract

This paper presents a staged deep-learning framework for robust modeling of noisy multi-input multi-output (MIMO) Hammerstein systems under correlated measurement disturbances. The proposed architecture preserves the Hammerstein decomposition by combining a gated recurrent nonlinear block for latent nonlinear transformation, an ARX-type recursive linear dynamic block for deterministic output reconstruction, and a correlated residual-noise model with heavy-tailed innovations. To improve training stability and robustness, the learning process is organized into three stages: robust initialization from step-response data, deterministic recovery using the simulator-provided noise-free target in a synthetic benchmark setting, and joint fine-tuning under measured noisy outputs.

The proposed method is validated on a synthetic 2-input/2-output benchmark with latent dimension two using a five-seed evaluation protocol. On the deterministic target, the framework achieves mean $\pm$ standard deviation test performance of RMSE 0.2168 ± 0.0454 , $R^2 = 0.9444 \pm 0.0240$, FIT = $76.99 \pm 5.14\%$, and wMAPE = $10.89 \pm 1.55\%$. Under measured noisy outputs, it remains stable and competitive, yielding RMSE 0.2948 ± 0.0318 , $R^2 = 0.9050 \pm 0.0233$, FIT = $69.40 \pm 3.76\%$, and wMAPE = $22.69 \pm 1.06\%$. Comparisons against a direct recurrent baseline and a linear ARX baseline further show that the proposed framework consistently improves over both unstructured deep prediction and purely linear modeling, while remaining stable across random seeds.

Overall, the results indicate that the proposed framework is best interpreted as a robust deep MIMO Hammerstein modeling strategy rather than an exact physical parameter-identification scheme. The method provides a practical compromise between structural interpretability, recurrent nonlinear representation, and robustness to noisy measurements.

Keywords: MIMO Hammerstein systems, GRU, robust deep modeling, correlated measurement noise, staged learning, residual-noise modeling.

1 Introduction

There are numerous practical engineering processes, which are nonlinear, multivariate, and highly coupled, whose analysis, discovery and control are further complicated by structured perturbations, uncertainty, and imprecise measurements. The difficulty is especially apparent in complicated interconnected systems, large-scale hydraulic and industrial processes, and data-driven control environments, where nonlinear interactions, as well as dynamic couplings, have to be accurately represented [1, 2, 3, 4, 5]. Simultaneously, the development of observer design, fault estimation, and robust control of uncertain nonlinear systems has demonstrated that structured models that can be viewed as interpretable and at the same time sufficiently expressive of noisy real-world data are important [6, 7, 8, 9, 10, 11].

Hammerstein systems are particularly popular among the block-based nonlinear models, as they maintain a physically interpretable split of a nonlinear block of stature and a linear block of dynamism. They have been applied in the nonlinear process modeling, prediction, and control of processes where mechanistic descriptions are unavailable or hard to calibrate due to their simplicity, interpretability, and ability to approximate [2, 12, 13]. The identification problem in the multi-input multi-output (MIMO) case is much more complicated, however, since the nonlinearities are cross-channel coupled, the intermediate variable is not observed, and the observed output is frequently contaminated by colored or correlated noise [14, 15, 16, 17]. These concerns drive systematized information-based frameworks which are able to concurrently retain the Hammerstein partition as well as take advantage of the presentation capability of contemporary learning strategies.

Recent progress in neural modeling has shown that deep recurrent, graph-based, reinforcement-learning, and deterministic-learning paradigms can greatly improve the representation of complex nonlinear dynamics and high-dimensional dependencies [18, 19, 20, 21, 22, 23]. More broadly, modern machine learning has demonstrated strong capability in large-scale nonlinear identification, relational reasoning, ensemble learning, and even entropy-based modeling of high-dimensional systems [24, 25, 26]. In the Hammerstein context, this trend has led to increasingly expressive nonlinear blocks, including multi-hidden-layer neural structures, deep recurrent architectures, and stacked sparse autoencoder networks [27, 16, 17]. At the optimization level, recent advances in robust gradient-based learning and fractional-order optimization further motivate alternative training strategies for difficult nonconvex models [28, 29].

The present work is closely related in theme to these developments but differs in both modeling objective and training design. Rather than centering the contribution on exact recovery of the true physical parameters, we formulate the problem as *robust deep modeling* of noisy MIMO Hammerstein systems under correlated measurement noise. Our framework combines a recurrent nonlinear block, an ARX-type recursive linear dynamic block, and a correlated residual-noise model with heavy-tailed innovations. In contrast to the signal-decoupled parameter-estimation logic adopted in recent Hammerstein identification studies, we use a three-stage pipeline consisting of robust initialization, deterministic recovery, and joint measured-output fine-tuning. This produces a strategy that is especially robust on deterministic recovery and also is robust to measured-output tracking.

This change of direction has two reasons behind it. To start with, the aims of most data-driven problems in practice are faithful reconstruction and prediction of trajectories in a noisy situation and not exact physical reconstruction of all latent matrices. Second, correlated residual noise can generally deteriorate the performance of pure deterministic recurrent models, particularly when the observed output has heavy-tails perturbations. The hampering of robustness can then be enhanced by a special residual-noise model at the expense of the Hammerstein decomposition. Such a view can also be explained by the more recent developing tendencies

towards learning-based modeling of constrained, interconnected, and noisy nonlinear systems, in which structural bias and robustness are becoming considered as complementary and not competing goals [1, 18, 20, 22].

The main contributions of this paper are as follows:

- In our proposal, we have suggested a well-organized framework of deep MIMO Hammerstein modeling that incorporates a nonlinear block based on GRU, a linear dynamical block, and a correlated residual-noise head.
- We present a three-step learning plan that comprises of strong Huber-based initialization, deterministic synthetic recovery and joint measured-output fine-tuning.
- We come up with a composite loss which combines strong pointwise prediction, derivative consistency, peak-sensitive penalties, one-step linear consistency, residual-noise likelihood and parameter anchoring.
- Our protocol-independent synthetic performance benchmark validation with a five seed synthetic protocol and both deterministic recovery and measured-output tracking performance are reported.
- We give a direct methodological analogy to new Hammerstein identification frameworks and elucidate the unique placement of the proposed method as a sound deep modeling framework as opposed to a precise recursive parameter-estimation scheme.

The rest of the paper is organized as follows. Section 2 reviews related work and positions the present contribution. Section 3 formulates the noisy MIMO Hammerstein modeling problem. Section 4 presents the proposed framework and its mathematical formulation. Section 5 describes the experimental protocol and benchmark configuration. Section 6 reports the quantitative and qualitative results and discusses the comparison with recent work. Section 7 concludes the paper.

2 Related Work and Positioning

Hammerstein and Wiener–Hammerstein systems have been studied extensively in nonlinear system identification, process modeling, and adaptive control. Classical approaches include polynomial or basis-function nonlinearities, recursive least-squares methods, covariance-based least squares, subspace identification, stochastic-gradient schemes, iterative decomposition, and hierarchical separation strategies [2, 14, 15, 12, 13]. In MIMO settings, the complexity of coupled nonlinearities and colored disturbances has led to signal-separated, two-stage, and recursive identification methods, particularly when the intermediate variable between the nonlinear and linear blocks cannot be measured directly [14, 15].

A first important line of work is represented by the 2025 MIMO Hammerstein framework of Li, Sun, and Cao, where the nonlinear block is modeled by a neural fuzzy model (NFM), the dynamic block is represented by ARMAX dynamics, and combined signals are used to decouple identification of the nonlinear and linear subsystems under measurement noise [14]. That study relies on correlation analysis and data filtering to improve recursive estimation accuracy and is validated on both a synthetic MIMO benchmark and a pH-neutralization process [14]. A related noisy-output identification strategy is presented by Zha, Li, and Liu, who combine covariance-function least squares for the linear block with improved particle swarm optimization for the nonlinear block under colored output noise [15]. These studies establish an important foundation for separation-identification of Hammerstein models under noise, but they remain rooted in shallow nonlinear structures and recursive parameter-learning objectives.

A second line of work moves toward deep neural Hammerstein identification. The 2026 DGRU-based framework of Li *et al.* replaces the nonlinear block by a deep GRU network,

while retaining an ARMA dynamic block and recursive extended least-squares estimation of the linear/noise parameters [16]. That work uses step signals for block decoupling, adaptive momentum estimation for the GRU parameters, and AMS-GWO for hyperparameter optimization, and it reports favorable results on a synthetic benchmark and a photovoltaic-cell application [16]. More recently, Li *et al.* also proposed an SSAE–ARMAX Hammerstein framework, where the nonlinear block is modeled by a stacked sparse autoencoder (SSAE), the dynamic/noise block is identified by multi-innovation recursive least squares, and validation is performed on a wind-power system [17]. These studies show that deep nonlinear approximators can improve expressive power beyond shallow neuro-fuzzy and polynomial models, and are consistent with broader trends in multi-hidden-layer Hammerstein identification [27].

In parallel with Hammerstein-specific research, modern learning-based nonlinear system modeling has expanded substantially. Deterministic learning-based neural identification has been used to improve model reusability and knowledge fusion in nonlinear dynamical systems [18]. Physics-informed convolutional recurrent neural networks have shown strong capability in forecasting and source identification for dynamical systems [19]. Reinforcement-learning and adaptive neural-control methods have also demonstrated promising results for interconnected and constrained nonlinear systems [20, 22, 23]. Graph-based reasoning, ensemble learning, and large-scale probabilistic modeling have likewise contributed to the general methodological landscape in which structured nonlinear identification is evolving [21, 25, 26, 24]. These advances motivate the use of structured deep architectures that can preserve system decomposition while exploiting modern sequence-learning capability.

The present work differs from the above Hammerstein identification studies in four important aspects. First, **objective**: our primary goal is robust deep modeling under correlated measurement noise, rather than exact physical parameter estimation. Second, **noise treatment**: instead of a standard moving-average or ARMAX output-noise formulation, we explicitly model correlated residual dynamics with a heavy-tailed innovation likelihood. Third, **training strategy**: we use a three-stage pipeline with robust initialization, deterministic synthetic recovery, and joint fine-tuning, rather than a purely signal-decoupled two-stage recursive estimator. Fourth, **evidence protocol**: we emphasize five-seed mean $\pm$ standard deviation evidence, separating deterministic recovery from measured-output tracking, rather than focusing exclusively on a single-run parameter-estimation result.

The other scientific claim is defined by them. The current approach is to be viewed as an effective deep MIMO Hammerstein framework of model structure with organised inductive bias and explicit residual-noise strong robustness, rather than as analogous to recursive physical-parameter-identification arrangements. The methods suggested above, in this respect, are complementary to the currently existing Li-family frameworks: the above works excel at signal-separated recursive identification and validation of applications, but the current work adds three-stage deep training, correlated residual-noise, and multi-seed robustness evidence.

3 Problem Formulation

Take a discrete-time multi-input multi-output (MIMO) Hammerstein system which consists of nonlinear transformation and then a linear dynamic block, which is corrupted by measurement disturbances. Its input output relationship is defined by

$$\mathbf{n}(t) = f(\mathbf{u}(1:t)), \quad (1)$$

$$\mathbf{x}(t) = - \sum_{i=1}^{n_a} \mathbf{A}_i \mathbf{x}(t-i) + \sum_{j=1}^{n_b} \mathbf{B}_j \mathbf{n}(t-j), \quad (2)$$

$$\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\eta}(t), \quad (3)$$

where:

- $\mathbf{u}(t) \in \mathbb{R}^p$ is the measured multi-channel input,
- $\mathbf{n}(t) \in \mathbb{R}^m$ is an unmeasurable intermediate (latent) variable generated by the nonlinear block,
- $\mathbf{x}(t) \in \mathbb{R}^s$ is the noise-free deterministic output of the Hammerstein system,
- $\mathbf{y}(t) \in \mathbb{R}^s$ is the measured output,
- $\boldsymbol{\eta}(t) \in \mathbb{R}^s$ denotes a correlated measurement disturbance.

The matrices $\mathbf{A}_i \in \mathbb{R}^{s \times s}$ and $\mathbf{B}_j \in \mathbb{R}^{s \times m}$ characterize the recursive linear dynamic block, while the nonlinear mapping $f(\cdot)$ transforms the measured input sequence into the latent variable $\mathbf{n}(t)$. The dependence on $\mathbf{u}(1:t)$ emphasizes that the nonlinear block is implemented through a recurrent architecture and may therefore encode sequence-dependent nonlinear effects rather than a purely memoryless static map.

For the benchmark considered in this work, the system dimensions are fixed to

$$p = 2, \quad m = 2, \quad s = 2, \quad n_a = 2, \quad n_b = 2.$$

The available dataset consists of measured input–output pairs

$$\mathcal{D} = \{(\mathbf{u}(t), \mathbf{y}(t))\}_{t=1}^T,$$

while the deterministic target $\mathbf{x}(t)$ is additionally available in the synthetic benchmark setting. The central challenge is that the intermediate variable $\mathbf{n}(t)$ is not observable, the output is corrupted by correlated noise, and the nonlinear and dynamic effects are coupled through the Hammerstein structure.

Accordingly, the objective of this work is to construct a structured predictor

$$\hat{\mathbf{x}}(t) = \mathcal{H}_\Theta(\mathbf{u}(1:t)),$$

parameterized by Θ , such that the following requirements are simultaneously satisfied:

1. **Deterministic recovery:** whenever the deterministic target is known, the predictor must be capable of the exact reconstruction of the noise-free output of the system, given by the value of $\mathbf{x}(t)$.
2. **Measured-output robustness:** the predictor must be predictable and of constant quality, in comparison with the actual noisy output, the observed and measured noisy output, denoted by $\mathbf{y}(t)$.
3. **Structural consistency:** the learned model must maintain a Hammerstein-compatible break down amongst nonlinear latent mapping and recursive linear dynamic evolution.

The learning problem is learned according to two complementary evaluations settings to represent these requirements:

$$\text{deterministic recovery:} \quad \hat{\mathbf{x}}(t) \approx \mathbf{x}(t), \quad (4)$$

$$\text{measured-output tracking:} \quad \hat{\mathbf{x}}(t) \approx \mathbf{y}(t). \quad (5)$$

The former is an assessment of the ability of the structured model to reconstruction of the underlying deterministic Hammerstein dynamics, and the latter is an assessment of robustness in the presence of correlated measurement noise. The method suggested aims at solving this issue by integrating. where (i) is a nonlinear block recurrent latent-variable generator, (ii) a linear dynamic block which is recursive and deterministic output reconstruction, and (iii) a special refining-noise modeling mechanism on measured-outputs.

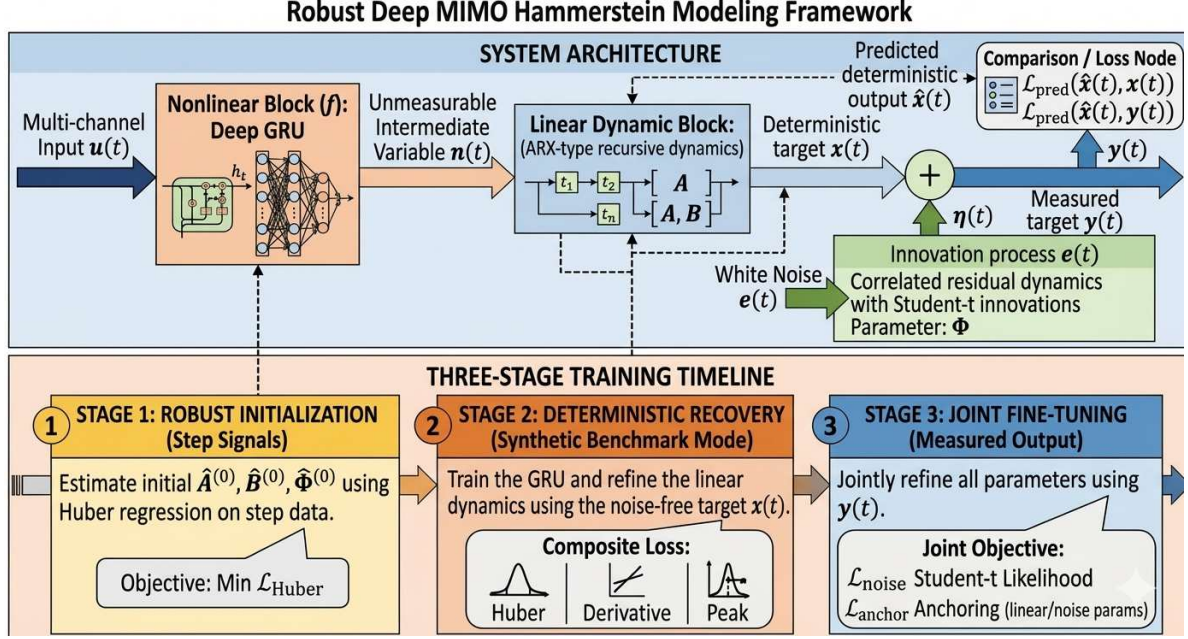


Figure 1: General architecture and training process of the presented robust deep MIMO Hammerstein modeling architect. The top panel has the hierarchical structure of the cascade, which consists of a deep GRU nonlinear block, a recursive linear dynamic block, and a correlated residual-noise element. The bottom panel has an overview of the training pipeline: Stage 1 to estimate initial linear/noise parameters with huber regression on step-response measurements, Stage 2 to recover deterministically with the noise-free target $x(t)$ provided by the simulator, Stage 3 to refine all the model parameters jointly with the measured output $y(t)$.

4 Mathematical Formulation of the Proposed Framework

Figure 1 gives a summary of the suggested robust deep MIMO Hammerstein modelling model. The top block recaps the hierarchical architecture which consists of the deep GRU nonlinear block, the ARX-type recursive linear dynamic block and the correlated residual-noise block with Student- t innovations. The bottom portion shows the three-step training plan that involves strong pre-training, conditional recovery with synthetic benchmark training, and joint refinement with measured outputs.

The proposed approach, as illustrated in Fig. 1, does not alter the structural decomposition of the Hammerstein system, but makes the system more robust to correlated measurement noise. The nonlinear transformation from the measured input $u(t)$ to the latent variable $n(t)$ is learned by a deep GRU block, the latent variable is propagated through a recursive linear dynamic block to produce the deterministic output estimate $\hat{x}(t)$, and the discrepancy between measured and predicted outputs is regularized through a correlated residual-noise component with Student- t innovations. The lower part of the figure highlights the staged training strategy: robust step-signal initialization, deterministic recovery using the noise-free target in the synthetic benchmark setting, and final joint refinement under measured noisy outputs.

4.1 Noisy MIMO Hammerstein model

Consider a discrete-time multi-input multi-output Hammerstein system composed of a nonlinear block followed by a linear dynamic block and corrupted by correlated measurement noise. The

system is written as

$$\mathbf{n}_t = f(\mathbf{u}_{1:t}), \quad (6)$$

$$\mathbf{x}_t = -\sum_{i=1}^{n_a} \mathbf{A}_i \mathbf{x}_{t-i} + \sum_{j=1}^{n_b} \mathbf{B}_j \mathbf{n}_{t-j}, \quad (7)$$

$$\mathbf{y}_t = \mathbf{x}_t + \boldsymbol{\eta}_t, \quad (8)$$

where:

- $\mathbf{u}_t \in \mathbb{R}^p$ is the measured input,
- $\mathbf{n}_t \in \mathbb{R}^m$ is an unmeasurable latent/intermediate variable,
- $\mathbf{x}_t \in \mathbb{R}^s$ is the noise-free deterministic output,
- $\mathbf{y}_t \in \mathbb{R}^s$ is the measured output,
- $\boldsymbol{\eta}_t \in \mathbb{R}^s$ is correlated measurement noise.

In the benchmark considered in this paper,

$$p = 2, \quad m = 2, \quad s = 2, \quad n_a = 2, \quad n_b = 2.$$

The linear-dynamic parameters satisfy

$$\mathbf{A}_i \in \mathbb{R}^{s \times s}, \quad \mathbf{B}_j \in \mathbb{R}^{s \times m}.$$

The objective is not to build an unconstrained black-box predictor, but rather to learn a structured model that (i) accurately reconstructs the deterministic output $\mathbf{x}_t$, (ii) remains robust under measured noisy outputs $\mathbf{y}_t$, and (iii) preserves a Hammerstein-compatible decomposition between nonlinear transformation and recursive linear dynamics.

4.2 GRU-based nonlinear block

The nonlinear block is implemented by a gated recurrent unit (GRU) network. Let $\mathbf{h}_t \in \mathbb{R}^{d_h}$ denote the hidden state at time t . For an input $\mathbf{u}_t$, the GRU update is defined as

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{u}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r), \quad (9)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{u}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z), \quad (10)$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h \mathbf{u}_t + \mathbf{U}_h (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h), \quad (11)$$

$$\mathbf{h}_t = (\mathbf{1} - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t, \quad (12)$$

where $\sigma(\cdot)$ is the logistic sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent, and $\odot$ denotes elementwise multiplication.

To obtain the latent nonlinear output, the hidden state is projected through an output layer and combined with an input skip mapping:

$$\hat{\mathbf{n}}_t = \mathbf{W}_o \mathbf{h}_t + \mathbf{W}_s \mathbf{u}_t + \mathbf{b}_o. \quad (13)$$

The use of the skip connection improves optimization stability and helps preserve low-order direct input-to-latent dependence. At the same time, the recurrent structure allows the nonlinear block to encode sequence-dependent nonlinear effects, making the latent representation flexible enough for coupled MIMO nonlinearities.

4.3 Recursive linear dynamic block

The latent estimate is propagated through an ARX-type recursive linear dynamic block:

$$\hat{\mathbf{x}}_t = - \sum_{i=1}^{n_a} \hat{\mathbf{A}}_i \hat{\mathbf{x}}_{t-i} + \sum_{j=1}^{n_b} \hat{\mathbf{B}}_j \hat{\mathbf{n}}_{t-j}. \quad (14)$$

Equation (14) preserves the recursive dynamic structure of the Hammerstein model while keeping the nonlinear transformation flexible and data-driven through the GRU block. This decomposition is crucial because it separates latent nonlinear representation from linear temporal evolution. In the present framework, $\hat{\mathbf{x}}_t$ is the primary predictive quantity used throughout both deterministic recovery and measured-output evaluation.

For notational convenience, define the predicted state history vector

$$\mathcal{X}_{t-1} = \left[\hat{\mathbf{x}}_{t-1}^\top, \dots, \hat{\mathbf{x}}_{t-n_a}^\top \right]^\top,$$

and the latent history vector

$$\mathcal{N}_{t-1} = \left[\hat{\mathbf{n}}_{t-1}^\top, \dots, \hat{\mathbf{n}}_{t-n_b}^\top \right]^\top.$$

Then the recursive linear block can be viewed as a structured affine map

$$\hat{\mathbf{x}}_t = \mathcal{F}_{\text{lin}}(\mathcal{X}_{t-1}, \mathcal{N}_{t-1}; \{\mathbf{A}_i, \mathbf{B}_j\}).$$

4.4 Correlated residual-noise component

To account for correlated measurement disturbances, we define the measured residual

$$\mathbf{r}_t = \mathbf{y}_t - \hat{\mathbf{x}}_t. \quad (15)$$

The residual-noise component models correlated residual dynamics through

$$\boldsymbol{\epsilon}_t = \mathbf{r}_t - \sum_{\ell=1}^{n_\eta} \hat{\boldsymbol{\Phi}}_\ell \mathbf{r}_{t-\ell}, \quad \hat{\boldsymbol{\Phi}}_\ell \in \mathbb{R}^{s \times s}, \quad (16)$$

where $\boldsymbol{\epsilon}_t$ is the innovation process. In this work, the residual model order is $n_\eta = 1$.

To improve robustness to heavy-tailed disturbances, each innovation channel is modeled with a Student- t negative log-likelihood. Let $\nu > 0$ denote the degrees of freedom and $\sigma_k > 0$ the channel-wise scale parameter. Then

$$\mathcal{L}_{\text{noise}} = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^s \left[\frac{\nu+1}{2} \log \left(1 + \frac{1}{\nu} \left(\frac{\epsilon_{t,k}}{\sigma_k} \right)^2 \right) + \log \sigma_k \right]. \quad (17)$$

Equation (17) penalizes large innovations more softly than a Gaussian negative log-likelihood and is therefore more appropriate under outliers or heavy-tailed measured disturbances.

Remark on the role of the residual-noise component. In the current formulation, the correlated residual-noise component is used primarily as a *training-time mechanism* that penalizes structured residual correlation under measured noisy outputs. Accordingly, the learned predictor remains the deterministic quantity $\hat{\mathbf{x}}_t$, while the residual-noise loss regularizes the model toward noise-robust behavior during optimization. The present framework therefore does not instantiate a full generative observation model at inference time; rather, it employs a noise-aware residual-correlation penalty to improve robustness of the deterministic predictor.

4.5 Stage 1: robust initialization from step data

The first stage estimates initial values for the linear block and residual-noise component using step-response data. Let $\{\mathbf{u}_t^{\text{step}}, \mathbf{y}_t^{\text{step}}\}$ denote the step-signal dataset. For each time index t , define the ARX regressor

$$\boldsymbol{\varphi}_t = [-\mathbf{y}_{t-1}^\top \quad \cdots \quad -\mathbf{y}_{t-n_a}^\top \quad \mathbf{u}_{t-1}^\top \quad \cdots \quad \mathbf{u}_{t-n_b}^\top]^\top. \quad (18)$$

For each output channel $k \in \{1, \dots, s\}$, the initial linear parameters are estimated by robust Huber regression:

$$\hat{\boldsymbol{\theta}}_k^{(0)} = \arg \min_{\boldsymbol{\theta}_k} \sum_t \rho_\delta(y_{k,t} - \boldsymbol{\varphi}_t^\top \boldsymbol{\theta}_k), \quad (19)$$

where $\rho_\delta(\cdot)$ is the Huber penalty

$$\rho_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \leq \delta, \\ \delta|e| - \frac{1}{2}\delta^2, & |e| > \delta. \end{cases} \quad (20)$$

By stacking the channel-wise coefficients $\hat{\boldsymbol{\theta}}_k^{(0)}$, the initial matrices

$$\hat{\mathbf{A}}_i^{(0)}, \quad \hat{\mathbf{B}}_j^{(0)}$$

are recovered from the appropriate coefficient blocks.

Next, let the stage-1 residual sequence be

$$\hat{\mathbf{r}}_t^{(0)} = \mathbf{y}_t - \hat{\mathbf{y}}_t^{(0)}. \quad (21)$$

The initial residual-correlation parameter is estimated by least squares:

$$\hat{\boldsymbol{\Phi}}^{(0)} = \arg \min_{\boldsymbol{\Phi}} \sum_{t=2}^T \left\| \hat{\mathbf{r}}_t^{(0)} - \boldsymbol{\Phi} \hat{\mathbf{r}}_{t-1}^{(0)} \right\|_2^2. \quad (22)$$

The closed-form solution is

$$\hat{\boldsymbol{\Phi}}^{(0)} = \left(\sum_{t=2}^T \hat{\mathbf{r}}_t^{(0)} \hat{\mathbf{r}}_{t-1}^{(0)\top} \right) \left(\sum_{t=2}^T \hat{\mathbf{r}}_{t-1}^{(0)} \hat{\mathbf{r}}_{t-1}^{(0)\top} \right)^{-1}. \quad (23)$$

This strongly informed initialize step is a structurally informed warm start of the subsequent prediction based optimization.

4.6 Stage 2: deterministic recovery

The noise-free target $\mathbf{x}_t$ is available in synthetic mode of benchmarking. The second step consequently maximises the GRU block and linear dynamics with reference to the deterministic target:

$$\hat{\mathbf{x}}_t \approx \mathbf{x}_t. \quad (24)$$

A channel-weighted Huber loss is the main predictive loss that is deterministic:

$$\mathcal{L}_{\text{pred}} = \sum_{k=1}^s \omega_k \frac{1}{T} \sum_{t=1}^T \rho_\delta(\hat{x}_{t,k} - x_{t,k}), \quad \sum_{k=1}^s \omega_k = 1. \quad (25)$$

This step is explicitly meant to enhance the quality of deterministic recovery and in itself, it does not ensure the precise recovery of the underlying physical parameters.

4.7 Derivative consistency and peak-aware regularization

To preserve temporal shape and transient behavior, define the first-order difference

$$\Delta \mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1}, \quad \Delta \hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1}.$$

The loss's derivative-consistency is

$$\mathcal{L}_{\text{diff}} = \sum_{k=1}^s \omega_k \frac{1}{T-1} \sum_{t=2}^T (\Delta \hat{x}_{t,k} - \Delta x_{t,k})^2. \quad (26)$$

To focus on high-dynamic regions, we define a peak mask

$$\mathcal{M}_{t,k} = \mathbf{1}(|\Delta x_{t,k}| \geq \tau_q), \quad (27)$$

where τ_q is the q -th percentile of the derivative magnitude, and $\mathbf{1}(\cdot)$ denotes the indicator function.

The peak-aware loss is then

$$\mathcal{L}_{\text{peak}} = \frac{\sum_{t=2}^T \sum_{k=1}^s \omega_k \mathcal{M}_{t,k} |\Delta \hat{x}_{t,k} - \Delta x_{t,k}|}{\sum_{t=2}^T \sum_{k=1}^s \omega_k \mathcal{M}_{t,k} + \varepsilon}. \quad (28)$$

4.8 One-step teacher-forced linear consistency

In order to enhance the identifiability of the recursive linear block, we add a one-step teacher-forced predictor:

$$\hat{\mathbf{x}}_t^{(1)} = - \sum_{i=1}^{n_a} \hat{\mathbf{A}}_i \mathbf{x}_{t-i} + \sum_{j=1}^{n_b} \hat{\mathbf{B}}_j \hat{\mathbf{n}}_{t-j}. \quad (29)$$

Unlike (14), the true deterministic history $\mathbf{x}_{t-i}$ is used instead of the recursively predicted history.

The corresponding loss is

$$\mathcal{L}_{\text{1step}} = \sum_{k=1}^s \omega_k \frac{1}{T - \max(n_a, n_b)} \sum_{t=\max(n_a, n_b)+1}^T \rho_\delta(\hat{x}_{t,k}^{(1)} - x_{t,k}). \quad (30)$$

This term gives a more direct identification signal to $\{\mathbf{A}_i, \mathbf{B}_j\}$ than pure free-run recursion, even though the overall training goal is essentially one that aims at prediction.

4.9 Anchoring and regularization

In order to stabilize the optimization process to the strong step-signal initialisation, we penalise the change in parameter estimates:

$$\mathcal{L}_{\text{anchor}} = \lambda_A \sum_{i=1}^{n_a} \|\mathbf{A}_i - \mathbf{A}_i^{(0)}\|_F^2 + \lambda_B \sum_{j=1}^{n_b} \|\mathbf{B}_j - \mathbf{B}_j^{(0)}\|_F^2 + \lambda_\Phi \sum_{\ell=1}^{n_\eta} \|\Phi_\ell - \Phi_\ell^{(0)}\|_F^2. \quad (31)$$

A standard quadratic parameter penalty is also used:

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{n_a} \|\mathbf{A}_i\|_F^2 + \sum_{j=1}^{n_b} \|\mathbf{B}_j\|_F^2 + \sum_{\ell=1}^{n_\eta} \|\Phi_\ell\|_F^2. \quad (32)$$

The anchoring term acts only as a soft structural bias and does not enforce exact recovery of the ground-truth physical parameters.

4.10 Overall objective and stage-wise optimization

The global training objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{peak}}\mathcal{L}_{\text{peak}} + \lambda_{\text{1step}}\mathcal{L}_{\text{1step}} + \lambda_{\text{noise}}\mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{anchor}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \quad (33)$$

The stage-wise objectives are:

$$\mathcal{L}_A = \mathcal{L}_{\text{pred}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{peak}}\mathcal{L}_{\text{peak}} + \lambda_{\text{1step}}\mathcal{L}_{\text{1step}} + \mathcal{L}_{\text{anchor}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \quad (34)$$

$$\mathcal{L}_B = \mathcal{L}_A + \lambda_{\text{noise}}\mathcal{L}_{\text{noise}}, \quad (35)$$

$$\mathcal{L}_C = \mathcal{L}_B. \quad (36)$$

Stage A emphasizes deterministic recovery, Stage B refines linear/noise parameters under residual-correlation regularization, and Stage C jointly fine-tunes all parameters. Because the dominant objective is predictive reconstruction, the optimization naturally favors representations that improve trajectory recovery rather than exact parameter fidelity.

4.11 Gradient structure

Let

$$\Theta = \{\theta_f, \{\mathbf{A}_i\}_{i=1}^{n_a}, \{\mathbf{B}_j\}_{j=1}^{n_b}, \{\Phi_\ell\}_{\ell=1}^{n_\eta}, \boldsymbol{\sigma}\}$$

denote the full parameter set. The total gradient is computed by back-propagation through time (BPTT) using the chain rule through the recurrent nonlinear block and recursive dynamic block.

For the linear dynamic matrices, the local gradient contributions satisfy

$$\nabla_{\mathbf{A}_i}\mathcal{L} = - \sum_{t=i+1}^T \boldsymbol{\delta}_t \hat{\mathbf{x}}_{t-i}^\top + 2\lambda_A(\mathbf{A}_i - \mathbf{A}_i^{(0)}) + 2\lambda_{\text{reg}}\mathbf{A}_i, \quad (37)$$

$$\nabla_{\mathbf{B}_j}\mathcal{L} = \sum_{t=j+1}^T \boldsymbol{\delta}_t \hat{\mathbf{n}}_{t-j}^\top + 2\lambda_B(\mathbf{B}_j - \mathbf{B}_j^{(0)}) + 2\lambda_{\text{reg}}\mathbf{B}_j, \quad (38)$$

where $\boldsymbol{\delta}_t = \partial\mathcal{L}/\partial\hat{\mathbf{x}}_t$ is the adjoint signal associated with the recursive deterministic output.

Similarly, for the residual-correlation parameter,

$$\nabla_{\Phi_\ell}\mathcal{L}_{\text{noise}} = - \sum_{t=\ell+1}^T \mathbf{g}_t \mathbf{r}_{t-\ell}^\top, \quad \mathbf{g}_t = \frac{\partial\mathcal{L}_{\text{noise}}}{\partial\boldsymbol{\epsilon}_t}. \quad (39)$$

The nonlinear-block parameters θ_f are updated through BPTT from the loss gradients propagated to the latent sequence $\hat{\mathbf{n}}_t$:

$$\nabla_{\theta_f}\mathcal{L} = \sum_{t=1}^T \frac{\partial\mathcal{L}}{\partial\hat{\mathbf{n}}_t} \frac{\partial\hat{\mathbf{n}}_t}{\partial\theta_f}. \quad (40)$$

In practice, the full gradient is computed automatically using reverse-mode differentiation, while the explicit forms (37)–(40) clarify the structure of the optimization problem.

4.12 Model-selection criterion

Hyperparameter search and post-Stage-A noise tuning use a validation score that emphasizes deterministic recovery rather than simply minimizing the total loss:

$$\text{Score} = \mathcal{L}_{\text{pred}} + 0.15\mathcal{L}_{\text{diff}} + 0.10\mathcal{L}_{\text{peak}} + 0.20\mathcal{L}_{\text{1step}}. \quad (41)$$

This choice avoids over-selecting configurations that reduce the total loss mainly by emphasizing the noise likelihood while degrading deterministic recovery quality.

4.13 Algorithm summary

Algorithm 1 summarizes the complete learning pipeline of the proposed framework. The process starts with a strong step-signal start stage, at which Huber-based regression is applied in arriving at structurally significant initial estimates of the linear dynamic and residual-correlation parameters. The model is subsequently trained in three consecutive steps. Stage A is concerned with deterministic recovery with the noisy target given in the synthetic benchmark environment and thus enhances the latent nonlinear representation and recursive linear dynamics. This is followed by the choice of the residual-noise weight based on the validation score in the formula (41) so that the model selection is performed based on the deterministic recoveries quality as opposed to the noise likelihood. Stage B, in its turn, fine-tunes the linear dynamic and residual-correlation parameters keeping the nonlinear block constant, but Stage C fine-tunes all the model parameters jointly. Lastly, the trained model is tested on two complementary conditions, that is, deterministic recovery with respect to the noise-free output and the measured-output tracking with respect to the noisy observations. Such staged design eases optimization effort, enhances robustness, and the resulting model is still structurally consistent despite the Hammerstein decomposition which it models, although the model need not be thought of as a precise parameter-identification process.

5 Experimental Setup

5.1 Synthetic Benchmark Configuration

The benchmark is an artificial 2-input/2-output MIMO Hammerstein system with a latent dimension of two, AR orders of $n_a = 2$, $n_b = 2$, and one-order correlated dynamics of the residual-noise. The following protocol is used in the experiment:

- 320 step-response training sequences and 80 step-response validation sequences for Stage 1,
- 512 random-signal training sequences, 128 validation sequences, and 128 test sequences for Stage 2 and Stage 3,
- sequence lengths $T_{\text{step}} = 80$ and $T_{\text{main}} = 100$,
- 12 search trials and 20 search epochs on CPU,
- batch size 32.

5.2 Representative Hyperparameters

For the representative seed 4040, the selected configuration is:

- hidden size: 64,
- number of GRU layers: 2,
- dropout: 0,
- learning rate: 0.001,
- $\lambda_{\text{diff}} = 0.1$,
- $\lambda_{\text{peak}} = 0.2$,
- $\lambda_{\text{1step}} = 0.05$,
- tuned residual-noise weight: $\lambda_{\text{noise}} = 0.002$.

The best validation epochs for the representative seed are:

Stage A: 53, Stage B: 59, Stage C: 11.

5.3 Evaluation Settings

Two complementary evaluation settings are used:

1. **Deterministic recovery:** compare the predicted deterministic output $\hat{\mathbf{x}}(t)$ against the simulator-provided noise-free target $\mathbf{x}(t)$.
2. **Measured-output tracking:** compare the same prediction against the measured noisy output $\mathbf{y}(t)$.

The main metrics reported in the paper are RMSE, R^2 , FIT, and wMAPE. Because the signals cross or approach zero, raw MAPE is not used as a headline metric.

5.4 Baseline Models

To contextualize the performance of the proposed framework, two additional baselines are considered. The first is a **Direct GRU baseline**, which maps the input sequence directly to the deterministic output without an explicit Hammerstein decomposition, recursive linear dynamic block, or residual-correlation component. This baseline evaluates whether the gains of the proposed method arise simply from using a recurrent architecture or from the structured staged design itself. The second is a **Linear ARX baseline**, fitted directly on the random-signal training data by Huber regression and evaluated through recursive rollout. This baseline provides a classical linear lower bound and helps quantify the extent to which the benchmark requires nonlinear modeling rather than purely linear dynamics.

6 Results and Discussion

6.1 Five-Seed Aggregate Results

The aggregate mean $\pm$ standard deviation over five independent seeds is summarized in Table 1. The use of multi-seed reporting is important here because it allows the proposed framework to be assessed not only in terms of raw predictive accuracy but also in terms of stability with respect to random initialization, stochastic optimization, and hyperparameter search variability.

Table 1: Five-seed aggregate performance (mean $\pm$ std).

Benchmark	RMSE	R^2	FIT (%)	wMAPE (%)
Validation on X	0.1818 ± 0.0227	0.9636 ± 0.0097	81.10 ± 2.60	10.62 ± 0.99
Test on X	0.2168 ± 0.0454	0.9444 ± 0.0240	76.99 ± 5.14	10.89 ± 1.55
Validation vs Y	0.2671 ± 0.0163	0.9257 ± 0.0108	72.82 ± 2.01	22.30 ± 0.64
Test vs Y	0.2948 ± 0.0318	0.9050 ± 0.0233	69.40 ± 3.76	22.69 ± 1.06

Several conclusions can be drawn from Table 1. First, the proposed framework is particularly strong on deterministic synthetic recovery, where the aggregate test performance on the noise-free target reaches $R^2 = 0.9444 \pm 0.0240$ and FIT = $76.99 \pm 5.14\%$. This indicates that the structured GRU–Hammerstein formulation can recover the underlying deterministic dynamics with high fidelity. Second, the measured-output benchmark remains stable under correlated noise, with aggregate test performance of RMSE 0.2948 ± 0.0318 , $R^2 = 0.9050 \pm 0.0233$, and

Table 2: Aggregate five-seed comparison with baseline models (mean $\pm$ std).

Setting	Model	RMSE	R^2	FIT (%)	wMAPE (%)
Test on X	Proposed staged GRU–Hammerstein	0.2168 ± 0.0454	0.9444 ± 0.0240	76.99 ± 5.14	10.89 ± 1.55
	Direct GRU baseline	0.2219 ± 0.0591	0.9403 ± 0.0316	76.46 ± 6.53	12.39 ± 2.38
	Linear ARX baseline	0.8944 ± 0.0175	0.1049 ± 0.0099	5.39 ± 0.52	85.96 ± 1.57
Test vs Y	Proposed staged GRU–Hammerstein	0.2948 ± 0.0318	0.9050 ± 0.0233	69.40 ± 3.76	22.69 ± 1.06
	Direct GRU baseline	0.2999 ± 0.0426	0.9008 ± 0.0305	68.88 ± 4.79	23.54 ± 1.83
	Linear ARX baseline	0.9154 ± 0.0177	0.1009 ± 0.0105	5.18 ± 0.55	86.60 ± 1.31

FIT = $69.40 \pm 3.76\%$. Although this setting is naturally more difficult, the relatively small standard deviations indicate that the proposed framework does not suffer from severe seed-to-seed instability. Third, the close agreement between validation and test performance suggests that the staged training design leads to well-behaved generalization rather than purely memorized synthetic trajectories.

6.2 Baseline Comparisons

To assess the value of the proposed structured staged design, we compare it against two additional baselines: a direct sequence-to-sequence GRU predictor and a linear ARX model. The Direct GRU baseline uses a recurrent architecture of comparable capacity but removes the explicit Hammerstein decomposition, recursive linear block, and residual-correlation component. The Linear ARX baseline provides a classical linear reference.

Table 2 leads to three main observations. First, the Linear ARX baseline performs very poorly in both evaluation settings, with R^2 values close to 0.10 and FIT values close to 5%, confirming that the benchmark is strongly nonlinear and cannot be explained adequately by a purely linear dynamic model. Second, the Direct GRU baseline is much stronger than the linear baseline and achieves performance close to that of the proposed method, indicating that a generic recurrent architecture can already capture a substantial portion of the deterministic dynamics. Third, the proposed staged GRU–Hammerstein framework consistently outperforms the Direct GRU baseline across all reported aggregate metrics, albeit by a moderate margin. In particular, the proposed model yields lower RMSE and wMAPE and slightly higher R^2 and FIT in both deterministic recovery and measured-output tracking.

An additional advantage of the proposed method is its lower seed-to-seed variability. Relative to the Direct GRU baseline, the proposed framework exhibits smaller standard deviations in RMSE and FIT in both test settings, suggesting that the structured decomposition and staged optimization improve not only predictive accuracy but also training stability. Therefore, the benefit of the proposed framework should not be interpreted as a dramatic raw-accuracy improvement over a strong recurrent baseline, but rather as a *systematic and more stable improvement* obtained by combining recurrent nonlinear representation with Hammerstein structure and staged optimization.

6.3 Channel-Wise Behavior

After establishing the global advantage of the proposed framework over both the linear and unstructured recurrent baselines, we next examine the aggregate per-channel test performance summarized in Table 3.

Table 3: Aggregate per-channel test performance (mean $\pm$ std).

Benchmark	Channel	R^2	FIT (%)
Test on X	Channel 1	0.9045 ± 0.0323	69.54 ± 5.29
	Channel 2	0.9496 ± 0.0271	78.36 ± 6.00
Test vs Y	Channel 1	0.7364 ± 0.0346	48.77 ± 3.46
	Channel 2	0.9319 ± 0.0261	74.38 ± 4.92

There is a significant asymmetry in the channel-wise analysis. Channels of the deterministic environment are rebuilt with reasonable precision, but Channel 2 is still a little easier than Channel 1. Conversely, with the noisy outputs under measured conditions the difference is far more evident: Channel 1 is reduced to $R^2 = 0.7364 \pm 0.0346$, but Channel 2 is able to reach $R^2 = 0.9319 \pm 0.0261$. It means that it is not the deterministic latent-to-output mapping that is the main challenge, but is the interplay of channel-specific dynamics with correlated measurement noise. On the modeling level, the finding implies that the suggested framework is generally robust, yet still, channel-sensitive in the case when the measured signal is heavily contaminated.

6.4 Representative Qualitative Results

Representative qualitative performance is presented of the seed 4040 which produced the best test performance of the five runs on deterministic recovery as well as measured-output tracking. This seed is then just a qualitative illustration seed only, and all headline quantitative conclusions are still founded on the five-seed aggregate statistics shown above.

For seed 4040:

- **Test on X :** RMSE = 0.1572, $R^2 = 0.9744$, FIT = 84.00%, wMAPE = 9.18%.
- **Test vs Y :** RMSE = 0.2563, $R^2 = 0.9347$, FIT = 74.44%, wMAPE = 21.40%.

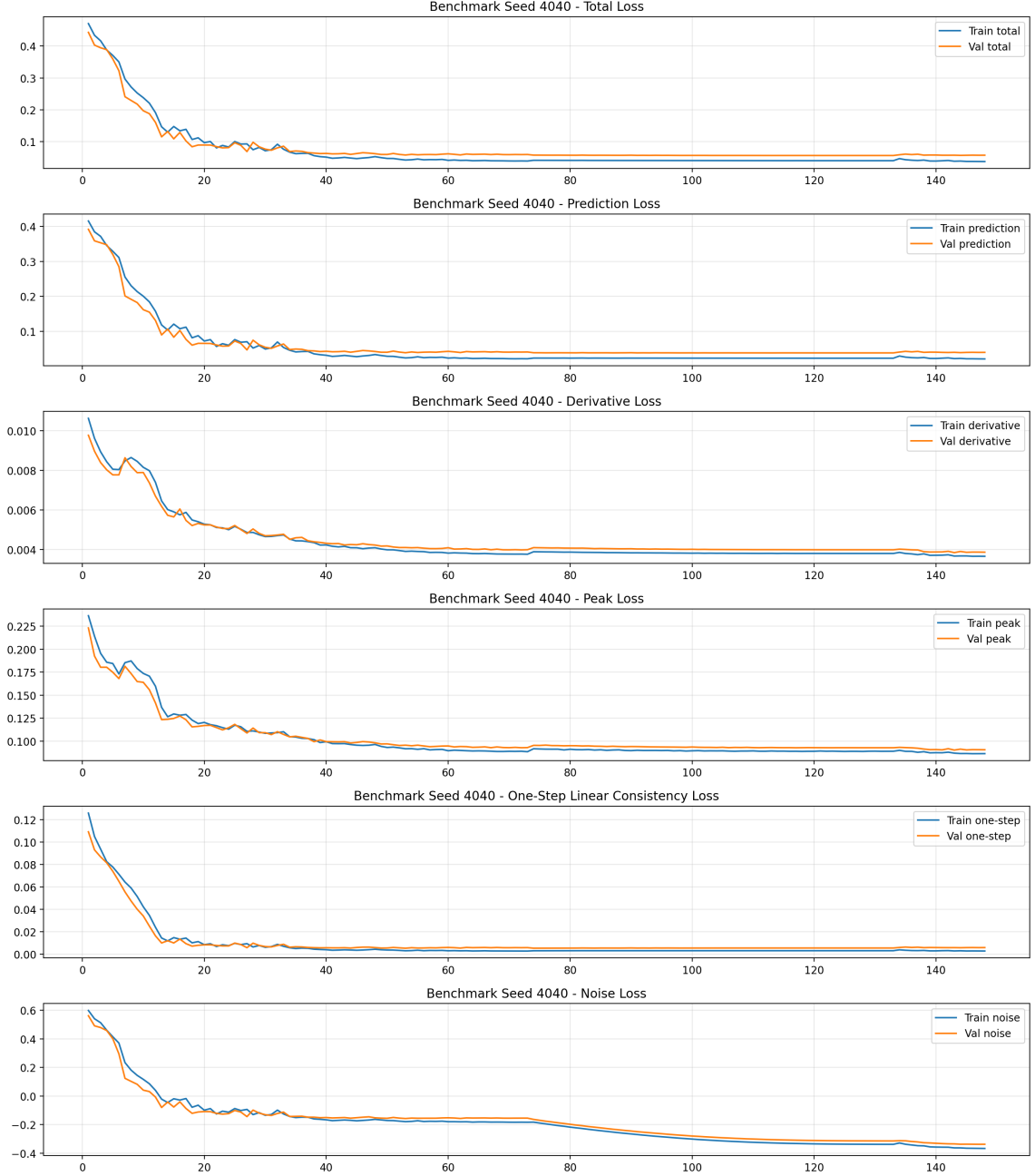
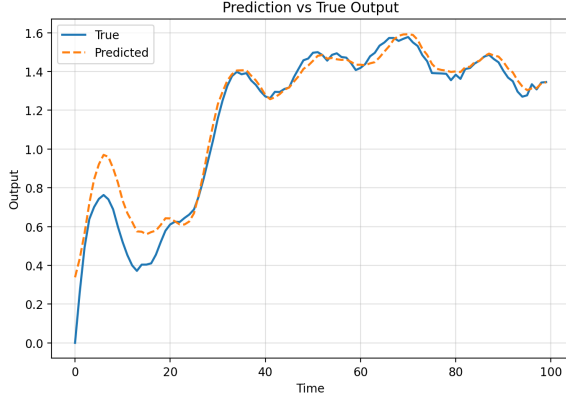
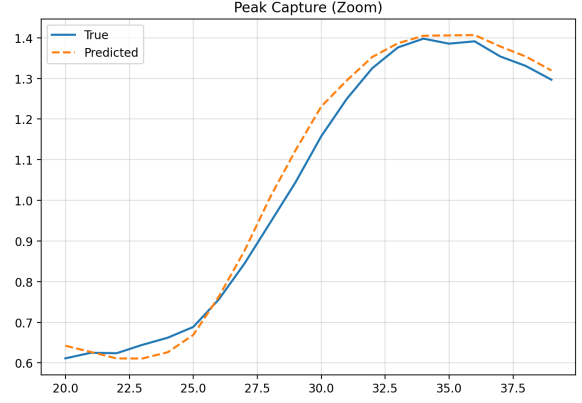


Figure 2: Representative training history for seed 4040. The losses remain stable throughout the three-stage training process.

The qualitative results support the aggregate statistics. In the deterministic setting, the proposed model provides close overlap with the underlying target and preserves local trajectory structure well, especially after the initial transient. In the measured-output setting, the model continues to follow the dominant trajectory trend but naturally appears smoother than the noisy measured signal in locally high-variance regions. This behavior is consistent with the role of the proposed framework as a structured robust predictor rather than a mechanism for reproducing every measurement-level fluctuation.

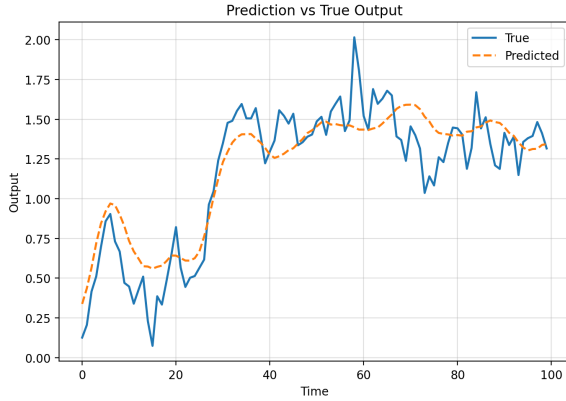


(a) Prediction vs. true deterministic output.

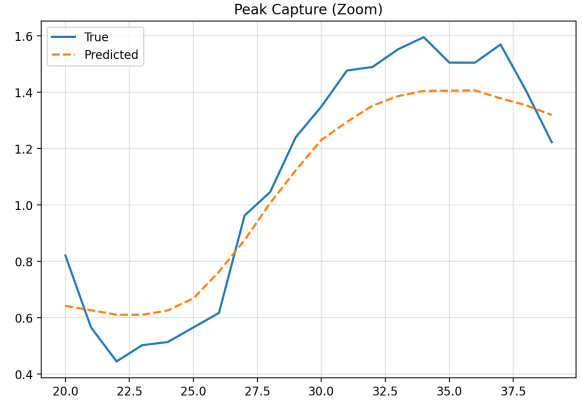


(b) Local zoom of deterministic peak capture.

Figure 3: Representative deterministic recovery for seed 4040 on Channel 2. The model closely follows the noise-free target and captures the dominant local trajectory structure.



(a) Prediction vs. measured noisy output.



(b) Local zoom of measured-output tracking.

Figure 4: Representative measured-output tracking for seed 4040 on Channel 2. The model captures the dominant trajectory trend under correlated measurement noise, while remaining smoother than the measured signal in locally irregular regions.

6.5 Parameter-Recovery Perspective

Although the proposed framework performs strongly as a robust modeling approach, the aggregate parameter-recovery errors remain relatively high:

$$\text{rel_error}(A) = 2.717 \pm 0.029, \quad \text{rel_error}(B) = 1.027 \pm 0.019, \quad \text{rel_error}(\Phi) = 0.826 \pm 0.094.$$

According to these values, the approach is significantly more powerful in deterministic recovery and noisy-output prediction than in deterministic recovery of the underlying linear and residual-noise parameters. Critically, the standard deviations are relatively small indicating that this restriction is systematic among seeds and not due to single optimization failures. The current study must therefore be understood as a powerful deep MIMO Hammerstein modeling and prediction framework and not as a strictly speaking physical parameter-identification approach.

6.6 Comparison with recent Hammerstein identification frameworks

The proposed method is related to a broader line of research on nonlinear and interconnected system modeling, large-scale Hammerstein estimation, and deep neural Hammerstein identifica-

tion [1, 2, 27, 12]. Within this general context, the closest comparison is with a recent sequence of Hammerstein identification studies by Li and co-authors, which differ from the present work in objective, architecture, and evidential emphasis [14, 15, 16, 17]. In particular, the attached literature includes: (i) a 2025 combined-signal MIMO Hammerstein identification framework in which the nonlinear block is modeled by a neural fuzzy model (NFM) and the dynamic block is described by an ARMAX formulation under measurement noise [14], (ii) a 2026 deep-GRU-based MIMO Hammerstein parameter-estimation framework with ARMA dynamics, recursive extended least squares (RELS), adaptive momentum estimation, and AMS-GWO hyperparameter optimization [16], and (iii) a 2026 SSAE-ARMAX Hammerstein identification framework with multi-innovation recursive extended least squares (MI-RELS), sparse deep feature extraction, and wind-power validation [17]. In addition, the 2024 noisy-output Hammerstein study of Zha *et al.* follows the same broad separation-identification philosophy through covariance-based least squares and improved particle-swarm optimization under colored output noise [15]. Taken together, these studies are primarily oriented toward *parameter estimation* and *block separation* using specially designed excitation signals, whereas the present work is positioned as a *robust deep modeling* framework for noisy MIMO Hammerstein systems with correlated measurement noise.

The comparison in Table 4 reveals both continuity and divergence. On the one hand, all of the above methods preserve the Hammerstein decomposition and exploit signal design to reduce the difficulty induced by the unmeasurable intermediate variable [14, 15, 16, 17]. On the other hand, the methodological emphasis is markedly different. The 2025 NFM-ARMAX framework is built around combined-signal separation, correlation analysis, and filtering-based recursive estimation under measurement noise [14]. The 2026 DGRU-ARMA framework extends this line toward deep recurrent nonlinear modeling, recursive estimation of the linear/noise parameters, and AMS-GWO-based hyperparameter optimization, while the 2026 SSAE-ARMAX framework emphasizes sparse deep feature extraction, MI-RELS estimation, and application-level validation on wind-power systems [16, 17]. The 2024 noisy-output Hammerstein study similarly adopts a separation-identification strategy using covariance-based least squares for the linear block and improved particle swarm optimization for the nonlinear block under colored output noise [15].

By contrast, the present work does not primarily claim exact identification of the true physical parameters. Instead, it focuses on a structured deep-learning formulation that is explicitly optimized for robust deterministic recovery and measured-output tracking under correlated measurement noise. This distinction is important. The attached studies are strongest in recursive parameter estimation and practical application validation, whereas the present study contributes a three-stage training framework, a correlated residual-noise model with heavy-tailed innovations, and five-seed stability evidence. Consequently, the present work should be interpreted as a *robust deep MIMO Hammerstein modeling* framework rather than as a direct replacement for recursive physical parameter-estimation schemes.

A direct numerical comparison with the attached studies should therefore be interpreted with caution. The synthetic targets, noise assumptions, excitation protocols, and application datasets are not identical. For example, the 2025 NFM-ARMAX framework reports improved identification accuracy over FF-SG, polynomial, and CS-TSRLS-type baselines under measurement noise and pH-process validation [14]; the 2026 DGRU study reports improved nonlinear-block approximation relative to neural fuzzy modeling together with photovoltaic prediction gains [16]; and the 2026 SSAE study reports improved wind-power prediction relative to NFM-H and LSTM-H baselines [17]. These results provide strong evidence of the maturity of signal-separated Hammerstein identification, but they are not directly comparable to the deterministic-recovery and multi-seed robustness protocol adopted here. Accordingly, the comparison in this paper is framed at the methodological and evidential level, while leaving matched benchmark replication and application-level validation as important directions for future work.

6.7 Limitations and Next Steps

The current study has two main limitations. First, although the proposed framework now includes internal baseline comparisons against both a linear ARX model and an unstructured direct GRU predictor, the benchmark remains synthetic-only and does not yet include matched replication of external recursive Hammerstein identification methods on the same dataset. Second, despite strong deterministic and measured-output performance, parameter recovery remains weak.

The most important next steps for a final high-quality submission are therefore:

1. to incorporate one or more matched external baselines from the Hammerstein identification literature under the same excitation, noise, and evaluation protocol,
2. and ideally to validate the framework on at least one additional real or public nonlinear process benchmark.

7 Conclusion

This paper has proposed a staged GRU-based framework for robust deep modeling of noisy MIMO Hammerstein systems under correlated measurement noise. The method combines a recurrent nonlinear block, a recursive linear dynamic block, and a correlated residual-noise model within a three-stage learning strategy consisting of robust initialization, deterministic recovery, and joint measured-output fine-tuning. In doing so, the framework preserves the structural bias of the Hammerstein model while improving robustness under noisy observations.

The experimental results demonstrate that the proposed method is particularly effective for deterministic synthetic recovery, achieving stable and strong performance across five independent seeds. In addition, the measured-output benchmark shows that the model remains robust under correlated noise, with limited seed-to-seed variability and competitive predictive accuracy. These findings indicate that the proposed framework is well suited for structured nonlinear modeling when reliable deterministic recovery and robust noisy-output tracking are both required.

At the same time, the parameter-recovery study reveals an important limitation: although predictive and deterministic reconstruction performance are strong, the exact recovery of the underlying linear and residual-noise parameters remains weak. Accordingly, the present approach should be interpreted as a robust deep MIMO Hammerstein modeling and prediction framework rather than as an exact physical parameter-identification method.

Future work will focus on three directions. First, beyond the internal comparisons reported here against linear and direct recurrent baselines, matched benchmarking against recursive Hammerstein identification methods from the literature will be incorporated to strengthen external comparison. Second, the framework will be extended to one or more real or public nonlinear process datasets in order to complement the current synthetic benchmark. Third, improved structural regularization, adaptive optimization strategies, and stronger parameter-anchoring mechanisms will be investigated to enhance interpretability and parameter consistency.

Data Availability

The full Python implementation used in this study is publicly available at:

<https://github.com/vegaws1/FGD/blob/18c5105bac3c46e657729dac525266e74fd9d38e/GRU>

Author Contributions

Essia Ben Alaia: Writing – original draft, software, data curation, visualization.

Slim Dhahri: Validation, analysis, writing–review and editing.

Mourad Elloumi: Methodology, formal analysis, writing–review and editing.

Afrah Alanazi: Investigation, resources, writing–review and editing.

Sahar Almenwer: Project administration, writing–review and editing.

Omar Naifar: Supervision, conceptualization, methodology, validation, writing–review and editing.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] M. Elloumi, H. Gassara, O. Naifar, An overview on modelling of complex interconnected nonlinear systems, *Mathematical Problems in Engineering* 2022 (1) (2022) 4789405.
- [2] M. Elloumi, S. Kamoun, An iterative parametric estimation method for hammerstein large-scale systems: a simulation study of hydraulic process, *International Journal of Simulation and Process Modelling* 11 (3–4) (2016) 207–219.
- [3] O. Naifar, Practical observer design for nonlinear systems using caputo fractional derivative with respect to another function, in: *Proceedings of the 2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD)*, IEEE, 2025, pp. 411–418.
- [4] O. Naifar, A. Ben Makhlouf, L. Mchiri, M. Rhaima, Finite time stability for hadamard fractional-order systems, *Ain Shams Engineering Journal* 16 (2) (2025) 103263.
- [5] O. Naifar, A. Ben Makhlouf, Synchronization of mutual coupled fractional order one-sided lipschitz systems, *Integration* 80 (2021) 41–45.
- [6] S. Dhahri, O. Naifar, Fault-tolerant tracking control for linear parameter-varying systems under actuator and sensor faults, *Mathematics* 11 (23) (2023) 4738.
- [7] S. Dhahri, O. Naifar, Robust fault estimation and tolerant control for uncertain takagi–sugeno fuzzy systems, *Symmetry* 15 (10) (2023) 1894.
- [8] A. Ben Makhlouf, O. Naifar, On the barbalat lemma extension for the generalized conformable fractional integrals: Application to adaptive observer design, *Asian Journal of Control* 25 (1) (2023) 563–569.
- [9] O. Naifar, A. Jmal, A. Ben Makhlouf, N. Derbel, M. A. Hammami, Observer-based control for fractional-order systems, in: *Fractional Order Systems—Control Theory and Applications: Fundamentals and Applications*, Springer, Cham, 2021, pp. 75–93.
- [10] O. Naifar, G. Boukettaya, On observer design of systems based on renewable energy, in: *Advances in Observer Design and Observation for Nonlinear Systems: Fundamentals and Applications*, Springer, Cham, 2022, pp. 135–176.
- [11] O. Naifar, A. Jmal, A. Ben Makhlouf, N. Derbel, M. A. Hammami, Fault estimation for nonlinear one-sided lipschitz systems, in: *Fractional Order Systems—Control Theory and Applications: Fundamentals and Applications*, Springer, Cham, 2021, pp. 95–122.

- [12] F. Li, X. Sun, T. Wang, R. Liu, Identification methodology for mimo hammerstein nonlinear model with process noise, *Transactions of the Institute of Measurement and Control* 47 (11) (2025) 2190–2202. [doi:10.1177/01423312241276072](https://doi.org/10.1177/01423312241276072).
- [13] M. Zhang, L. Jia, F. Li, Identification of the hammerstein system with time delay and colored noise, in: *Proceedings of the 2025 IEEE 14th Data Driven Control and Learning Systems (DDCLS)*, IEEE, 2025, pp. 1815–1819.
- [14] F. Li, X. Sun, Q. Cao, Parameter learning of multi-input multi-output hammerstein system with measurement noises utilizing combined signals, *International Journal of Adaptive Control and Signal Processing* 39 (7) (2025) 1416–1433. [doi:10.1002/acs.3857](https://doi.org/10.1002/acs.3857).
- [15] Q. Zha, F. Li, R. Liu, Identification of the hammerstein nonlinear system with noisy output measurements, *Control Theory and Technology* 22 (2024) 203–212. [doi:10.1007/s11768-024-00196-9](https://doi.org/10.1007/s11768-024-00196-9).
- [16] F. Li, Y. Zhu, C. Zhao, R. Liu, Parameter estimation of multi-input multi-output hammerstein nonlinear system with deep gru networks, *International Journal of Robust and Nonlinear Control* (2026). [doi:10.1002/rnc.70461](https://doi.org/10.1002/rnc.70461).
- [17] F. Li, L. Song, T. Wang, R. Liu, Parameter identification for nonlinear hammerstein models with stacked sparse autoencoder network, *Engineering Applications of Artificial Intelligence* 163 (2026) 113002. [doi:10.1016/j.engappai.2025.113002](https://doi.org/10.1016/j.engappai.2025.113002).
- [18] W. Wu, J. Hu, Z. Zhu, F. Zhang, J. Xu, C. Wang, Deterministic learning-based neural identification and knowledge fusion, *Neural Networks* 169 (2024) 165–180. [doi:10.1016/j.neunet.2023.10.004](https://doi.org/10.1016/j.neunet.2023.10.004).
- [19] P. Saha, S. Dash, S. Mukhopadhyay, Physics-incorporated convolutional recurrent neural networks for source identification and forecasting of dynamical systems, *Neural Networks* 144 (2021) 359–371. [doi:10.1016/j.neunet.2021.07.004](https://doi.org/10.1016/j.neunet.2021.07.004).
- [20] G. Yang, X. Yang, Data-based decentralized control of nonlinear-constrained interconnected systems using reinforcement learning, *Neural Networks* 187 (2025) 107780. [doi:10.1016/j.neunet.2024.107780](https://doi.org/10.1016/j.neunet.2024.107780).
- [21] B. Li, H. Wang, X. Tan, Q. Li, J. Chen, X. Qiu, Adaptive heterogeneous graph reasoning for relational understanding in interconnected systems, *The Journal of Supercomputing* 81 (1) (2025) 112. [doi:10.1007/s11227-024-06623-7](https://doi.org/10.1007/s11227-024-06623-7).
- [22] R. Meng, C. Hua, K. Li, Q. Li, Dynamic events-based adaptive nn output feedback control of interconnected nonlinear systems under general output constraint, *Neural Networks* 188 (2025) 107452. [doi:10.1016/j.neunet.2025.107452](https://doi.org/10.1016/j.neunet.2025.107452).
- [23] A. Sahoo, V. Narayanan, Differential-game for resource aware approximate optimal control of large-scale nonlinear systems with multiple players, *Neural Networks* 124 (2020) 95–108. [doi:10.1016/j.neunet.2019.12.031](https://doi.org/10.1016/j.neunet.2019.12.031).
- [24] Y. Tang, J. Ding, Y. Chen, Y. Gao, A. Jiang, C. Wang, Anxiety disorder identification with biomarker detection through subspace-enhanced hypergraph neural network, *Neural Networks* 187 (2025) 107293. [doi:10.1016/j.neunet.2025.107293](https://doi.org/10.1016/j.neunet.2025.107293).
- [25] J. P. Ntayagabiri, Y. Bentaleb, J. Ndikumagenge, H. El Makhtoum, Omic: A bagging-based ensemble learning framework for large-scale iot intrusion detection, *Journal of Future Artificial Intelligence and Technologies* 1 (4) (2025) 401–416. [doi:10.62411/faith.3048-3719-63](https://doi.org/10.62411/faith.3048-3719-63).

- [26] C. W. Lynn, Q. Yu, R. Pang, S. E. Palmer, W. Bialek, Exact minimax entropy models of large-scale neuronal activity, *Physical Review E* 111 (5) (2025) 054411. [doi:10.1103/PhysRevE.111.054411](https://doi.org/10.1103/PhysRevE.111.054411).
- [27] M. F. da Silva, M. T. da Silva, P. R. Barros, G. A. Junior, Enhanced hammerstein models identification using multiple hidden layers neural networks, *Circuits, Systems, and Signal Processing* 44 (2025) 4669–4703. [doi:10.1007/s00034-025-03029-5](https://doi.org/10.1007/s00034-025-03029-5).
- [28] O. Naifar, Tempered fractional gradient descent: Theory, algorithms, and robust learning applications, *Neural Networks* 160 (2025) 108005.
- [29] E. B. Alaia, S. Dhahri, O. Naifar, A gradient-based optimization algorithm for optimal control problems with general conformable fractional derivatives, *IEEE Access* 13 (2025) 3595958.

Algorithm 1 Three-stage training of the proposed robust deep MIMO Hammerstein model

Require: Step-response dataset $\mathcal{D}_{\text{step}} = \{(\mathbf{u}_t, \mathbf{y}_t)\}_{t=1}^{T_s}$, random-signal dataset $\mathcal{D}_{\text{rand}} = \{(\mathbf{u}_t, \mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{T_r}$, model orders n_a, n_b, n_η , hyperparameter search space Ω , candidate noise-weight set Λ_{noise} , early-stopping patience P

Ensure: Learned parameter set $\Theta^* = \{\theta_f, \{\mathbf{A}_i\}_{i=1}^{n_a}, \{\mathbf{B}_j\}_{j=1}^{n_b}, \{\Phi_\ell\}_{\ell=1}^{n_\eta}, \sigma\}$

- 1: Construct ARX regressors from $\mathcal{D}_{\text{step}}$ according to (18).
 - 2: Estimate $\{\mathbf{A}_i^{(0)}\}_{i=1}^{n_a}$ and $\{\mathbf{B}_j^{(0)}\}_{j=1}^{n_b}$ by solving the robust Huber regression problem in (19).
 - 3: Compute stage-1 residuals $\hat{\mathbf{r}}_t^{(0)}$ using (21) and estimate $\{\Phi_\ell^{(0)}\}_{\ell=1}^{n_\eta}$ using (23).
 - 4: Split $\mathcal{D}_{\text{rand}}$ into training, validation, and test subsets; normalize all channels using statistics computed from the training subset only.
 - 5: Initialize the GRU block, recursive linear dynamic block, and residual-noise component using $\{\mathbf{A}_i^{(0)}\}$, $\{\mathbf{B}_j^{(0)}\}$, and $\{\Phi_\ell^{(0)}\}$.
- ▷ **Hyperparameter search**
- 6: **for** each candidate configuration $\omega \in \Omega$ **do**
 - 7: Train the model for a short Stage-A run using the deterministic objective (34).
 - 8: Compute the validation score defined in (41).
 - 9: **end for**
 - 10: Select the best configuration ω^* with minimum validation score and reinitialize the model accordingly.
- ▷ **Stage A: deterministic recovery**
- 11: Optimize (34) on the training set using the noise-free target $\mathbf{x}_t$.
 - 12: Save the model parameters with the lowest validation loss under early stopping.
- ▷ **Noise-weight tuning**
- 13: **for** each candidate $\lambda_{\text{noise}} \in \Lambda_{\text{noise}}$ **do**
 - 14: Starting from the best Stage-A model, run a short Stage-B refinement using (35).
 - 15: Compute the validation score in (41).
 - 16: **end for**
 - 17: Select the optimal residual-noise weight λ_{noise}^* with minimum validation score.
- ▷ **Stage B: linear/noise refinement**
- 18: Freeze the nonlinear GRU block and optimize (35) with λ_{noise}^* to refine the linear dynamic and residual-correlation parameters.
 - 19: Save the best validation model under early stopping.
- ▷ **Stage C: joint fine-tuning**
- 20: Unfreeze all parameters and jointly optimize (36) using λ_{noise}^* .
 - 21: Save the final parameter set Θ^* with the best validation performance.
- ▷ **Evaluation**
- 22: Compute deterministic-recovery metrics by comparing $\hat{\mathbf{x}}_t$ against $\mathbf{x}_t$.
 - 23: Compute measured-output tracking metrics by comparing $\hat{\mathbf{x}}_t$ against $\mathbf{y}_t$.
 - 24: Report RMSE, R^2 , FIT, and wMAPE for both evaluation settings.
-

Table 4: Comparison with recent Hammerstein identification frameworks.

Aspect	Li et al. (2025): NFM-ARMAX	Li et al. (2026): DGRU-ARMA / SSAE-ARMAX	This work
Primary objective	Parameter learning of MIMO Hammerstein systems under measurement noise using combined signals [14]	Parameter estimation / identification of deep Hammerstein models, including DGRU-ARMA and SSAE-ARMAX formulations [16, 17]	Robust deep modeling of noisy MIMO Hammerstein systems with explicit deterministic recovery and measured-output tracking
Nonlinear block	Neural fuzzy model (NFM), coupled nonlinear subsystem [14]	Deep GRU network (DGRU) or stacked sparse autoencoder (SSAE), both used as deep nonlinear approximators [16, 17]	GRU-based nonlinear block with latent-dimension alignment and structured staged training
Linear/noise block	ARMAX with moving-average measurement-noise model and filtering-based recursive estimation [14]	ARMA / ARMAX dynamic block with recursive extended least squares or multi-innovation recursive least squares [16, 17]	ARX-type recursive dynamic block with correlated residual-noise model and Student- t innovations
Signal strategy	Combined signals (separable/Gaussian + random signals) for decoupled identification [14]	Step signals for linear-block decoupling, random signals for nonlinear-block learning [16, 17]	Step-signal robust initialization followed by random-signal three-stage learning
Optimization / estimation	Correlation analysis + data filtering + recursive least squares [14]	RELS + adaptive momentum + AMS-GWO (DGRU), or MI-RELS + layer-wise pretraining / fine-tuning (SSAE) [16, 17]	Hyperparameter search + Huber initialization + deterministic recovery + noise-weight tuning + joint fine-tuning
Evidence protocol	Synthetic 2×2 numerical case + pH-neutralization process [14]	Synthetic benchmark + photovoltaic-cell case (DGRU) or wind-power case (SSAE) [16, 17]	Five-seed synthetic benchmark with separate deterministic and measured-output evaluation
Reported strengths	Effective block separation, improved robustness to measurement noise, better nonlinear approximation than polynomial / CS-TSRLS baselines [14]	Improved nonlinear approximation over shallow models; application-level predictive gains on photovoltaic and wind-power benchmarks [16, 17]	Stable robust modeling under correlated measurement noise and strong deterministic synthetic recovery with explicit multi-seed evidence
Main limitation	Dependence on signal design, filtering assumptions, and noise-model structure [14]	Computational burden, extensive tuning, and continued emphasis on exact parameter estimation [16, 17]	Weak exact parameter recovery and no real-process benchmark yet