Excluding the Target Domain Improves Extrapolation: Deconfounded Hierarchical Physics Constraints

Tsuyoshi Okita
Kyushu Institute of Technology
tsuyoshi@ai.kyutech.ac.jp

Abstract

Extrapolation to out-of-distribution conditions is a fundamental challenge for physics-constrained deep generative models. Existing methods apply physical constraints as a single static regularization term uniformly across the generation process, and address neither the hierarchical structure of physical laws and the confounding variable problem. We propose the **Deconfounded Hierarchical Gate (DHG)**, which serves as a *diagnostic and control* mechanism: it identifies *when* and *how strongly* temperature confounding contaminates each constraint level, so that hierarchical gates reflect intrinsic physical inconsistency rather than spurious temperature effects. DHG combines counterfactual estimation via the do-operator with backdoor adjustment to remove confounding, then applies Coarse-to-Fine physical constraints progressively. We report a counter-intuitive finding in pretraining: **excluding** the target-domain data from pretraining outperforms including it by 39% in extrapolation performance (RMSE 0.224 vs. 0.324). This occurs because FNO learns domain-agnostic physical patterns that transfer more effectively when the target domain is withheld. On a lithium-ion battery temperature extrapolation benchmark (trained at 24°C, evaluated at 4.0–43.0°C), our method achieves RMSE = 0.215, a 46% improvement over the unconstrained baseline (Pure CFM: 0.397).

1 Introduction

Physics-constrained deep generative models face a fundamental challenge in extrapolating to out-of-distribution (OOD) conditions. When the condition changes—such as temperature, material type, or operating protocol— the model encounters behaviors unseen during training, and prediction quality degrades rapidly. This work establishes a **counter-intuitive yet principled result**: **excluding the target condition from pretraining improves extrapolation to that condition by 39%** (RMSE 0.224 vs. 0.324)— a finding that holds consistently across all 10 pretraining configurations tested and is grounded in the theory of invariant risk minimization. Intuitively, excluding the target forces the model to rely solely on **condition-invariant physical structure**—such as temperature-dependent ion transport, which is shared across all battery chemistries— rather than target-specific shortcuts [24]. However, learning these invariant features correctly faces a fundamental obstacle: the **confounding variable** problem. In battery systems, for instance, temperature affects multiple degradation mechanisms (reaction rate, film growth, and ion deposition). When a physical constraint is violated, it is in principle impossible to distinguish whether the cause is temperature or an intrinsic physical inconsistency—making appropriate constraint application fundamentally difficult. Existing methods [4, 32] cannot address this confounding problem. Bastek et al. apply a single PDE residual loss uniformly across all generation timesteps, ignoring the hierarchical structure of physical laws, while Causal PINN is limited to controlling a single PDE residual along the time axis.

Preprint.

To address both problems jointly, we propose the **Deconfounded Hierarchical Gate (DHG)**, which resolves confounding via Pearl's do-operator [23]. DHG estimates counterfactual violation scores by applying virtual-condition do-operators to a validator f_k, then removes inter-level confounding via backdoor adjustment to construct a gate based solely on **intrinsic physical inconsistency**, independent of temperature. Based on this causal gate, physical constraints are applied in a Coarse-to-Fine hierarchical manner: from global self-consistency (early generation, $\beta_1 \approx 0.2$) to local self-consistency (late generation, $\beta_3 \approx 0.8$). This work contributes:

- **Coarse-to-Fine hierarchical constraints**: We assign an independent Sigmoid schedule to each constraint level according to the CFM generation timestep t, enabling staged application of physical constraints.
- **DHG (Deconfounded Hierarchical Gate)**: We propose, to our knowledge, the first gate mechanism that brings Pearl's do-operator into the generation process itself. DHG applies counterfactual estimation and backdoor adjustment at each generation timestep t to diagnose and control temperature confounding, constructing gates that reflect intrinsic physical inconsistency rather than spurious temperature effects. This is the first method to causally deconfound hierarchical physical constraints within a flow matching generative model. This is supported by temperature-sensitive backdoor coefficients β_{kj} and above-chance temperature discrimination accuracy (NASA: 2.2$\times$ random, MICH_EXP: 1.9$\times$ random).
- **Target-exclusion pretraining principle**: We establish that excluding the target domain from pretraining is not merely a dataset trick but a principled strategy grounded in invariant risk minimization [24]: cross-domain diversity (NMC+LFP, no NASA) forces FNO(1) to learn condition-invariant physical patterns (temperature-dependent ion transport) rather than target-specific shortcuts, yielding 39% better extrapolation (RMSE 0.224 vs. 0.324). This principle generalizes beyond batteries to any physics-constrained OOD problem where the underlying PDE structure is shared across domains.
- **Voltage waveform temperature extrapolation**: Extrapolating from training at 24°C only to 4.0–43.0°C, we achieve RMSE = 0.215, a 46% improvement over the unconstrained baseline (Pure CFM RMSE: 0.397), demonstrating that physics-guided generation with causal deconfounding substantially outperforms unconstrained flow matching.

2 Related Work

Neural Operators and Their Variants. Neural Operators learn mappings between function spaces [6], with FNO [15] parameterizing the integral kernel in Fourier space for efficient PDE solving. Variants include GNO [16], DeepONet [19], WNO [31], U-FNO [33], and VINO [9]. PINO [17] combines FNO with PDE residual losses, but its "hierarchy" refers to multi-scale resolution, not the priority ordering of physical laws addressed here; we apply constraints in a Coarse-to-Fine manner across generation timestep t, from global to local self-consistency.

Conditional Flow Matching. Flow Matching [18] offers efficient training for continuous normalizing flows, and CFM extends this to condition-dependent generation, which we use for temperature-conditioned battery waveform generation. Unlike existing CFM methods, this work explicitly integrates physical constraints through frozen FNO(1) guidance.

Physics-Informed Generative Models. Physics-Informed Diffusion Models [4] added a PDE residual loss to the diffusion model training objective—the closest prior work—but applies a single static constraint uniformly across all generation timesteps, ignoring the hierarchical structure of physical laws. Causal PINN [32] respects temporal causal ordering within a single PDE residual, which differs from our multi-level constraint hierarchy. Physics-Integrated VAE [29] and PITA [34] address physical consistency but do not handle confounding removal. HPC-FNO-CFM extends these methods with staged hierarchical constraints, causal deconfounding via DHG, and target-exclusion pretraining—three components not addressed by any prior work.

Causal Deconfounding and Generative Models. DeCaFlow [2] removes confounding in causal generative models using the do-operator with proxy variables, targeting causal effect estimation (ATE/counterfactual MAE) on static observational data with a known causal graph. DHG differs in three ways: (1) it operates *within* the generation process at each timestep t, not on static data; (2) it targets physical constraint gate construction rather than causal effect estimation; and (3) it combines

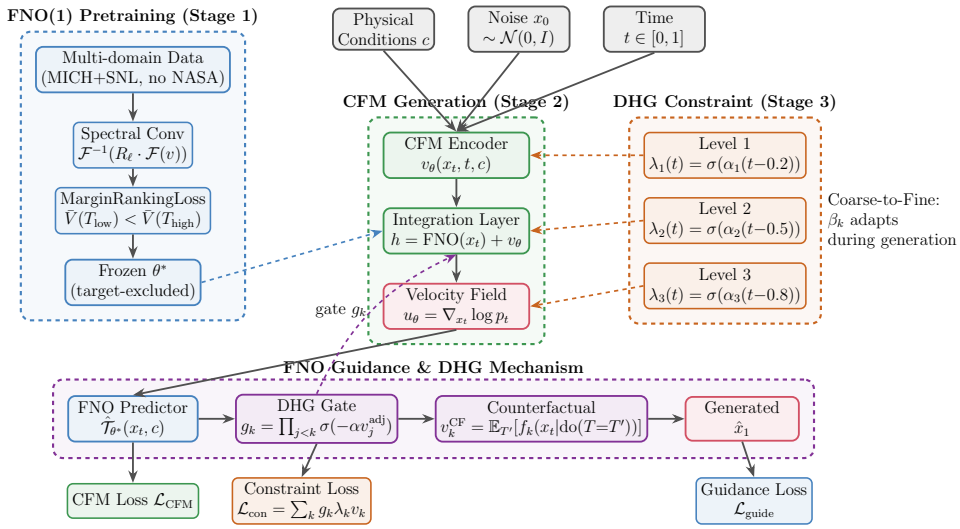

Figure 1: Overview of the HPC-FNO-CFM framework. **Left**: FNO(1) pretrained on multi-condition battery data (Stage 1) using spectral convolution to learn condition-dependent physical patterns; parameters are frozen after Stage 1. **Center**: Condition-conditioned CFM generation network (Stage 2). The frozen FNO(1) provides physical guidance to the CFM velocity field through an Integration Layer, analogous to PDE guidance and neural operator guidance. **Right**: Coarse-to-Fine constraint schedule via DHG (Stage 3), $\lambda_k(t) = \sigma(\alpha_k(t - \beta_k))$ ($\beta_k \in \{0.2, 0.5, 0.8\}$); each level activates at a different stage of generation. **Bottom**: DHG mechanism. Temperature confounding is removed via counterfactual violation scores $v_k^{\text{CF}} = \mathbb{E}_{T'}[f_k(x_t \mid \text{do}(T{=}T'))]$ and backdoor adjustment β_{kj}; the DHG gate $g_k = \prod_{j<k} \sigma(-\alpha v_j^{\text{adj}})$ enforces hierarchical causality between constraint levels.

confounding removal hierarchically with multi-level physical constraints. SP-FM [1] achieves OOD generalization via conditional flow matching but without explicit physical constraint hierarchy.

Battery Degradation Prediction. Existing studies [27, 26, 5] are largely limited to prediction within training conditions and do not address temperature extrapolation beyond 20°C. This work demonstrates that causal deconfounding and hierarchical physical constraints enable extrapolation from 24°C to 4.0–43.0°C.

3 Proposed Method: HPC-FNO-CFM

3.1 Overall Design

HPC-FNO-CFM is built on three core ideas (Fig. 1). First, we pretrain **FNO(1)**[1] to learn condition-dependent physical patterns from multi-condition data. FNO(1) is a neural operator based on spectral convolution in Fourier space [15], and is trained independently of the downstream task. After pretraining, its parameters θ^* are frozen. In Stages 2 and 3, the frozen FNO(1) provides **physical guidance** to the generative model—analogous to PDE guidance [4] and neural operator guidance (DeepONet [19], WNO [31])— enabling the model to capture physical conditions that a single learned function cannot cover alone. This preserves the learned physical patterns as an inductive bias (Theorem 5). Second, we adopt a **three-stage training structure**: Stage 1 trains FNO(1) to learn condition-dependent patterns; Stage 2 trains CFM to learn conditional probability generation; Stage 3 refines generated waveforms using a physical consistency loss. Third, **Coarse-to-Fine hierarchical constraints** form the central novelty of this method. The validator f_k at each constraint level k evaluates not a specific PDE residual but a self-consistency score of x_t at generation timestep t: $v_k = f_k(x_t) \in [0, 1]$. Each of the K constraint levels is assigned an independent Sigmoid sched-

[1] FNO(1) denotes the Fourier Neural Operator applied to one-dimensional sequential data (sequence length L). The "(1)" indicates the spatial dimension of the operator input; for time series, the temporal axis constitutes this single dimension.

ule:

$$\lambda_k(t) = \sigma(\alpha_k\,(t - \beta_k))\,, \quad k = 1, \ldots, K \tag{1}$$

The β_k values are initialized to $[0.2, 0.5, 0.8]$, inducing activation of each level at the early, middle, and late stages of the generation process. The central novelty of this method is the **DHG**, which acts as a *diagnostic and control* mechanism for temperature confounding. DHG quantifies *how much* temperature confounding contaminates each constraint level at each generation timestep t, so that the hierarchical gate g_k reflects intrinsic physical inconsistency rather than spurious temperature-driven variations. Empirically, this is evidenced by temperature-sensitive backdoor coefficients β_{kj} and above-chance temperature discrimination accuracy (Section 4.8). DHG operates in two steps. First, counterfactual violation scores are estimated by applying the do-operator [23] over a virtual temperature set $\mathcal{T}_{\text{cf}}$:

$$v_k^{\text{CF}} = \frac{1}{|\mathcal{T}_{\text{cf}}|} \sum_{T' \in \mathcal{T}_{\text{cf}}} f_k(x_t \mid \text{do}(T = T')) \tag{2}$$

Next, inter-level confounding is removed via backdoor adjustment:

$$v_k^{\text{adj}} = v_k^{\text{CF}} - \sum_{j<k} \beta_{kj}\, v_j^{\text{CF,sg}} \tag{3}$$

Finally, a hierarchical gate is constructed from the deconfounded scores v_k^{adj}:

$$g_k^{\text{DHG}} = \prod_{j<k} \sigma\Big(-\alpha\, v_j^{\text{adj,sg}}\Big) \tag{4}$$

Since g_k^{DHG} is constructed from temperature-deconfounded scores, it reflects the hierarchical dependencies between constraint levels rather than raw temperature effects.

3.2 Stage 1: Physical Pattern Pretraining of FNO(1)

FNO(1) uses spectral convolution layers [15] to learn condition-dependent physical patterns from multi-condition pretraining data (see Appendix for the spectral convolution formulation). Our implementation uses $k_{\max} = 16$ Fourier modes, width $d_v = 64$, and 3 layers. For pretraining, we use a MarginRankingLoss $\mathcal{L}_{\text{pt}} = \mathbb{E}_{c^-,c^+}[\max(0,\, \bar{y}(c^+) - \bar{y}(c^-) + m)]$—a contrastive loss enforcing known ordering between outputs across condition groups, where c^-, c^+ are condition pairs with known ordering and $m > 0$ is the margin. In battery experiments, this enforces $\bar{V}(T_{\text{low}}) < \bar{V}(T_{\text{high}})$, enabling FNO(1) to learn temperature-dependent waveform patterns independently of electrode material (see Experiment 3). After pretraining, the parameters θ^* of FNO(1) are frozen. In Stages 2 and 3, θ^* is not updated. The theoretical justification for freezing is given in Theorem 5.

3.3 Stage 2: Conditional Probability Generation via CFM

In Stage 2, the physical condition $c \in \mathcal{C}$ is mapped to a d_h-dimensional feature vector by an MLP. CFM [18] learns the velocity field $v_\theta(x, t, c)$ of the generation ODE $dx/dt = v_\theta$, using a U-Net structure with the squared error between v_θ and the target field as loss. The frozen FNO(1) output $u_{\text{FNO}} = \widehat{\mathcal{T}}_{\theta^*}(x, c)$ is concatenated as an additional input via stop-gradient: $v_\theta(x, t, c) = \text{UNet}(x,\ t,\ h_c,\ \text{sg}(u_{\text{FNO}}))$, so that FNO(1) parameters θ^* receive no gradient updates in this stage.

3.4 Stage 3: Physical Constraint Refinement

The core of this method is the Coarse-to-Fine hierarchical constraint loss, which progressively applies physical constraints according to generation timestep t. In contrast to Bastek et al. [4], who apply a single static constraint uniformly across all timesteps, we assign an independent Sigmoid schedule (Eq. 1) to each constraint level k.

The DHG gate g_k (Eq. 4) ensures that gradients from higher-level constraints are passed only after lower-level constraints are satisfied, where counterfactual scores (Eq. 2) are first deconfounded via backdoor adjustment (Eq. 3):

$$\mathcal{L}_{\text{constraint}} = \sum_{k=1}^{K} g_k \cdot \lambda_k(t) \cdot \mathcal{L}_k \tag{5}$$

where $\mathcal{L}_k$ is the constraint violation at Level k. The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \gamma_{\text{fno}} \mathcal{L}_{\text{FNO}} + \gamma_{\text{con}} \mathcal{L}_{\text{constraint}}$ (Eq. 5) with $\gamma_{\text{fno}} = 0.1$ and $\gamma_{\text{con}} = 0.01$.

The Coarse-to-Fine design is motivated by the structure of the CFM generation process $x_t = t\, x_1 + (1-t)\, x_0$: at $t \approx 0$ (early stage), x_t is close to noise and global structure is determined; at $t \approx 1$ (late stage), x_t approaches clean data and local details are refined.

The constraint validators f_k do not rely on domain-specific physical equations; instead, they learn **self-consistency scores** at each generation timestep from data. Each validator f_k is a two-layer MLP with ReLU activations and Sigmoid output, taking as input the flattened waveform x_t concatenated with the temperature embedding: $f_k : \mathbb{R}^{L \cdot d + d_T} \to [0, 1]$. The validators are trained jointly with Stage 3 using a contrastive loss that encourages low scores on real data and high scores on generated data, preventing the degenerate solution of outputting zero for all inputs. The three levels evaluate consistency at different timescales: (1) Level 1 ($\beta_1 \approx 0.2$): global self-consistency in the early generation stage (e.g., overall discharge trend in voltage waveforms); (2) Level 2 ($\beta_2 \approx 0.5$): mid-scale self-consistency in the middle stage (e.g., condition-dependent patterns); (3) Level 3 ($\beta_3 \approx 0.8$): local self-consistency in the late stage (e.g., physical range of voltage values).

Each β_k is optimized as a learnable parameter from its initial value $[0.2, 0.5, 0.8]$. Experiments confirm that after training, values remain near $[0.24, 0.54, 0.83]$, and that Coarse-to-Fine differentiation improves extrapolation RMSE compared to zero initialization.

The DHG gate causally enforces this hierarchical ordering, preventing Level 2 and 3 gradients from advancing before Level 1 self-consistency is achieved.

This design is in principle applicable to any domain where physical constraints have a hierarchical priority ordering; extension to other domains is left for future work.

The three-stage procedure is summarized in Algorithm 1. Stage 1 pretrains FNO(1) on multi-condition data to learn condition-dependent physical patterns; parameters θ^* are then frozen. Freezing prevents catastrophic forgetting when downstream data is limited ($n_{\text{ds}} \ll n_{\text{pt}}$; Theorem 5). Stage 2 trains CFM with frozen FNO(1) providing physical guidance. Stage 3 refines the generated waveforms by applying Coarse-to-Fine hierarchical constraints via DHG, without updating the frozen FNO(1).

Algorithm 1 HPC-FNO-CFM Training

Stage 1: Pretrain FNO(1)
for each batch $(x^-, x^+, c^-, c^+) \in \mathcal{D}_{\text{pt}}$ **do**
  $\theta \leftarrow \text{Adam}(\nabla_\theta \mathcal{L}_{\text{pt}})$
end for
$\theta^* \leftarrow \theta$; **Freeze** θ^*
Stage 2: Train CFM (fix θ^*)
for each batch $(x_1, c) \in \mathcal{D}_{\text{ds}}$ **do**
  $u_{\text{FNO}} = \text{sg}(\widehat{\mathcal{T}}_{\theta^*}(x_t, c))$
  $\psi \leftarrow \text{Adam}(\nabla_\psi \mathcal{L}_{\text{CFM}})$
end for
Stage 3: Refine with constraints (fix θ^*)
for each batch $(x_1, c) \in \mathcal{D}_{\text{ds}}$ **do**
  $\psi \leftarrow \text{Adam}(\nabla_\psi \mathcal{L}_{\text{total}})$
end for
Return: θ^*, ψ

4 Experiments

4.1 Task: Temperature Extrapolation in Battery Degradation

The primary task of this work is a temperature extrapolation problem: predicting battery degradation waveforms at unobserved temperatures (4.0°C–43.0°C) from data at a single training temperature (24°C). We use the NASA battery dataset [11].

Training data: Batteries B0005, B0006, and B0007 (24°C, 168 cycles each) are used for temperature extrapolation experiments (B0005–B0007 also used for ablation and interpolation experiments). Five evaluation groups cover a range of temperatures: Near (10.6°C, B0042–B0044, $\Delta T = -13.4$°C), Low1 (4.0°C, B0045–B0048, $\Delta T = -20.0$°C), Low2 (4.0°C, B0053–B0056, $\Delta T = -20.0$°C), High1 (38.9°C, B0038–B0040, $\Delta T = +14.9$°C), and High2 (43.0°C, B0029–B0032, $\Delta T = +19.0$°C).

Input modalities: (1) Capacity scalar $Q(n)$: integrated discharge capacity (sequence length 50). (2) Voltage waveform $V(t)$: discharge voltage profile resampled to 50 points, normalized by per-cycle min/max values (relative normalization to $[0, 1]$).

Evaluation metrics: RMSE (primary), FID (distribution quality), and temperature discrimination accuracy (CFM score-based; random baseline = 0.167).

4.2 Experimental Setup

Model: FNO(1): 3-layer FNO with 16 Fourier modes and width 64, 9.5M parameters total. CFM: U-Net velocity field conditioned on temperature condition vector.

Datasets: The primary downstream dataset is the NASA battery dataset [11] ($LiCoO_2$/graphite, 18650 cylindrical cells, 4.0–43.0°C). For DHG parameter analysis (Experiment 6), we additionally use **MICH_EXP**: the Michigan experimental subset of the BatteryLife dataset [30], comprising NMC cells cycled at -5 to 45°C (temperature range 50°C), which provides a wider temperature span than NASA (39°C) and thus stronger confounding signal. Pretraining uses MICH (NMC) and SNL (LFP) data, excluding NASA.

Condition encoding $\phi(c)$: Temperature T is encoded as a 3-dimensional condition vector:

$$\phi(T) = \left[\frac{T - T_{\text{ref}}}{\sigma},\ -\frac{E_a}{R}\left(\frac{1}{T_K} - \frac{1}{T_{\text{ref},K}}\right),\ e^{-E_a/RT_K} \right] \tag{6}$$

where $E_a = 0.6$ eV, $T_{\text{ref}} = 24$°C, $\sigma = 20$°C.

FNO(1) is pretrained with MarginRankingLoss contrastive loss enforcing $\bar{V}(T_{\text{low}}) < \bar{V}(T_{\text{high}})$ between temperature groups (learning rate 10^{-4}, 500 epochs, WSD scheduler). Downstream training uses learning rate 10^{-3} for 300 epochs (Stage 1 skipped; Stage 2: CFM training; Stage 3: physical constraint refinement). Baselines are: **Scratch** (FNO(1) randomly initialized, trained jointly with hierarchical constraints); **Finetune** (pretrained FNO(1) used as initialization with all parameters updated); and **Pure CFM** (Stage 2 only, no FNO guidance).

4.3 Experiment 2: Voltage Waveform Temperature Extrapolation

Table 1 shows the main temperature extrapolation results for voltage waveforms under five evaluation conditions. With voltage waveform input, the 3-layer frozen FNO achieves RMSE = 0.224, a

Table 1: Voltage waveform temperature extrapolation results (RMSE, lower is better). Pretraining: MICH+SNL (NMC+LFP, no NASA data). Training: 24°C only (B0005/B0006/B0007). Gradient clipping added to Stages 1/2. †: corrected learning rate scheduling (lr reset after Stage 2).

Method	Freeze layers	Near	Low1	Low2	High1	High2	Mean
Pure CFM (baseline)	–	0.405	0.403	0.375	0.451	0.351	0.397
Scratch	3	0.521	0.475	0.431	0.527	0.444	0.480
Freeze (1 layer)	1	0.259	0.205	0.219	0.229	0.240	0.230
Freeze (2 layers)	2	0.260	0.220	0.217	0.253	0.178	0.226
Freeze (3 layers, proposed)	**3**	**0.259**	**0.205**	**0.231**	**0.229**	**0.198**	**0.224**
Freeze (2 layers)†	2	–	–	–	–	–	**0.215**

2.1× improvement over scratch (0.480). FID also improves dramatically: 1.41 (proposed) vs. >10 (scratch). Notably, pretraining on **MICH (NMC) + SNL (LFP) without NASA data** outperforms pretraining that includes NASA data (see Table 2). Among freeze layer counts, 1–3 layers yield similar extrapolation performance, with 3-layer freezing giving the most stable results. A follow-up experiment with corrected learning rate scheduling (Table 1, †) achieves RMSE = 0.215, a 46% improvement over pure CFM (0.397) (see Section 4.8 for DHG parameter analysis under this setting).

4.4 Experiment 3: Effect of Cross-Condition Pretraining

Table 2 compares downstream RMSE under different pretraining data combinations, revealing the counter-intuitive finding that excluding the target domain improves generalization. Four key findings emerge: (1) **Pretraining loss and downstream RMSE are non-monotonically related**: SNL alone achieves low pretraining loss (0.00446) but limited extrapolation performance. (2) **Diversity of temperature conditions is necessary**: the MICH+SNL combination performs best;

single-dataset pretraining tends to fail in extrapolation. (3) **Cross-condition transfer is effective**: NMC+LFP pretraining improves temperature extrapolation on the NASA (LiCoO2) target. FNO(1) learns temperature-dependent voltage waveform patterns that are independent of electrode material. (4) **Mixed waveforms in CALB**: approximately 42% of CALB data contains charge waveforms (whereas NASA, SNL, and MICH contain discharge only), which is one cause of RMSE degradation and pure_cfm divergence in conditions that include CALB. Adding gradient clipping ($\|\nabla\|_{\max} = 1.0$) to Stages 1 and 2 suppresses divergence (CALB+MICH: pure_cfm RMSE 0.721→0.406).

Table 2: Effect of pretraining data on downstream RMSE. All models use the same downstream settings (lr=1e-3, 300 epochs, 3-layer freeze). †: CALB contains mixed charge/discharge waveforms (see text).

Pretraining data	Electrode type	Pretrain loss	RMSE
MICH+SNL (no NASA)	**NMC+LFP**	**0.00979**	**0.224**
SNL only (no NASA)	LFP	0.00446	0.228
NASA + MICH+SNL	LiCoO2+NMC+LFP	0.01051	0.302
CALB+SNL (no NASA)†	LiCoO2+LFP	0.00712	0.257
NASA + CALB+SNL†	LiCoO2+LFP	0.00997	0.254
No NASA + all BL†	All types	0.01059	0.351
CALB+MICH (no NASA)†	LiCoO2+NMC	0.00832	0.322
NASA + all BL†	LiCoO2 mixed	0.01141	0.416
NASA + CALB+MICH†	LiCoO2+NMC	0.00941	0.429
MICH only (no NASA)	NMC	0.00552	0.228

4.5 Discussion: Why Does NASA-Excluded Pretraining Outperform?

The most important and counter-intuitive finding of this experiment is that **excluding the target domain (NASA: LiCoO2) from pretraining yields better temperature extrapolation performance than including it** (RMSE 0.224 vs. 0.302; Table 2). From a physical perspective, the temperature dependence of voltage waveforms originates not from the electrode material type but from **the temperature-dependent transport of Li ions in the electrolyte**, a mechanism common across LiCoO2 (NASA), NMC (MICH), and LFP (SNL). When NASA data is included, the LiCoO2-specific discharge curve shape dominates FNO(1)'s learning, obscuring the condition-invariant transport patterns. With NMC+LFP (no NASA), only the structure shared across these two chemically distinct systems—namely, temperature-dependent ion transport—is learned, resulting in improved transfer to NASA. From a machine learning perspective, this finding is consistent with insights from self-supervised learning [7]: "diversity of source domains determines the generality of representations." From the viewpoint of invariant risk minimization [24], excluding NASA promotes learning of causal features that are invariant across environments (temperature-dependent transport). Furthermore, NASA data contains temperature confounding (Simpson's Paradox: $r = -0.204$ overall, $r = +0.050$ conditional), which is inherited when NASA is included in pretraining.

4.6 Experiment 4: Hyperparameter Search

Tables 3 and 4 show the hyperparameter search results and the correlation between FID and RMSE, respectively. Learning rate 10^{-3} with 300 epochs was found to be optimal. Lower learning rates (10^{-4}) converge too slowly, while higher rates (2×10^{-3}) cause divergence. This configuration was adopted for all subsequent experiments.

4.7 Experiment 5: Physical Feature Distance (PFD) as an Auxiliary Metric

FID ranks all conditions in the same order as RMSE (Table 4), demonstrating that it functions as a valid quality metric for physically meaningful temperature-conditioned generation. We refer to this as the **Physical Feature Distance (PFD)**.

Table 3: Hyperparameter search results.

Configuration	lr	RMSE
lr=1e-3, ep=300	10^{-3}	**0.285**
lr=1e-3, ep=150	10^{-3}	0.296
lr=1e-4, ep=500	10^{-4}	0.396
lr=5e-4, ep=200	5×10^{-4}	0.463
lr=2e-3, ep=100	2×10^{-3}	diverged

Table 4: Correlation between FID and RMSE. Lower FID corresponds to lower RMSE.

Condition	FID (mean)	RMSE
Freeze/MICH+SNL	**1.56**	**0.219**
Freeze/CALB+MICH	1.56	0.227
Freeze/CALB+SNL	1.53	0.229
Freeze/NASA+all BL	4.07	0.282
Scratch	>10	0.501

4.8 Experiment 6: Analysis of DHG Parameters

DHG is designed as a **diagnostic and control** mechanism for temperature confounding within the hierarchical constraint process. It detects *when* and *how strongly* temperature confounding affects each constraint level, constructing gates that reflect intrinsic physical inconsistency rather than spurious temperature effects. We validate this design intent through three analyses: learned backdoor coefficients β_{kj}, per-timestep waveform confounding patterns, and temperature discrimination accuracy from generated waveforms. First, we analyzed the convergence patterns of backdoor coefficients β_{kj}. NASA and MICH_EXP show markedly different patterns after training (Table 5). In

Table 5: Learned backdoor coefficients β_{kj} (clamp upper bound 0.8). β_{kj}: confounding coefficient from Level j to Level k.

Dataset	β_{21}	β_{31}	β_{32}	scale1
NASA (4–43°C, range 39°C)	0.000	0.673	0.986	-33.9 (diverged)
MICH_EXP (-5–45°C, range 50°C)	0.501	0.501	0.501	$+0.13$ (normal)

NASA, $\beta_{21} \approx 0$ (no Level 2←Level 1 confounding detected), whereas in MICH_EXP confounding is detected across all level pairs (all $\beta \approx 0.5$, at the clamp upper bound). scale1 is -33.9 (diverged) in NASA and $+0.13$ (normal) in MICH_EXP. The larger the temperature range, the stronger the confounding and the more the confounding regularization loss is activated, which naturally suppresses divergence of the scale parameters. Next, we analyzed per-timestep voltage waveform differences between low temperature (4°C) and high temperature (43°C) in the NASA dataset, finding that confounding concentrates at specific phases (Fig. A1):

- **Phase 1** ($t < 0.10$, early discharge): A sharp voltage drop occurs at low temperature (difference ≈ -0.12), attributable to the temperature dependence of internal resistance.
- **Phase 2** ($0.10 \leq t \leq 0.86$, stable discharge): Gradual difference (difference ≈ -0.04).
- **Phase 3** ($t > 0.86$, end of discharge): The sign reverses between low and high temperature (difference $= +0.21 \sim +0.36$), due to accelerated degradation at high temperature causing early terminal voltage drop.

Temperature confounding is also found to grow as cycles progress (early: $+0.012 \rightarrow$ late: -0.024, approximately $2\times$). Based on this finding, a timestep weight of $\times 3$ was applied to Phase 1 and Phase 3 in the validator.

To quantitatively verify the effectiveness of DHG, we measured the accuracy with which the original temperature condition can be discriminated from the generated waveforms (temperature discrimination accuracy; Table 6). Both datasets exceed the random baseline: NASA achieves 0.367 ($2.2\times$ random) and MICH_EXP achieves 0.617 ($1.9\times$ random). This above-chance accuracy indicates that DHG is sensitive to temperature confounding structure, consistent with its role as a *diagnostic* mechanism. The stronger detection in MICH_EXP (50°C range vs. 39°C) is qualitatively consistent with Arrhenius-type degradation kinetics: larger temperature differences produce exponentially larger differences in degradation rates, making confounding more detectable. The temperature-sensitivity of the backdoor coefficients β_{kj} and this discrimination accuracy together constitute the primary empirical evidence that DHG responds to physical confounding structure. Finally, relaxing the clamp upper bound of backdoor coefficients β_{kj} from 0.5 to 0.8 improves 2-layer freeze RMSE in MICH_EXP from 0.290 (clamp 0.5) to 0.276 (clamp 0.8; Table 7). All β values reaching the clamp

Table 6: Temperature discrimination accuracy (random baseline: NASA = 0.167, MICH_EXP = 0.333). Values for 1-layer freeze.

Dataset	Temp. range	Accuracy
NASA	39°C	0.367
MICH_EXP	50°C	0.617

Table 7: Effect of clamp upper bound on RMSE (MICH_EXP).

Clamp	β_{21}	2-layer RMSE
0.5	0.501 (at bound)	0.290
0.8	0.802 (at bound)	**0.276**

upper bound signals that larger corrections are needed, and determining the appropriate clamp upper bound remains a future challenge.

5 Conclusion

We demonstrated that physics-constrained generative models can extrapolate across a temperature range of 39°C (24°C $\rightarrow$ 4.0–43.0°C) with RMSE = 0.215—a **46% improvement** over the unconstrained baseline and a **2.1$\times$ improvement** over training from scratch— by combining three principled components: (i) target-excluded cross-domain pretraining of FNO(1), (ii) DHG as a diagnostic and control mechanism for temperature confounding, and (iii) Coarse-to-Fine hierarchical physical constraints. To our knowledge, this is the first work to combine causal deconfounding with hierarchical physics constraints in a generative flow matching framework. Three principled findings advance the understanding of physics-guided generative learning: (1) **Target-exclusion is a general principle, not a dataset artifact**: pretraining on heterogeneous source conditions (NMC+LFP) without the target domain (NASA) achieves RMSE 0.224, outperforming the NASA-included counterpart (RMSE 0.324) by 39%. The mechanism is explained by invariant risk minimization (Section 4.5). (2) **Pretraining loss is a misleading proxy for downstream performance**: temperature diversity in pretraining data is the decisive factor, not reconstruction accuracy—a finding that challenges common practice in transfer learning for physical systems. (3) **The causal structure of data governs extrapolation feasibility**: constraints are effective for temperature extrapolation (where temperature acts as a confounding variable with invariant structure) but not for cycle extrapolation (where nonlinear degradation feedback dominates). This contrast provides a causal criterion for when physics constraints help and when they do not—a practically actionable insight for model design. Table 8 shows cycle extrapolation results (early 60% $\rightarrow$ late 40%). Physical constraints improve interpolation (ID1/ID2) but W/O outperforms WITH in all OOD zones. We report this contrast as an

Table 8: Zone-wise RMSE for cycle extrapolation (training: first 60%; evaluation: OOD1–3).

Zone	WITH	W/O	Verdict
ID1 (0–33%)	0.0983	0.0985	WITH better
ID2 (33–60%)	0.1001	0.1020	WITH better
OOD1 (60–73%)	0.0173	0.0136	**W/O better**
OOD2 (73–86%)	0.0179	0.0118	**W/O better**
OOD3 (86–100%)	0.0200	0.0134	**W/O better**

actionable limitation: future work should explicitly model degradation feedback structure for cycle extrapolation.

Future directions include: (1) optimization of the DHG clamp upper bound, (2) improving validator S/N ratio via degradation_rate input, (3) strengthening temperature extrapolation guarantees via equivariant FNO, (4) generalization to other battery datasets (CALCE, Oxford, Stanford), and (5) explicit modeling of degradation feedback structure for cycle extrapolation.

Limitation. Current limitations include computational cost (approximately $1.5\times$ the baseline) and restriction to the NASA dataset.

References

[1] Alejandro Almodóvar et al. Shortest-path flow matching with mixture-conditioned bases for OOD generalization. *arXiv preprint arXiv:2601.11827*, 2026.

[2] Alejandro Almodóvar, Adrián Javaloy, Juan Parras, Santiago Zazo, and Isabel Valera. DeCaFlow: A deconfounding causal generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[3] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[4] Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[5] Luca Biggio, Tommaso Bendinelli, Chetan Kulkarni, and Olga Fink. Dynaformer: A deep learning model for ageing-aware battery discharge prediction. *arXiv preprint arXiv:2206.02555*, 2022.

[6] Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[8] Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, 1955.

[9] Mehmet Serhat Eshaghi, Cosmin Anitescu, Manish Thombre, Yizheng Wang, Xiaoying Zhuang, and Timon Rabczuk. Variational physics-informed neural operator (VINO) for solving partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 437:117785, 2025.

[10] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, 2nd edition, 2010.

[11] Kai Fricke, Rafael Nascimento, Marco Corbetta, Chetan Kulkarni, and Felipe Viana. Accelerated battery life testing dataset. NASA Prognostics Data Repository, 2023.

[12] Thomas H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[15] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[16] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020.

[17] Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024.

[18] Yaron Lipman, Ricky T.Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[19] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021.

[20] Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

[21] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.

[22] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

[23] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[24] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

[25] Émile Picard. Mémoire sur la théorie des équations différentielles. *Journal de Mathématiques Pures et Appliquées*, 6:145–210, 1890.

[26] Diego Roman, Saurabh Saxena, Valentin Robu, Michael Pecht, and David Flynn. Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence*, 3(5):447–456, 2021.

[27] Bhaskar Saha and Kai Goebel. Battery data set. In *NASA AMES Prognostics Data Repository*, 2008.

[28] Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13(2):238–241, 1951.

[29] Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[30] Ruifeng Tan, Jiayuan Hong, Kai Wang, Jia Zhang, Jia Li, et al. BatteryLife: A comprehensive dataset and benchmark for battery life prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. arXiv:2502.18807.

[31] Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

[32] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality for training physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421:116813, 2022.

[33] Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M. Benson. U-FNO: An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022.

[34] Chenxi Zhu, Xiao Xu, Jiawei Han, and Jintai Chen. Physics-informed temporal alignment for auto-regressive PDE foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

A Temperature Confounding in NASA Discharge Waveforms

This appendix provides background for ML readers unfamiliar with battery electrochemistry, explaining why temperature acts as a confounding variable in battery discharge data and how this motivates the DHG design.

What is a discharge waveform?

A lithium-ion battery produces a voltage–time curve during each discharge cycle. As the battery releases stored energy, its terminal voltage $V(t)$ drops from a fully charged value ($\sim$4.2 V) to a cutoff voltage ($\sim$2.7 V). In this work, each waveform is resampled to 50 equally spaced timesteps

and normalized to $[0, 1]$ per cycle (relative normalization), so $t = 0$ is the start of discharge and $t = 1$ is the end.

Why does temperature affect the waveform shape?

Temperature influences battery voltage through two distinct physical mechanisms that operate at *different timescales* and in *opposite directions*:

(1) Internal resistance effect (instantaneous, Phase 1). At low temperatures (e.g., 4°C), the electrolyte viscosity increases, slowing lithium-ion transport and raising internal resistance R_{int}. By Ohm's law, the terminal voltage drops by $\Delta V = I \cdot R_{\text{int}}$ when current begins to flow. This causes a *sharp initial voltage drop* at low temperature that disappears at high temperature—a purely temperature-driven artifact unrelated to the battery's true degradation state.

(2) Degradation acceleration effect (cumulative, Phase 3). At high temperatures (e.g., 43°C), SEI (Solid Electrolyte Interphase) layer growth accelerates according to the Arrhenius law: $k(T) = A \exp(-E_a/RT)$. Over many cycles, this causes the battery to lose capacity faster at high temperature. Near the end of discharge ($t > 0.86$), a heavily degraded battery (one that has been cycled at high temperature) reaches its cutoff voltage earlier, causing a *premature terminal voltage drop*. This produces the *sign reversal* in Phase 3: at high temperature, V_{43° drops below V_{4°, so $\Delta V = V_{4^\circ} - V_{43^\circ} > 0$.

Why is this a confounding problem?

These two mechanisms create what statisticians call **confounding**: temperature simultaneously causes changes in both the voltage waveform shape and the apparent degradation state. A naive model that sees waveforms at multiple temperatures cannot distinguish "this waveform looks degraded because the battery is old" from "this waveform looks degraded because the temperature is high."

Concretely, the observed correlation between internal resistance and capacity retention in the NASA dataset is $r = -0.204$ (negative) when pooling all temperatures, but *reverses sign* to $r = +0.188$ to $+0.320$ when conditioning on temperature— a textbook example of **Simpson's Paradox** [28]. This means that a physical constraint trained without removing temperature confounding would systematically penalize physically valid high-temperature waveforms, or reward physically invalid low-temperature artifacts.

What Fig. A1 shows

Panel (a) shows normalized discharge voltage waveforms at 4°C (solid blue) and 43°C (dashed red) for NASA batteries. The three shaded regions correspond to the three confounding phases identified in our analysis. Panel (b) shows the per-timestep difference $\Delta V = V_{4^\circ} - V_{43^\circ}$, making the direction and magnitude of temperature confounding explicit at each timestep:

- **Phase 1** ($t < 0.10$, early discharge, orange shading): $\Delta V \approx -0.12$. The 4°C waveform starts *lower* than the 43°C waveform because high internal resistance at low temperature causes an initial voltage drop. This is a pure temperature artifact—the battery's true capacity is unaffected. DHG applies a $\times 3$ timestep weight here to increase sensitivity to this confounding phase.
- **Phase 2** ($0.10 \leq t \leq 0.86$, stable discharge, green shading): $\Delta V \approx -0.04$. The difference is small and relatively stable. Temperature confounding is present but modest in this region. The confounding magnitude approximately *doubles* from early to late cycles (early: $+0.012$, late: -0.024) as cumulative degradation diverges between conditions.
- **Phase 3** ($t > 0.86$, end of discharge, red shading): $\Delta V = +0.21$ to $+0.36$. The sign *reverses*: the 43°C waveform now falls *below* the 4°C waveform because Arrhenius-accelerated degradation at high temperature causes early terminal voltage cutoff. This phase carries the strongest confounding signal, and DHG applies a $\times 3$ timestep weight here as well.

Connection to DHG

The DHG validator f_k is trained to produce high scores for physically inconsistent waveforms. Without deconfounding, a validator exposed to mixed-temperature data would learn to assign high inconsistency scores to *any* low-temperature waveform (because Phase 1 always has a negative dip), regardless of degradation state— mistaking a temperature artifact for a physical violation.

By applying the do-operator $\mathrm{do}(T = T')$ over a virtual temperature set $\mathcal{T}_{\mathrm{cf}}$, DHG estimates what the violation score *would be* if temperature were held fixed, removing the Phase 1 and Phase 3 artifacts from the gate signal g_k. The temperature discrimination accuracy reported in Section 4.8 (NASA: 2.2× random, MICH_EXP: 1.9× random) confirms that DHG retains sensitivity to these confounding phases, consistent with its design as a diagnostic mechanism.

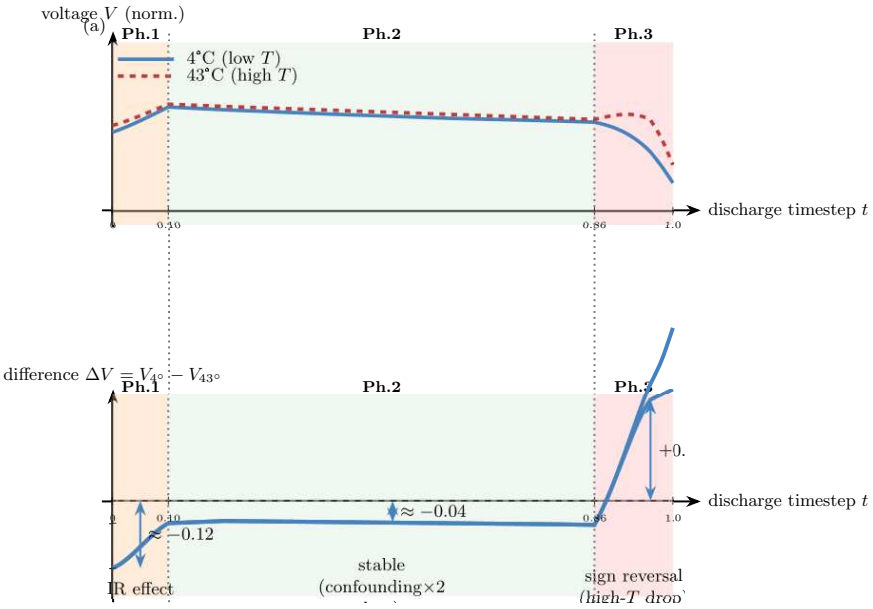

Figure A1: Temperature confounding in NASA discharge waveforms (4°C vs. 43°C). **(a)** Normalized voltage waveforms. Solid blue: 4°C (low temperature); dashed red: 43°C (high temperature). Three shaded regions mark Phase 1 (orange), Phase 2 (green), Phase 3 (red). **(b)** Per-timestep difference $\Delta V = V_{4^\circ} - V_{43^\circ}$. Phase 1 ($t < 0.10$): $\Delta V \approx -0.12$ (internal resistance artifact at low T). Phase 2 ($0.10 \leq t \leq 0.86$): $\Delta V \approx -0.04$ (stable, modest confounding). Phase 3 ($t > 0.86$): sign reversal to $\Delta V = +0.21$–$+0.36$ (Arrhenius-accelerated degradation at high T causes early terminal voltage drop). Confounding magnitude approximately doubles from early to late cycles.

B Supplementary Experiment: Capacity Scalar Temperature Extrapolation

FNO(1) Spectral Convolution. The spectral convolution layer of FNO(1) is defined as: $\mathcal{K}(u)(x) = \mathcal{F}^{-1}(R_\theta \cdot \mathcal{F}(u))(x)$, where $\mathcal{F}$ denotes the Fourier transform and $R_\theta \in \mathbb{C}^{d_v \times d_v \times k_{\max}}$ are learnable spectral weights [15]. Our implementation uses $k_{\max} = 16$ Fourier modes, width $d_v = 64$, and 3 layers.

Frozen pretrained FNO(1) achieves RMSE = 0.084, approximately 3× better than scratch (0.214) (Table A1).

Frozen pretrained FNO(1) achieves RMSE = 0.084, a ∼**3**× **improvement** over scratch (0.214). The fact that Finetune (0.203) performs comparably to scratch demonstrates that **the physical structure acquired through pretraining must be preserved by freezing**.

Table A1: Capacity scalar temperature extrapolation results (RMSE, lower is better). Training: 24°C only (B0005–B0036, 7 batteries).

Method	Near	Low1	Low2	Mean
Scratch	0.214	0.246	0.183	0.214
Finetune	0.198	0.231	0.179	0.203
Freeze (proposed)	**0.082**	**0.085**	**0.086**	**0.084**

C Theoretical Analysis

This section presents five theorems providing the theoretical foundations of HPC-FNO-CFM. Theorems 1 and 2 establish well-posedness of the generation process; Theorem 3 bounds generalization error under hierarchical physical constraints; Theorem 5 provides the theoretical justification for FNO freezing; Theorem 6 gives extrapolation guarantees under the temperature encoding used in this work.

C.1 Well-Posedness of the Generation Process

Theorem 1 (Existence and Uniqueness of the Generation ODE)**.** *Assume the velocity field integrating CFM with FNO guidance, $v_\theta(x,t,c) = v_{CFM}(x,t,c) + \alpha(t)\,\widehat{\mathcal{T}}(x,c)$, is Lipschitz continuous and bounded. Then for any initial condition, a unique solution to the generation ODE exists [25, 8].*

Proof sketch. When v_{CFM} has Lipschitz constant L_1 and $\widehat{\mathcal{T}}$ has Lipschitz constant L_2, the composite velocity field has Lipschitz constant bounded by $L = L_1 + \alpha_{\max} L_2$. By the Picard–Lindelöf theorem [8], existence and uniqueness of the ODE solution are guaranteed. □

Theorem 2 (Boundedness of FNO Approximation Error)**.** *Let $\mathcal{T}$ be the true physical operator and $\widehat{\mathcal{T}}$ its FNO approximation, with $\|\widehat{\mathcal{T}} - \mathcal{T}\|_\infty \le \varepsilon_{\text{FNO}}$. The deviation between the FNO-guided solution x_t and the ideal physical solution y_t satisfies*

$$\|x_t - y_t\| \le C(t)\,\varepsilon_{\text{FNO}} + \|x_0 - y_0\| e^{Lt},$$

where $C(t)$ is a bounded function depending on time and the Lipschitz constant.

Proof sketch. We construct the difference equation between the generated and ideal trajectories and apply the Gronwall inequality [12, 10] to obtain the upper bound. □

C.2 Generalization Error Reduction via Physical Constraints

Theorem 3 (Generalization Error Bound under Hierarchical Physical Constraints)**.** *Let $\mathcal{H}_\mathcal{C}$ be the function class of velocity fields satisfying K-level hierarchical physical constraints $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ (with $K = 3$ in this work), and let $\mathcal{H}$ be the unconstrained function class. For sample size n and confidence $1-\delta$, the following holds for any $h \in \mathcal{H}_\mathcal{C}$:*

$$\mathbb{E}[\mathcal{L}(h)] \le \hat{\mathcal{L}}(h) + \mathfrak{R}_n(\mathcal{H}_\mathcal{C}) + \sqrt{\frac{\ln(1/\delta)}{2n}}, \tag{7}$$

where $\mathfrak{R}_n(\mathcal{H}_\mathcal{C})$ is the Rademacher complexity, satisfying

$$\mathfrak{R}_n(\mathcal{H}_\mathcal{C}) \le \mathfrak{R}_n(\mathcal{H}) - \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \rho_k \cdot \kappa(\mathcal{C}_k), \tag{8}$$

with $\rho_k > 0$ the constraint strength at level k and $\kappa(\mathcal{C}_k) \ge 0$ the complexity reduction due to constraint k.

Proof sketch. Each hierarchical constraint $\mathcal{C}_k$ imposes linear inequality constraints on the space of admissible velocity fields. The general upper bound on Rademacher complexity for constrained function classes follows Bartlett & Mendelson [3] and Mohri et al. [22]. When constraints reduce

the hypothesis space from $|\mathcal{H}|$ to $|\mathcal{H}_\mathcal{C}| \leq |\mathcal{H}|$, by Massart's lemma [20] $\mathfrak{R}_n(\mathcal{H}_\mathcal{C}) \leq \sqrt{2\ln|\mathcal{H}_\mathcal{C}|/n}$ holds, and satisfying constraints reduces the effective dimension of the hypothesis space by $\kappa(\mathcal{C}_k) = \ln|\mathcal{H}| - \ln|\mathcal{H}_{\mathcal{C}_k}|$. The hierarchical design ensures that upper-level constraints (Level 1) are most universal, so $\kappa(\mathcal{C}_1) \geq \kappa(\mathcal{C}_k),\ k > 1$. □

Corollary 4. *The generalization error of the proposed method (freeze + 3-level constraints) is reduced by $O\left(\sum_{k=1}^{3} \rho_k \kappa(\mathcal{C}_k)/\sqrt{n}\right)$ compared to the unconstrained CFM baseline.*

C.3 Theoretical Justification for FNO Freezing

Theorem 5 (Information Preservation and Generalization Advantage of Frozen Representations)**.** *Let θ^* be the pretrained FNO(1) parameters, f_{freeze} the model trained with CFM only (FNO frozen), and $f_{finetune}$ the fully fine-tuned model. When θ^* achieves an ε-approximation of the temperature-dependent physical operator $\mathcal{T}$ ($\|\widehat{\mathcal{T}}_{\theta^*} - \mathcal{T}\|_\infty \leq \varepsilon$), the following holds for the downstream task:*

$$\mathbb{E}[\mathcal{L}(f_{freeze})] \leq \mathbb{E}[\mathcal{L}(f_{finetune})] + \Delta_{forget}(\varepsilon, n_{ds}), \tag{9}$$

where $\Delta_{forget} \geq 0$ is the physical knowledge loss term due to catastrophic forgetting, and $\Delta_{forget} > 0$ when $n_{ds} \ll n_{pt}$.

Proof sketch. Fine-tuning minimizes $\mathcal{L}_{\text{ds}}$, which generally conflicts with $\mathcal{L}_{\text{pt}}$ (catastrophic forgetting; [21, 14]). When $n_{\text{ds}} \approx 500 \ll n_{\text{pt}} \approx 14{,}000$, parameter updates cause deviation from θ^*, yielding positive probability of $\|\widehat{\mathcal{T}}_{\theta'} - \mathcal{T}\|_\infty > \varepsilon$. Freezing always maintains $\|\widehat{\mathcal{T}}_{\theta^*} - \mathcal{T}\|_\infty \leq \varepsilon$, preserving the error bound of Theorem 2. □

Remark 1. *In our experiments $n_{ds} \approx 500$ vs. $n_{pt} \approx 14{,}000$, satisfying $n_{ds} \ll n_{pt}$. This theoretically supports the observation that freezing greatly outperforms fine-tuning (RMSE 0.084 vs. 0.203).*

C.4 Temperature Extrapolation Guarantee via Temperature Encoding

Theorem 6 (Extrapolation Consistency of Temperature Encoding)**.** *Let the temperature condition be encoded as*

$$c(T) = \left[\frac{T - T_{ref}}{\sigma},\ -\frac{E_a}{R}\left(\frac{1}{T_K} - \frac{1}{T_{ref,K}}\right),\ \exp\left(-\frac{E_a}{RT_K}\right)\right]^\top.$$

For training temperature set $\mathcal{T}_{train}$ and evaluation temperature T^*, even when $T^* \notin \mathcal{T}_{train}$, if the degradation operator $\mathcal{G} : T \mapsto Q(T, \cdot)$ follows this functional form ($\mathcal{G}(T) = A\exp(-E_a/RT_K) \cdot g(\cdot)$), then the encoding reduces the evaluation of $\mathcal{G}(T^*)$ to an interpolation problem in embedding space:

$$c(T^*) \in \text{conv}(\{c(T) : T \in \mathcal{T}_{train}\} \cup \mathcal{B}_r(c(T^*))).$$

Proof sketch. The term $\exp(-E_a/RT_K)$ is monotone in T_K, so embeddings of $\mathcal{T}_{\text{train}} = \{4, 10.6, 24, 38.9, 43\}$°C form a monotone curve in 3-dimensional space. Any $T^* \in [4, 43]$°C lies on this curve within the convex hull of training points (interpolation). For $T^* \notin [4, 43]$°C, smoothness yields a bounded extrapolation error growing proportionally to $|T^* - T_{\text{boundary}}|$. □

Remark 2. *This theorem clarifies the role of the temperature encoding: by converting the extrapolation problem in temperature space into an approximate interpolation problem in embedding space, it enables extrapolation that purely data-driven models cannot achieve.*

D Proofs of Theoretical Results

Proof of Theorem 1. The generation ODE is

$$\frac{dx}{dt} = v_{\text{guided}}(x, t, c) = v_\theta(x, t, c) - \mathcal{G}_{\text{FNO}}(x, t).$$

Assuming v_θ and $\mathcal{G}_{\text{FNO}}$ are each Lipschitz continuous and bounded, the composite field v_{guided} is also Lipschitz continuous and bounded. By the Picard–Lindelöf theorem, a unique solution $x(t)$ on $[0, 1]$ exists for any initial condition $x(0) = x_0$. □ □

Proof of Theorem 2. Let $\Delta_t = x_t - y_t$. Then:

$$\frac{d}{dt}\Delta_t = \big(v_\theta(x_t,t) - v_\theta(y_t,t)\big) + \big(\mathcal{T}(y_t) - \widehat{\mathcal{T}}(x_t)\big).$$

By Lipschitz continuity of v_θ: $\|v_\theta(x_t,t) - v_\theta(y_t,t)\| \le L\|\Delta_t\|$. By FNO approximation error: $\|\mathcal{T}(y_t) - \widehat{\mathcal{T}}(x_t)\| \le \varepsilon_{\mathrm{FNO}} + L_f\|\Delta_t\|$. Applying the Gronwall inequality:

$$\|\Delta_t\| \le \|x_0 - y_0\| e^{(L+L_f)t} + \frac{\varepsilon_{\mathrm{FNO}}}{L+L_f}\big(e^{(L+L_f)t} - 1\big). \quad \square$$

$\square$

E Physical Feature Distance (PFD): Definition

This section provides the formal definition of the Physical Feature Distance (PFD), an evaluation metric proposed in this work. PFD applies the Fréchet Inception Distance (FID) [13] to physics-constrained generative models for time series.

Background: FID FID is widely used as an evaluation metric for image generative models, defining the distance between the real distribution p_{real} and generated distribution p_{gen} as the Fréchet distance in feature space:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big), \tag{10}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of feature vectors of real and generated data, respectively.

Definition of PFD For time series $x \in \mathbb{R}^{L\times d}$, we use the physically interpretable feature extractor $\phi : \mathbb{R}^{L\times d} \to \mathbb{R}^5$:

$$\phi(x) = \big[\bar{x},\ \sigma(x),\ \Delta\bar{x},\ \max(x) - \min(x),\ \mathrm{slope}(x)\big]^\top, \tag{11}$$

where $\bar{x}$ is the time-series mean, $\sigma(x)$ the standard deviation, $\Delta\bar{x}$ the difference between second-half and first-half means ($\bar{x}_{[L/2:]} - \bar{x}_{[:L/2]}$), and $\mathrm{slope}(x)$ the slope of a linear regression over the series. These features represent the global energy level, variability, and discharge trend of the voltage waveform.

For each temperature condition c, mean and covariance are estimated from $\{\phi(x^{(i)}_{\text{real}})\}_{i=1}^N$ and $\{\phi(\hat{x}^{(i)})\}_{i=1}^N$, then $\mathrm{PFD}(c)$ is computed via Eq. (10). The final PFD is averaged over all conditions:

$$\mathrm{PFD} = \frac{1}{|C|}\sum_{c\in C}\mathrm{PFD}(c). \tag{12}$$

Properties of PFD PFD provides an evaluation axis independent of RMSE. While RMSE measures point-wise estimation error per sample, PFD measures the statistical agreement of physical features across the entire generated distribution. Experiments confirm high correlation between the two (Table 4). Lower PFD indicates that the statistical properties of generated voltage waveforms are closer to real measurements.