UNITED: RESOLVING SUB-GRID HETEROGENEITY FOR ROBUST GRID CELL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning (ML) has seen widespread adoption in the physical sciences for developing data-driven emulators of grid-based simulations. Despite this rapid progress, physical systems characterized by sub-grid heterogeneity remain largely underexplored. In domains ranging from land surface modeling in Earth sciences to multi-scale stress analysis in materials science, a grid cell's state is far from uniform. Instead, the cell's state is highly heterogeneous: distinct sub-grid "tiles" exhibit intrinsic processes that must be resolved before computing the grid cell's aggregate behavior. Directly applying standard ML models, which lack the inductive bias to account for this heterogeneity or resolve these underlying processes, can lead to structural entanglement and poor generalization when sub-grid fractions shift. To address this gap, we introduce UniTeD (Unified Tiled-Expert Demixer), a sub-grid-aware ML architecture that strictly enforces physical aggregation logic. UniTeD isolates latent constitutive functions purely from aggregate supervision, enabling robust generalization across varying sub-grid fraction compositions. Evaluated on synthetic benchmarks and global ERA5 surface energy flux inference, UniTeD achieves zero-shot compositional generalization, reducing prediction errors by up to 28% relative to the strongest baseline.

1 INTRODUCTION

In the physical sciences, continuous domains are discretized into grid cells for numerical modeling. However, in most physical systems, these individual cells are rarely homogeneous. For instance, geographical cells encompass diverse land covers, such as forests, grasslands, and water bodies (Chaney et al., 2018; Torres-Rojas et al., 2022), while finite elements in materials science contain varied crystal grains (Roters et al., 2010). These partitions are commonly referred to as tiles or sub-grids. Each tile possesses unique physical properties and exhibits a distinct response to external forcings. For example, within a geographical cell containing both lawn and pavement, the sunlit lawn feels cooler to the touch than the pavement because each tile responds to solar radiation differently. The macroscopic grid cell state is ultimately determined by enforcing conservation laws across these varied responses. This conservation constraint results in aggregation, typically via area- or volume-weighted averaging of the sub-grid responses, as in sub-grid tiling schemes for land surface models (Balsamo et al., 2009; Torres-Rojas et al., 2022) and in mean-field homogenization of solids (Roters et al., 2010). Standard datasets, however, record only the final cell aggregate, leaving the underlying sub-grid responses unavailable as training labels.

Standard ML models, ranging from deep neural networks (DNN) (Goodfellow et al., 2016) to Transformers (Vaswani et al., 2017; Gorishniy et al., 2021), lack the inductive bias to resolve this aggregation structure. When applied directly to cell aggregates, these architectures operate as monolithic models, treating the grid cell as a single unit that maps cell-level inputs directly to aggregate outputs. This approximation makes these methods highly susceptible to structural entanglement (Section 3.2), in which they conflate distinct physical regimes across tiles. Thus, they struggle to make grid cell-level predictions when the sub-grid mixture distribution shifts, such as in novel land covers or rare crystallographic textures.

To address this limitation, we introduce UniTeD (Unified Tiled-Expert Demixer), a heterogeneity-aware architecture that explicitly embeds the physical aggregation logic of sub-grid processes. This allows the model to structurally disentangle latent sub-grid processes, ensuring generalization across

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

varying sub-grid fractions and ultimately yielding more robust grid cell-level predictions. UniTeD has two core mechanisms: (i) a shared expert conditioned via FiLM (Perez et al., 2018) to recover the latent constitutive functions of individual tiles, and (ii) a physics-prescribed aggregation layer that uses known constituent fractions to strictly enforce the conservative linear superposition.

The key contributions of our work are as follows:

- **Analysis of structural entanglement**: We provide a theoretical analysis of why monolithic models are prone to learning spurious correlations between independent tiles, which could compromise their robustness when sub-grid compositions shift.
- **The UniTeD framework**: We propose a heterogeneity-aware neural architecture that explicitly resolves the underlying physical structure of heterogeneous grid cells. By combining physics-prescribed aggregation with a conditioned shared expert, the model recovers latent constitutive functions from aggregate supervision while ensuring parameter efficiency.
- **Validation on surface thermodynamics**: We provide empirical evidence that structural disentanglement enables zero-shot compositional generalization. On ERA5, UniTeD reduces error by up to 28% on unseen pure tiles and by up to 22% on unseen mixtures, relative to the strongest baseline in each setting.

2 Related Work

Data-driven Emulators. Data-driven emulators have achieved remarkable success in approximating complex physical systems governed by partial differential equations. In fluid dynamics and continuum mechanics, a vast array of architectures has emerged, including mesh- and graph-based neural solvers (De Avila Belbute-Peres et al., 2020; Pfaff et al., 2021; Brandstetter et al., 2022; Sharma & Fink, 2026), neural operators and functional diffusion models (Li et al., 2023; Liu et al., 2026; Wang et al., 2026), and latent-space transformers (Han et al., 2022; Alkin et al., 2024). This rapid progress is further propelled by large-scale foundational simulation benchmarks, such as The Well (Ohana et al., 2024). At the same time, data-driven modeling has revolutionized global weather forecasting and climate simulation (Lam et al., 2023; Kochkov et al., 2024; Bodnar et al., 2025; Lang et al., 2024), as well as specialized geophysical systems such as in hydrology (Kazadi et al., 2024). Despite differences in architectures and spatial discretizations, these approaches share a common assumption: the computational cell or mesh element is treated as homogeneous. While this bulk approximation is effective, for instance, in tracking water elevation above ground or mesoscale-to-synoptic atmospheric processes, it becomes a critical bottleneck for phenomena governed by fine-scale spatial heterogeneities, such as surface energy fluxes, where sub-grid composition is paramount.

Sub-Grid Heterogeneity. Existing approaches to sub-grid modeling generally fall into two extremes. On one hand, models such as Deep Material Networks (Liu et al., 2019; Jean et al., 2025) or differentiable Earth system models (Feng et al., 2022) explicitly retain sub-grid tiling structure, but they restrict the neural network's role to simple parameter calibration (e.g., estimating static stiffness or roughness scalars). On the other hand, highly expressive neural emulators designed for sub-grid parameterization, such as Paraformer (Wang et al., 2024) and aiLand (Raoult et al., 2026), learn complex non-linear response functions but remain monolithic. They treat sub-grid fractions merely as static, concatenated input features, leaving them susceptible to generalization failures when sub-grid compositions shift. UniTeD bridges this gap by enforcing strict aggregation logic while employing highly expressive neural networks to learn the full latent constitutive functions.

Mixture-of-Experts (MoE) vs. Physical Aggregation. Architecturally, predicting an aggregate grid cell state from diverse sub-grid tiles resembles a Mixture-of-Experts (MoE) (Jacobs et al., 1991; Puigcerver et al., 2024), where a gating network routes inputs to specialized sub-networks. However, in physical domains, relying on a learned routing mechanism is similar to unsupervised regime identification. Additionally, MoEs often suffer from expert co-adaptation or "routing collapse" (Shazeer et al., 2017; Fedus et al., 2022), where a small subset of experts dominates the routing while others remain underutilized. UniTeD departs from standard MoEs in two critical ways. First, it replaces the learned router with a physics-prescribed aggregation layer driven by exact constituent fractions, strictly enforcing conservative linear superposition. Second, rather than using disjoint sub-networks, UniTeD utilizes a unified, shared expert conditioned via FiLM (Perez et al., 2018). This shared parameterization leverages the underlying physical manifold, preventing representation collapse.

3 METHODOLOGY

3.1 PROBLEM FORMULATION: LINEAR AGGREGATION FROM PHYSICAL CONSTRAINTS

In systems where sub-grid heterogeneity governs macroscopic behavior, the aggregate cell state is often derived from the spatial integration of conserved physical quantities (e.g., mass, energy, momentum). For instance, in order to compute a grid cell's temperature (a non-conserved variable), conserved variables such as energy fluxes are first aggregated over the cell to derive that temperature.

Consider a grid cell representing a spatial domain Ω with total area (or volume) A. The domain is partitioned into K disjoint sub-domains (tiles) Ω_k with sub-areas A_k, such that $\sum_{k=1}^{K} A_k = A$. Let $q(\mathbf{s})$ be the local rate of transfer of a conserved quantity (e.g., a localized flux density or stress tensor) at spatial coordinate $\mathbf{s}$. To satisfy macroscopic balance laws across spatial scales for conservation, the macroscopic rate of transfer Q is defined as the spatial average over Ω:

$$Q = \frac{1}{A}\int_{\Omega} q(\mathbf{s})\, d\mathbf{s} = \frac{1}{A}\sum_{k=1}^{K}\int_{\Omega_k} q(\mathbf{s})\, d\mathbf{s} = \sum_{k=1}^{K}\left(\frac{A_k}{A}\right)\left(\frac{1}{A_k}\int_{\Omega_k} q(\mathbf{s})\, d\mathbf{s}\right)$$

Defining the grid cell response as $y = Q$, the fractional area $\alpha_k = A_k/A$, and the tile-level response as $\phi_k = \frac{1}{A_k}\int_{\Omega_k} q(\mathbf{s})\, d\mathbf{s}$, the physics directly yields a linear superposition of sub-grid responses:

$$y = \sum_{k=1}^{K} \alpha_k \phi_k \tag{1}$$

Our goal is to predict y by strictly embedding this physical linear superposition, where α_k are the known sub-grid area fractions. We approximate unknown sub-grid responses ϕ_k with $\hat{\phi}_k(\mathbf{x}; \tau_k)$ driven by shared macroscopic forcing $\mathbf{x}$ and tile-specific attributes τ_k, assuming weak or negligible inter-tile interactions. This is an instance of supervised learning from aggregates (Musicant et al., 2007; Wei et al., 2023). By structurally restricting our estimator to $\hat{y} = \sum_{k=1}^{K} \alpha_k \hat{\phi}_k(\mathbf{x}; \tau_k)$, we gain the distinct advantage of discovering hidden sub-grid constitutive behaviors from aggregate data.

3.2 STRUCTURAL ENTANGLEMENT IN MONOLITHIC MODELS

Consider a monolithic model $f(\mathbf{x}, \boldsymbol{\alpha})$ taking fractions $\boldsymbol{\alpha}$ as input, with its Taylor expansion around a composition $\boldsymbol{\alpha}^*$ given by:

$$f(\mathbf{x}, \boldsymbol{\alpha}) = c(\mathbf{x}) + \sum_k g_k(\mathbf{x})(\alpha_k - \alpha_k^*) + \underbrace{\frac{1}{2}\sum_{j,k} H_{jk}(\mathbf{x})(\alpha_j - \alpha_j^*)(\alpha_k - \alpha_k^*)}_{\text{Spurious entanglement}} + \mathcal{O}(\|\Delta\boldsymbol{\alpha}\|^3)$$

In the physical system (Eq. 1), the response is strictly linear in $\boldsymbol{\alpha}$, implying the term with the Hessian H_{jk} should be identically zero. While a monolithic model could theoretically learn this linear aggregation, it lacks the structural constraint to guarantee it, making it prone to entangling independent tiles and inducing systematic biases whenever sub-grid compositions shift.

3.3 THE UNITED ARCHITECTURE

We introduce the **Uni**fied **T**iled-**E**xpert **D**emixer (UniTeD) (Fig. 1) as a grid cell estimator $F_\Theta : \mathcal{X} \times \Delta^{K-1} \times \mathcal{T}^K \to \mathcal{Y}$ defined as a composition of a *trunk encoder*, a *modulated shared expert*, and a *physics-prescribed aggregator*. UniTed strictly enforces the aggregation scheme of Eq. 1.

The Trunk Encoder The trunk encoder $T_\psi : \mathcal{X} \to \mathcal{Z}$ maps forcing $\mathbf{x} \in \mathcal{X}$ to $\mathbf{z}$ in a high-dimensional latent space $\mathcal{Z}$, isolating shared global conditions $\mathbf{x}$ from tile-specific properties τ_k.

$$\mathbf{z} = T_\psi(\mathbf{x}) \tag{2}$$

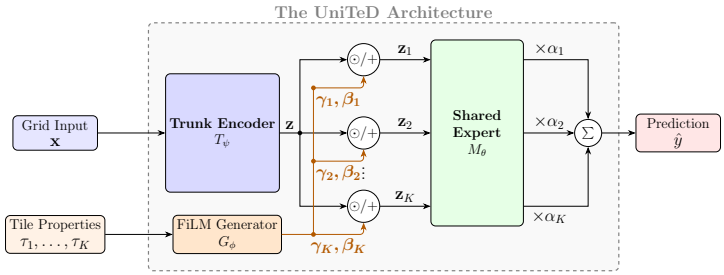

Figure 1: The UniTeD Architecture. Grid cell inputs $\mathbf{x}$ are mapped by the trunk encoder (T_ψ) to a latent context $\mathbf{z}$. Parallel FiLM generators (G_ϕ) modulate the shared expert (M_θ) using tile attributes τ_k to recover distinct constitutive functions. The aggregate prediction is synthesized via physics-prescribed aggregation using tile fractions α_k at the summation node (Σ).

The Modulated Shared Expert To capture constitutive functions, we employ a shared expert $M_\theta : \mathcal{Z} \to \mathcal{Y}$ conditioned via Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). A generator network $G_\phi : \mathcal{T} \to \mathbb{R}^{d_{lat}} \times \mathbb{R}^{d_{lat}}$ maps tile attributes $\tau_k \in \mathcal{T}$ to adaptive parameters $(\boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) = G_\phi(\tau_k)$. These parameters inject tile context into the latent stream via the element-wise transformation $\mathbf{z}_k = \boldsymbol{\gamma}_k \odot T_\psi(\mathbf{x}) + \boldsymbol{\beta}_k$. The expert processes this tile-specific manifold to produce the approximated latent response:

$$\hat{\phi}_k = M_\theta(\boldsymbol{\gamma}_k \odot T_\psi(\mathbf{x}) + \boldsymbol{\beta}_k) \tag{3}$$

Unlike standard MoEs using K disjoint networks, UniTeD utilizes a single shared expert. By conditioning M_θ on tile context, the model learns a continuous representation of physical behaviors rather than discrete modes, promoting inter-tile transfer learning and parameter efficiency.

Physics-Prescribed Aggregation The final estimator $\hat{y}$ aggregates latent tile responses $\hat{\phi}_k$ weighted by fractions $\boldsymbol{\alpha} \in \Delta^{K-1}$:

$$\hat{y} = \sum_{k=1}^{K} \alpha_k \hat{\phi}_k \tag{4}$$

By introducing $\boldsymbol{\alpha}$ only at this final summation, UniTeD strictly enforces functional independence ($\frac{\partial \hat{\phi}_k}{\partial \alpha_j} = 0$). This guarantees that the spurious Hessian cross-terms vanish ($H_{jk} \equiv 0$), structurally disentangling the learned constitutive functions from the mixture configuration (Section 3.2).

3.4 Theoretical Identifiability Analysis

Unmixing aggregate data is an inherently ill-posed inverse problem. We demonstrate that latent constitutive functions can be recovered by coupling physical data properties with architectural linear aggregation constraints, transforming an under-determined problem into a uniquely solvable system.

Theorem 3.1 *Let the true responses be $\mathbf{y} = \sum_{k=1}^{K} \alpha_k \boldsymbol{\phi}_k^*(\mathbf{x}) \in \mathbb{R}^D$. Assuming a realizable model class and optimization success $\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2 \leq \delta$ within a local neighborhood $\mathcal{N}_r = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_2 \leq r\}$, the latent functions $\{\boldsymbol{\phi}_k^*\}$ are structurally identifiable at $\mathbf{x}^*$ as $r, \delta \to 0$, provided:*

1. ***Regularity:** The input space $\mathcal{X}$ is compact; $\boldsymbol{\phi}_k^*$ and $\hat{\boldsymbol{\phi}}_k$ are Lipschitz continuous.*
2. ***Structural Exclusion:** The estimator $\hat{\boldsymbol{\phi}}_k$ is independent of fractions $\boldsymbol{\alpha}$ (i.e., $\partial \hat{\boldsymbol{\phi}}_k / \partial \boldsymbol{\alpha} \equiv \mathbf{0}$).*
3. ***Local Non-Degeneracy:** A batch $\mathcal{B}$ of size $N \geq K$ exists in $\mathcal{N}_r$ where the fraction matrix $A_\mathcal{B}$ has full column rank ($\sigma_{\min}(A_\mathcal{B}) \geq c > 0$).*

Proof Sketch. For a batch $\mathcal{B}$ sharing identical forcing $\mathbf{x}^*$, minimizing the aggregation loss yields the linear system $A_\mathcal{B} \hat{\boldsymbol{\Phi}} = A_\mathcal{B} \boldsymbol{\Phi}^*$. *Structural Exclusion* ensures $\hat{\boldsymbol{\Phi}}$ is constant across the batch, while

Local Non-Degeneracy ensures $A_{\mathcal{B}}$ is left-invertible, yielding the unique solution $\hat{\mathbf{\Phi}} = \mathbf{\Phi}^*$. For $r > 0$, *Regularity* bounds the perturbation error, ensuring stable convergence as $r, \delta \to 0$.

See Appendix A for a detailed proof. These conditions naturally align with physical data: *Regularity* holds because physical states are bounded, while *Local Non-Degeneracy* reflects that similar environmental conditions frequently apply to diverse mixture configurations.

4 EXPERIMENTS

We evaluate whether structurally enforcing the linear superposition of sub-grid processes enables robust generalization to unseen fractional compositions. We validate this hypothesis on Earth's surface energy balance, a domain that provides an ideal testbed for modeling heterogeneous mixtures. Here, macroscopic grid cells comprise diverse tiles (e.g., forests, lakes) governed by distinct physical laws, yet only aggregate responses are available for supervision. The task is framed as a multi-target regression for latent heat (LE), sensible heat (H), and ground heat (G) fluxes, with the surface energy balance enforced via physical regularization during training (see Appendix B). Our experiments are motivated by the increasing need for large-scale climate monitoring under anthropogenic and natural shifts, where models must generalize from geographically biased measurements to rare or entirely unseen landscapes (Chaney et al., 2018; Torres-Rojas et al., 2022; Yin et al., 2023).

We benchmark UniTeD against the strongest possible monolithic and modular approximations for surface energy flux modeling. To ensure our baselines represent the current state-of-the-art in domain-specific emulation, all models are re-implemented with the same transformer backbone capacity for fair comparison. Detailed architectural definitions are provided in Appendix B.

- **Monolithic DNNs**: We evaluate a Deep Neural Network (DNN) representative of recent emulators like ECMWF's aiLand (Raoult et al., 2026), as well as approaches by McCandless et al. (2022) and Wulfmeyer et al. (2023). These map inputs directly to the aggregate state using multiple fully connected layers, treating sub-grid fractional composition merely as concatenated static input features.
- **Monolithic Transformers:** We evaluate FT-Transformer (Gorishniy et al., 2021) as an adaption of recent attention-based parameterization models like Paraformer (Wang et al., 2024) for grid cell prediction. While highly expressive, these treat the grid cell as a monolithic sequence, lacking the structural inductive bias to separate tiles.
- **Soft-MoE (adapted from Puigcerver et al. (2024)):** A Mixture-of-Experts using learned soft routing, tested to determine if disentanglement emerges naturally without the strict structural constraint of known constituent fractions.
- **Frac-MoE (Disjoint Experts):** Employs physics-prescribed aggregation but utilizes disjoint parameters for each expert, isolating the specific benefit of UniTeD's unified shared expert in promoting data efficiency and inter-tile transfer learning.

We conduct experiments via a two-tier protocol: (i) **Synthetic Validation** with a controlled ground truth verifies that UniTeD's architecture accurately recovers latent sub-grid processes and achieves zero-shot compositional generalization to novel mixtures; and (ii) **Large-Scale Evaluation** on the ERA5 dataset demonstrates how explicit aggregation logic improves performance and robustness on unseen land-cover distributions where sub-grid responses remain strictly latent.

4.1 SYNTHETIC VALIDATION OF STRUCTURAL IDENTIFIABILITY

To empirically validate our theoretical claims (Section 3.4), this phase verifies the models' ability to accurately recover the underlying, unobservable sub-grid processes before evaluating their generalization to unseen mixture configurations.

4.1.1 EXPERIMENTAL SETUP: THE SYNTHETIC ENVIRONMENT

We simulate a thermodynamic system where grid cells comprise Vegetation, Water, and Urban tiles. To rigorously test identifiability, we focus on latent heat (LE) and design orthogonal tile physics: Vegetation follows a sigmoidal temperature response ($\sigma(T_{air})$), while Water scales linearly with

wind speed (U_{wind}) (see Appendix C). This decoupling ensures that a successful model must attribute aggregate signals to their specific latent sources rather than relying on statistical correlations.

We select baselines to isolate specific architectural failure modes: (i) For **Disentanglement & Generalization** (Exp. I & II), we compare UniTeD against FT-Transformer (monolithic) and Soft-MoE (unconstrained). We omit Frac-MoE here because it shares the physics-prescribed aggregation of UniTeD and thus achieves similar disentanglement performance. (ii) For **Few-Shot Transfer** (Exp. III), Frac-MoE serves as a structural ablation to test the necessity of UniTeD's shared expert for parameter efficiency and inter-tile transfer learning.

4.1.2 Experiment I: Disentanglement of Mixed Signals

We assess the model's capacity to unmix the aggregate signal into sub-grid responses. Trained exclusively on heterogeneous mixtures ($\alpha_k \leq 0.5$) and tested on pure tiles ($\alpha_k = 1.0$), we verify if the model has learned to recover latent sub-grid signals or simple correlations to the aggregate data.

Scenario A: Wind-Driven Physics. We isolate the aerodynamic response by varying wind speed (U_{wind}) while holding other drivers constant. As shown in Fig. 2, UniTeD correctly identifies that wind speed is an exclusive driver of the Water tile. It recovers the linear response (Fig. 2(B)) while remaining invariant for Vegetation and Urban tiles ($\partial\hat{\phi}_k/\partial U_{wind} \approx 0$). The baselines fail this orthogonality test. The monolithic FT-Transformer displays significant leakage, producing physically inconsistent curves that couple independent tiles. Similarly, Soft-MoE exhibits clear expert bleeding, as it erroneously attributes wind sensitivity to the Vegetation tile (Fig. 2(A)) and overestimates the Water response, confirming that learned routing alone is insufficient to isolate causal drivers without the structural constraint of the aggregation.

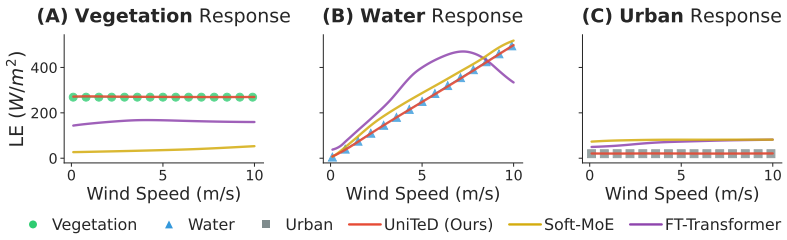

Figure 2: Disentanglement of Mixed Signals: Wind-Driven Physics (Exp. I-A). Recovery of pure cell responses ($\alpha_k = 1.0$) after training solely on aggregate mixtures ($\alpha_k \leq 0.5$). UniTeD (red) perfectly tracks the trend for Water while correctly identifying the invariance of Vegetation and Urban tiles. In contrast, Soft-MoE (gold) and FT-Transformer (purple) exhibit expert bleeding, erroneously leaking wind sensitivity into the invariant tiles.

Scenario B: Temperature-Driven Physics. We isolate the thermal response by varying temperature (T_{air}) to probe the non-linear vegetation activation. As illustrated in Fig. 3, UniTeD recovers the latent sigmoidal activation of the Vegetation tile (Fig. 3(A)), capturing the ground-truth non-linearity. In contrast, FT-Transformer produces a response that smoothes over the activation threshold, and Soft-MoE learns a nearly linear representation for Vegetation (Fig. 3(A)). Both baselines leak this thermal sensitivity into the Water and Urban tiles (Fig. 3(B) and (C)), which should be invariant to temperature. This confirms that without explicit structural disentanglement, high-capacity models default to learning an entangled representation.

4.1.3 Experiment II: Zero-Shot Compositional Generalization

To test if the disentanglement shown in Exp. I enables zero-shot compositional generalization, we train exclusively on pure tiles and evaluate on a novel mixture configuration (40% Veg, 30% Urb, 30% Wat) never encountered during training (see Appendix C.7 for additional configurations).

Scenario A: Aerodynamic Generalization. We vary wind speed (U_{wind}) to assess the model's response within this unseen mixture. As shown in Fig. 4(A), UniTeD accurately tracks the grid cell aggregate ground truth (black triangles). Sub-grid analysis reveals that UniTeD correctly isolates the wind-driven Water response while remaining invariant for Vegetation and Urban tiles (Fig. 4(B)).

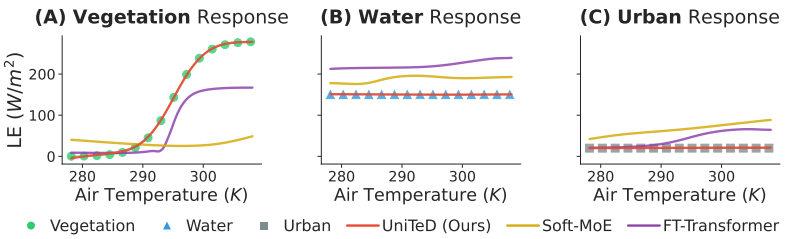

Figure 3: Disentanglement of Mixed Signals: Temperature-Driven Physics (Exp. I-B). Recovery of pure cell responses ($\alpha_k = 1.0$) after training exclusively on aggregate mixtures ($\alpha_k \leq 0.5$). UniTeD (red) successfully recovers the latent sigmoidal activation for Vegetation. In contrast, FT-Transformer (purple) and Soft-MoE (gold) produce smoothed responses that fail to resolve the non-linearity and erroneously leak the thermal signal into the invariant Water and Urban tiles.

In contrast, baselines fail to generalize. Fig. 4(C) shows Soft-MoE suffering from expert collapse, where its Urban expert erroneously reacts to the wind-driven Water response.

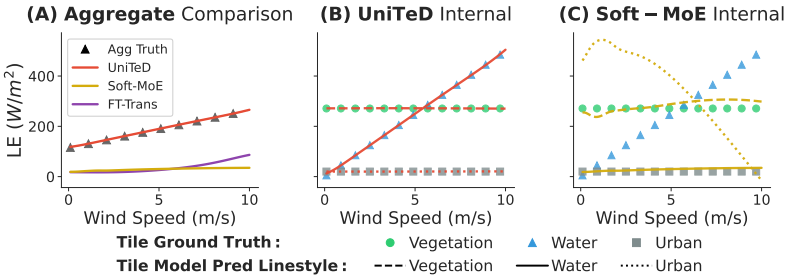

Figure 4: Compositional Generalization: Aerodynamic Response (Exp. II-A). Trained on pure tiles, evaluated on a novel $40/30/30$ mixture. (A) Aggregate: UniTeD (red) accurately tracks the ground truth (black triangles), whereas baselines drift. (B) UniTeD Internal: The model correctly disentangles latent physics. The Water output (solid line) recovers the aerodynamic signal, while Vegetation and Urban correctly remain invariant. (C) Soft-MoE Internal: Latent representations collapse. For instance, the Urban output (dotted line) violates invariance, erroneously reacting to wind speed (grey squares). *Colored markers denote Ground Truth; Line styles denote Model Predictions: Solid=Water, Dashed=Vegetation, Dotted=Urban.*

Scenario B: Thermal Generalization. Next, we vary air temperature (T_{air}). UniTeD successfully resolves this latent activation within the unseen mixture (Fig. 5(B)). Soft-MoE fails to capture the constitutive function (Fig. 5(C)), collapsing the sigmoidal curve into a linear approximation.

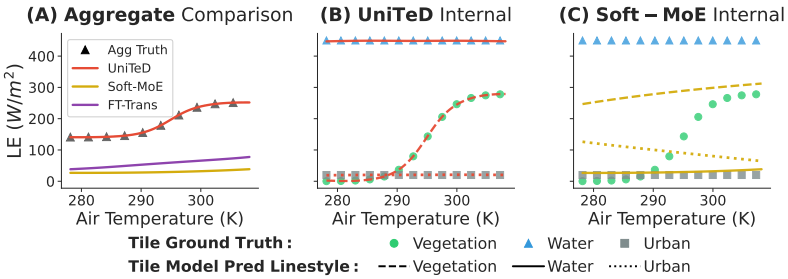

Figure 5: Compositional Generalization: Thermal Response (Exp. II-B). (A) Aggregate: UniTeD resolves the aggregate flux even as sigmoidal signals are diluted. (B) UniTeD Internal: Recovers the latent tile ground truth. (C) Soft-MoE Internal: Fails to recover the underlying tile processes.

4.1.4 EXPERIMENT III: FEW-SHOT PHYSICS TRANSFER

Finally, we isolate the data efficiency benefits of the shared expert by simulating a scenario where distinct tile variants share the underlying physics but differ drastically in data availability (details in Appendix C.4). For each tile type, we generate an *Abundant* variant ($N = 7,500$) and a *Rare* variant ($N = 10$). Both variants follow the exact same functional form $f_k(\mathbf{x})$ (e.g., sigmoidal temperature response), differing only by a scaling factor. Fig. 6 reveals the failure mode of disjoint experts (Frac-MoE) in this data-scarce regime. Because parameters are isolated, the expert for the rare variant cannot leverage the physical structure learned by the abundant expert and collapses into unphysical linear overfitting (purple line) when attempting to relearn the non-linearity from just 10 samples. In contrast, UniTeD (red) successfully transfers the constitutive function. By structurally coupling the variants via the shared representation M_θ, the abundant data constrains the global functional form f_k, allowing the model to generalize to the rare variant by simply adapting the embedding to capture the scaling factor rather than relearning the underlying physics.

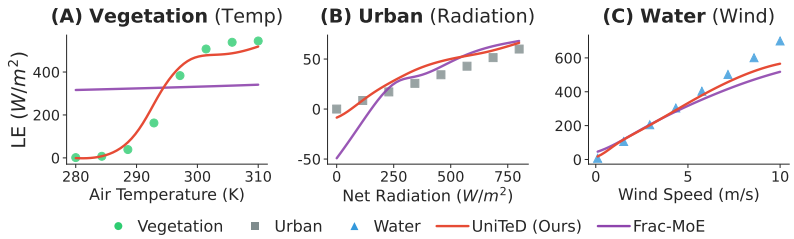

Figure 6: Few-Shot Physics Transfer (Exp. III). Models are evaluated on *Rare* variants ($N = 10$) that share physics with Abundant variants ($N = 7,500$) but differ by a scaling factor. While both models solve the Abundant task (not shown), Frac-MoE (purple) fails to generalize to rare tiles due to parameter isolation, producing linear artifacts. UniTeD (red) transfers non-linear physics (e.g., Sigmoidal Veg) from abundant to rare tiles via the shared expert, aligning with the Ground Truth.

Robustness to Non-Linear Coupling and Noisy Fractions. While the current experimental setup is based on linear superposition and perfectly known constituent fractions, we rigorously stress-test these assumptions in the Appendix. Against non-linear sub-grid coupling (Appendix C.5), UniTeD maintains structural identifiability even under severe lateral interactions ($\gamma = 1.0$), exhibiting stable degradation where unconstrained baselines suffer from pronounced representation collapse. Furthermore, when exposed to observation noise in the tile fractions (Appendix C.6), UniTeD maintains a significant performance margin over baselines, only losing its comparative advantage at extreme noise levels ($\sigma_f \geq 0.25$) when the prescribed aggregation logic becomes heavily corrupted.

4.2 VALIDATION ON LARGE-SCALE GLOBAL DATA

We next perform our evaluation on the ECMWF ERA5 reanalysis dataset (Hersbach et al., 2020) (see Appendix D), which provides a complex environment to stress-test methods at scale. This setting incorporates an additional prediction target (upward thermal radiation, L_{up}) and scales up to six physical tiles to test robustness in higher-dimensional systems. We adopt a temporal splitting strategy, training on the 7th of even months (2024) and evaluating on the 22nd. This protocol stratifies samples across the annual cycle to capture the seasonality of dynamic tiles (e.g., transient snow cover) while maintaining a $>$ 15-day buffer to mitigate synoptic autocorrelation.

Models are evaluated on two generalization tasks: **(i) Task 1 (Mixture $\rightarrow$ Pure):** Recovering pure tiles (test set) from aggregated mixtures ($\alpha_k \leq 0.5$, training set). **(ii) Task 2 (Pure $\rightarrow$ Mixture):** Generalizing to heterogeneous mixtures (test set) from pure-tile training ($\alpha_k \approx 1$, training set).

Task 1 (Table 1a) tests structural identifiability. While DNN performs well in-distribution test (60.4), it collapses on pure tiles (139.5), confirming it learned statistical shortcuts rather than physical laws. UniTeD reduces this pure-tile error by 32% relative to Soft-MoE and 38% relative to Frac-MoE. These gains validate UniTeD's dual benefits: the structural constraint enforces disentanglement (overcoming Soft-MoE's routing collapse), while the shared expert ensures robust learning for naturally rare tiles via transfer from abundant ones (overcoming Frac-MoE's parameter isolation).

Table 1: ERA5 Surface Energy Flux. Values represent the summation of RMSE ($W\,m^{-2}$) across G, LE, H, L_{up}. Results are aggregated across three independent runs and reported as mean $\pm$ std. Per-variable RMSE breakdowns are provided in Appendix D.4, showing similar performance trends.

(a) Disentanglement of Mixed Signals (Mixture→Pure). Trained on mixtures ($\alpha_k \leq 0.5$). UniTeD achieves significant gains by identifying constituent physics. Baselines suffer from routing collapse or parameter isolation.

		Purity Threshold ($\exists \alpha_k \geq \dots$)			
Model	**Test (Mixture)**	**.7**	**.8**	**.9**	**1.0 (Pure)**
DNN	$\mathbf{60.4 \pm 1.6}$	117.4 ± 10.1	122.3 ± 11.4	130.9 ± 15.6	139.5 ± 26.5
FT-Transformer	63.0 ± 1.1	118.2 ± 5.6	123.6 ± 6.3	131.6 ± 9.1	128.1 ± 13.9
Soft-MoE	66.7 ± 1.3	122.6 ± 3.9	127.5 ± 3.8	137.1 ± 3.4	134.5 ± 7.1
Frac-MoE	64.3 ± 1.5	129.1 ± 3.7	135.7 ± 4.6	147.6 ± 9.3	148.1 ± 35.1
UniTeD (Ours)	62.4 ± 0.8	$\mathbf{113.4 \pm 2.3}$	$\mathbf{118.2 \pm 2.6}$	$\mathbf{122.7 \pm 4.0}$	$\mathbf{92.1 \pm 11.4}$

(b) Zero-Shot Compositional Generalization (Pure→Mixture). Trained on pure tiles ($\alpha_k \approx 1$). UniTeD demonstrates superior performance. Modular baselines (Soft/Frac-MoE) degrade as mixture complexity increases.

		Mixture Threshold ($\forall \alpha_k \geq \dots$)		
Model	**Test (Pure)**	**.01**	**.05**	**.08**
DNN	$\mathbf{32.5 \pm 0.4}$	116.2 ± 13.3	124.3 ± 18.7	130.8 ± 24.9
FT-Transformer	36.1 ± 0.6	92.8 ± 3.7	98.2 ± 5.6	95.9 ± 2.5
Soft-MoE	39.6 ± 2.5	82.4 ± 1.6	86.9 ± 1.9	89.8 ± 2.5
Frac-MoE	33.5 ± 1.3	87.1 ± 1.9	89.9 ± 3.1	104.1 ± 1.9
UniTeD (Ours)	34.5 ± 0.3	$\mathbf{77.1 \pm 0.7}$	$\mathbf{78.2 \pm 0.3}$	$\mathbf{70.3 \pm 0.9}$

In Task 2 (Table 1b), UniTeD exhibits robust compositional results. Even at the highest evaluated mixture complexity (threshold 0.08), it bounds its error at 70.3, minimizing degradation relative to its pure-tile performance (34.5). In contrast, the modular baselines fail; both Soft-MoE and Frac-MoE see errors explode (89.8 and 104.1, respectively), meaning UniTeD achieves a 22% error reduction relative to the strongest baseline. For Soft-MoE, this indicates a failure of learned routing to generalize to unseen distributions. For Frac-MoE, the failure suggests that even with perfect physical routing, disjoint experts might fail to capture the shared underlying physics required to stabilize predictions in complex, noisy mixtures. Beyond compositional shifts, we also evaluate using geographical splits (Appendix D.5). When trained on the Northern Hemisphere and evaluated on the Southern Hemisphere, UniTeD achieves up to a 36% error reduction relative to baselines.

5 Conclusion

Physical systems with sub-grid heterogeneity benefit from resolving the processes of distinct sub-grid tiles to accurately model the macroscopic grid cell state. Therefore, monolithic models, which treat a grid cell as a uniform unit, are prone to failure when generalizing to novel mixture compositions. We propose UniTeD, a heterogeneity-aware framework that enforces structural linear superposition derived from physical conservation laws. By combining this explicit aggregation logic with a conditioned shared expert, UniTeD enables zero-shot recovery of latent constitutive functions and robust compositional generalization. UniTeD reduces grid cell-level prediction error by up to 28% compared to the strongest baselines on ERA5 surface flux inference, demonstrating that embedding physical structure into neural architectures is essential for achieving reliable scientific emulators.

Limitations and Future Work. Identifiability currently relies on the independence of sub-grid fractions from the physical forcing; if sub-grid mixtures are static or perfectly correlated with the forcing, the inverse problem is under-determined. Furthermore, while UniTeD effectively handles linear superposition, it is naturally extensible to non-linear aggregation (e.g., hydrodynamic crowding) or strong inter-tile interactions (e.g., lateral advection), opening avenues for future work. These include: (i) incorporating known non-linear physical interactions directly into the aggregation operator or employing specialized modules to learn unknown interaction logic, and (ii) integrating Graph Neural Networks to capture spatial inter-tile interactions. These extensions can enable fully coupled multiscale modeling without reintroducing the entanglement of monolithic approximations.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on hypotheses and proofs, edit the paper to improve readability, identify relevant literature, and modify scientific figures. We have reviewed all AI-assisted content through careful proofreading. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our empirical results, we provide the complete codebase as supplementary material attached to this submission, which includes all data processing scripts, UniTeD and baseline re-implementations, and the full training and evaluation pipelines. Furthermore, to ensure the reproducibility of our theoretical claims, a complete algebraic proof of structural identifiability, along with a detailed discussion of all underlying physical assumptions, is provided in Appendix A. Comprehensive details regarding model hyperparameters, architectural configurations, and ERA5 data processing are detailed in Appendix B and D.

REFERENCES

Benedikt Alkin, Andreas Fürst, Simon Lucas Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers: A framework for efficiently scaling neural operators. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Gianpaolo Balsamo, Anton Beljaars, Klaus Scipal, Pedro Viterbo, Bart van den Hurk, Martin Hirschi, and Alan K. Betts. A revised hydrology for the ecmwf model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, 10(3):623 – 643, 2009.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

David Bolton. The computation of equivalent potential temperature. *Monthly weather review*, 108 (7):1046–1053, 1980.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

N. W. Chaney, M. H. J. Van Huijgevoort, E. Shevliakova, S. Malyshev, P. C. D. Milly, P. P. G. Gauthier, and B. N. Sulman. Harnessing big data to rethink land heterogeneity in earth system models. *Hydrology and Earth System Sciences*, 22(6):3311–3330, 2018.

Filipe De Avila Belbute-Peres, Thomas Economon, and J. Zico Kolter. Combining differentiable PDE solvers and graph neural networks for fluid flow prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2402–2411, 2020.

ECMWF. *IFS Documentation CY41R2 - Part IV: Physical Processes*. Number 4. ECMWF, 2016. URL `https://www.ecmwf.int/node/16648`.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Dapeng Feng, Jiangtao Liu, Kathryn Lawson, and Chaopeng Shen. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, 58(10):e2022WR032404, 2022.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT Press, 2016.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Xu Han, Han Gao, Tobias Pfaff, Jian-Xun Wang, and Liping Liu. Predicting physics in mesh-reduced space with temporal attention. In *International Conference on Learning Representations*, 2022.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jimmy Gaspard Jean, Tung-Huan Su, Szu-Jui Huang, Cheng-Tang Wu, and Chuin-Shan Chen. Graph-enhanced deep material network: multiscale materials modeling with microstructural informatics. *Computational Mechanics*, 75(1):113–136, 2025.

Arnold Kazadi, James Doss-Gollin, and Arlei Silva. Pluvial flood emulation with hydraulics-informed message passing. In *Forty-first International Conference on Machine Learning*, 2024.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. AIFS – ECMWF's data-driven forecasting system, 2024.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-informed neural operator for large-scale 3D PDEs. In *Advances in Neural Information Processing Systems*, pp. 35836–35854, 2023.

Shouyi Liu, Xiaokang Yang, and Yuntian Chen. Riesz neural operator for solving partial differential equations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zeliang Liu, CT Wu, and M Koishi. A deep material network for multiscale topology learning and accelerated nonlinear modeling of heterogeneous materials. *Computer Methods in Applied Mechanics and Engineering*, 345:1138–1168, 2019.

Tyler McCandless, David John Gagne, Branko Kosović, Sue Ellen Haupt, Bai Yang, Charlie Becker, and John Schreck. Machine learning for improving surface-layer-flux estimates. *Boundary-Layer Meteorology*, 185(2):199–228, 2022.

David R Musicant, Janara M Christensen, and Jamie F Olson. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 252–261, 2007.

Ruben Ohana, Michael McCabe, Lucas Thibaut Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond Buschman Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah M. Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024.

N. Raoult, E. Pinnington, M. Santa Cruz, F. Pinault, B. Raoult, N. Zelenka, G. Arduini, G. Balsamo, S. Boussetta, M. Chantry, P. de Rosnay, P. Dueben, and C. Rüdiger. ailand v1: Physics-based land surface emulator with observational fine-tuning. *EGUsphere*, 2026:1–39, 2026.

Franz Roters, Philip Eisenlohr, Luc Hantcherli, Denny Dharmawan Tjahjanto, Thomas R Bieler, and Dierk Raabe. Overview of constitutive laws, kinematics, homogenization and multiscale methods in crystal plasticity finite-element modeling: Theory, experiments, applications. *Acta materialia*, 58(4):1152–1211, 2010.

Vinay Sharma and Olga Fink. A physics-informed graph neural network conserving linear and angular momentum for dynamical systems. *Nature Communications*, 17, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

L Torres-Rojas, N Vergopolan, JD Herman, and NW Chaney. Towards an optimal representation of sub-grid heterogeneity in land surface models. *Water Resources Research*, 58(12): e2022WR032233, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Shuochen Wang, Nishant Yadav, and Auroop R. Ganguly. Paraformer: Parameterization of sub-grid scale processes using transformers, 2024. URL `https://arxiv.org/abs/2412.16763`.

Sifan Wang, Zehao Dou, Siming Shan, Tong-Rui Liu, and Lu Lu. Fundiff: diffusion models over function spaces for physics-informed generative modeling. *Nature Communications*, 17:5749, 2026.

Zixi Wei, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Xiaofeng Zhu, and Heng Tao Shen. A universal unbiased method for classification from aggregate observations. In *International Conference on Machine Learning*, pp. 36804–36820, 2023.

Volker Wulfmeyer, Juan Manuel Valencia Pineda, Sebastian Otte, Matthias Karlbauer, Martin V Butz, Temple R Lee, and Verena Rajtschan. Estimation of the surface fluxes for heat and momentum in unstable conditions with machine learning and similarity approaches for the lafe data set. *Boundary-Layer Meteorology*, 186(2):337–371, 2023.

M. Yin, Y. Han, Y. Wang, W. Sun, J. Deng, D. Wei, Y. Kong, and B. Wang. Climate impacts of parameterizing subgrid variation and partitioning of land surface heat fluxes to the atmosphere with the ncar cesm1.2. *Geoscientific Model Development*, 16(1):135–156, 2023.

A Formal Proof of Structural Identifiability

We provide the algebraic derivation for Theorem 3.1. Before presenting the formal proof, we first contextualize how the mathematical assumptions of the theorem naturally emerge from the physical systems we target. We then prove exact identifiability under ideal conditions (identical forcing), and finally demonstrate stability under input perturbations.

A.1 Physical Grounding of Theoretical Assumptions

The assumptions underpinning Theorem 3.1 were not devised to force a mathematical proof, but rather to leverage the properties of physical systems characterized by sub-grid heterogeneity and linear superposition derived from conservation laws.

- **Regularity:** This condition requires a compact input space. This is naturally satisfied, as physical state variables (e.g., temperature, solar radiation) are inherently bounded.
- **Structural Exclusion:** This dictates that the estimator $\hat{\boldsymbol{\phi}}_k$ must be functionally independent of the constituent fraction $\boldsymbol{\alpha}$. Physically, this mirrors reality: a water tile's energy flux responds to atmospheric forcing in the exact same manner regardless of whether it covers 10% or 90% of a grid cell.
- **Local Non-Degeneracy:** This requires diverse mixture configurations to ensure the fraction matrix achieves full column rank. This requirement is directly satisfied by the spatio-temporal dynamics of Earth's climate system captured by datasets like ERA5. Meteorological forcings operate at the synoptic scale, often spanning hundreds or thousands of kilometers, meaning a single, uniform meteorological condition can simultaneously blanket many distinct grid cells. Furthermore, the same weather conditions experienced in one region today can occur in a completely different geographical region tomorrow. In contrast, land-cover geography (mixture fractions, $\boldsymbol{\alpha}$) varies at much finer micro-scales independently of these transient weather patterns. Consequently, a local neighborhood of identical meteorological states ($\mathbf{x}$) will naturally encompass highly diverse fractional compositions ($\boldsymbol{\alpha}$). This spatio-temporal variance guarantees that the fraction matrices across local batches are well-conditioned, providing the necessary mathematical rank to stably unmix the latent physical signals.
  For a quantitative empirical validation of this local non-degeneracy assumption using the ERA5 dataset, see Appendix D.6.

A.2 Exact Identifiability (Ideal Case)

Consider a local batch of $N \geq K$ observations $\mathcal{B} = \{i_1, \ldots, i_N\}$ sharing identical forcing $\mathbf{x}_i = \mathbf{x}^*$. Condition 2 (Structural Exclusion) ensures the model's latent response is invariant across the batch: $\hat{\boldsymbol{\phi}}(\mathbf{x}_i^*) = \hat{\boldsymbol{\phi}}(\mathbf{x}^*)$. Similarly, the ground truth $\boldsymbol{\phi}^*(\mathbf{x}^*)$ is invariant.

Minimizing the reconstruction loss implies matching the model predictions to the ground truth observations:

$$A_{\mathcal{B}}\hat{\boldsymbol{\Phi}} = \mathbf{Y}_{\mathcal{B}} = A_{\mathcal{B}}\boldsymbol{\Phi}^* \tag{5}$$

where $\mathbf{Y}_{\mathcal{B}}$ is the matrix of aggregate observations for the batch. The *Local Non-Degeneracy* condition asserts that the constituent fraction matrix $A_{\mathcal{B}}$ has full column rank (K). Left-multiplying by the pseudo-inverse $A_{\mathcal{B}}^{\dagger} = (A_{\mathcal{B}}^{\top}A_{\mathcal{B}})^{-1}A_{\mathcal{B}}^{\top}$ isolates the target variables and yields the unique solution $\hat{\boldsymbol{\Phi}} = \boldsymbol{\Phi}^*$. Because the matrix is full rank, the null space is trivial; no non-zero combination of component errors can sum to zero.

A.3 Stability Analysis (Real-World Case)

We analyze the stability of the recovery for vector-valued outputs $\mathbf{y} \in \mathbb{R}^D$ when sample inputs lie within a neighborhood $\mathcal{N}_r = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_2 \leq r\}$.

System Setup and Criteria:

1. **Realizability:** The shared expert M_θ possesses sufficient capacity to represent the ground truth physical functions $\boldsymbol{\phi}^*$.
2. **Empirical Optimization Success:** The training procedure minimizes the empirical root mean squared error (RMSE) within the neighborhood $\mathcal{N}_r$ to a tolerance δ. That is, for a batch of N samples, $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2} \leq \delta$.

Regularity Conditions:

1. **Ground Truth:** The true physical functions $\boldsymbol{\phi}_k^*(\mathbf{x})$ are L^*-Lipschitz with respect to the Euclidean norm.
2. **Latent Tile Response:** The sequence of full estimators $\hat{\boldsymbol{\phi}}_k(\mathbf{x})$ produced during optimization is uniformly $\hat{L}$-Lipschitz with respect to the Euclidean norm. That is, as the optimization error $\delta \to 0$, the Lipschitz constant $\hat{L}$ remains strictly bounded ($\sup \hat{L} < \infty$).

Perturbation Bounds (Vector Case): For a sample i with input $\mathbf{x}_i \in \mathcal{N}_r$, we define the variations from the reference state $\mathbf{x}^*$ as:

$$\boldsymbol{\delta}_{i,k}^* = \boldsymbol{\phi}_k^*(\mathbf{x}_i) - \boldsymbol{\phi}_k^*(\mathbf{x}^*) \quad \text{and} \quad \hat{\boldsymbol{\delta}}_{i,k} = \hat{\boldsymbol{\phi}}_k(\mathbf{x}_i) - \hat{\boldsymbol{\phi}}_k(\mathbf{x}^*) \tag{6}$$

Their Euclidean norms are bounded by $\|\boldsymbol{\delta}_{i,k}^*\|_2 \leq L^* r$ and $\|\hat{\boldsymbol{\delta}}_{i,k}\|_2 \leq \hat{L} r$ respectively.

Let $\boldsymbol{\zeta}_i = \mathbf{y}_i - \hat{\mathbf{y}}_i$ be the optimization residual. We equate the model and truth:

$$\hat{\mathbf{y}}_i + \boldsymbol{\zeta}_i = \mathbf{y}_i \implies \sum_{k=1}^{K} \alpha_{i,k}(\hat{\boldsymbol{\phi}}_k(\mathbf{x}^*) + \hat{\boldsymbol{\delta}}_{i,k}) + \boldsymbol{\zeta}_i = \sum_{k=1}^{K} \alpha_{i,k}(\boldsymbol{\phi}_k^*(\mathbf{x}^*) + \boldsymbol{\delta}_{i,k}^*) \tag{7}$$

Rearranging to isolate the parameter error at the reference state:

$$\sum_{k=1}^{K} \alpha_{i,k}(\hat{\boldsymbol{\phi}}_k(\mathbf{x}^*) - \boldsymbol{\phi}_k^*(\mathbf{x}^*)) = \underbrace{\sum_{k=1}^{K} \alpha_{i,k}(\boldsymbol{\delta}_{i,k}^* - \hat{\boldsymbol{\delta}}_{i,k})}_{\mathbf{p}_i} - \boldsymbol{\zeta}_i \tag{8}$$

Because the constituent fractions represent a partition of unity ($\sum_{k=1}^{K} \alpha_{i,k} = 1$), the Lipschitz perturbation vector $\mathbf{p}_i$ is bounded by $\|\mathbf{p}_i\|_2 \leq (L^* + \hat{L})r$.

Matrix Formulation: Let $\Delta\Phi = \hat{\boldsymbol{\Phi}}(\mathbf{x}^*) - \boldsymbol{\Phi}^*(\mathbf{x}^*)$ be the error matrix at the reference state. We stack the equations for the batch $\mathcal{B}$ into a matrix system:

$$A_{\mathcal{B}} \Delta\Phi = \mathbf{E} \tag{9}$$

where $\mathbf{E} = \mathbf{P} - \mathbf{Z}$ is the matrix of stacked residuals, with $\mathbf{P}$ representing the stacked perturbations $\mathbf{p}_i$ and $\mathbf{Z}$ the stacked optimization errors $\boldsymbol{\zeta}_i$. Constrained by our empirical RMSE condition, $\|\mathbf{Z}\|_F \leq \sqrt{N}\delta$. By applying the triangle inequality, the Frobenius norm of $\mathbf{E}$ is bounded by:

$$\|\mathbf{E}\|_F \leq \|\mathbf{P}\|_F + \|\mathbf{Z}\|_F \leq \sqrt{N}(L^* + \hat{L})r + \sqrt{N}\delta \tag{10}$$

Inverting the system using the pseudo-inverse $A_{\mathcal{B}}^{\dagger}$ (provided local non-degeneracy holds with fixed N and $\sigma_{\min}(A_{\mathcal{B}}) \geq c(\mathbf{x}^*) > 0$):

$$\|\Delta\Phi\|_F \leq \frac{\sqrt{N}}{\sigma_{\min}(A_{\mathcal{B}})} \cdot \Big((L^* + \hat{L})r + \delta\Big) \tag{11}$$

Conclusion: The recovery error is bounded linearly by the neighborhood size r and the optimization tolerance δ. For a fraction matrix $A_{\mathcal{B}}$ whose rows are sampled from a non-degenerate distribution over the simplex, the minimum singular value $\sigma_{\min}(A_{\mathcal{B}})$ scales asymptotically as $\Theta(\sqrt{N})$. This ensures the prefactor $\frac{\sqrt{N}}{\sigma_{\min}(A_{\mathcal{B}})}$ is $\mathcal{O}(1)$, meaning the theoretical bound remains strictly stable and does not diverge for arbitrarily large batch sizes. Since $\sigma_{\min}(A_{\mathcal{B}}) > 0$, component errors cannot cancel each other out; any deviation in latent parameters $\Delta\Phi$ strictly induces a non-zero aggregate error. Thus, as $r \to 0$ (locality) and $\delta \to 0$ (successful training), the component error vanishes ($\|\Delta\Phi\|_F \to 0$).

B Implementation Details and Hyperparameters

We provide the complete architectural specifications, baseline definitions, and training protocols to ensure reproducibility of the results presented in Section 5. The following configurations were consistent across all ERA5 experiments.

B.1 Architectural Specifications

All models were trained using a composite loss function comprising Mean Squared Error (MSE) on the 4 output fluxes and a physical constraint term:

$$\mathcal{L} = \text{MSE}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda \cdot \mathcal{L}_{energy} \tag{12}$$

where $\mathcal{L}_{energy}$ penalizes violations of surface energy conservation, implemented as a soft penalty weighted by $\lambda = 10^{-4}$. Depending on the dataset, this penalty is defined as the mean squared energy balance residual.

For the synthetic dataset:

$$\mathcal{L}_{energy} = \frac{1}{N}\sum_{i=1}^{N}\left(R_{net}^{(i)} - \hat{LE}^{(i)} - \hat{H}^{(i)} - \hat{G}^{(i)}\right)^2 \tag{13}$$

For the ERA5 dataset:

$$\mathcal{L}_{energy} = \frac{1}{N}\sum_{i=1}^{N}\left(SW_{net}^{(i)} + LW_{\downarrow}^{(i)} - \hat{L}_{up}^{(i)} - \hat{LE}^{(i)} - \hat{H}^{(i)} - \hat{G}^{(i)}\right)^2 \tag{14}$$

Prior to computing this residual, all normalized network outputs ($\hat{L}_{up}, \hat{LE}, \hat{H}, \hat{G}$) and normalized inputs ($SW_{net}, LW_{\downarrow}$) are un-scaled back to their native physical units ($W\,m^{-2}$).

Shared Trunk (T_ψ). For all Transformer-based models (UniTeD, FT-Transformer, and MoE baselines), we utilize the standard Transformer encoder for tabular data as implemented in the original FT-Transformer Gorishniy et al. (2021). We configure it as a lightweight backbone:

- **Structure:** 1-layer Transformer Encoder.
- **Dimensions:** Token embedding dimension of 128.
- **Attention:** 4 heads. The Key, Query, and Value matrices are all 128×128.
- **Feed-Forward Network (FFN):** A 2-layer block projecting the concatenated head outputs (128×4 signals) $\rightarrow 512 \rightarrow 128$.

UniTeD (Shared Expert). The core architecture augments the shared trunk with two lightweight modules:

- **FiLM Generator:** A modulation network where the γ (scale) and β (shift) components are each generated by a single linear layer (dim 128). The condition vector τ is the tile identity.
- **Shared Expert (M_θ):** A dense network with 2 hidden layers ($128 \rightarrow 256 \rightarrow 256 \rightarrow 4$).
- **Input:** The trunk receives only the physical state $\mathbf{x}$. The mixing fractions $\boldsymbol{\alpha}$ are used solely for the physics-prescribed aggregation step.

B.2 Baseline Configurations

1. FT-Transformer. A monolithic baseline using the exact same trunk architecture as UniTeD but without structural disentanglement.

- **Structure:** Identical 1-layer Transformer trunk and FFN.
- **Difference:** The input includes both the physical state and the mixing fractions $[\mathbf{x}, \boldsymbol{\alpha}]$. It lacks the FiLM and Aggregation mechanism.

2. Frac-MoE (Disjoint Experts). An ablation replacing the single shared expert with $K = 6$ independent networks.

- **Structure:** Uses the same 1-layer Transformer trunk.
- **Experts:** Instead of one shared expert, it instantiates 6 separate experts, each matching the UniTeD expert architecture (2 hidden layers, $128 \to 256 \to 256 \to 4$).
- **Aggregation:** Like UniTeD, it aggregates outputs using the true mixing fractions $\boldsymbol{\alpha}$, but parameters are not shared across regimes.

3. Soft-MoE (Learned Routing). Identical to Frac-MoE but ignores the known physical mixing fractions.

- **Routing:** Adds a learnable 1-layer MLP router ($128 \to 6$) to predict mixing weights from the latent state.
- **Input:** Receives $[\mathbf{x}, \boldsymbol{\alpha}]$ to allow the model to potentially infer the mixing ratios.

4. Monolithic DNN. A purely dense baseline replacing the Transformer trunk with a deep MLP encoder.

- **Structure:** A 4-layer MLP encoder with a wide hidden dimension of 768.
- **Output Head:** A block with 2 hidden layers ($768 \to 512 \to 512 \to 4$).
- **Input:** Concatenated vector $[\mathbf{x}, \boldsymbol{\alpha}]$.

B.3 Summary of Model Complexity

Table 2 contrasts the parameter efficiency and input strategies of all evaluated models. Note that despite having $\sim 3\times$ fewer parameters than the disjoint MoE baselines, UniTeD achieves superior generalization.

Table 2: **Model Architecture and Complexity Comparison.** All models predict 4 target fluxes.

Model	Input	Core Mechanism	Routing / Agg.	Parameters
UniTeD (Ours)	$\mathbf{x}$	Shared Expert + FiLM	Physics-Prescribed ($\boldsymbol{\alpha}$)	666,756
FT-Transformer	$[\mathbf{x}, \boldsymbol{\alpha}]$	Monolithic Transformer	Implicit (Black Box)	665,988
Frac-MoE	$\mathbf{x}$	6 Disjoint Experts	Physics-Prescribed ($\boldsymbol{\alpha}$)	1,825,560
Soft-MoE	$[\mathbf{x}, \boldsymbol{\alpha}]$	6 Disjoint Experts	Learned Soft Router	1,827,102
Monolithic MLP	$[\mathbf{x}, \boldsymbol{\alpha}]$	Deep MLP Encoder	Implicit (Black Box)	5,947,183

B.4 Training Protocol and Hyperparameters

To ensure reproducibility across all experiments, models were trained using the following consistent optimization protocol:

- **Optimizer:** AdamW with an initial learning rate of 10^{-4}.
- **Learning Rate Scheduler:** A Reduce-on-Plateau scheduler was employed, which decayed the learning rate by a factor of 0.5 if the total validation loss did not improve for 10 consecutive epochs (with the minimum learning rate bounded at 10^{-6}).
- **Batch Size & Duration:** Models were trained for a maximum of 400 epochs with a batch size of 1920. To ensure optimal generalization, the best model weights were saved based on the lowest RMSE evaluated on the validation set.
- **Hardware:** All experiments were executed on a single NVIDIA A100 GPU.

C Synthetic Benchmark: Generation Protocols and Extended Results

To rigorously evaluate structural identifiability and compositional generalization, we constructed a "Physics-lite" synthetic environment governed by the surface energy balance equation. This appendix details the constitutive functions, sampling distributions, and aggregation logic used to generate the dataset.

C.1 Governing Physical Equations

The system represents a thermodynamic grid cell where the net radiation (R_{net}) is partitioned into Latent Heat (LE), Sensible Heat (H), and Ground Heat (G) fluxes according to the conservation law:

$$R_{net} = LE + H + G \tag{15}$$

Each grid cell is composed of at most three distinct tile types: **Vegetation** (veg), **Water** (wat), and **Urban** (urb). The grid-level aggregate flux ($\bar{\phi}$) for any variable $\phi \in \{LE, H, G\}$ is the area-weighted average of the sub-grid components:

$$\bar{\phi} = \alpha_{veg}\phi_{veg} + \alpha_{wat}\phi_{wat} + \alpha_{urb}\phi_{urb} \tag{16}$$

where $\sum \alpha_k = 1$.

C.2 Constitutive Functions for Sub-Grid Fluxes

To enforce a double dissociation of causal drivers, we defined distinct constitutive equations for the Latent Heat Flux (LE) of each regime. The Sensible Heat (H) and Ground Heat (G) fluxes were derived to satisfy energy conservation.

1. Vegetation Regime (Temperature Driven) Vegetation fluxes are driven by a non-linear biological activation function dependent on Air Temperature (T_{air}), simulating stomatal conductance.

$$LE_{veg} = 0.7 \cdot R_{net} \cdot \sigma\left(\frac{T_{air} - 295}{2.5}\right) \tag{17}$$

$$G_{veg} = 0.05 \cdot R_{net} \tag{18}$$

$$H_{veg} = R_{net} - LE_{veg} - G_{veg} \tag{19}$$

where $\sigma(\cdot)$ is the sigmoid function.

2. Water Regime (Aerodynamically Driven) Water fluxes are driven linearly by aerodynamic terms, specifically Wind Speed (U_{wind}) and Vapor Pressure Deficit (VPD).

$$LE_{wat} = 25 \cdot U_{wind} \cdot \text{VPD} \tag{20}$$

$$G_{wat} = 0.5 \cdot R_{net} \tag{21}$$

$$H_{wat} = R_{net} - LE_{wat} - G_{wat} \tag{22}$$

3. Urban Regime (Radiatively Driven) Urban fluxes are driven primarily by available radiation, with minimal latent heat partitioning (simulating impervious surfaces).

$$LE_{urb} = 0.05 \cdot R_{net} \tag{23}$$

$$G_{urb} = 0.3 \cdot R_{net} \tag{24}$$

$$H_{urb} = R_{net} - LE_{urb} - G_{urb} \tag{25}$$

C.3 Sampling Distributions and Noise

To ensure adequate coverage of the state space, input variables were sampled uniformly from physically plausible ranges appropriate for a diurnal cycle.

Table 3: Sampling ranges for meteorological forcing variables.

Variable	Symbol	Distribution	Unit
Net Radiation	R_{net}	$\mathcal{U}(0, 800)$	W/m^2
Air Temperature	T_{air}	$\mathcal{U}(280, 310)$	K
Wind Speed	U_{wind}	$\mathcal{U}(0.1, 10)$	m/s
Vapor Pressure Deficit	VPD	$\mathcal{U}(0, 4)$	hPa

Fractional Coverages The tile fractions $\boldsymbol{\alpha} = [\alpha_{veg}, \alpha_{urb}, \alpha_{wat}]$ were sampled from a Dirichlet distribution to ensure $\sum \alpha_k = 1$. For the Generalization Tests (Experiment II), fixed ratios were used (e.g., $[0.4, 0.3, 0.3]$) that were explicitly held out from the training distribution.

Observation Noise To simulate realistic sensor error and measurement uncertainty, zero-centered Gaussian noise ϵ was added to the aggregate target variables ($\bar{Y}_{obs} = \bar{Y}_{true} + \epsilon$). The standard deviation of the noise was set to 10% of the standard deviation of the clean target signal:

$$\epsilon \sim \mathcal{N}(0, (0.1 \cdot \sigma_{target})^2) \tag{26}$$

The models were trained only on the noisy aggregate targets $\bar{Y}_{obs}$ and the aggregate forcings $\bar{X}$, never observing the individual tile fluxes ϕ_k.

C.4 Protocol for Few-Shot Transfer (Experiment III)

To evaluate the transfer of physical laws from data-rich to data-starved regimes, we modified the generation protocol to create paired "Abundant" and "Rare" classes for all three regimes (Vegetation, Urban, Water).

Shared Physics, Distinct Efficiency The core hypothesis of Exp. III is that rare classes share the same underlying constitutive functions (functional forms) as abundant classes but differ in specific efficiency parameters. To simulate this, we introduced a regime-specific efficiency coefficient C_{eff} into the flux equations:

$$LE_k = C_{eff}^{(k)} \cdot f_k(\mathbf{x}) \tag{27}$$

where $f_k(\mathbf{x})$ is the base physical function defined in Section B.2. We generated two variations for each regime:

- **Abundant Class (Standard):** Represents the baseline response ($N = 7,500$). We assigned a standard coefficient $C_{eff} = 1.0$.
- **Rare Class (Modified):** Represents a distinct variant governed by the same physics but with altered magnitude (e.g., higher stomatal density or roughness). We assigned varying coefficients $C_{eff} > 1.0$ (specifically 1.3 for Vegetation, 1.5 for Urban, and 1.4 for Water) to test extrapolation.

For example, the Rare Vegetation flux follows $LE_{rare} = 1.3 \cdot [0.7 \cdot R_{net} \cdot \sigma(T_{air})]$, retaining the sigmoidal temperature dependence but with amplified amplitude.

Data Scarcity Protocol To simulate the long-tail distribution, we imposed extreme data imbalance between these paired classes:

- **Abundant Training Set:** $N = 7,500$ samples. This provides sufficient density for the model to learn the complex non-linear shape of $f_k(\mathbf{x})$.
- **Rare Training Set:** $N = 10$ samples. This is insufficient to learn non-linear constitutive functions from scratch, forcing the model to transfer the structural knowledge (f_k) from the abundant class and merely adapt the scalar coefficient (C_{eff}).

C.5 Robustness to Non-Linear Sub-Grid Coupling (Experiment IV)

We extended our synthetic benchmark to include non-linear inter-tile coupling to evaluate model robustness. We modified the grid-level aggregation (Equation 16) for Latent Heat (LE) by introducing a quadratic interaction penalty governed by a coupling strength parameter γ:

$$LE_{grid} = \sum_k \alpha_k LE_k(\mathbf{x}) + \kappa \left(\alpha_{veg}\alpha_{wat}\psi_{vw} + \alpha_{veg}\alpha_{urb}\psi_{vu} + \alpha_{urb}\alpha_{wat}\psi_{uw}\right) \tag{28}$$

where the physical cross-terms ψ_{ij} explicitly model distinct mechanisms of lateral heat transfer between specific tile pairs:

- **Vegetation-Water (Oasis Effect):** Adjacent water bodies advectively boost the latent heat flux of vegetation. This interaction scales with the latent heat of the water tile: $\psi_{vw} = 0.5 \cdot LE_{wat}$.
- **Vegetation-Urban (Urban Heat Island):** Sensible heat dissipating from impervious urban tiles provides additional thermal energy that drives further vegetation transpiration: $\psi_{vu} = 0.2 \cdot H_{urb}$.
- **Urban-Water (Evaporative Interaction):** A generic local interaction driven directly by available net radiation: $\psi_{uw} = 0.1 \cdot R_{net}$.

To rigorously enforce the surface energy balance constraint ($R_{net} = LE + H + G$), any additional latent heat generated by the coupling perturbation ($\kappa \sum \alpha_i \alpha_j \psi_{ij}$) was strictly deducted from the aggregate Sensible Heat (H_{grid}), while Ground Heat (G_{grid}) remained unaffected.

Experimental Setup We generated datasets featuring highly mixed grid cells across an extreme spectrum of interaction strengths, ranging from perfectly linear ($\gamma = 0.0$) to severely coupled ($\gamma = 1.0$), where lateral interactions act with equal strength to the base physics. Models were trained on these highly distorted mixtures and evaluated on their ability to isolate and recover the pure, uncoupled tile responses ($\gamma = 0.0$), representing the true latent constitutive functions.

Results As illustrated in Figure 7(A), UniTeD exhibits highly stable, monotonic degradation, robustly accommodating the non-linear perturbations without severe error divergence. In contrast, unconstrained monolithic models (FT-Transformer) and data-driven routers (Soft-MoE) demonstrate acute instability leading to sharp error peaks across the training runs, with FT-Transformer reaching an RMSE of $\sim$169 at $\gamma = 0.45$. Figure 7(B) illustrates the practical impact at the extreme limit ($\gamma = 1.0$). Even under this severe lateral interaction scenario, UniTeD closely approximates the underlying physics and preserves the sigmoidal functional form of the vegetation tile. This confirms that UniTeD provides a robust, mathematically bounded approximation even under extreme violations of the linear superposition.

C.6 Robustness to Noisy Tile Fractions (Experiment V)

The theoretical guarantees of UniTeD rely on physics-prescribed aggregation using known constituent fractions α. To evaluate the framework's robustness to real-world classification errors and sensor uncertainty, we simulated a scenario where the true physical environment is governed by clean fractions, but the models are trained using noisy observed fractions.

Experimental Setup We injected zero-centered Gaussian noise ($\sigma_f \in [0.0, 0.30]$) into the training fractions, subsequently clipping and re-normalizing them to ensure they represent a valid probability simplex. The models were trained on these corrupted mixtures and then evaluated on their ability to recover the pure latent constitutive functions using a clean, 0% noise test set.

Results As illustrated in Figure 8, even under the regime of noisy fraction values, UniTeD maintains a significant margin of performance compared to the baselines, and only loses its comparative advantage at extreme noise levels ($\sigma_f \geq 0.25$), where the prescribed aggregation logic becomes too heavily corrupted by the noisy inputs.

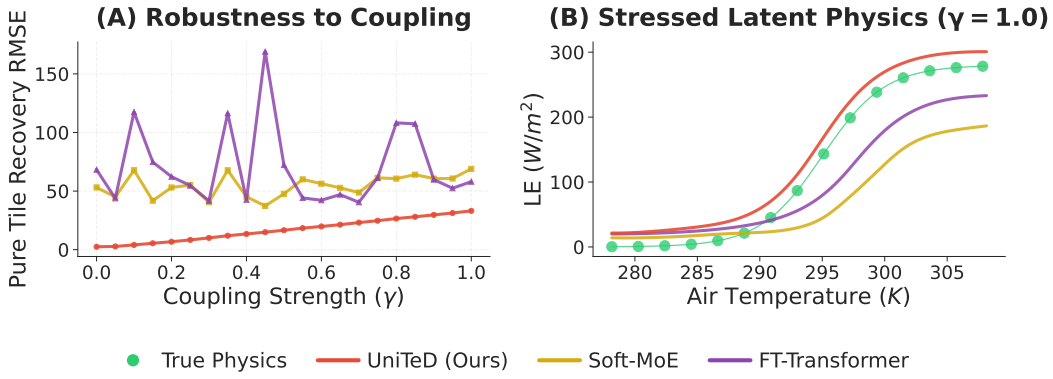

Figure 7: **Robustness to Non-Linear Sub-Grid Interactions (Exp. IV).** (A) Pure tile recovery RMSE as a function of the coupling strength (γ). While unconstrained baselines exhibit severe fluctuations and sharp error divergence (e.g., FT-Transformer at $\gamma = 0.45$), UniTeD demonstrates stable, monotonic degradation. (B) Sub-grid recovery at the extreme limit ($\gamma = 1.0$). UniTeD preserves the functional form (the sigmoidal curve) of the true vegetation physics, while baselines yield dampened representations.

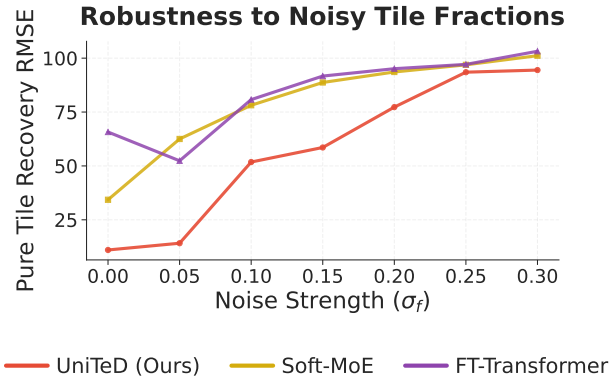

Figure 8: **Robustness to Noisy Tile Fractions (Exp. V).** Pure tile recovery RMSE evaluated across varying levels of Gaussian noise (σ_f) injected into the observed constituent fractions during training. Even under noisy conditions, UniTeD (red) maintains a significant disentanglement advantage, converging with the baselines only when the physical aggregation logic is severely corrupted at extreme noise levels ($\sigma_f \geq 0.25$).

C.7 Robustness to Compositional Variance on Synthetic Data (Experiment II Extended)

To verify that the zero-shot success of UniTeD is not localized to the specific 40/30/30 mixture analyzed in the main text, we evaluate the model across a wide spectrum of compositional ratios.

As shown in Figure 9 and Figure 10, UniTeD maintains near-perfect sub-grid recovery across three distinct edge-case configurations:

1. **Dominant Tile (80/10/10):** UniTeD correctly resolves the minority signals (e.g., 10% Urban) even when they are heavily diluted by the dominant Vegetation flux, proving it does not suffer from "minority collapse."
2. **Sparse Mixture (50/50/0):** UniTeD accurately recovers the latent response of the absent Water tile while mathematically scaling its aggregate contribution to zero, demonstrating strict structural identifiability.
3. **Balanced Interference (33/33/34):** UniTeD successfully de-mixes the aggregate signal when all three tiles contribute nearly equal weights, which represents the maximum entropy state for the demixing problem.

In all cases, UniTeD's conditioned responses remain physically stable, following the ground-truth sigmoidal and linear profiles regardless of their tile fraction. In contrast, the Soft-MoE experts exhibit significant instability; as a tile's fraction decreases, its expert profile often collapses or hallucinates trends from other tiles (e.g., negative fluxes in the 33/33/34 wind scenario). This confirms that the disjoint experts in Soft-MoE suffer from structural entanglement, whereas UniTeD has learned a unified physical basis set that is truly invariant to the constituent fractions.

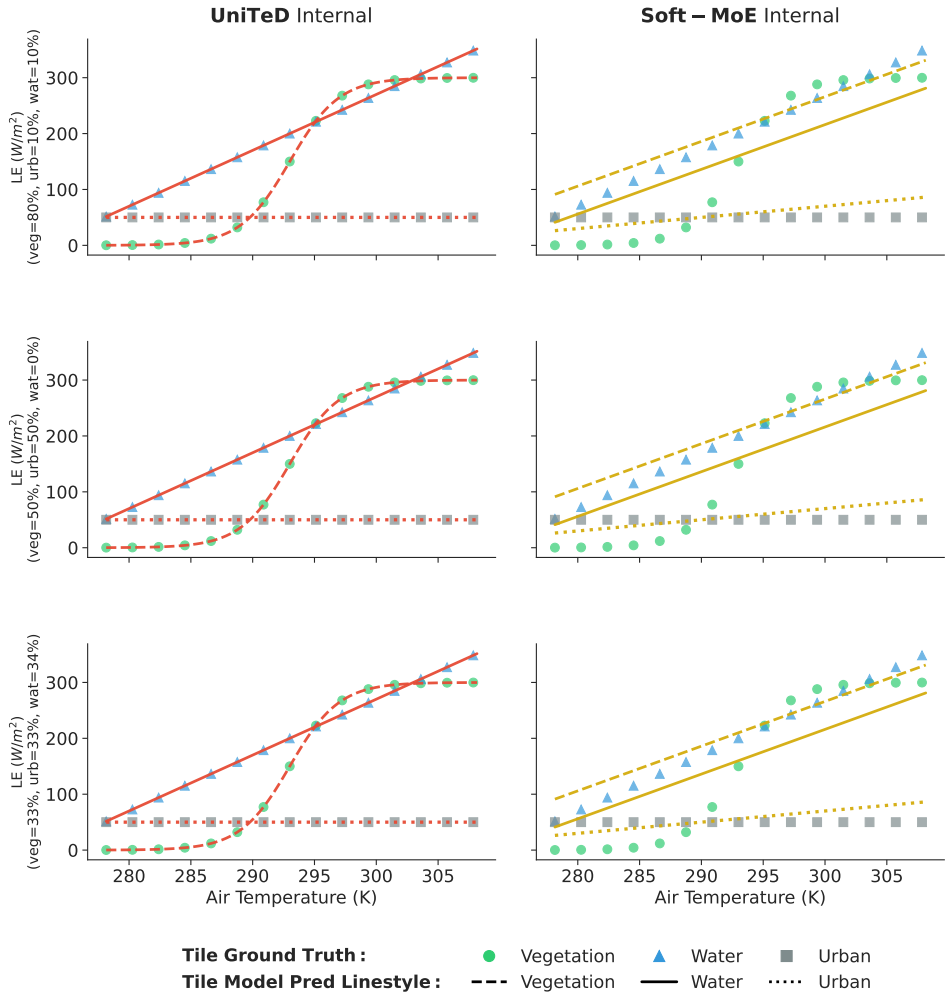

Figure 9: Thermal Stability across Mixtures. Comparison of sub-grid recovery across varying air temperatures for different land cover mixtures. UniTeD (left, red lines) consistently recovers the sigmoidal vegetation response across all mixing ratios. In contrast, Soft-MoE (right, gold lines) suffers from linearization and representation collapse. *(Colored markers denote Ground Truth; Line styles denote Model Predictions: Solid=Water, Dashed=Vegetation, Dotted=Urban).*

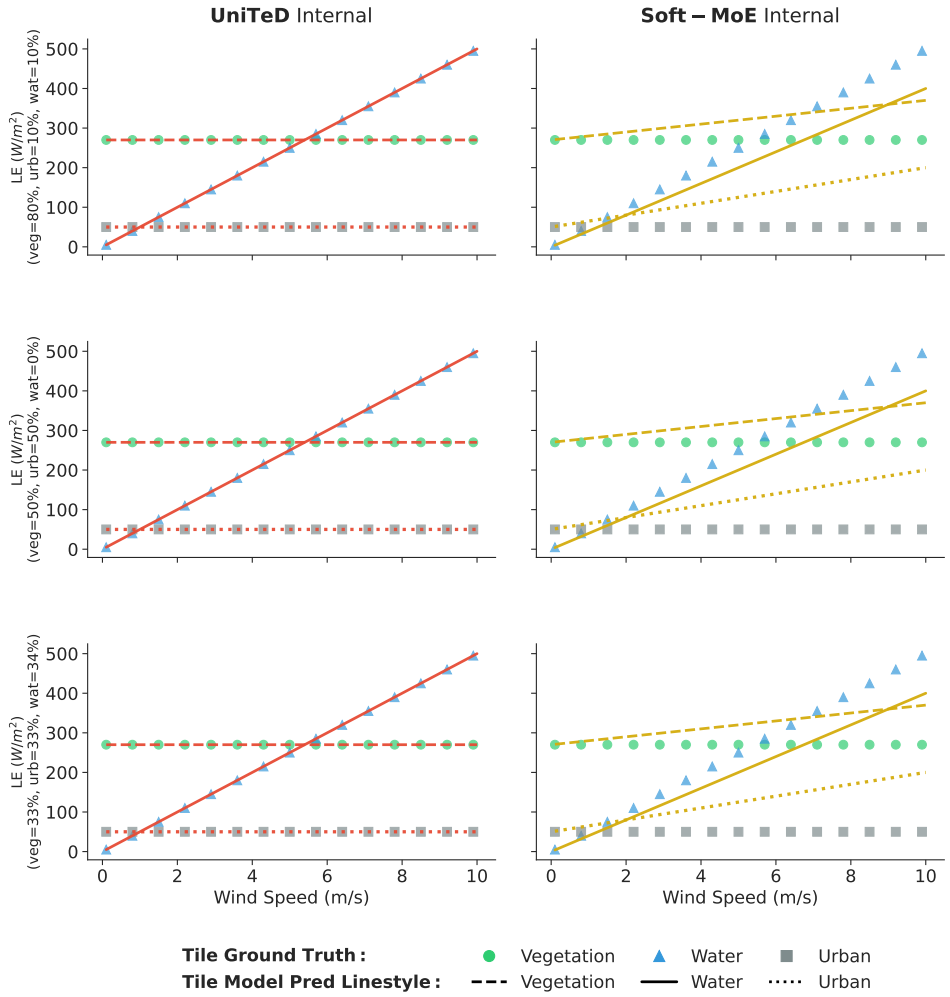

Figure 10: Aerodynamic Stability across Mixtures. Comparison of sub-grid recovery across varying wind speeds. UniTeD (left, red lines) maintains perfect linear scaling for Water and invariance for other tiles. In contrast, Soft-MoE (right, gold lines) experts fail to isolate the wind dependency, particularly in the balanced mixture. *(Colored markers denote Ground Truth; Line styles denote Model Predictions: Solid=Water, Dashed=Vegetation, Dotted=Urban).*

D ERA5 BENCHMARK: DATASET CURATION AND EXTENDED RESULTS

To evaluate the UniTeD framework in a high-dimensional, global setting, we utilize the ECMWF ERA5 reanalysis dataset. The data curation follows the physics schema of the Integrated Forecasting System (IFS).

Sub-Grid Tile Taxonomies. We adopt the IFS tiling formulation, treating each grid cell as a linear composition of $K = 6$ distinct physical regimes. The taxonomy is derived directly from the IFS Cycle 41r2 documentation on Physical Processes (ECMWF, 2016). These tiles represent distinct thermodynamic manifolds: (i) Intercepted Water, (ii) Dry Low Vegetation, (iii) Exposed Snow, (iv) Dry High Vegetation, (v) Shaded Snow, and (vi) Dry Bare Ground.

D.1 INPUT FEATURE SPACE ($\mathbf{X}$)

The model ingests a high-dimensional feature vector $\mathbf{x} \in \mathbb{R}^{36}$, partitioned into dynamic states, aerodynamic forcings, and static physical constants (Table 4).

Table 4: **ERA5 Input Feature Space.** Partition of the 36-dimensional input vector into physical groups.

Category	Feature Description	Unit
Thermodynamic State	2m Temperature (T_{2m})	K
	Surface Pressure (P_s)	Pa
	Vapor Pressure (e_a)	Pa
	Vapor Pressure Deficit (D)	Pa
	Specific Humidity (q)	kg/kg
	Skin Reservoir Content (W_{skin})	m
Aerodynamic Forcing	10m U-Wind Component (u_{10})	m/s
	10m V-Wind Component (v_{10})	m/s
	Boundary Layer Height (h_{bl})	m
Radiative Forcing	Net Solar Radiation (SW_{net})	W/m^2
	Downward Thermal Radiation ($LW_{\downarrow}$)	W/m^2
	Solar Time Cosine ($\cos(t)$)	-
	Solar Time Sine ($\sin(t)$)	-
Sub-surface State	Soil Temperature (Levels 1-3)	K
	Volumetric Soil Water (Layers 1-4)	m^3/m^3
Dynamic Surface	Leaf Area Index High Veg (LAI_H)	m^2/m^2
	Leaf Area Index Low Veg (LAI_L)	m^2/m^2
Vegetation Constants	**High Veg:** $r_{s,min}, c_{veg}, a_r, b_r$	-
	Low Veg: $r_{s,min}, c_{veg}, a_r, b_r$	-
	(4 parameters per type × 2 types)	
	Min. resist., cover, aero. coeffs.	
Soil Constants	Saturation Moisture (θ_{sat})	m^3/m^3
	Field Capacity (θ_{cap})	m^3/m^3
	Wilting Point (θ_{pwp})	m^3/m^3
	Thermal Conductivity (K_{dry}, K_{sat})	$W/m \cdot K$
	Heat Capacity (C_{dry})	$J/m^3 \cdot K$

D.2 DERIVED FEATURES AND NORMALIZATION

Vapor Pressure (e_a) and Deficit (VPD). We derive moisture potentials explicitly from 2-meter air temperature (T) and dewpoint temperature (T_d). Actual Vapor Pressure (e_a) is computed using the standard Magnus formula approximation by Bolton (1980):

$$e_a = 611.2 \cdot \exp\left(\frac{17.67 \cdot T_{d,c}}{T_{d,c} + 243.5}\right) \tag{29}$$

where $T_{d,c} = T_d - 273.15$. The Vapor Pressure Deficit (D) is computed similarly utilizing the saturation vapor pressure e_s derived from T, such that $D = \max(0, e_s - e_a)$.

Fixed-Factor Normalization. To ensure stable gradient descent dynamics while preserving the physical semantics of the input space (e.g., non-negativity of humidity), we employ fixed-factor scaling rather than statistical standardization. All features x_i are normalized by dividing by a fixed physical scalar s_i chosen to approximate the maximum expected magnitude of that variable (e.g., $s_P = 10^5$ Pa, $s_{LAI} = 10$).

D.3 Predictive Targets ($\mathbf{Y}$) and Target Scaling

The target vector comprises four components of the surface energy balance, representing the area-weighted sum of tile-specific responses. In the raw ERA5 dataset, these variables are provided as accumulated energy over a one-hour forecast step, with units of Joules per square meter (J/m^2).

To ensure stable gradient descent and physical interpretability, we apply a precise two-step unit transformation:

1. **Accumulation to Power:** We first convert the raw accumulated energy (J/m^2) to an average instantaneous power flux (W/m^2) by dividing by the accumulation period (3,600 seconds).
2. **Magnitude Scaling:** Surface fluxes can reach magnitudes of several hundred W/m^2. To stabilize neural network training, we scale this power flux by dividing it by $10^6/3600$.

Thus, the exact normalization factor applied to the raw ERA5 targets prior to training is $s_{flux} = 10^6$. The scaled target fluxes are:

- **Ground Heat Flux (G):** Conductive heat exchange with the soil.
- **Surface Latent Heat Flux (LE):** Energy flux from evaporation and sublimation.
- **Surface Sensible Heat Flux (H):** Turbulent heat exchange with the atmosphere.
- **Upward Thermal Radiation (L_{up}):** Longwave radiation emitted by the surface.

D.4 ERA5 Extended Results: Per-Variable RMSE Breakdown

As discussed in Section 4.2, Table 5 provides the per-variable RMSE breakdown for the ERA5 evaluation to confirm that UniTeD's performance gains hold across all individual physical components (G, LE, H, L_{up}).

D.5 ERA5 Extended Results: Spatial Generalization

To rigorously evaluate out-of-distribution spatial generalization, we implemented a macro-regional split by training all models exclusively on the Northern Hemisphere and evaluating them on the held-out Southern Hemisphere. This protocol exposes the models to simultaneous, substantial shifts in geography, climate, and land-cover distributions.

When trained on mixtures and evaluated on pure tiles, UniTeD consistently outperforms the baselines across all purity thresholds (Table 6). In the extreme pure-tile setting ($\alpha_k = 1.0$), monolithic and modular baselines exhibit severe performance degradation and high variance, struggling to extrapolate latent physics to novel geographic coordinates. In contrast, UniTeD maintains a bounded prediction error (RMSE = 93.53), achieving a 12% error reduction relative to the strongest baseline (DNN) and up to a 36% reduction against modular baselines. These results confirm that UniTeD's structural disentanglement extends beyond compositional shifts to provide robust spatial generalization across entirely unseen global domains.

D.6 Empirical Validation of Local Non-Degeneracy

To empirically verify the local non-degeneracy assumption (Theorem 3.1) in real-world heterogeneous regions, we analyzed the condition number of the local fraction matrix $A_{\mathcal{B}}$ within the ERA5 dataset.

Table 5: Detailed Per-Variable Breakdown for ERA5 Surface Energy Flux. RMSE values ($W\ m^{-2}$) are reported as mean $\pm$ std across three independent runs for individual flux components (G, LE, H, L_{up}).

(a) Task 1: Disentanglement of Mixed Signals (Mixture $\rightarrow$ Pure)

			Purity Threshold ($\exists \alpha_k \geq \dots$)			
Model	**Flux**	**Test (Mix)**	**.7**	**.8**	**.9**	**1.0 (Pure)**
DNN	G	10.4 ± 0.4	24.5 ± 3.2	25.9 ± 3.2	29.5 ± 5.6	39.1 ± 10.4
	LE	22.9 ± 0.9	41.1 ± 5.3	42.6 ± 5.2	45.0 ± 8.6	49.3 ± 13.1
	H	21.6 ± 0.8	37.3 ± 4.8	38.5 ± 4.7	39.1 ± 7.5	34.4 ± 9.1
	L_{up}	5.5 ± 0.2	14.5 ± 1.9	15.4 ± 1.9	17.3 ± 3.3	16.8 ± 4.5
FT-Transformer	G	11.0 ± 0.3	21.4 ± 1.4	22.4 ± 1.7	24.1 ± 2.8	24.8 ± 3.5
	LE	23.7 ± 0.7	43.7 ± 2.9	45.8 ± 3.5	48.9 ± 5.8	48.3 ± 6.8
	H	22.3 ± 0.6	39.1 ± 2.6	40.6 ± 3.1	41.9 ± 4.9	34.0 ± 4.8
	L_{up}	6.1 ± 0.2	14.0 ± 0.9	14.8 ± 1.1	16.8 ± 2.0	21.0 ± 3.0
Soft-MoE	G	12.1 ± 0.4	22.9 ± 1.2	24.0 ± 1.0	26.9 ± 0.9	33.8 ± 2.7
	LE	25.1 ± 0.7	46.2 ± 2.4	48.0 ± 2.0	51.4 ± 1.7	47.7 ± 3.8
	H	23.5 ± 0.7	39.2 ± 2.0	40.3 ± 1.7	41.0 ± 1.3	30.6 ± 2.4
	L_{up}	6.0 ± 0.2	14.4 ± 0.7	15.3 ± 0.6	17.9 ± 0.6	22.4 ± 1.8
Frac-MoE	G	11.3 ± 0.3	23.6 ± 1.2	24.8 ± 1.3	27.5 ± 2.4	39.6 ± 14.1
	LE	24.4 ± 0.7	47.3 ± 2.3	49.8 ± 2.7	54.1 ± 4.8	45.5 ± 16.2
	H	22.7 ± 0.7	43.8 ± 2.1	46.0 ± 2.5	49.1 ± 4.3	40.8 ± 14.5
	L_{up}	6.0 ± 0.2	14.4 ± 0.7	15.2 ± 0.8	16.9 ± 1.5	22.3 ± 7.9
UniTeD (Ours)	G	10.4 ± 0.2	20.7 ± 0.6	21.7 ± 0.7	23.5 ± 1.2	21.5 ± 3.7
	LE	23.9 ± 0.5	40.1 ± 1.2	41.5 ± 1.3	41.7 ± 2.2	30.8 ± 5.4
	H	22.3 ± 0.5	39.6 ± 1.2	41.2 ± 1.3	42.3 ± 2.2	26.2 ± 4.6
	L_{up}	5.9 ± 0.1	13.1 ± 0.4	13.8 ± 0.4	15.2 ± 0.8	13.5 ± 2.3

(b) Task 2: Zero-Shot Compositional Generalization (Pure $\rightarrow$ Mixture)

			Mixture Threshold ($\forall \alpha_k \geq \dots$)		
Model	**Flux**	**Test (Pure)**	**.01**	**.05**	**.08**
DNN	G	9.6 ± 0.2	29.6 ± 4.8	32.1 ± 7.7	33.3 ± 9.5
	LE	7.8 ± 0.2	30.3 ± 4.9	30.8 ± 7.4	32.4 ± 9.2
	H	10.7 ± 0.2	35.6 ± 5.7	36.9 ± 8.9	40.1 ± 11.5
	L_{up}	4.4 ± 0.1	20.6 ± 3.3	24.5 ± 5.9	25.1 ± 7.2
FT-Transformer	G	11.0 ± 0.3	24.5 ± 1.5	27.4 ± 2.2	28.3 ± 1.3
	LE	8.7 ± 0.2	25.7 ± 1.6	26.5 ± 2.1	25.3 ± 1.1
	H	11.6 ± 0.3	30.3 ± 1.8	31.6 ± 2.5	29.0 ± 1.3
	L_{up}	4.9 ± 0.1	12.2 ± 0.7	12.6 ± 1.0	13.3 ± 0.6
Soft-MoE	G	12.5 ± 1.1	24.6 ± 0.8	25.2 ± 0.8	28.6 ± 1.0
	LE	9.1 ± 0.8	21.8 ± 0.7	23.6 ± 0.8	23.9 ± 0.9
	H	12.4 ± 1.1	27.3 ± 0.9	28.6 ± 0.9	28.0 ± 1.0
	L_{up}	5.6 ± 0.5	8.7 ± 0.3	9.5 ± 0.3	9.3 ± 0.3
Frac-MoE	G	10.4 ± 0.7	22.0 ± 0.6	23.0 ± 1.2	24.6 ± 0.7
	LE	7.5 ± 0.5	22.9 ± 0.7	23.5 ± 1.2	31.4 ± 0.9
	H	10.7 ± 0.7	30.2 ± 0.9	32.0 ± 1.7	37.5 ± 1.1
	L_{up}	4.9 ± 0.3	12.0 ± 0.3	11.4 ± 0.6	10.5 ± 0.3
UniTeD (Ours)	G	10.6 ± 0.1	20.7 ± 0.3	20.6 ± 0.1	18.1 ± 0.4
	LE	8.0 ± 0.1	22.0 ± 0.3	22.4 ± 0.1	20.0 ± 0.4
	H	11.1 ± 0.2	25.6 ± 0.4	26.4 ± 0.1	24.1 ± 0.5
	L_{up}	4.8 ± 0.1	8.9 ± 0.1	8.8 ± 0.0	8.2 ± 0.2

Table 6: Spatial Generalization (Northern Hemisphere Train → Southern Hemisphere Test). RMSE scores ($W\,m^{-2}$) for the Mixture → Pure task. UniTeD maintains superior stability when extrapolating to an entirely unseen hemisphere.

		Purity Threshold ($\exists \alpha_k \geq \dots$)			
Model	**Test (Mix)**	**.7**	**.8**	**.9**	**1.0 (Pure)**
DNN	$\mathbf{92.3 \pm 1.3}$	138.1 ± 6.1	141.0 ± 8.7	150.6 ± 12.6	106.3 ± 24.1
FT-Transformer	96.3 ± 2.8	137.6 ± 5.4	140.0 ± 9.2	146.5 ± 12.4	114.7 ± 18.2
Soft-MoE	99.5 ± 1.9	142.4 ± 3.8	147.3 ± 5.9	156.4 ± 7.7	124.5 ± 14.9
Frac-MoE	96.0 ± 0.8	147.3 ± 10.0	157.7 ± 12.0	161.4 ± 14.1	147.5 ± 10.6
UniTeD (Ours)	94.4 ± 2.5	$\mathbf{128.3 \pm 2.3}$	$\mathbf{133.1 \pm 2.7}$	$\mathbf{138.9 \pm 2.1}$	$\mathbf{93.5 \pm 12.8}$

Neighborhood Definition: We defined neighborhoods using a K-Nearest Neighbors search ($K = 1920$, matching our batch size) on the 36-dimensional feature vectors x. To empirically verify that these neighborhoods contain clustered, approximately identical environmental conditions, we calculated the local standard deviation of the features within each neighborhood relative to the global dataset standard deviation. This ratio demonstrates the tightness of the neighborhoods: we observed a minimum ratio of 0.12, a mean of 0.27, and a maximum of 0.65.

Condition Number Distribution: For 10,000 randomly sampled anchor cells, we retrieved their local neighborhood of $K = 1920$, constructed the local fraction matrix $A_{\mathcal{B}} \in \mathbb{R}^{K \times T}$, and computed the condition number $\kappa(A) = \sigma_{max}/\sigma_{min}$. The local fraction matrix condition numbers follow a long-tail distribution:

- Median κ: 81.7
- 90th Percentile: $\approx 1.16 \times 10^4$
- 95th Percentile: $\approx 8.31 \times 10^4$
- 99th Percentile: 3.20×10^{18} (effectively singular)

A median condition number of $\sim$82 confirms that for the vast majority of local weather regimes, there is robust, mathematically stable sub-grid variance. The fraction matrix is well-conditioned, guaranteeing that the local unmixing problem is highly identifiable during a standard gradient update. This long-tail distribution is physically expected. We observe that extreme condition numbers occur almost exclusively in highly homogeneous geographical areas (e.g., vast forests). In these localized regimes, the condition number approaches infinity because the grid cells are effectively pure. While the matrix mathematically approaches rank-deficiency in these extreme cases, the unmixing problem is trivially solved during training: because there is only one dominant tile occupying the entire cell, the aggregate ground-truth signal is directly assigned to this single tile.

Relationship to Empirical Error: To establish that these condition number spikes do not destabilize the model, we analyzed the empirical prediction RMSE for these neighborhoods. Our quantitative analysis reveals that the model's prediction error remains stable across the condition number spectrum. For instance, neighborhoods near the median condition number (81.7) exhibit an error of 115.35 ± 4.59 W/m^2 (across the three-seed runs), and the extreme tail ($\geq$ 90th percentile) shows an error of 94.83 ± 8.25 W/m^2. Therefore, these spikes in condition number in ERA5 are not inducing unbounded error divergence. This analysis provides empirical evidence that the ERA5 dataset possesses the necessary sub-grid compositional diversity to enable stable identifiability from aggregate supervision, fulfilling the prerequisites of Theorem 3.1.