CONSERVATIVE LOCAL-GLOBAL TRANSPORT: MESH-LOCAL CONVOLUTION AND SCALABLE LINEAR ATTENTION FOR NEURAL LAGRANGIAN SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many neural Lagrangian simulators rely on Graph Neural Networks (GNNs) with spatially constructed interaction graphs, introducing considerable topology-processing and message-passing costs as the particle count grows. Moreover, these architectures are typically agnostic to the underlying conservation laws. In this work, we propose a conservative local–global transport mechanism in latent space: mesh convolution captures strongly coupled near-field interactions, while separable linear attention enables dense long-range interactions at linear complexity, both enforcing conservation laws by construction. Beyond global conservation, we introduce sign-split latent flux limiters that strictly preserve prescribed physical bounds without breaking the conservative latent update, and connect these discrete limiters to a unified family of latent advection dynamics. It demonstrates advantages over GNN-based baselines in computational efficiency (supports up to $38\times$ more particles) and physical consistency while maintaining competitive predictive accuracy. Our code is available at https://anonymous.4open.science/r/CONSERVATIVE-LOCAL-GLOBAL-TRANSPORT-I16S.

1 INTRODUCTION

Lagrangian neural simulators represent a physical system as a collection of moving particles and therefore provide a natural interface for learning complex continuum dynamics. Most existing approaches build particle interactions with Graph Neural Networks (GNNs), where edges are constructed according to spatial proximity and messages are repeatedly propagated along the resulting graph (Sanchez-Gonzalez et al., 2020; Li et al., 2019; Toshev et al., 2024; Brandstetter et al., 2021). Continuous-kernel alternatives similarly model particle interactions through localized neighborhoods (Ummenhofer et al., 2020; Winchenbach & Thuerey, 2024). Such locality is effective for short-range physical interactions, but becomes increasingly expensive when the particle number grows or when long-range dependencies must be captured. Hierarchical graphs and spatial clustering enlarge the receptive field at the cost of additional topology construction and multilevel processing (Li et al., 2019; Janny et al., 2023), while attention-based particle models provide more direct long-range interactions but remain constrained by quadratic attention cost (Wang et al., 2026). Linear attention offers an appealing alternative by factorizing query–key interactions into separable feature maps and reducing dense interaction to linear complexity (Katharopoulos et al., 2020; Choromanski et al., 2021; Chen et al., 2024). However, computational scalability alone does not make the resulting neural interaction operator physically structured.

A central difficulty is that generic neural interaction operators are largely agnostic to conservation laws governing the underlying continuum system. Physical quantities such as mass, momentum, or elemental composition should be transported among particles rather than arbitrarily created or destroyed. Although physical consistency can be encouraged through architectural inductive biases such as equivariance (Brandstetter et al., 2021) or through auxiliary numerical corrections (Toshev et al., 2024), such mechanisms do not by themselves guarantee exact conservation at every network update. Moreover, global conservation alone is insufficient: a conservative update may still generate locally inadmissible states like negative mass if the outgoing transport exceeds the available quantity.

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.

In this work, we introduce a *conservative latent linear* framework that interprets self-attention as transport in latent space. While antisymmetric messages have been used to enforce conservation on predefined graph edges (Horie & Mitsume, 2024), we instead skew-symmetrize dense linear-attention interactions, yielding exact latent conservation without explicit pairwise graph construction and retaining $\mathcal{O}(N)$ complexity through query–key separability. Any physical quantity obtained through a linear projection inherits this conservation, while the construction remains largely orthogonal to the particular linear-attention feature map. To complement the predominantly low-frequency global interactions of attention (Park & Kim, 2022), we further introduce an MPM-inspired (Sharabi et al., 2025) conservative MeshConv to capture local high-frequency dynamics. Beyond global conservation, we introduce a latent flux limiter strictly enforcing physical bounds while preserving conservation and $\mathcal{O}(N)$ complexity. We further interpret self-attention as latent transport, with different limiters inducing a unified family of nonlinear advection–diffusion dynamics.

Our contributions are summarized as follows:

- A separable linear attention as a skew-symmetric latent transport operator, achieving exact conservation with linear complexity, complemented by a conservative local MeshConv.
- Extending the same conservative transport structure with a sign-split flux limiter that enforces local physical bounds without sacrificing conservation or linear-time aggregation.
- The proposed framework allows the particle amount to scale up to roughly **38x** larger than baselines under the same GPU memory budget while remaining competitive accuracy.

2 Related Work

Neural Lagrangian Simulation. Continuum mechanics is commonly modeled using Eulerian or Lagrangian approaches. Neural Lagrangian simulation primarily employs GNNs to learn particle dynamics. Early mesh-free solvers represent particles as graph vertices, with edges constructed dynamically or fixed according to material types and interactions (Li et al., 2019; Sanchez-Gonzalez et al., 2020). Subsequent works improve stability through SPH relaxation (Toshev et al., 2024) and physical consistency through equivariance (Brandstetter et al., 2021). Alternatively, continuous convolutions (Ummenhofer et al., 2020) reconstruct fluid fields using learnable interpolation kernels, later refined as linear combinations of symmetric bases (Winchenbach & Thuerey, 2024). However, graph construction (e.g., hierarchical partitions (Li et al., 2019)) and long-range feature extraction (e.g., spatial clustering (Janny et al., 2023)) incur substantial topology processing costs, limiting particle scalability. Standard attention captures long-range interactions (Wang et al., 2026) but suffers from quadratic complexity. Our linear transformer bypasses both topology construction and attention complexity bottlenecks.

Linear Attention. Standard Transformers incur quadratic complexity, which linear attention reduces using separable kernels $S(\mathbf{q}, \mathbf{k}) = \varphi(\mathbf{q})^\top \varphi(\mathbf{k})$ (Katharopoulos et al., 2020), enabling linear complexity via a shared KV state $\sum_j \varphi(\mathbf{k}_j)\mathbf{v}_j^\top$. Early methods approximate exponential kernels using random Fourier features (Choromanski et al., 2021), second-order truncation (Arora et al., 2024), or discrete cosine transforms (Chen et al., 2024). Recent works recover softmax properties through power enhancement for sharper distributions (Han et al., 2023), rectification of neglected query magnitudes (Fan et al., 2025a; Meng et al., 2025a), polarity-aware interactions recovering negative features (Meng et al., 2025b), and nonlinear post-processing addressing low-rank limitations (Fan et al., 2025b). Alternatively, agent attention (Han et al., 2024; Li et al., 2023a; Wu et al., 2024) aggregates global context into bottleneck tokens before broadcasting, reducing complexity to $\mathcal{O}(\#\text{Token} \cdot \#\text{Agent})$. Householder transforms (Yang et al., 2026) further enable learnable similarity embeddings beyond prior positional embeddings. Our conservative linear attention remains compatible with most existing variants.

Scalability on PDE Domains. Scaling neural solvers to massive complex domains remains computationally challenging. Schwarz alternating methods decompose domains into subregions (Snyder et al., 2023), solve them separately, and iteratively synchronize boundaries, incurring synchronization costs. Geo-FNO (Li et al., 2023b) maps irregular meshes onto regular grids through difficult-to-obtain mesh parameterizations. Deep learning approaches instead extract geometric features using point cloud encoders (Hao et al., 2023), but cross-attention incurs quadratic complexity. Geometry pre-training (Wu et al., 2026) accelerates downstream learning yet struggles to generalize to scenes

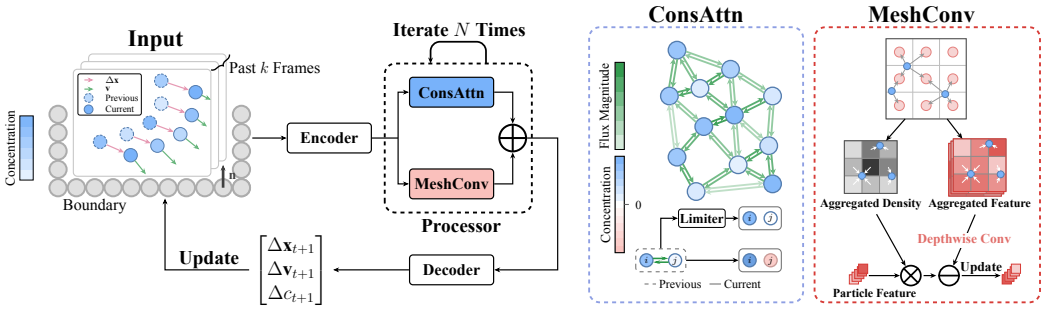


Figure 1: Architecture overview. Encoded particle states are iteratively updated by conservative global attention and local transport, with latent flux limiters enforcing physical bounds.

with numerous objects and complex boundaries. In contrast, our framework theoretically supports infinitely large domains without global resolution degradation.

3 Methodology

Problem Formulation. Particle-based continuum mechanics discretizes a domain Ω into N moving particles with position $\mathbf{x}_i$, velocity $\mathbf{v}_i$, concentration $\mathbf{c}_i$, and volume V_i, $i \in [N]$. Given historical states and boundary conditions $\mathcal{B}$, an encoder maps $\mathcal{B}$ and particle states $\mathbf{u}_i = (\mathbf{x}_i, \mathbf{v}_i, \mathbf{c}_i)$ to latent features $\mathbf{z}_i$, which are transformed by a processor and decoded into physical increments $\Delta\mathbf{u}_i = (\Delta\mathbf{x}_i, \Delta\mathbf{v}_i, \Delta\mathbf{c}_i)$. The entire pipeline is denoted by a neural simulator $\mathcal{G}_\theta$:

$$\{\Delta\mathbf{u}_i^t\}_{i=1}^N = \mathcal{G}_\theta(\{\mathbf{u}_i^t\}_{i=1}^N, \mathcal{B}), \qquad \mathbf{u}_i^{t+1} = \mathbf{u}_i^t + \Delta\mathbf{u}_i^t, \tag{1}$$

where the predicted increments $\Delta\mathbf{u}_i^t$ are used to update the states at time step t to produce the states at time step $t+1$ (the pipeline is presented in the left of Figure 1). However, a naive implementation of unconstrained transformation has no guarantee on physical conservation laws and may lead to inferior simulation, which motivates our conservative latent framework.

The remainder of this section is organized as follows. We start by establishing latent conservative linear attention (ConsAttn in Figure 1) for coupled conservation laws in Section 3.1, whereas Section 3.2 introduces flux limiters for physical admissibility and their underlying PDE correspondence. Lastly, Section 3.3 integrates advances in linear attention with conservative local convolution (MeshConv in Figure 1).

3.1 Conservative Linear Attention

In particle-based continuum mechanics, a fundamental requirement is satisfying macroscopic conservation laws across the domain Ω. Given a conserved physical field $\tilde{\mathbf{u}}$ (we use $\tilde{\mathbf{u}}$ to represent the conserved quantities in $\mathbf{u}$) and a source term $\mathbf{f}$, the governing equation dictates

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{\mathbf{u}}\mathrm{d}V = \int_{\partial\Omega} \nabla\tilde{\mathbf{u}}\cdot\mathbf{n}\mathrm{d}S + \int_\Omega \mathbf{f}\mathrm{d}V. \tag{2}$$

When discretizing the domain without external boundary fluxes or source terms, the system must strictly maintain $\sum_i V_i\Delta\tilde{\mathbf{u}}_i = 0$. Motivated by this, we propose a conservative linear attention by injecting inductive bias to enforce the structural constraint in the neural architecture, so as to natively achieve the conservation. The idea is that instead of computing $\Delta\mathbf{u}_i$ directly, we construct the pairwise flux by design a skew-symmetric function and use it to produce latent increments that guarantee conservation in latent space, and the conservation translates into physical space naturally.

Given the latent representations $\mathbf{z}_i = \mathsf{Encoder}(\mathbf{u}_i)$, we propose a conservative linear attention ConsAttn by skew-symmetrizing the canonical normalized linear attention components as follows,

$$\mathsf{ConsAttn}(\mathbf{z}_i, \mathbf{z}_j; \{\mathbf{z}_k\}) := \frac{\phi_Q(\mathbf{z}_i)\cdot\phi_K(\mathbf{z}_j)\phi_V(\mathbf{z}_j)^\top}{\sum_j \phi_Q(\mathbf{z}_i)^\top\phi_K(\mathbf{z}_j)} - \frac{\phi_Q(\mathbf{z}_j)\cdot\phi_K(\mathbf{z}_i)\phi_V(\mathbf{z}_i)^\top}{\sum_i \phi_Q(\mathbf{z}_j)^\top\phi_K(\mathbf{z}_i)}. \tag{3}$$

Next, we use it to construct an skew-symmetric latent flux vector $\mathbf{F}_{ij} = \mathsf{ConsAttn}(\mathbf{z}_i, \mathbf{z}_j)$, which can be used to calculate the increments $\Delta\mathbf{z}_i = -\sum_j \mathbf{F}_{ij}$ in latent space. Since ConsAttn is a skew-symmetric function, we can natively guarantee $\mathbf{F}_{ij} = -\mathbf{F}_{ji}$. In such a manner, the flux is conservative under state updating $V_i\Delta\mathbf{z}_i = -\Delta t \sum_j \mathbf{F}_{ij}$ because

$$\sum_{i,j} \mathbf{F}_{ij} = \frac{1}{2}\sum_{i,j} \mathbf{F}_{ij} + \mathbf{F}_{ji} = \frac{1}{2}\sum_{i,j} \mathbf{F}_{ij} - \mathbf{F}_{ij} = 0. \tag{4}$$

It is clear that the dense flux interaction can be rigorously evaluated in $\mathcal{O}(N)$ time. To align this latent conservation with actual physical quantities (e.g., mass or momentum), we can linearly transforming it back to physical space with a learned projection matrix $\mathbf{W}_P$. Since linear transformations commute with summation, any conservation guarantees established in the latent space are perfectly inherited by the projected physical space, i.e., $\sum_i V_i\Delta\tilde{\mathbf{u}}_i = \sum_i V_i\mathbf{W}_P\Delta\mathbf{z}_i = \mathbf{W}_P\sum_i V_i\Delta\mathbf{z}_i = 0$.

For environments with source terms $\mathbf{f}_i$, the discrete update requires an iterative scheme to maintain stability over large time steps Δt. Instead of a simplistic explicit addition in low-dimension physical space, which becomes inaccurate for large Δt, we couple the source dynamically into the latent space via a learnable source mapping ϕ in an implicit manner:

$$V_i\Delta\mathbf{z}_i = -\sum_j \mathsf{ConsAttn}(\mathbf{z}_i + \Delta\mathbf{z}_i, \mathbf{z}_j + \Delta\mathbf{z}_j; \{\mathbf{z}_k\})\Delta t + \phi(\mathbf{z}_i, \mathbf{z}_i + \Delta\mathbf{z}_i, \mathbf{f}_i). \tag{5}$$

Since the implicit update couples all particle states, directly solving Eq. 5 at each inference step is computationally expensive. To reduce the computational cost, we reformulate Eq. 5 as a fixed-point problem $\Delta\mathbf{Z} = g_\theta(\mathbf{Z}, \Delta\mathbf{Z})$ and solve it via Picard iteration, where $\mathbf{Z}$ and $\Delta\mathbf{Z}$ denote the concatenations of $\{\mathbf{z}_i\}$ and $\{\Delta\mathbf{z}_i\}$, respectively. By imposing spectral constraints on the networks in $\mathbf{F}_{ij}$ and ϕ, the Lipschitz constant of g_θ with respect to $\Delta\mathbf{Z}$ can be bounded by $\gamma \in (0, 1)$, ensuring exponential convergence $\mathcal{O}(\gamma^n)$.

Remark 3.1. (**Multiple Conservation Laws**) The proposed ConsAttn trivially scales to complex multi-phase or multi-physics systems. To conserve K distinct physical quantities simultaneously, we simply define unique linear projections $\mathbf{W}_{P,k}$ and augment the source mapping ϕ to $\sum_{k=1}^{K} \phi_k(\mathbf{z}_i, \mathbf{z}_i + \Delta\mathbf{z}_i, \mathbf{f}_{i,k})$ in Eq. 5, seamlessly enforcing multiple coupled conservation laws within a single $\mathcal{O}(N)$ latent space. Details are offered in Appendix A.2.

Simulating Reacting Dilute Solutions. As an illustrative example, consider incompressible fluid particles carrying dilute reacting solutes, whose feedback to the carrier flow is negligible. We split the latent state as $\mathbf{z}_i = \mathrm{concat}(\mathbf{z}_i^F, \mathbf{z}_i^C)$, where $\mathbf{z}_i^F$ describes the fluid and $\mathbf{z}_i^C$ describes the solute, with momentum $\mathbf{p}_i = \mathbf{W}_F\mathbf{z}_i^F$ and concentration $\mathbf{c}_i = \mathbf{W}_C\mathbf{z}_i^C$. The one-way coupling (Figure 2) is directly imposed through the solute attention – query and key from the fluid state, while value from the solute state:

$$\mathbf{q}_i^C = \phi_Q^C(\mathbf{z}_i^F), \mathbf{k}_i^C = \phi_K^C(\mathbf{z}_i^F), \mathbf{v}_i^C = \phi_V^C(\mathbf{z}_i^C), \mathbf{F}_{ij}^C = \frac{\mathbf{q}_i^C \cdot \mathbf{k}_j^C \mathbf{v}_j^{C\top}}{\sum_j \mathbf{q}_i^C \cdot \mathbf{k}_j^C} - \frac{\mathbf{q}_j^C \cdot \mathbf{k}_i^C \mathbf{v}_i^{C\top}}{\sum_i \mathbf{q}_j^C \cdot \mathbf{k}_i^C}. \tag{6}$$

Therefore, the carrier flow determines how the solute is transported, while the solute does not affect the fluid dynamics. Chemical reactions are incorporated through a fixed stoichiometric matrix $\mathbf{S} \in \mathbb{R}^{N_s \times N_r}$, where each column represents one balanced reaction, with negative coefficients for reactants and positive coefficients for products. The network predicts only the reaction-rate $\mathbf{f}_i^{\mathrm{r}} \in \mathbb{R}^{N_r}$, and the resulting concentration source is $\dot{\mathbf{c}}_i = \mathbf{S}\mathbf{f}_i^{\mathrm{r}}$. Since every column of $\mathbf{S}$ corresponds to a balanced chemical reaction, any source generated in this form automatically preserves the atomic stoichiometry. This *separates fluid dynamics, solute advection, and chemical reactions while preserving conservation structures explicitly*, and naturally extends to other single coupled systems.

Limitations. Conserved quantities, such as momentum and concentration, are retrieved through linear projections $\mathbf{W}_F$ and $\mathbf{W}_C$ (instances of $\mathbf{W}_P$), while other physical outputs, such as positions, are predicted by the standard decoder. However, it still has negative concentrations issues as addressed in Section 3.2.

3.2 Linear Latent Flux Limiter

While the anti-symmetric flux rigorously guarantees global latent conservation, projecting the updated states back to the physical domain (e.g., retrieving particle mass m_i) may yield non-physical

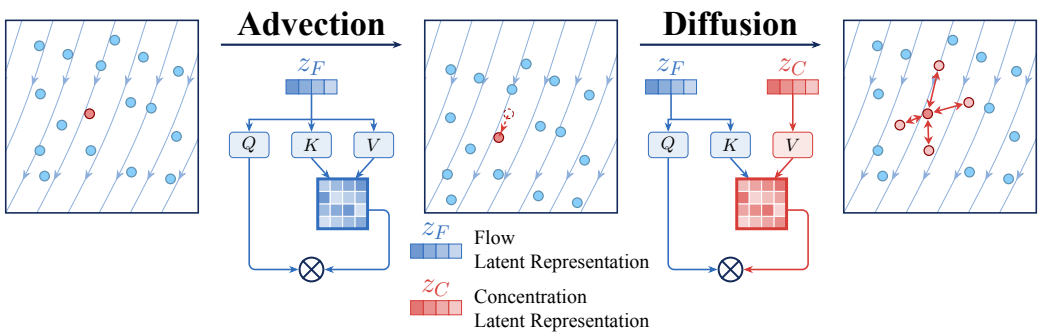


Figure 2: One-way coupling of fluid advection and solute transport through latent attention. The dependencies in latent implementation corresponds to the physical coupling order exactly.

negative values if the local net outflow exceeds the available capacity. Thankfully, one can scale the flux magnitude to avoid such negativity without breaking conservation laws.

Directly limiting the net flux may underestimate gross outflow due to cancellation between positive and negative interactions, motivating a sign-split decomposition. Each channel of the directed linear-attention transport takes the form $\mathbf{A}_{ijh} = L_{ih}k_{jh}\mathbf{z}_j$ where L_{ih} is the normalization of query vector q_{ih}. For the projected scalar $m_j = \mathbf{W}_P\mathbf{W}_V\mathbf{z}_j$, define $R_{jh} := k_{jh}m_j$ and $x_{ijh} := \mathbf{W}_P\mathbf{A}_{ijh} = L_{ih}R_{jh}$. Using identity $x = x^+ - x^-$, the projected skew flux f_{ij} thus can be decomposed into non-negative inflow I_{ij} and outflow O_{ij} by:

$$f_{ij} := \mathbf{W}_P\mathbf{F}_{ij} = \sum_h (x_{ijh} - x_{jih}) = \sum_h \left(x^+_{ijh} + x^-_{jih}\right) - \sum_h \left(x^-_{ijh} + x^+_{jih}\right) =: O_{ij} - I_{ij}. \quad (7)$$

satisfying $O_{ij} = I_{ji} \geq 0$. The corresponding gross outgoing and incoming fluxes are $O_i = \sum_j O_{ij}$ and $I_i = \sum_j I_{ij}$. Importantly, these quantities can still be evaluated in $\mathcal{O}(N)$ time because the positive/negative decomposition preserves the separable linear-attention structure:

$$O_i = \sum_h L^+_{ih} \sum_j R^+_{jh} + \sum_h L^-_{ih} \sum_j R^-_{jh} + \sum_h R^+_{ih} \sum_j L^-_{jh} + \sum_h R^-_{ih} \sum_j L^+_{jh}, \quad (8)$$

$$I_i = \sum_h L^-_{ih} \sum_j R^+_{jh} + \sum_h L^+_{ih} \sum_j R^-_{jh} + \sum_h R^+_{ih} \sum_j L^+_{jh} + \sum_h R^-_{ih} \sum_j L^-_{jh}. \quad (9)$$

As verified in Theorem A.2, to make the constraint $m_i \in [m_{\min}, m_{\max}]$ enforced strictly, we can set $\tau_i = \Delta t/V_i$ and define $\alpha_i = \min\{1, \frac{m_i - m_{\min}}{\tau_i O_i}, \frac{m_{\max} - m_i}{\tau_i I_i}\}$. The limited latent flux is $\alpha_i\alpha_j\mathbf{F}_{ij}$, remaining skew-symmetric and therefore preserving the conservation law since $\alpha_i\alpha_j = \alpha_j\alpha_i$.

Correspondence with Underlying PDEs. To characterize how limiters affect latent transport, we analyze a local 1D upwind specialization of self-linear attention. When ϕ_V is identity, each directed self-attention interaction locally transports the donor latent state itself, motivating a unified donor–receiver form $\mathbf{J}_{i+1/2} = v\, a_i b_{i+1}\mathbf{z}_i$, where a_i and b_i control the donor and receiver gates, respectively. The symmetric limiter used above corresponds to $a = b = \alpha$, the single-sided donor limiter to $(a, b) = (\alpha, 1)$, and a repulsive occupancy limiter to $(a, b) = (\rho, 1 - \rho)$. As derived in Appendix B, their semi-discrete continuum limit takes the unified form

$$\frac{\partial \mathbf{z}}{\partial t} + v\mathbf{D}_{a,b}(\mathbf{z})\frac{\partial \mathbf{z}}{\partial x} = \frac{v\Delta x}{2}\frac{\partial}{\partial x}\left(\mathbf{M}_{a,b}(\mathbf{z})\frac{\partial \mathbf{z}}{\partial x}\right) + O(\Delta t) + O(\Delta x^2). \quad (10)$$

Thus the limiter acts as a state-dependent modulation of latent advection and numerical diffusion. In particular, for $a = b = \alpha$, $\mathbf{D}_{\text{sym}} = \alpha^2\mathbf{I} + 2\alpha\mathbf{z} \otimes \nabla_{\mathbf{z}}\alpha$ and $\mathbf{M}_{\text{sym}} = \alpha^2\mathbf{I}$, so $\alpha = 0$ freezes both transport and diffusion. Detailed derivations and other limiter variants are deferred to Appendix B.

3.3 Conservative Local Transport

The proposed ConsAttn captures global interactions through factorized latent transport. To complement these interactions with near-field details, we introduce a second conservative module,

MeshConv. This follows hybrid attention designs that combine local interactions with linear attention for low-rank long-range dependencies (Chen et al., 2021), and is motivated by the complementary low-frequency and high-frequency behavior of attention and convolution (Park & Kim, 2022). Crucially, we extend the pairwise cancellation principle underlying ConsAttn to local convolution, allowing both modules to preserve the same latent conservation laws.

Local Interactions through a Background Grid. Inspired by material point methods (Sharabi et al., 2025), we implement local particle interactions through a fixed background grid: particle features are aggregated onto grid nodes, exchanged by a local convolution, and interpolated back to particles. Let $\omega_{i\alpha} \geq 0$ denote the interpolation weight between particle i and grid node α, satisfying $\sum_\alpha \omega_{i\alpha} = 1$. The particle-to-grid and grid-to-particle operations are

$$\mathbf{H}_\alpha = \mathsf{P2G}(\{\mathbf{z}_j\})_\alpha := \sum_j \omega_{j\alpha}\mathbf{z}_j, \qquad \mathsf{G2P}(\{\mathbf{H}_\alpha\})_i := \sum_\alpha \omega_{i\alpha}\mathbf{H}_\alpha. \tag{11}$$

A local kernel $K_{\alpha,\beta} = k_\theta(\mathbf{x}_\alpha - \mathbf{x}_\beta)$ then transfers features from grid node β to α through $(\mathbf{K} * \mathbf{H})_\alpha := \sum_\beta K_{\alpha,\beta}\mathbf{H}_\beta$, in which kernel multiplication is understood component-wise when using channel-wise weights.

Conservative Mesh Convolution. Simply composing these operations aggregates neighboring features without balancing the corresponding outgoing contributions. Following the construction of ConsAttn, we subtract the reverse interaction to obtain

$$\mathsf{MeshConv}(\mathbf{z}_i; \{\mathbf{z}_j\}) := \mathsf{G2P}(\mathbf{K} * \mathsf{P2G}(\{\mathbf{z}_j\}))_i - \mathbf{z}_i \odot \mathsf{G2P}(\mathbf{K}^\top * \mathsf{P2G}(\{\mathbf{1}\}))_i, \tag{12}$$

where $(\mathbf{K}^\top)_{\alpha,\beta} := K_{\beta,\alpha}$ reverses the grid interaction. The first term collects neighboring particle features through the grid. In the second term, $\mathsf{P2G}(\{\mathbf{1}\})_\alpha = \sum_j \omega_{j\alpha}$ measures the interpolation weight accumulated at each grid node; applying the reversed kernel and G2P yields the total coefficient with which particle i contributes to other particles. Specifically, particle j contributes $\omega_{i\alpha}K_{\alpha,\beta}\omega_{j\beta}\mathbf{z}_j$ to particle i through grid nodes β and α. The correction subtracts this same contribution from particle j, ensuring pairwise cancellation and hence $\sum_i \mathsf{MeshConv}(\mathbf{z}_i; \{\mathbf{z}_j\}) = 0$, as proved in Theorem A.1. Compactly supported interpolation and a fixed-size convolution stencil give $\mathcal{O}(N + N_g)$ cost for N_g grid nodes, retaining linear complexity when $N_g = \mathcal{O}(N)$. Therefore, MeshConv can equivalently be viewed as locally constructed anti-symmetric particle fluxes, allowing the same sign-split limiting strategy as the latent linear attention above. See Appendix A.3.

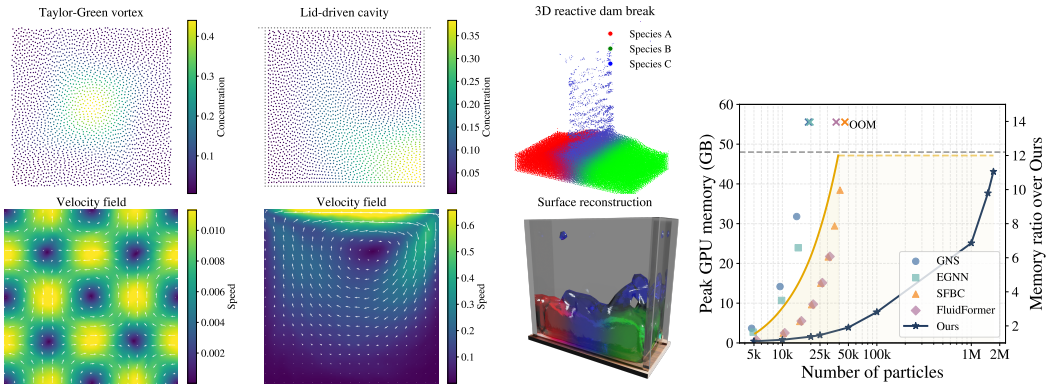
Hybrid Conservative Transport. With coefficients $\gamma_1, \gamma_2 \geq 0$, the two modules jointly produce the transport update by Eq. 13. Both terms sum to zero, so their combination preserves the volume-weighted latent quantities and their linear physical projections. This hybrid transport replaces the attention-only transport term in Eq. 5, with the source mapping retained. For reacting dilute solutions, we use both modules for the carrier flow and set $\gamma_2 = 0$ for solute transport since high-frequency advection has been tackled in flow latent representation, allowing the attention flux limiter in Section 3.2 to enforce concentration bounds.

$$V_i\Delta\mathbf{z}_i = -\Delta t \left[\gamma_1 \sum_j \mathbf{F}_{ij} + \gamma_2 \mathsf{MeshConv}(\mathbf{z}_i; \{\mathbf{z}_j\})\right]. \tag{13}$$

Remark 3.2. (**Compatibility with Prior Advances in Linear Attention**) The conservation guarantee of ConsAttn depends only on the anti-symmetry of $\mathbf{F}_{ij}$, remaining agnostic to the specific feature maps ϕ_Q, ϕ_K. Modifications that improve similarity concentration or information awareness can therefore be incorporated without changing Eq. 3. We also augment query/key features with a separable trigonometric spatial bias to introduce distance-aware interactions while retaining $\mathcal{O}(N)$ complexity, as detailed in Appendix C.1. Empirically, this bias substantially improves concentration transport but has little effect on convective prediction. More broadly, pairing ConsAttn with MeshConv extends the local–global design principle of softmax–linear hybrid attention (Chen et al., 2021): conservative local convolution captures high-frequency details that complement global attention (Park & Kim, 2022), while preserving latent conservation.

4 Experiments

We evaluate physical consistency, computational cost, and prediction accuracy on two 2D benchmarks, and scalability on a larger 3D reactive transport setting where graph-based baselines run out



(a) (b)

Figure 3: (a) Representative states of the 2D TGV, 2D LDC, and 3D reactive dam-break datasets with the reaction $A + B = C$; (b) Memory scalability with increasing particle counts. Our method allows around **38x** more particles to simulate in one go than baselines on an A6000 GPU with 48GB memory. The shaded region gives the lower envelope of memory ratio.

of memory. We compare our complete model against baselines on 2D tasks, then ablate flux limiting, spatial bias, MeshConv, and iterative latent updates to identify each component's contribution.

4.1 Experimental Setup

Dataset. We generate Taichi GPU-SPH datasets using weakly compressible SPH (Becker & Teschner, 2007), Delta-SPH density diffusion (Marrone et al., 2011), and layered boundaries Akinci et al. (2012). Particles store position, velocity, density, pressure, and solute concentration c, where $\frac{\partial c}{\partial t} + \mathbf{v} \cdot \nabla c = D\Delta c$. Figure 3a shows 2D incompressible Navier–Stokes benchmarks: periodic Taylor–Green vortex (TGV) and solid-boundary lid-driven cavity (LDC), each with one passive one-way-coupled solute, plus a harder 3D dam-break with more particles, multiple solutes, and reactions. See Appendix C.2.

Baselines. Compared with representative simulators spanning graph-based, equivariant, and continuous interaction paradigms. Following Takamoto et al. (2023), we include GNS (Sanchez-Gonzalez et al., 2020) and EGNN (Satorras et al., 2021), as well as continuous convolution and transformer-based methods SFBC (Winchenbach & Thuerey, 2024) and FluidFormer (Wang et al., 2026).

Evaluation Metrics. We report mean squared error (MSE) for fluid and concentration predictions. Physical consistency is measured by relative momentum and mass errors, $E_{\text{mom}} = \|\widehat{\mathbf{P}} - \mathbf{P}\|_2 / \|\mathbf{P}\|_2$ and $E_{\text{mass}} = |\widehat{M} - M|/M$, with hats denoting predictions. Since our flow is incompressible, mass is proportional to concentration, and the two terms are used interchangeably below. Non-negativity is measured by relative negative mass $M_- = \sum_i V_i \max(-\widehat{c}_i, 0)/M$ and negative-particle ratio $r_- = N^{-1} \sum_i \chi_{\widehat{c}_i < 0}$. Runtime is the single-step forward time. All are averaged over the test set.

4.2 Better Physical Consistency and Particle Amount Scalability

Table 1 summarizes the comparison on the 2D TGV benchmark, with an emphasis on physical consistency, computational cost and prediction accuracy. Implementation details of the proposed model is provided in Appendix C.3. The prediction errors should be interpreted together with the architectural constraints of different models. Most graph-based baselines employ expressive nonlinear message passing or Transformer-style interaction blocks, allowing them to fit complex particle dynamics with relatively few structural restrictions. In contrast, our model deliberately restricts the interaction operator to separable linear attention, supplemented only by lightweight mesh convolution, also introducing global conservative constraint implemented by latent limiters. This constrained hypothesis class naturally has lower approximation capacity than highly expressive black-box models, particularly at the moderate particle counts used in this benchmark. Still, it remains considerably competitive learning capability demonstrated by Figure 4b.

Table 1: Comparison on 2D-TGV and 2D-LDC with respect to physical consistency, inference time, and prediction losses. The best results are highlighted in bold and the second best underlined.

	Physical consistency				Runtime	Prediction loss (10^{-3})	
Model	E_{mom} (10^{1})	E_{mass} (10^{-3})	M_{-} (10^{-3})	r_{-} (10^{-2})	ms/step	$\mathcal{L}_{\text{fluid}}$	$\mathcal{L}_{\text{conc}}$
2D-TGV							
GNS	5.211	<u>1.711</u>	0.014	0.985	38.28	**0.211**	**0.014**
EGNN	<u>4.358</u>	12.46	<u>0.008</u>	<u>0.651</u>	39.01	1.051	0.071
SFBC	23.51	6.659	0.254	3.325	<u>14.74</u>	2.352	0.150
FluidFormer	467.1	10130	20270	53.90	65.37	76.70	5907
Ours	**0.004**	**0.005**	**0.000**	**0.000**	**11.98**	<u>0.775</u>	<u>0.064</u>
2D-LDC							
GNS	<u>0.130</u>	47.80	<u>0.853</u>	<u>5.640</u>	35.79	**0.383**	**1.040**
EGNN	0.507	<u>34.30</u>	0.854	8.860	37.92	3.390	<u>1.150</u>
SFBC	2.197	129.5	3.868	6.407	<u>18.60</u>	46.55	2.451
FluidFormer	6.367	107.1	163.1	15.44	73.59	16092	30.23
Ours	**0.126**	**0.103**	**0.000**	**0.000**	**9.92**	<u>0.415</u>	1.842

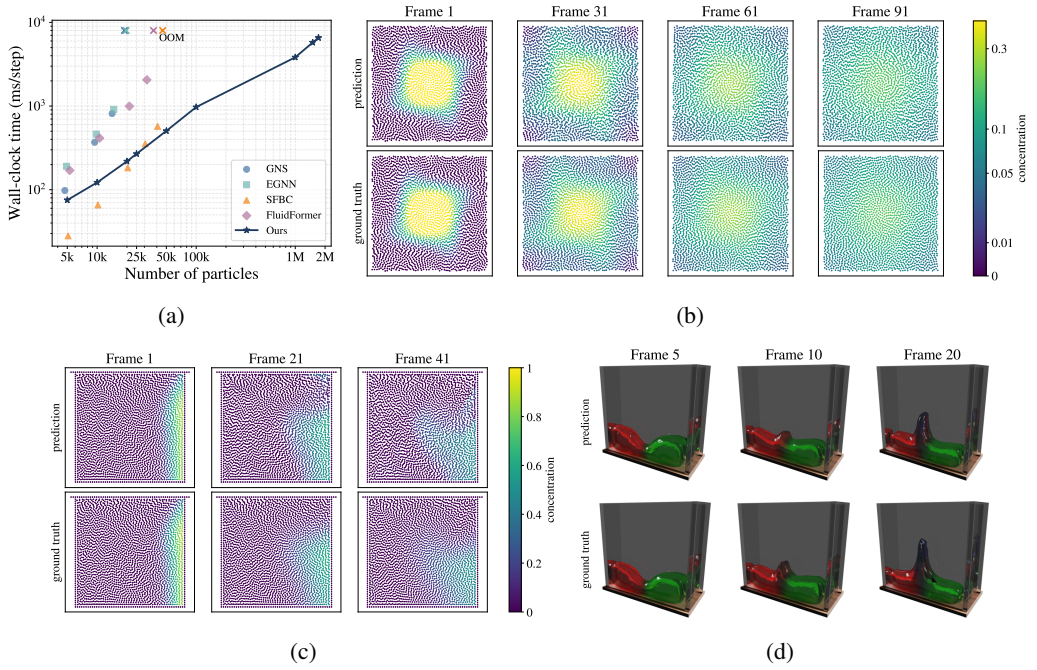
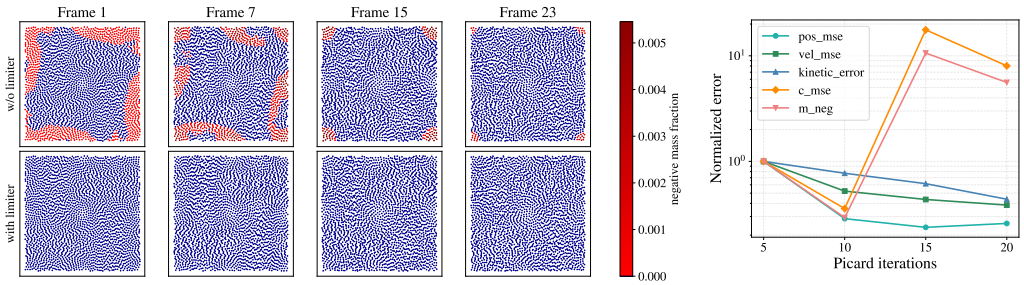


Figure 4: (a) Runtime scalability with increasing particle counts; (b–d) Visualized concentration fields under advection and diffusion in 2D-TGV, 2D-LDC and 3D Dam-break of the prediction of the proposed model and ground truth.

Such advantage over physics consistency remains in 2D-LDC with momentum sources as in Table 1 and Figure 4c. We omit direct 3D accuracy comparisons because standard GNS-style 3D benchmarks contain only less than 10K particles, whereas our 3D setting contains about 60K, making a scale-matched comparison unavailable. We nevertheless evaluate computational scalability at this scale, where the graph-based baselines exceed GPU memory during inference while our model remains feasible as mentioned below. The result is visualized in Figure 4d.

Consequently, our model trades a moderate amount of raw fitting capacity for substantially stronger physical consistency and lower inference cost, while retaining prediction errors of the same practical scale as the more expressive baselines. The advantage becomes increasingly explicit as the system size grows, as shown in Figure 3b: while graph-based baselines quickly exhaust GPU memory, our model supports roughly 38× more particles in a single run, with a noticeably slower runtime growth trend even before competing methods become out of memory (OOM) in Figure 4a.



(a) (b)

Figure 5: (a) Negative-mass violations during rollout. Red denotes negative concentration and blue feasible states; the limiter eliminates violations observed without limiting; (b) Picard iteration ablation on TGV. Dynamic and quantity errors are normalized by their values at 5 iterations.

Table 2: Ablation studies. Lower is better. In (a), $\mathcal{L}_{\text{conc}}$, E_{mass} and r_- are reported in units of 10^{-4}, 10^{-6} and 10^{0}, respectively; losses in (b) and (c) are reported in units of 10^{-3}.

Limiter	$\mathcal{L}_{\text{conc}}$	r_-
None	34.061	4.917
Single-sided	8.793	**0.000**
Symmetric	8.793	**0.000**
Repulsive	**8.587**	**0.000**

(a) Latent flux limiter.

Spatial Bias	$\mathcal{L}_{\text{fluid}}$	$\mathcal{L}_{\text{conc}}$
None	**0.782**	3.406
F	1.211	7.860
C	**0.782**	**0.057**
F+C	1.211	0.087

(b) Spatial-bias ablation.

MeshConv	$\mathcal{L}_{\text{fluid}}$	$\mathcal{L}_{\text{conc}}$
W/O	26.655	73.421
16×16	26.147	58.545
32×32	0.921	16.102
64×64	**0.782**	**3.406**

(c) Ablations on MeshConv.

4.3 Effect of Conservative and Spatial Structures

Latent Flux Limiters. Three types of donor-receiver limiters are tested based on the same learned latent dynamics in the left of Table 2a. Interestingly, they achieve nearly identical prediction and conservation errors while completely eliminating negative concentrations as Figure 5a. This suggests that their dominant empirical role is to enforce local physical feasibility, rather than to increase predictive capacity, and that the limiter family is robust to the specific gate parameterization. Nevertheless, the variants are not dynamically equivalent – they induce distinct state-dependent advection and numerical viscosity structures in the continuum limit (Appendix B).

Mesh Convolution and Spatial Inducted Bias. Table 2b shows task-dependent effects of cosine distance attenuation, where F/C denote bias on fluid/concentration. For fluid dynamics, attenuation restricts the receptive field and weakens nonlocal momentum and pressure interactions, slightly degrading prediction; for local and gradient-driven diffusion, suppressing long-range interactions strongly reduces concentration error. Applying bias only to concentration thus preserves nonlocal fluid interactions while imposing diffusion locality. Table 2c further shows that higher MeshConv resolution progressively reduces both errors: fluid MSE drops sharply at 32×32, while concentration further improves at 64×64, confirming the importance of near-field interactions for concentration transport, which inherits velocity errors through advection.

Accuracy and Cost of Latent Iteration. As Figure 5b shows, in the 2D-TGV benchmark, increasing Picard iterations without limiters improve the convergence of momentum-related quantities with diminishing marginal gains, which is consistent with the contraction behavior of the implicit latent solver. However, the concentration variables exhibit non-monotonic errors as the iteration proceeds. It can be attributed to the absence of explicit positivity constraints in the latent equilibrium operator: while additional iterations drive the solution closer to the learned fixed point, the equilibrium itself is not guaranteed to lie in the physically feasible region. This motivates the latent limiting mechanism.

5 Conclusion

We propose scalable conservative latent linear attention ConsAttn combining skew-symmetric transport, sign-split limiting, and local MeshConv. Fluid and reactive transport experiments reveal improved physical consistency and scaling with competitive accuracy. Future work targets strongly coupled multiphysics and adaptive representations for heterogeneous spatial dynamics.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with language polishing, grammar checking, and improving the clarity of manuscript presentation. We have not used generative AI tools for generating research results, designing experiments, performing data analysis, or making scientific claims. Additionally, we used generative AI tools for limited assistance in debugging and improving code implementation. We have reviewed all AI-assisted work, manually verified the correctness of the generated suggestions and code modifications, and independently validated all experimental results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Gizem Akinci, Markus Ihmsen, Nadir Akinci, and Matthias Teschner. Parallel surface reconstruction for particle-based fluids. In *Computer graphics forum*, volume 31, pp. 1797–1809. Wiley Online Library, 2012.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Re. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840. PMLR, 21–27 Jul 2024.

Markus Becker and Matthias Teschner. Weakly compressible sph for free surface flows. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 209–217, 2007.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik Bekkers, and Max Welling. Se(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18126–18138, 2021.

Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention approximation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17413–17426, 2021.

Hanting Chen, Liuzhi Cheng, Xutao Wang, Yuchuan Tian, and Yunhe Wang. Dijiang: Efficient large language models through compact kernelization. In *Forty-first International Conference on Machine Learning, ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7103–7117, 2024.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Qihang Fan, Huaibo Huang, Yuang Ai, and Ran He. Rectifying magnitude neglect in linear attention. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21505–21514. IEEE, 2025a.

Qihang Fan, Huaibo Huang, and Ran He. Breaking the low-rank dilemma of linear attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25271–25280. IEEE, 2025b.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5938–5948. IEEE, 2023.

Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *European conference on computer vision*, pp. 124–140. Springer, 2024.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569, 2023.

Masanobu Horie and Naoto Mitsume. Graph neural PDE solvers with conservation and similarity-equivariance. In *Forty-first International Conference on Machine Learning, ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18785–18814, 2024.

Steeven Janny, Aurélien Béneteau, Madiha Nadri, Julie Digne, Nicolas Thome, and Christian Wolf. EAGLE: large-scale learning of turbulent fluid dynamics with mesh transformers. In *The Eleventh International Conference on Learning Representations,ICLR*, 2023.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5156–5165, 2020.

Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18278–18289. IEEE, 2023a.

Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *7th International Conference on Learning Representations, ICLR*, 2019.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research (JMLR)*, 2023b.

Salvatore Marrone, MAGD Antuono, A Colagrossi, G Colicchio, D Le Touzé, and Giorgio Graziani. δ-sph model for simulating violent impact flows. *Computer Methods in Applied Mechanics and Engineering*, 200(13-16):1526–1542, 2011.

Weikang Meng, Yadan Luo, Liangyu Huo, Yingjian Li, Yaowei Wang, Xin Li, and Zheng Zhang. Norm $\times$ direction: Restoring the missing query norm in vision linear attention. *arXiv preprint arXiv:2506.21137*, 2025a.

Weikang Meng, Yadan Luo, Xin Li, Dongmei Jiang, and Zheng Zhang. Polaformer: Polarity-aware linear attention for vision transformers. In *The Thirteenth International Conference on Learning Representations,ICLR*, 2025b.

Namuk Park and Songkuk Kim. How do vision transformers work? In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Ruihan Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8459–8468, 2020.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Learning Representations*, 2021.

Omer Rochman Sharabi, Sacha Lewin, and Gilles Louppe. A neural material point method for particle-based emulation. *Trans. Mach. Learn. Res.*, 2025, 2025.

Will Snyder, Irina Tezaur, and Christopher Wentland. Domain decomposition-based coupling of physics-informed neural networks via the schwarz alternating method. *arXiv preprint arXiv:2311.00224*, 2023.

Makoto Takamoto, Krishnan Pradip, et al. Lagrangebench: A benchmarking suite for physics-informed machine learning. In *Advances in Neural Information Processing Systems*, 2023.

Artur Toshev, Jonas A. Erbesdobler, Nikolaus A. Adams, and Johannes Brandstetter. Neural SPH: Improved neural modeling of lagrangian fluid dynamics. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48428–48452. PMLR, 21–27 Jul 2024.

Benjamin Ummenhofer, Lukas Prantl, Nils Thuerey, and Vladlen Koltun. Lagrangian fluid simulation with continuous convolutions. In *International Conference on Learning Representations*, 2020.

Nianyi Wang, Shuai Zheng, Yu Chen, Hai Zhao, and Zhou Fang. Fluidformer: Transformer with continuous convolution for particle-based fluid simulation. *Neural Networks*, pp. 108631, 2026.

Rene Winchenbach and Nils Thuerey. Symmetric basis convolutions for learning lagrangian fluid mechanics. In *International Conference on Learning Representations*, volume 2024, pp. 4677–4730, 2024.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705, 2024.

Haixu Wu, Minghao Guo, Zongyi Li, Zhiyang Dou, Mingsheng Long, Kaiming He, and Wojciech Matusik. Geopt: Scaling physics simulation via lifted geometric pre-training. *arXiv preprint arXiv:2602.20399*, 2026.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. Path attention: Position encoding via accumulating householder transformations. *Advances in Neural Information Processing Systems*, 38:62220–62247, 2026.

A Algebraic Conservation of the Network Architecture

A.1 Proofs

Theorem A.1. *The operation* MeshConv *is conservative, i.e.,* $\sum_i$ MeshConv$(\mathbf{h}_i; \{\mathbf{h}_j\}) = 0$.

Proof.

$$\sum_i \Delta\mathbf{h}_i = \sum_i \mathsf{MeshConv}(\mathbf{h}_i; \{\mathbf{h}_j\}) \tag{14}$$

$$= \sum_i \sum_\alpha \omega_{i\alpha}(K * \mathbf{H})_\alpha - \sum_i \mathbf{h}_i \odot \sum_\alpha \omega_{i\alpha}(K^\top * \rho)_\alpha \tag{15}$$

$$= \sum_\alpha (K * \mathbf{H})_\alpha \underbrace{\sum_i \omega_{i\alpha}}_{=:\rho_\alpha} - \sum_\alpha (K^\top * \rho)_\alpha \odot \underbrace{\sum_i \omega_{i\alpha}\mathbf{h}_i}_{\mathbf{H}_\alpha} \tag{16}$$

$$= \sum_\alpha (K * \mathbf{H})_\alpha \rho_\alpha - \sum_\alpha (K^\top * \rho)_\alpha \odot \mathbf{H}_\alpha \tag{17}$$

$$= \sum_{\alpha,\beta} K_{\alpha,\beta} \odot \mathbf{H}_\beta \rho_\alpha - \sum_{\alpha,\beta} \mathbf{H}_\alpha \odot K_{\beta,\alpha}\rho_\beta \tag{18}$$

$$= \sum_{\alpha,\beta} K_{\alpha,\beta} \odot \mathbf{H}_\beta \rho_\alpha - \sum_{\beta,\alpha} \rho_\alpha K_{\alpha,\beta} \odot \mathbf{H}_\beta = \mathbf{0} \tag{19}$$

□

Theorem A.2. *Under the source-free gated self-attention skew-symmetric updating rule, for* $m_i^{(n)} \in [m_{\min}, m_{\max}]$*, we have* $m_i^{(n+1)} \in [m_{\min}, m_{\max}]$.

Proof. The source-free gated Eq. 5 gives

$$m_i^{(n+1)} = m_i^{(n)} - \tau_i \sum_j \alpha_i \alpha_j f_{ij} = m_i^{(n)} - \tau_i \alpha_i \sum_j \alpha_j O_{ij} + \tau_i \alpha_i \sum_j \alpha_j I_{ij} \tag{20}$$

which gives

$$\begin{aligned} m_i^{(n+1)} &\geq m_i^{(n)} - \tau_i \alpha_i O_i \geq m_i^{(n)} - (m_i^{(n)} - m_{\min}) = m_{\min} \\ m_i^{(n+1)} &\leq m_i^{(n)} + \tau_i \alpha_i I_i \leq m_i^{(n)} + (m_{\max} - m_i^{(n)}) = m_{\max}. \end{aligned} \tag{21}$$

□

A.2 Extension to Multiple Conservation Laws

Suppose that K physical quantities are represented by linear projections of the same latent state,

$$\tilde{\mathbf{u}}_i^{(k)} = W_{P,k} \mathbf{z}_i, \qquad k = 1, \ldots, K. \tag{22}$$

We augment Eq. 5 with one source encoder for each physical law:

$$V_i \Delta \mathbf{z}_i = -\Delta t \sum_j \mathbf{F}_{ij} + \sum_{k=1}^{K} \phi_k(\mathbf{z}_i, \mathbf{z}_i + \Delta \mathbf{z}_i, \mathbf{f}_{i,k}). \tag{23}$$

Since $\mathbf{F}_{ij} = -\mathbf{F}_{ji}$, its contribution vanishes globally under every linear projection:

$$\sum_i V_i \Delta \tilde{\mathbf{u}}_i^{(k)} = W_{P,k} \sum_i V_i \Delta \mathbf{z}_i = W_{P,k} \sum_i \sum_{\ell=1}^{K} \phi_\ell(\mathbf{z}_i, \mathbf{z}_i + \Delta \mathbf{z}_i, \mathbf{f}_{i,\ell}). \tag{24}$$

Hence all K projected quantities are simultaneously conserved in the source-free case. With sources, their global changes are determined solely by the projected source encoders; for a prescribed source balance $\mathbf{s}_{i,k}$, one may impose

$$W_{P,k} \sum_{\ell=1}^{K} \phi_\ell(\mathbf{z}_i, \mathbf{z}_i + \Delta \mathbf{z}_i, \mathbf{f}_{i,\ell}) = \mathbf{s}_{i,k}, \tag{25}$$

which gives

$$\sum_i V_i \Delta \mathbf{u}_i^{(k)} = \sum_i \mathbf{s}_{i,k}, \qquad k = 1, \ldots, K. \tag{26}$$

Thus multiple conservation laws share the same anti-symmetric latent transport and differ only through their linear projections and structured source terms, preserving the $\mathcal{O}(N)$ complexity.

A.3 Sign-Split Limiter for MeshConv

The conservative MeshConv can also be equipped with a sign-split limiter. Define the effective particle-to-particle coefficient

$$C_{ij} := \sum_{\alpha,\beta} \omega_{i\alpha} K_{\alpha,\beta} \omega_{j\beta}. \tag{27}$$

Then Eq. 12 can be equivalently written as

$$\mathsf{MeshConv}(\mathbf{h}_i; \{\mathbf{h}_j\}) = \sum_j (C_{ij} \mathbf{h}_j - C_{ji} \mathbf{h}_i) =: \sum_j \mathbf{F}_{ij}^M, \tag{28}$$

where $\mathbf{F}_{ij}^M = -\mathbf{F}_{ji}^M$. Hence MeshConv admits the same pairwise conservative-flux interpretation as our linear attention.

For a bounded scalar quantity $m_i = W_P \mathbf{h}_i$, define the directed interpolation-path contribution

$$x_{ij}^{\alpha\beta} := \omega_{i\alpha} K_{\alpha,\beta} \omega_{j\beta} m_j. \tag{29}$$

The projected pairwise flux becomes

$$f_{ij}^M := W_P \mathbf{F}_{ij}^M = \sum_{\alpha,\beta} \left(x_{ij}^{\alpha\beta} - x_{ji}^{\beta\alpha} \right) = O_{ij}^M - I_{ij}^M, \tag{30}$$

where

$$O_{ij}^M = \sum_{\alpha,\beta} \left[(x_{ij}^{\alpha\beta})^+ + (x_{ji}^{\beta\alpha})^- \right], \tag{31}$$

$$I_{ij}^M = \sum_{\alpha,\beta} \left[(x_{ij}^{\alpha\beta})^- + (x_{ji}^{\beta\alpha})^+ \right]. \tag{32}$$

Therefore $O_{ij}^M = I_{ji}^M \geq 0$, and the gross outgoing and incoming fluxes are $O_i^M = \sum_j O_{ij}^M$ and $I_i^M = \sum_j I_{ij}^M$.

Importantly, these quantities need not be evaluated by explicitly forming C_{ij}. Let $K = K^+ - K^-$ and $m = m^+ - m^-$, with all positive and negative parts non-negative. Using the non-negative interpolation weights $\omega_{i\alpha}$,

$$\begin{aligned} O_i^M &= \mathsf{G2P}\big(K^+ * \mathsf{P2G}(m^+) + K^- * \mathsf{P2G}(m^-)\big)_i \\ &\quad + m_i^- \, \mathsf{G2P}\big((K^\top)^+ * \mathsf{P2G}(1)\big)_i + m_i^+ \, \mathsf{G2P}\big((K^\top)^- * \mathsf{P2G}(1)\big)_i, \end{aligned} \tag{33}$$

$$\begin{aligned} I_i^M &= \mathsf{G2P}\big(K^+ * \mathsf{P2G}(m^-) + K^- * \mathsf{P2G}(m^+)\big)_i \\ &\quad + m_i^+ \, \mathsf{G2P}\big((K^\top)^+ * \mathsf{P2G}(\{1\})\big)_i + m_i^- \, \mathsf{G2P}\big((K^\top)^- * \mathsf{P2G}(\{1\})\big)_i. \end{aligned} \tag{34}$$

Thus the sign split preserves the original P2G–convolution–G2P factorization and costs $\mathcal{O}(N + N_g)$ for a fixed-size convolution stencil.

For the update

$$V_i \Delta \mathbf{h}_i = -\gamma_2 \Delta t \sum_j \mathbf{F}_{ij}^M, \tag{35}$$

let $\eta_i = \gamma_2 \Delta t / V_i$ and define

$$\alpha_i^M = \min\left\{1, \frac{m_i - m_{\min}}{\eta_i O_i^M}, \frac{m_{\max} - m_i}{\eta_i I_i^M}\right\}. \tag{36}$$

The limited pairwise flux is

$$\widetilde{\mathbf{F}}_{ij}^M := \alpha_i^M \alpha_j^M \mathbf{F}_{ij}^M. \tag{37}$$

Because $\alpha_i^M \alpha_j^M = \alpha_j^M \alpha_i^M$, $\widetilde{\mathbf{F}}_{ij}^M$ remains skew-symmetric, preserving global conservation while guaranteeing $m_i \in [m_{\min}, m_{\max}]$.

Moreover, the limited operator can still be evaluated without explicit particle pairs:

$$\begin{aligned} \widetilde{\mathsf{MeshConv}}(\mathbf{h}_i) &= \alpha_i^M \, \mathsf{G2P}\left(K * \mathsf{P2G}(\{\alpha_j^M \mathbf{h}_j\})\right)_i \\ &\quad - \alpha_i^M \mathbf{h}_i \odot \mathsf{G2P}\left(K^\top * \mathsf{P2G}(\{\alpha_j^M\})\right)_i. \end{aligned} \tag{38}$$

Hence flux limiting preserves both the conservation structure and linear computational complexity of MeshConv.

B Continuous Limit of Self-Linear Attention and the Donor–Receiver Limiter Family

This section establishes the connection between self-linear attention and latent transport, and then derives a unified continuum representation for several discrete limiter constructions.

The key structural assumption is $\phi_V = \mathrm{id}$, which means that self-attention does not transform the quantity being transported: every directed attention contribution is a scalar coefficient multiplying the donor latent state itself. The following result can be generalized beyond linear self-attention.

B.1 Self-Linear Attention as Latent Transport

For normalized linear attention, the directed contribution from particle j to particle i is

$$\mathbf{A}_{ij} = \omega_{ij} \mathbf{z}_j, \omega_{ij} := \sum_h L_{ih} k_{jh}. \tag{39}$$

Hence the skew-symmetric conservative interaction is

$$\mathbf{F}_{ij} = \omega_{ij}\mathbf{z}_j - \omega_{ji}\mathbf{z}_i. \tag{40}$$

To identify its continuum transport velocity, consider a uniform 1D grid with spacing Δx and locally freeze the attention kernel over one time step, i.e., $\omega_{i,i-k} \approx p_k, \omega_{i-k,i} \approx p_{-k}$, then a local attention update can be written as

$$\begin{aligned}
\mathbf{z}_i^{(n+1)} - \mathbf{z}_i^{(n)} &= \Delta t \sum_{k\in\mathbb{Z}} \omega_{i,i-k}\mathbf{z}_{i-k}^{(n)} - \omega_{i-k,i}\mathbf{z}_i^{(n)} && (41)\\
&\approx \sum_{k\neq 0}(p_k\Delta t)\mathbf{z}_{i-k}^{(n)} - \sum_{k\neq 0}(p_{-k}\Delta t)\mathbf{z}_i^{(n)} && (42)\\
&= \sum_{k\neq 0}(p_k\Delta t)\mathbf{z}_{i-k}^{(n)} - \sum_{k\neq 0}(p_k\Delta t)\mathbf{z}_i^{(n)} && (43)\\
&= \sum_{k\neq 0}(p_k\Delta t)(\mathbf{z}_{i-k}^{(n)} - \mathbf{z}_i^{(n)}) && (44)
\end{aligned}$$

Thus for Δt small enough, define $\omega_k := p_k\Delta t$ for $k \neq 0$ and $\omega_0 := 1 - \sum_{k\neq 0}\omega_k$, then we have:

$$\mathbf{z}_i(t+\Delta t) = \sum_k \omega_k \mathbf{z}_{i-k}(t), \sum_k \omega_k = 1. \tag{45}$$

Expanding

$$\begin{aligned}
\mathbf{z}(x, t+\Delta t) &= \mathbf{z} + \Delta t\mathbf{z}_t + O(\Delta t^2)\\
\mathbf{z}(x-k\Delta x, t) &= \mathbf{z} - k\Delta x\,\mathbf{z}_x + \frac{k^2\Delta x^2}{2}\mathbf{z}_{xx} + O(\Delta x^3)
\end{aligned} \tag{46}$$

into Eq. 45 yields

$$\mathbf{z}_t = -\underbrace{\frac{\Delta x}{\Delta t}\left(\sum_k k\omega_k\right)}_{=:v}\mathbf{z}_x + \underbrace{\frac{\Delta x^2}{2\Delta t}\left(\sum_k k^2\omega_k\right)}_{=:\nu}\mathbf{z}_{xx} + O(\Delta t) + O\left(\Delta x^2\right). \tag{47}$$

if the moments of ω_k are all assumed $\mathcal{O}(\Delta t/\Delta x)$. Thus the first and second spatial moments of the local attention kernel can be identified as the latent advection velocity v and self-attention numerical viscosity ν. Specifically, for the local one-cell upwind specialization used below, the directional attention transport across the interface $x_{i+\frac{1}{2}}$ takes the form $\mathbf{J}_{i+\frac{1}{2}} = v\mathbf{z}_i$ and $\nu = v\Delta x/2$.

B.2 Donor–Receiver Limiter Family

Because the quantity carried by the self-attention flux $\mathbf{J}_{i+\frac{1}{2}} = v\mathbf{z}_i$ is exactly the donor latent state $\mathbf{z}_i$, a scalar limiter modifies the transport coefficient without changing the transported variable.

We introduce two scalar gates where a_i acts on the donor and b_i acts on the receiver $a_i, b_i \in [0, 1]$. And the general limited interface flux is

$$\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = v\, a_i b_{i+1}\mathbf{z}_i. \tag{48}$$

The corresponding conservative finite-volume update is

$$\frac{\mathbf{z}_i(t+\Delta t) - \mathbf{z}_i(t)}{\Delta t} = -\frac{\widetilde{\mathbf{J}}_{i+\frac{1}{2}} - \widetilde{\mathbf{J}}_{i-\frac{1}{2}}}{\Delta x}, \tag{49}$$

Let $\mathbf{u} := a\mathbf{z}$, and then around $x_i = x$, we have

$$\begin{aligned}
\mathbf{u}(x-\Delta x) =& \mathbf{u} - \Delta x\,\mathbf{u}_x + \frac{\Delta x^2}{2}\mathbf{u}_{xx} + O(\Delta x^3),\\
b(x+\Delta x) =& b + \Delta x\, b_x + \frac{\Delta x^2}{2}b_{xx} + O(\Delta x^3).
\end{aligned} \tag{50}$$

Substituting these expansions into Eq. equation 49 gives the modified equation

$$\frac{\partial \mathbf{z}}{\partial t} + v\frac{\partial}{\partial x}(ab\mathbf{z}) = \frac{v\Delta x}{2}\frac{\partial}{\partial x}\left[b\frac{\partial(a\mathbf{z})}{\partial x} - a\mathbf{z}\frac{\partial b}{\partial x}\right] + O(\Delta t) + O(\Delta x^2). \tag{51}$$

Eq. 51 is the continuum counterpart of the general donor–receiver limiter. It is obtained directly from the self-attention transport $\mathbf{J} = v\mathbf{z}$ rather than postulating an advection equation a priori.

If $a = a(\mathbf{z})$ and $b = b(\mathbf{z})$ are locally differentiable functions of the latent state, then

$$\frac{\partial}{\partial x}(ab\mathbf{z}) = \mathbf{D}_{a,b}(\mathbf{z})\frac{\partial \mathbf{z}}{\partial x}, \tag{52}$$

where

$$\mathbf{D}_{a,b} = ab\mathbf{I} + \mathbf{z} \otimes (b\nabla_{\mathbf{z}}a + a\nabla_{\mathbf{z}}b)\,. \tag{53}$$

Similarly,

$$b\frac{\partial(a\mathbf{z})}{\partial x} - a\mathbf{z}\frac{\partial b}{\partial x} = \mathbf{M}_{a,b}(\mathbf{z})\frac{\partial \mathbf{z}}{\partial x}, \tag{54}$$

with

$$\mathbf{M}_{a,b} = ab\mathbf{I} + \mathbf{z} \otimes (b\nabla_{\mathbf{z}}a - a\nabla_{\mathbf{z}}b)\,. \tag{55}$$

Therefore,

$$\frac{\partial \mathbf{z}}{\partial t} + v\mathbf{D}_{a,b}(\mathbf{z})\frac{\partial \mathbf{z}}{\partial x} = \frac{v\Delta x}{2}\frac{\partial}{\partial x}\left(\mathbf{M}_{a,b}(\mathbf{z})\frac{\partial \mathbf{z}}{\partial x}\right) + O(\Delta t) + O(\Delta x^2). \tag{56}$$

When a and b vary slowly with the latent state, $\nabla_{\mathbf{z}}a$ and $\nabla_{\mathbf{z}}b$ are small and

$$\mathbf{D}_{a,b} \approx \mathbf{M}_{a,b} \approx ab\mathbf{I}. \tag{57}$$

Hence the latent features are approximately advected at the effective velocity

$$v_{\text{eff}} \approx ab\,v, \tag{58}$$

with a corresponding local numerical viscosity proportional to $ab\,v\Delta x/2$ for the one-cell upwind specialization.

B.3 Special Cases

- **Single-sided donor limiter**. Setting $a = \min\left\{1, \frac{m_i - m_{\min}}{\tau_i O_i}\right\}, b = 1$ gives $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = \alpha_i v\mathbf{z}_i$ which is precisely the donor-cell form and ensures $m_i \geq m_{\min}$. Eq. 53 and Eq. 55 reduce to $\mathbf{D}_{\text{donor}} = \mathbf{M}_{\text{donor}} = \alpha\mathbf{I} + \mathbf{z} \otimes \nabla_{\mathbf{z}}\alpha$. Consequently, even when $\alpha = 0$, the transport Jacobian does not necessarily vanish because the term $\mathbf{z} \otimes \nabla_{\mathbf{z}}\alpha$ may remain nonzero.
- **Double-sided symmetric limiter**. For the aforementioned symmetric limiter, $a_i = b_i = \min\left\{1, \frac{m_i - m_{\min}}{\tau_i O_i}, \frac{m_{\max} - m_i}{\tau_i I_i}\right\}$. The general expressions become $\mathbf{D}_{\text{sym}} = \alpha^2\mathbf{I} + 2\alpha\mathbf{z} \otimes \nabla_{\mathbf{z}}\alpha$ and $\mathbf{M}_{\text{sym}} = \alpha^2\mathbf{I}$. Thus when $\alpha = 0$, both $\mathbf{D}_{\text{sym}}$ and $\mathbf{M}_{\text{sym}}$ vanishes. Both the convective transport and its leading numerical viscosity therefore collapse at the interface. This is the structural *topological deadlock* induced by double-sided symmetric gating.
- **Repulsive occupancy limiter**. If $\rho \in [0, 1]$ denotes a physical occupation fraction, choosing $a = \rho, b = 1 - \rho$ gives $\widetilde{\mathbf{J}}_{i+\frac{1}{2}} = v\rho_i(1 - \rho_{i+1})\mathbf{z}_i$, so that transport is suppressed either when the donor is empty or when the receiver is fully occupied. The corresponding transport Jacobian is $\mathbf{D}_{\text{rep}} = \rho(1 - \rho)\mathbf{I} + (1 - 2\rho)\mathbf{z} \otimes \nabla_{\mathbf{z}}\rho$ while the diffusion matrix is $\mathbf{M}_{\text{rep}} = \rho(1 - \rho)\mathbf{I} + \mathbf{z} \otimes \nabla_{\mathbf{z}}\rho$. The interpretation of $1 - \rho$ as receiver capacity specifically requires ρ to represent an occupation variable; it should not be identified in general with a donor restriction ratio such as m_i/O_i.

Remark B.1. The max/min operations used to construct the discrete limiter are piecewise differentiable. Therefore, the Jacobians above describe the continuum equation inside each fixed active-set regime. At switching surfaces where the active limiter branch changes, the modified equation should be understood piecewise (or in the corresponding weak sense).

C Implementation Details

C.1 Separable Spatial Bias for Linear Attention

Although linear attention provides dense long-range interactions, not all particle pairs should contribute equally to physical transport. For many transported quantities, especially concentrations, it is desirable to introduce a spatial inductive bias that favors nearby particles while suppressing excessively distant interactions. A direct pairwise modulation

$$\mathbf{A}_{ij} \propto \phi_Q(\mathbf{z}_i)^\top \phi_K(\mathbf{z}_j)\phi_V(\mathbf{z}_j)^\top g(\mathbf{x}_i, \mathbf{x}_j) \tag{59}$$

would in general destroy the linear-attention factorization unless the spatial kernel admits a finite-dimensional separable form

$$g(\mathbf{x}_i, \mathbf{x}_j) = \psi_Q(\mathbf{x}_i)^\top \psi_K(\mathbf{x}_j). \tag{60}$$

We therefore construct a low-dimensional trigonometric spatial feature map that preserves linear-time aggregation.

One-dimensional Construction. Consider first a one-dimensional bounded domain. After rescaling the particle coordinates to $x \in [0, \pi/2]$, we choose $g(x_i, x_j) = \cos(x_i - x_j)$. Since $|x_i - x_j| \in [0, \pi/2]$, this kernel is non-negative and monotonically decreases with the particle distance within the simulation domain. The trigonometric identity $\cos(x_i - x_j) = \cos x_i \cos x_j + \sin x_i \sin x_j$ implies its separability. And define the two-dimensional positional feature

$$\varphi(x) := \begin{pmatrix} \cos x \\ \sin x \end{pmatrix} \implies g(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j). \tag{61}$$

The spatially modulated attention term can therefore be written as

$$\phi_Q(\mathbf{z}_i)^\top \phi_K(\mathbf{z}_j)\varphi(x_i)^\top \varphi(x_j) = \left[\phi_Q(\mathbf{z}_i) \otimes \varphi(x_i)\right]^\top \left[\phi_K(\mathbf{z}_j) \otimes \varphi(x_j)\right]. \tag{62}$$

Hence it is sufficient to replace the original query and key features by

$$\widehat{\phi}_Q(\mathbf{z}_i, \mathbf{x}_i) = \phi_Q(\mathbf{z}_i) \otimes \varphi(\mathbf{x}_i), \qquad \widehat{\phi}_K(\mathbf{z}_j, \mathbf{x}_j) = \phi_K(\mathbf{z}_j) \otimes \varphi(\mathbf{x}_j). \tag{63}$$

The resulting attention remains separable and therefore retains $\mathcal{O}(N)$ complexity in the number of particles, while the feature dimension is increased only by the dimension of the spatial feature map.

Sharper spatial suppression. If a stronger locality bias is desired, one may replace the basic kernel by $g_K(x_i, x_j) = \cos^K(x_i - x_j)$ where increasing K produces a more concentrated kernel around $x_i = x_j$ on the normalized domain. For integer K, $\cos^K \theta$ admits a finite Fourier expansion,

$$\cos^K \theta = \frac{1}{2^K} \sum_{\ell=0}^{K} \binom{K}{\ell} \cos\left((K - 2\ell)\theta\right), \tag{64}$$

after combining conjugate frequency pairs. Each non-zero cosine frequency can again be separated. Consequently, $\cos^K(x_i - x_j)$ admits an exact positional feature representation using $K + 1$ real features. This provides a controllable trade-off between spatial concentration and feature dimension.

Extension to multiple spatial dimensions A direct Cartesian product of one-dimensional kernels, $\prod_{d=1}^{D} \cos(x_{i,d} - x_{j,d})$ is exactly separable, but its positional feature dimension grows as 2^D and its level sets inherit the coordinate axes, which introduces an undesirable anisotropic bias. Instead, we use a small set of approximately uniformly distributed unit directions $\{\mathbf{w}_m\}_{m=1}^{M}$ and apply the one-dimensional construction to their projections:

$$g_M(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{M} \sum_{m=1}^{M} \cos\left(\mathbf{w}_m^\top(\mathbf{x}_i - \mathbf{x}_j)\right). \tag{65}$$

The corresponding positional feature is

$$\varphi_M(\mathbf{x}) = \frac{1}{\sqrt{M}} \begin{pmatrix} \cos(\mathbf{w}_1^\top \mathbf{x}) \\ \sin(\mathbf{w}_1^\top \mathbf{x}) \\ \vdots \\ \cos(\mathbf{w}_M^\top \mathbf{x}) \\ \sin(\mathbf{w}_M^\top \mathbf{x}) \end{pmatrix} \in \mathbb{R}^{2M}, \tag{66}$$

for which

$$g_M(\mathbf{x}_i, \mathbf{x}_j) = \varphi_M(\mathbf{x}_i)^\top \varphi_M(\mathbf{x}_j). \tag{67}$$

Thus the multidimensional spatial modulation remains exactly factorizable with only $2M$ positional features, independently of the particle number. Sampling the directions approximately uniformly mitigates the axis-aligned anisotropy of a Cartesian-product construction.

In practice, coordinates are scaled such that the projected displacements remain primarily within the decreasing branch of the cosine function over the simulated domain. We therefore interpret Eq. 65 as a *locality-promoting spatial modulation*, rather than a globally monotone radial kernel.

An alternative is to approximate a conventional decaying kernel, such as a Gaussian, using random Fourier features. For example,

$$\phi_{\mathrm{RFF}}(\mathbf{x}) = \sqrt{\frac{2}{D}} \left[\cos(\boldsymbol{\omega}_1^\top \mathbf{x} + b_1), \ldots, \cos(\boldsymbol{\omega}_D^\top \mathbf{x} + b_D)\right]^\top, \tag{68}$$

with $\boldsymbol{\omega}_l \sim \mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I})$, provides a Monte Carlo approximation to the Gaussian kernel $\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$. While this also preserves linear-attention factorization, the kernel is only approximated and its accuracy increases with the random feature dimension D. Since the positional features are coupled to the original attention features through a tensor product, a large D directly multiplies the effective query/key dimension. We therefore use the deterministic low-dimensional trigonometric construction above, which gives an exact separable modulation for a fixed set of directions.

C.2 Dataset Construction

Solver Details. All datasets are generated with a custom weakly-compressible SPH (WCSPH) solver implemented in Taichi. Each particle i is a Lagrangian fluid parcel carrying position $\mathbf{x}_i$, velocity $\mathbf{v}_i$, density ρ_i, pressure p_i, and concentration c_i of one or more solutes. The behavior of fluids is primarily governed by two fundamental equations.

$$\frac{D\mathbf{v}}{Dt} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \mathbf{v} + \mathbf{g}. \tag{69}$$

$$\frac{D\rho}{Dt} = -\rho(\nabla \cdot \mathbf{v}). \tag{70}$$

where ν is the kinematic viscosity , $\mathbf{g}$ is the gravitational acceleration and $\frac{D(\cdot)}{Dt}$ denotes the material derivative. The core idea of SPH is using kernel function $W : \mathbb{R}^d \to \mathbb{R}$ to approximate continuous fields. In our dataset generation, we use a normalized quintic spline kernel with smoothing length $h = \Delta x$ and compact support radius $H = 3h$, where Δx is the initial particle spacing. Defining $q = \|\mathbf{r}\|/h$, the kernel is given by

$$W(q) = \frac{\alpha_d}{h^d} \begin{cases} (3-q)^5 - 6(2-q)^5 + 15(1-q)^5, & 0 \le q < 1, \\ (3-q)^5 - 6(2-q)^5, & 1 \le q < 2, \\ (3-q)^5, & 2 \le q < 3, \\ 0, & q \ge 3. \end{cases} \tag{71}$$

where d is the spatial dimension and α_d is the dimensionless normalization constant. In the context of SPH, for a field f, we approximate its gradient and its Laplacian using formula as

$$(\nabla f)_i \approx \rho_i \sum_{j \in \mathcal{N}_i} m_j \left(\frac{f_i}{\rho_i^2} + \frac{f_j}{\rho_j^2}\right) \nabla_i W_{ij}, \tag{72}$$

$$(\nabla^2 f)_i \approx 2 \sum_{j \in \mathcal{N}_i} V_j (f_j - f_i) G_{ij}. \tag{73}$$

where $\mathcal{N}_i$ denotes the fluid neighbors of particle i, m_i is its mass, $V_i = m_i/\rho_i$, and

$$G_{ij} = -\frac{\mathbf{r}_{ij} \cdot \nabla_i W_{ij}}{\|\mathbf{r}_{ij}\|^2 + \eta H^2} \qquad \mathbf{r}_{ij} = \mathbf{x}_i - \mathbf{x}_j, \quad \eta = 0.01. \tag{74}$$

Thus, we discretize the pressure term in Eq. 69 by applying the gradient operator to p and multiplying by $-1/\rho_i$. The viscous term is obtained by applying the Laplacian operator to $\mathbf{v}$ and multiplying by ν.

We adopt one-way coupling between the fluid and solutes: the solutes are advected by the fluid velocity but do not affect the fluid dynamics. For each solute species s, diffusion is governed by diffusion equation

$$\frac{Dc_s}{Dt} = D_s \nabla^2 c_s. \tag{75}$$

where D_s denotes the diffusivity of species s. We discretize Eq. 75 using pairwise interactions between neighboring fluid particles

$$c^*_{i,s} = c^n_{i,s} + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i} D_s(\rho_i + \rho_j) V_i V_j G_{ij} \left(c^n_{j,s} - c^n_{i,s}\right). \tag{76}$$

where $c^*_{i,s}$ denotes the concentration after the diffusion step and before the subsequent reaction update. This pairwise discretization conserves the total amount of each solute species during diffusion.

Taylor–Green Vortex. The domain is the unit square $[0,1]^2$, with periodic boundary conditions in both directions, and contains $50 \times 50 = 2{,}500$ particles with initial spacing $\Delta x = 0.02$. The initial velocity field is

$$\mathbf{v}(x,y) = \big(-\cos(2\pi x)\sin(2\pi y),\ \sin(2\pi x)\cos(2\pi y)\big). \tag{77}$$

and the solute is initialized as a smooth circular patch

$$c(\mathbf{x}) = \frac{1}{2}\left[1 - \tanh\left(\frac{\|\mathbf{x} - \mathbf{x}_c\| - 0.18}{0.015}\right)\right], \quad \mathbf{x}_c = (0.5, 0.5). \tag{78}$$

We set up $\nu = D = 0.01$ and neglect gravity.

To introduce variability across trajectories, particle positions are initialized on a Cartesian lattice and independently perturbed by $\boldsymbol{\epsilon}_i \sim \mathcal{N}\big(\mathbf{0}, (0.25\Delta x)^2 I_d\big)$, followed by 5,000 damped relaxation steps before imposing the initial velocity and concentration fields. The same perturbation and relaxation procedure is used for the two datasets described below. Each trajectory contains 100 frames sampled at intervals of $0.04\,\mathrm{s}$, spanning $3.96\,\mathrm{s}$. The training, test, and validation sets contain 100, 50, and 50 trajectories, respectively, differing only in the random seed used for the initial particle perturbations.

Lid-Driven Capacity Flow. The domain is a unit-square cavity containing $50 \times 50 = 2{,}500$ fluid particles with initial spacing $\Delta x = 0.02$, enclosed by three layers of boundary particles. The top lid moves at $\mathbf{v}_{\mathrm{lid}} = (1, 0)$, while the other walls remain stationary. We set $\nu = D = 0.01$. Starting from rest, the fluid undergoes a $5\,\mathrm{s}$ warm-up before the solute is initialized as a smooth layer adjacent to the right wall

$$c(\mathbf{x}) = \frac{1}{2}\left[1 - \tanh\left(\frac{(1-x) - 0.10}{0.015}\right)\right]. \tag{79}$$

Dam-Break with Chemical Reaction. The domain is an open-top tank of size $2.0 \times 1.0 \times 0.6$, bounded by four layers of wall particles along the bottom and lateral surfaces. Two fluid columns of size $0.6 \times 0.8 \times 0.6$ are initialized at rest at opposite ends of the tank, containing a total of 52,488 fluid particles. The columns collapse under gravity and collide near the center of the tank. We set $\nu = 0.025$ and $\mathbf{g} = (0, -1, 0)$.

The two fluid columns initially carry solutes A and B, respectively, with no solute C present. As the columns collide and mix, A and B react irreversibly to form C:

$$A + B \to C. \tag{80}$$

All three solutes share a diffusivity of $D = 0.01$. The reaction rate is proportional to the product of the molar concentrations of A and B, with rate constant $k = 0.001$.

C.3 Network Architecture Details

Dynamics Input. For each particle i, the dynamics input consists of its past T positions $\{\mathbf{x}_i^{t-T+1}, \ldots, \mathbf{x}_i^t\}$, its current velocity $\mathbf{v}_i^t$, and its particle type $\tau_i \in \{0, 1, 2\}$, where $\tau_i = 0$ denotes a fluid particle, $\tau_i = 1$ a static wall particle, and $\tau_i = 2$ a moving wall particle. Absolute positions are

not used as features; only the displacements between consecutive frames, $\Delta\mathbf{x}_i^k = \mathbf{x}_i^{t-k+1} - \mathbf{x}_i^{t-k}$, $k = 1, \ldots, T-1$, are used. The velocity is normalized as $\hat{\mathbf{v}}_i^t = (\mathbf{v}_i^t - \boldsymbol{\mu}_\mathbf{v})/\boldsymbol{\sigma}_\mathbf{v}$ with dataset statistics $\boldsymbol{\mu}_\mathbf{v}, \boldsymbol{\sigma}_\mathbf{v}$. In periodic domains, $\Delta\mathbf{x}_i^k$ is computed under the minimum-image convention so that particles crossing the boundary do not produce spurious jumps. Each component $\Delta x_{i,a}^k$ is then embedded with Fourier features $\big[\cos(2\pi n\,\Delta x_{i,a}^k/L_a),\ \sin(2\pi n\,\Delta x_{i,a}^k/L_a)\big]_{n=1,2}$, where L_a is the domain length along axis a. All particles are fluid, so a single encoder gives the initial dynamics latent

$$\mathbf{z}_i^{(0)} = \mathrm{MLP}\Big(\big[\cos(2\pi n\,\Delta x_{i,a}^k/L_a),\ \sin(2\pi n\,\Delta x_{i,a}^k/L_a)\big]_{a,k,n} \oplus \hat{\mathbf{v}}_i^t\Big). \tag{81}$$

In wall-bounded domains, the displacements are normalized as velocities, and fluid particles additionally receive three wall features: the signed distances $\mathbf{d}_i$ to all wall surfaces, clipped at the kernel support radius H; the inward normal $\mathbf{n}_i$ of the nearest wall; and the lid velocity $\mathbf{v}_i^{\mathrm{lid}}$ for particles near a moving lid (zero otherwise) in Lid-Driven Capacity Flow dataset. Fluid and wall particles are encoded by two separate encoders

$$\mathbf{z}_i^{(0)} = \begin{cases} \mathrm{MLP}\Big(\big[\Delta\mathbf{x}_i^k/(\Delta t\,\boldsymbol{\sigma}_v)\big]_k \oplus \hat{\mathbf{v}}_i^t \oplus \mathbf{d}_i \oplus \mathbf{n}_i \oplus \mathbf{v}_i^{\mathrm{lid}}\Big), & \tau_i = 0, \\ \mathrm{MLP}\big(\hat{\mathbf{v}}_i^t \oplus \mathbf{n}_i \oplus \mathbf{e}_{\tau_i}\big), & \tau_i = 1, 2. \end{cases} \tag{82}$$

where $\mathbf{e}_{\tau_i}$ is a learned embedding that distinguishes static walls from moving walls. $\mathbf{z}_i^{(0)} \in \mathbb{R}^d$.

Concentration Input. For each particle i, the concentration input consists only of its past T concentrations $\{c_i^{t-T+1}, \ldots, c_i^t\}$. We represent the history by differences relative to the current value, append the current value, and encode the input with a two-layer MLP

$$\mathbf{s}_i^{(0)} = \begin{cases} \mathrm{MLP}\Big(\big[c_i^{t-k} - c_i^t\big]_k \oplus c_i^t\Big) & \tau_i = 0, \\ 0, & \tau_i = 1, 2. \end{cases} \tag{83}$$

Since boundary particles do not carry concentration, we set their $\mathbf{s}$ to a constant value of 0. In wall-bounded domains, we also add the fluid particles' signed distances $\mathbf{d}_i$ to all wall surfaces into the encoder's input.

Iteration in Latent Space. The dynamics latent $\mathbf{z}$ and the concentration latent $\mathbf{s}$ are updated together over L iterations. In each iteration the dynamics latent is updated by ConsAttn followed by MeshConv. The concentration latent is updated by ConsAttn only, without MeshConv, and only fluid particles take part in the concentration exchange.

The coupling between the two branches follows the physics of solute transport. The solute evolves by advection–diffusion: it is carried by the fluid velocity and spreads between neighboring fluid particles. However, the solute does not change the density, viscosity, or body force of the fluid, so it does not enter the momentum equation and has no influence on the flow. Therefore, we build the network to match this one-way dependence. The concentration branch does not compute its own queries and keys from $\mathbf{s}^{(l)}$. It builds them from the dynamics latent $\mathbf{z}^{(l)}$ of the same iteration, with gradients stopped, and only the values are computed from $\mathbf{s}^{(l)}$. Information thus flows from $\mathbf{z}$ to $\mathbf{s}$ but never in the opposite direction. Because of the stop-gradient, the concentration loss can't modify $\mathbf{z}$ or the dynamics parameters in the backward pass.

Update. Each physical quantity has its own decoder. We use three bias-free linear decoders: $W_{\bar{\mathbf{v}}}$ for the interval-averaged velocity, $W_\mathbf{v}$ for the instantaneous velocity, and W_c for the concentration. The first two decode the dynamics latent and the last decodes the concentration latent. Each decoder is applied to the latent increment accumulated over the L iterations, so the network predicts only a residual correction of the current state. The decoders predict only the contribution of particle interactions. For a fluid particle, the update formula for the physical quantity it carries is as follows

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \Delta t\,\mathbf{g} + \mathbf{W}_\mathbf{v}(\mathbf{z}_i^{(L)} - \mathbf{z}_i^{(0)}), \tag{84}$$

$$\bar{\mathbf{v}}_i^{t+1} = \bar{\mathbf{v}}_i^t + \Delta t\,\mathbf{g} + \mathbf{W}_{\bar{\mathbf{v}}}(\mathbf{z}_i^{(L)} - \mathbf{z}_i^{(0)}), \tag{85}$$

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \Delta t\,\bar{\mathbf{v}}_i^{t+1}, \tag{86}$$

$$c_i^{t+1} = c_i^t + \mathbf{W}_c(\mathbf{s}_i^{(L)} - \mathbf{s}_i^{(0)}). \tag{87}$$

In periodic domains, $\mathbf{x}_i^{t+1}$ is wrapped back into the box. The states of wall particles remain unchanged.