
FLOW MATCHING MEETS PDES: A UNIFIED FRAMEWORK FOR PHYSICS-CONSTRAINED GENERATION

Giacomo Baldan

Politecnico di Milano

giacomo.baldan@polimi.it

Qiang Liu

Technical University of Munich

qiang7.liu@tum.de

Alberto Guardone

Politecnico di Milano

alberto.guardone@polimi.it

Nils Thuerey

Technical University of Munich

nils.thuerey@tum.de

ABSTRACT

Generative machine learning methods, such as diffusion models and flow matching, have shown great potential in modeling complex system behaviors and building efficient surrogate models. However, these methods typically learn the underlying physics implicitly from data. We propose Physics-Based Flow Matching (PBFM), a novel generative framework that explicitly embeds physical constraints, both PDE residuals and algebraic relations, into the flow matching objective. We also introduce temporal unrolling at training time that improves the accuracy of the final, noise-free sample prediction. Our method jointly minimizes the flow matching loss and the physics-based residual loss without requiring hyperparameter tuning of their relative weights. Additionally, we analyze the role of the minimum noise level, σ_{\min} , in the context of physical constraints and evaluate a stochastic sampling strategy that helps to reduce physical residuals. Through extensive benchmarks on three representative PDE problems, we show that our approach yields up to an $8\times$ more accurate physical residuals compared to FM, while clearly outperforming existing algorithms in terms of distributional accuracy. PBFM thus provides a principled and efficient framework for surrogate modeling, uncertainty quantification, and accelerated simulation in physics and engineering applications.

1 Introduction

Partial differential equations (PDEs) constitute the foundational mathematical framework for describing the spatial and temporal evolution of physical systems (Evans, 2010). The numerical discretization of PDEs typically leads to high-dimensional systems, whose solution can become computationally prohibitive, particularly in the presence of nonlinearities or multiscale features (Haber et al., 2018; Valencia et al., 2025). To address these challenges, machine learning-based models have been proposed as an efficient alternative, capable of approximating the solution at a fraction of the computational cost (Chen et al., 2021; Fresca et al., 2021; Baldan et al., 2021; Brunton and Kutz, 2024). The framework of physics-informed neural networks (Raissi et al., 2019) rigorously embeds PDE constraints into the training process through automatic differentiation. However, this well-established approach is able to provide only a unique solution of the underlying PDE. Focusing only on deterministic systems is limiting if stochastic effects play a critical role, such as in the context of uncertainty quantification (Roy and Oberkampf, 2011; Abdar et al., 2021; Liu and Thuerey, 2024). Among generative modeling approaches, *denoising diffusion probabilistic models* (DDPMs) (Ho et al., 2020; Nichol and Dhariwal, 2021), and implicit DDIM (Song et al., 2022), have been widely adopted across domains including image (Rombach et al., 2022), video (Ho et al., 2022), audio (Kong et al., 2021), and graph tasks (Chamberlain et al., 2021), demonstrating remarkable capabilities in reproducing complex data distributions. As an alternative to diffusion models, *flow matching* has emerged as a promising direction, providing a more direct and computationally efficient framework for generative modeling (Lipman et al., 2023). Furthermore, flow matching approach has shown the potential to achieve high-quality sample generation with a significantly reduced number of function evaluations (Esser et al., 2024).

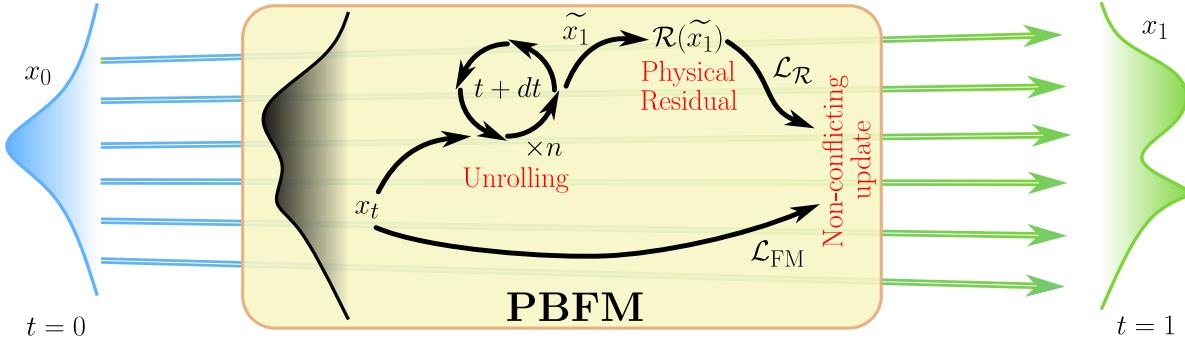


Figure 1: The *Physics-Based Flow Matching* (PBFM) framework: During training, the sample x_t at time t is evolved to $t = 1$ over n time steps to compute the residual $\mathcal{R}(\tilde{x}_1)$. The flow matching loss \mathcal{L}_{FM} and residual loss $\mathcal{L}_{\mathcal{R}}$ are combined in a conflict-free manner.

The key challenge in integrating physical constraints into flow matching and diffusion-based generative models is resolving the fundamental conflict between two objectives: adhering to physical laws and accurately modeling the data distribution. The *physical constraint term* drives the model toward solutions that satisfy governing equations, typically favoring low-variance, deterministic outputs. In contrast, the *generative objective* promotes diversity and coverage of the data distribution, encouraging the model to increase variability. These goals are often in conflict: enforcing strict physical consistency can restrict the model’s capacity to represent the data distribution, while maximizing expressiveness may lead to physically invalid samples.

We propose *Physics-Based Flow Matching* (PBFM), a generative framework that addresses these shortcomings: our method retains the advantage of satisfying PDE constraints, while also enabling the reproduction of complex data distributions. It explicitly incorporates physical knowledge into the learning process by embedding it into the probability distributions. Our approach leverages temporal unrolling during training to refine the final state from which the physical residual is computed. This makes it possible to fully leverage the accurate gradient information from the embedded physical models. Importantly, the substantial gains in accuracy are achieved without negatively affecting inference performance. To the best of our knowledge, this represents the first integration of flow matching in a framework that embeds physical principles with high accuracy. Figure 1 provides a visual overview of the key training stages.

The key contributions of our work can be summarized as follows: **(I)** We introduce a novel framework that leverages flow matching to incorporate physical constraints, minimizing both PDE and algebraic residuals in a non-conflicting manner. Through numerical experiments across diverse test cases, we demonstrate that our approach outperforms state-of-the-art methods without the need to tune hyperparameters for balancing the loss terms. **(II)** We investigate the impact of temporal unrolling during training, showing that it enhances final predictions and reduces residual errors, without increasing inference time. **(III)** We highlight the effects of added Gaussian noise in flow matching when imposing a physical constraint. **(IV)** We compare the commonly used deterministic flow matching sampler with a stochastic alternative, revealing the latter’s superior performance in reproducing complex distributions.

2 Methodology

2.1 Physical residuals

A general time-dependent PDE in n -spatial dimensions can be expressed as $\mathbf{u}_t(\mathbf{x}, t) = \mathcal{L}[\mathbf{u}(\mathbf{x}, t)] + \mathbf{f}(\mathbf{x}, t)$, $\mathbf{x} \in \Omega \subseteq \mathbb{R}^n$, where $\mathbf{u}(\mathbf{x}, t)$ denotes the unknown solution field, \mathcal{L} is a spatial differential operator, and $\mathbf{f}(\mathbf{x}, t)$ represents external forcing. To obtain numerical approximations, the spatial domain Ω , along with its boundary $\partial\Omega$, is typically discretized, and continuous operators are replaced by their discrete counterparts (Karniadakis and Sherwin, 2005).

Any formulation that can be reduced to a residual for minimization can, in principle, be used to enforce the underlying physics. Residuals can be divided into three macro categories. In the case of *steady-state* PDEs, where the distribution of solution data typically stems from uncertainties in the underlying physics, the residuals can be directly computed from the governing equations as $\mathcal{R}(\mathcal{L}(\mathbf{u}) + \mathbf{f}) = 0$. For *time-dependent* PDEs, where the learned solution evolves over time and consists of temporal snapshots, residuals are often formulated to enforce the conservation of physical quantities such as mass, momentum, or energy. Finally, *algebraic constraints* between different physical fields can also be incorporated as residuals, thereby further enforcing consistency with the underlying physics.

2.2 Physics-based flow matching

Integrating physical constraints into generative models often poses an inherently conflicting goal, leading to suboptimal distributional accuracy and trivially satisfied physical constraints. Existing, diffusion-based methods (Shu et al., 2023; Bastek et al., 2025) use the following optimization objective:

$$\arg \max_{\theta} \mathbb{E}_{x_1 \sim q(x_1)} [\log p_{\theta}(x_1)] + \mathbb{E}_{x_1 \sim p_{\theta}(x_1)} [\log q_{\mathcal{R}}(\hat{r} = 0 \mid x_1)] \quad (1)$$

where \hat{r} are virtual observables (Rixner and Koutsourelakis, 2021) of the residual $\mathcal{R}(x_1)$.

In the context of flow matching, the objective function to be minimized reduces to (Bastek et al., 2025):

$$\mathcal{L} = w_{\text{FM}} \mathcal{L}_{\text{FM}} + w_{\mathcal{R}} \mathcal{L}_{\mathcal{R}} = w_{\text{FM}} \|u_t^{\theta}(x_t, t) - u_t(x_t)\|_2 + w_{\mathcal{R}} \|\mathcal{R}(x_1(x_t, t))\|_2, \quad (2)$$

where w_{FM} and $w_{\mathcal{R}}$ are weights balancing the two losses. Specifically, increasing the weight of the physical residual term tends to degrade the quality of the learned distribution, while prioritizing the generative term undermines physical fidelity. Interestingly, multi-task learning has addressed the challenges of conflicting learning goals by formulating update directions based on gradients from the individual loss terms. We derive our method from the *ConFIG* method for computing conflict-free updates (Liu et al., 2025a), which in the context of physics-based flow matching yields:

$$\mathbf{g}_{\text{update}} = (\mathbf{g}_{\text{FM}}^\top \mathbf{g}_v + \mathbf{g}_{\mathcal{R}}^\top \mathbf{g}_v) \mathbf{g}_v \quad (3)$$

$$\mathbf{g}_v = \mathcal{U}[\mathcal{U}(\mathcal{O}(\mathbf{g}_{\text{FM}}, \mathbf{g}_{\mathcal{R}})) + \mathcal{U}(\mathcal{O}(\mathbf{g}_{\mathcal{R}}, \mathbf{g}_{\text{FM}}))] \quad (4)$$

where \mathbf{g}_{FM} and $\mathbf{g}_{\mathcal{R}}$ are the gradients for the flow matching loss \mathcal{L}_{FM} and residual loss $\mathcal{L}_{\mathcal{R}}$, respectively. $\mathcal{O}(\mathbf{g}_1, \mathbf{g}_2) = \mathbf{g}_2 - \frac{\mathbf{g}_1^\top \mathbf{g}_2}{\|\mathbf{g}_1\|^2} \mathbf{g}_1$ is the orthogonality operator and $\mathcal{U}(\mathbf{g}) = \mathbf{g}/\|\mathbf{g}\|$ is the unit vector operator. The obtained $\mathbf{g}_{\text{update}}$ satisfies $\mathbf{g}_{\text{update}}^\top \cdot \mathbf{g}_{\text{FM}} = \mathbf{g}_{\text{update}}^\top \cdot \mathbf{g}_{\mathcal{R}} > 0$, which ensures that the gradient descent optimization along $\mathbf{g}_{\text{update}}$ always minimizes both losses simultaneously.

Crucially, this solves the gradient-level conflict between the objectives: rather than allowing one loss to dominate or oppose the other, this formulation ensures that both objectives contribute constructively to each optimization step adaptively. This prevents the optimizer from getting trapped in regions where one loss is minimized at the expense of the other, a common failure mode in standard weighted objectives. Consequently, our method enables the model to approximate a distribution close to pure flow matching while simultaneously reducing the enforced residual. Furthermore, we jointly optimize both terms by aligning their gradients into a unified update direction without relying on manual hyperparameter tuning. A comprehensive analysis is provided in Appendix E.

2.3 Improving physical and distributional accuracy

An important requirement for generative models in the context of physics applications is the necessity to produce high-accuracy samples that are on-par with traditional methods. We identify the reconstruction of the final, noise-free, sample during training as a crucial aspect for accurately evaluating the physical residual. The integration of the intermediate sample at time t forward to the final time is performed with ODE integration (Lipman et al., 2024). Employing the predicted velocity field u_t^{θ} by integrating over a single time step of size $1 - t$. While trajectories are straight in theory, relying on just one integration step limits prediction accuracy, particularly when t is close to zero. To improve the final prediction, we incorporate unrolling (Monga et al., 2021), by performing n ODE integration steps of size $(1 - t)/n$. The integration over multiple time steps has an averaging effect, which naturally promotes trajectories that are closer to the theoretically straight paths. Unrolling is applied via a curriculum, increasing the number of unrolling steps over the course of training. We additionally down-weight less accurate predictions near $t = 0$ by applying a scaling factor of t^p , $p_{\text{opt}} = 1$ in the residual loss computation. A detailed analysis of different power laws is provided in Appendix D. The more accurate prediction of x_1 substantially improves the evaluation of the physics residual in eq. 2, which in turn yields a better learning direction.

A final aspect of key importance in physics based flow matching is the value of σ_{\min} , which denotes the amount of Gaussian noise added to the training data. In computer vision tasks, σ_{\min} is typically set to 10^{-3} (Lipman et al., 2023; Esser et al., 2024). However, the introduction of Gaussian noise perturbs the physical residuals, degrading the model’s performance. As a result, the amount of added noise establishes a lower bound on the residuals that the model can achieve, thereby directly influencing its ability to satisfy physical constraints. Thus, the σ_{\min} value has to be chosen carefully. We also found that $\sigma_{\min} = 0$ does not negatively affect the quality of the distribution and the underlying physics. Our explanation is that the numerical noise inherent to the physics residuals compensates for the lack of Gaussian mixing as required by flow matching theory.

Inspired by natural image generation (Esser et al., 2024), we additionally explore sampling the time variable t from a logit-normal distribution with zero mean and unit standard deviation during training, instead of the commonly used

Algorithm 1 Training step for Physics Based Flow Matching

```

 $n \leftarrow$  number of unrolling steps
 $dt \leftarrow (1 - t)/n$ 
 $\tilde{t} \leftarrow t$ 
 $u_t^\theta \leftarrow \text{model}(x_t, t)$ 
 $\tilde{x}_1 \leftarrow x_t + dt \cdot u_t^\theta$ 
for  $i = 1, i < n$  do
     $\tilde{t} = \tilde{t} + dt$ 
     $\tilde{u}_t^\theta \leftarrow \text{model}(\tilde{x}_1, \tilde{t})$ 
     $\tilde{x}_1 \leftarrow \tilde{x}_1 + dt \cdot \tilde{u}_t^\theta$ 
end for
 $\mathcal{R} \leftarrow \text{compute residual}(\tilde{x}_1)$ 
 $\mathcal{L}_{\mathcal{R}} \leftarrow \|t^p \cdot \mathcal{R}\|_2$ 
 $\mathcal{L}_{\text{FM}} \leftarrow \|u_t^\theta - u_t\|_2$ 
 $\nabla_\theta \leftarrow \text{compute } g_{\text{update}} \text{ via Eq. 3}$ 
AdamW optimizer step with  $\nabla_\theta$ 

```

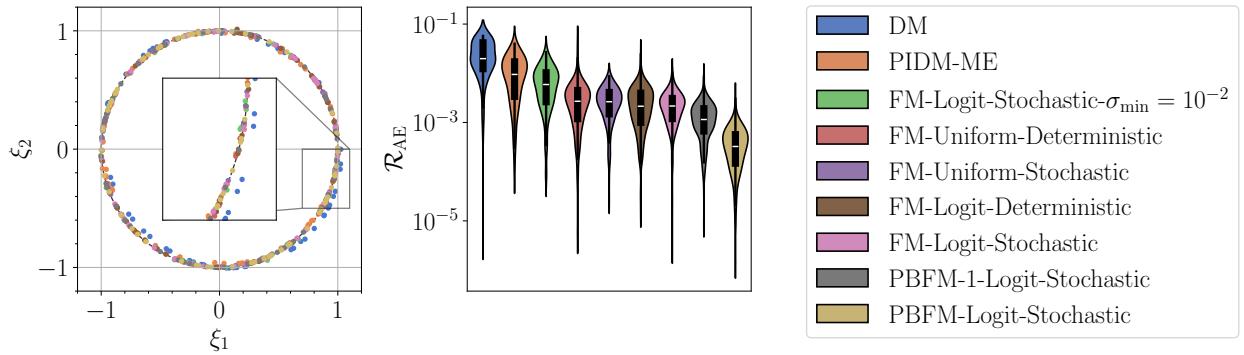


Figure 2: Point distribution and absolute error of the physical residual (circle radius squared) for SOTA reference DM, PIDM-ME (Bastek et al., 2025), and all proposed approaches.

uniform distribution. This choice is motivated by the observation that flow matching tends to exhibit higher errors around $t = 0.5$.

Algorithm 1 details the resulting training procedure. The FM loss is not affected by the unrolling and is computed for accurately evaluating the physics at time t . The initial time is stored to weight the residual loss.

At inference time, the final sample is typically obtained by evolving the initial noise, $\mathcal{N}(0, I)$ with ODE integration. This represents a deterministic sampler in which all the stochasticity is embedded in the initial noise sample (Gao et al., 2025). In contrast, the sampling process in DDPM is stochastic (Ho et al., 2020). Given that both diffusion models and flow matching can be seen generative modeling variants under arbitrary Markov processes (Holderrieth et al., 2025), we explore the use of a stochastic sampler within the physically based flow matching framework. The central idea is to evolve from time t to $t = 1$ and then return to $t + dt$ using a different noise sample. This step backwards in time with additional noise increases the stochasticity in the sampling process and improves distributional accuracy. The resulting procedure is outlined in Algorithm 2.

2.4 Ablation with a toy problem

To provide an intuition for the proposed improvements, we consider an ablation with a toy problem where the neural network outputs are the x and y coordinates of points on a circumference subject to the physical constraint of unit radius. Figure 2 shows the resulting point distribution along with the corresponding absolute physical error through mean estimation of the final sample. We compare with state-of-the-art models DM, as a DDPM representative (Song et al., 2022), and PIDM-ME, the best performing physics-informed algorithm (Bastek et al., 2025), and a PBFM-1 variant of our approach that omits unrolling. It is apparent that each extension yields a noticeable gain, and the final PBFM model exhibits a residual error that is 61.8 and 27.5 times lower than the DM and PIDM-ME baselines.

3 Experimental Setup

We evaluate our method on three benchmark problems. Across all experiments, we employ a diffusion transformer (DiT) backbone architecture (Peebles and Xie, 2023), with minor modifications detailed in Appendix F. For completeness, we also provide a comparison to the UNet from previous work (Ronneberger et al., 2015; Bastek et al., 2025), shown in Figure 11. The problems differ in difficulty and were chosen to cover the three types of physics residuals (steady-state, transient, and analytic) as outlined above. A complete description of datasets is provided in Appendix B.

Darcy flow We begin with the two-dimensional Darcy flow problem to compare with previous work Bastek et al. (2025). The Darcy equations which models steady-state fluid flow through a porous medium. The solution comprises pressure p and permeability K . We use the corresponding public dataset (Bastek, 2024), which contains 10k training and 1k validation samples, each of size 64×64 . It is worth noting that, while common in literature (Zhu and Zabaras, 2018; Jacobsen et al., 2025), this dataset lacks conditioning inputs, making it less representative of real-world application scenarios. The residual directly corresponds to the governing PDE:

$$\mathcal{R} = \nabla \cdot (K \nabla p) + f = 0$$

Kolmogorov flow The second benchmark is the two-dimensional Kolmogorov flow over a 128×128 spatial domain with periodic boundary conditions. The dataset is generated using a spectral solver and consists of velocity field snapshots for various Reynolds numbers in the range [100; 500]. The Reynolds number conditions the data generation, influencing the turbulence scales and flow complexity. The training set includes data for 32 Reynolds numbers with 1024 temporal snapshots each. The test set additionally includes 16 unseen Reynolds numbers. The data distribution reflects the temporal variation within each flow regime. The prediction target consists of the two velocity components, which are expected to satisfy the conservation of mass via:

$$\mathcal{R} = \nabla \cdot \mathbf{U} = 0$$

Dynamic Stall The final and most complex setup involves spatio-temporal fields over a pitching NACA0012 airfoil, capturing the effects of dynamic stall in compressible flows. Each sample consists of 128×128 distributions of absolute pressure, temperature, density, skin friction, and tangential velocity gradients. This physical model is highly relevant for real world cases such as the design and control of helicopter and wind turbine blades, where dynamic stall plays a critical role. Conditioning is performed on four parameters that define the operating conditions and airfoil motion. The training set comprises 128 base configurations, each perturbed 32 times to model uncertainty, while the validation set includes 16 unseen configurations. Physical consistency across fields is enforced through two analytical, point-wise residual constraints: \mathcal{R}_{ig} imposes the ideal gas law, while \mathcal{R}_τ minimizes the skin friction starting from the tangential velocity gradients and the dynamic viscosity computed with Sutherland's law.

$$\begin{aligned} \mathcal{R}_{ig} &= P - \rho RT \\ \mathcal{R}_\tau &= \tau_w - \mu_0 \frac{T_0 + S}{T + S} \left(\frac{T}{T_0} \right)^{\frac{3}{2}} \sqrt{\left(\frac{\partial u}{\partial x} \right)_{n=0}^2 + \left(\frac{\partial u}{\partial y} \right)_{n=0}^2} \end{aligned}$$

4 Results

The key findings of this study highlight the framework’s effectiveness in reducing physical residual errors while closely matching the data distribution. For the conditioned cases, the Kolmogorov flow and dynamic stall, we further assess the mean and standard deviation of each conditioned distribution.

4.1 Darcy flow

We benchmark our proposed framework, PBPM, against five alternative models: the FM baseline trained without incorporating physical constraints, the PBPM-1 model that includes the additional loss term but omits unrolling. Additionally, we compare with DM as DDPM baseline, the state-of-the-art PIDM-ME model Bastek et al. (2025), and the *CoCoGen* model of Jacobsen et al. (2025). The latter is an interesting variant, as it, contrary to the other algorithms, employs residual gradients during inference.

We begin by evaluating overall performance, focusing on residual mean absolute error (MAE), along with pressure and permeability distributions. The results are summarized in Figure 3. The first panel illustrates how residual error evolves

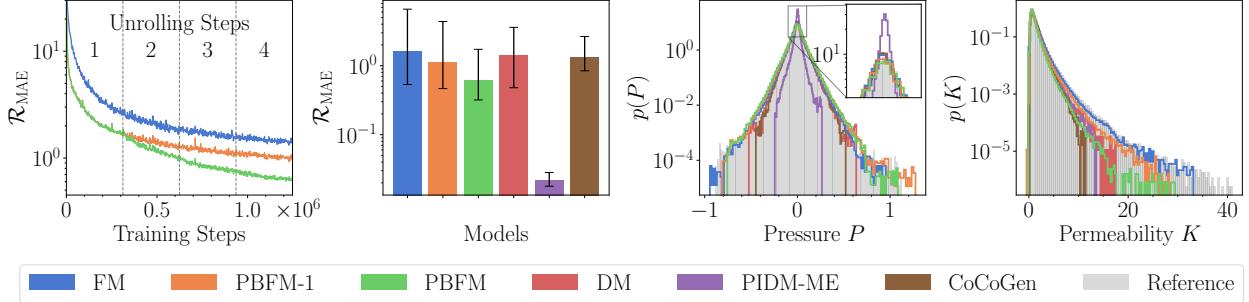


Figure 3: Darcy flow validation over 1024 samples using 20 FM steps. The left panel shows the residual MAE as a function of training steps, followed by visualizations of the residual error (error bars refer to min-max values within the validation dataset samples), pressure, and permeability distributions. Importantly, the low residual error of PIDM-ME is attributed to its partial coverage of pressure, and permeability, distributions (shown right). Both graphs highlight the high distributional accuracy of PBFM.

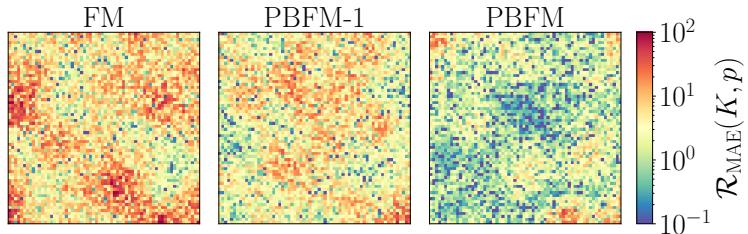


Figure 4: Physical residual of Darcy flow examples for the proposed method with 20 FM steps.

with the number of unrolling steps used during training. The bar plot shows that MAE decreases significantly from approximately 1.7 for FM to about 0.6 for PBFM. Notably, the improvement achieved by introducing the additional loss term is comparable in magnitude to the improvement gained via unrolling. When comparing across methods, we observe that the DM baseline performs similarly to FM. CoCoGen produces an $\approx 10\%$ lower residual error than FM but ca. 20% higher than PBFM-1, at the same time incurring a higher compute cost for inference. While PIDM-ME yields the lowest residual error, this performance comes at a steep cost: As shown in the last two panels, the residual scores for PIDM-ME are associated with a severely degraded pressure distribution, characterized by values constrained within the narrow range ± 0.2 . This artificially reduces the residual due to the lower magnitude of P , rather than improved fidelity. In comparison, all variants of our framework more faithfully reproduce the true pressure distribution. Regarding permeability, the DM baseline, as well as CoCoGen, exhibit a loss of distributional fidelity, particularly at the extrema. While our framework also shows a slight degradation for the least probable values, it successfully recovers the maximum K values around 30, which are approximately six orders of magnitude less frequent than the central mode of the distribution. Figure 4 visualizes the trend of decreasing residual error as the FM model is progressively improved.

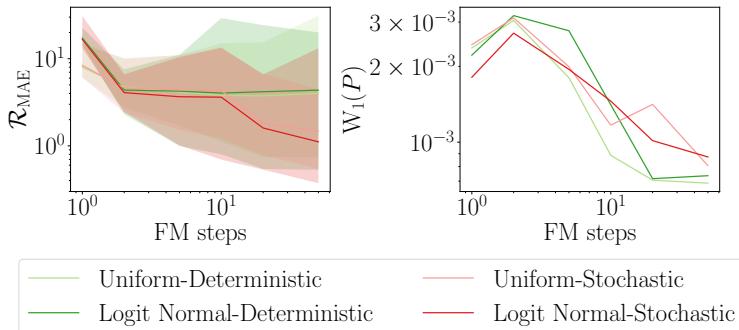


Figure 5: Darcy flow MAE residual (error bars refer to min-max values over validation set) and Wasserstein distance over FM steps for 1024 samples.

To complete the analysis on the Darcy dataset, we evaluated the PBFM-1 framework for sampling time t during training with the logit-normal distribution, comparing it to the standard uniform distribution. We also analyze the effect of different samplers: deterministic and stochastic. Figure 5 reports the residual error along with the Wasserstein distance (Panaretos and Zemel, 2019; SciPy) as a function of the number of FM steps. The use of the logit-normal distribution does not appear to offer any particular advantage over the uniform distribution, in terms of either residual error or distribution. In contrast, switching to a stochastic sampler reduces the residual error if the number of function evaluations exceeds 10. However, this improvement diminishes when the number of FM steps becomes too large, as the distributional quality worsens. Particularly, with 20 integration steps, the proposed sampler reduces the residual error by a significant factor of $3\times$, while the Wasserstein distance is only 30% lower for the deterministic sampler. Further results showing pressure and permeability samples and an additional UNet are reported in Figure 10 and 11 of the appendix.

4.2 Kolmogorov flow

Kolmogorov flow differs from the Darcy case as it is conditioned on the Reynolds number, and in this case the physical residual only matches parts of the underlying PDE. Here, the PBFM model reduces the average residual MSE by a factor of $1.68\times$ compared to the baseline, while the PBFM-1 approach without unrolling offers only limited improvement of $1.05\times$, as shown Figure 6. This difference is clearly reflected in the panel on the left, which visually depict the residual MSE. When examining the velocity distributions, all FM variants yield nearly overlapping profiles, although none fully capture the extrema of the reference data. In the appendix, Figure 13 and 14 report the instantaneous velocity fields along the divergence and an example of mean and standard deviation, respectively. This test case highlights the importance of accurate samples via unrolling to fully leverage the physics residuals at training time.

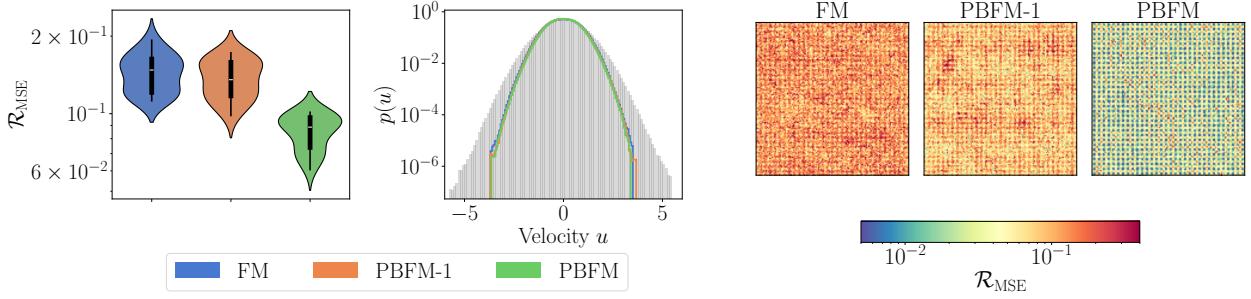


Figure 6: Kolmogorov flow residual and distribution comparison over 128 samples using 20 FM steps. Examples of residual MSE for the three analyzed approaches. Violin plot distributions refer to the different conditioning values within the validation dataset.

4.3 Dynamic Stall

The dynamic stall case presents additional challenges, including two physical constraints of increased complexity and six predicted fields. The underlying phenomena are highly non-linear and governed by a system of four time-dependent PDEs.

Figure 7 summarizes the key results, including the residual MSE, Wasserstein distance. Focusing first on the residual error, we observe a consistent improvement when progressing from the FM baseline to PBFM, with the PBFM-1 providing a $2\times$ reduction compared to the baseline. The incorporation of residual loss and unrolling not only reduces the average error of a $8\times$ factor but also sharpens the residual error distribution, concentrating values closer to the mean. In terms of data distribution, the Wasserstein distance reveals that while the PBFM-1 framework alone may degrade distributional accuracy, the addition of unrolling effectively regularizes the predictions, narrowing the gap with the ground truth distribution. It is worth noting that MSEs of the predicted mean and standard deviation are on par for all methods (details are given in the appendix), indicating that the improvements of PBFM in residual and distributional accuracy are achieved without compromising first- and second-order statistics.

Figure 8 presents a visual comparison: the left panel illustrates the mean and standard deviation of the pressure field P and wall shear stress τ_w , while the right panel shows the residual error magnitude. PBFM accurately captures both the average behavior and the associated fluctuations, maintaining consistency with the minimum and maximum bounds of each variable. Regarding residuals, the PBFM-1 model demonstrates a reduction in the average error relative to the baseline. However, it is worth noting that the high-error regions do not show significant improvement. In contrast, incorporating the unrolling strategy leads to a notable reduction even in zones where the error is largest.

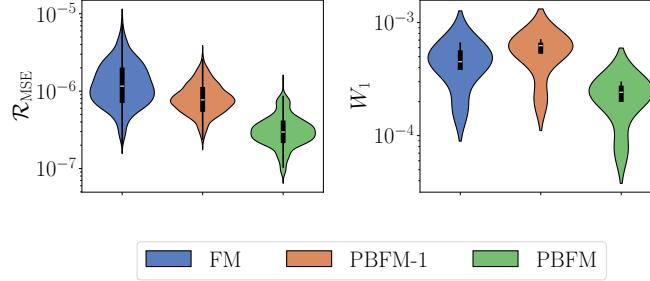


Figure 7: Comparison of the proposed approaches of physical residual MSE, Wasserstein distance, for dynamic stall. Each condition includes 128 samples with 20 FM steps. Violin plots show distributions over the validation dataset.

Performance At training time, evaluating physical residuals and unrolling both lead to slight increases in computational load. Importantly, the increased accuracy of PBFM does not negatively affect performance at inference time. While reliably comparing computational performance is difficult, the dynamic stall scenario roughly indicates benefits of the proposed generative model: the CPU-based data generating solver takes 76 minutes on average for a single simulation, while the trained PBFM model requires only 4 minutes to generate 128 samples on the same CPU architecture and 0.2 seconds on an A100 GPU. Detailed measurements are provided in Table 4.

5 Related work

Numerous deep learning methods have been developed to address complex problems in physics and engineering (Morton et al., 2018; Wang et al., 2020; Sanchez-Gonzalez et al., 2020; Thuerey et al., 2020). More recently, there has been growing interest in bringing the foundation model paradigm to scientific machine learning. Efforts such as PDE-former (Ye et al., 2024), Poseidon (Herde et al., 2024), Aurora (Bodnar et al., 2024), and Unisolver (Zhou et al., 2025) aim to build models with broad generalization capabilities across diverse physical systems. In parallel, specialized transformer architectures tailored for PDEs have also been proposed, including OFormer (Li et al., 2023), Transolver (Wu et al., 2024), Fengbo (Pepe et al., 2025), and PDE-Trasformer (Holzschuh et al., 2025).

Focusing on diffusion models in the physics and engineering fields, representative applications include the generation and design of new molecules and drugs (Guo et al., 2024; Schneuing et al., 2024; Bose et al., 2024), the simulation of particle trajectories in collider experiments (Mikuni et al., 2023), solving inverse PDE-problems (Holzschuh et al., 2023), and the modeling of particle motion in turbulent flows (Li et al., 2024). Alongside these efforts, several works have sought to embed physical knowledge or constraints directly into the diffusion process to enhance model performance. For instance, Huang et al. (2024) presented a framework to solve PDEs from partial observations by

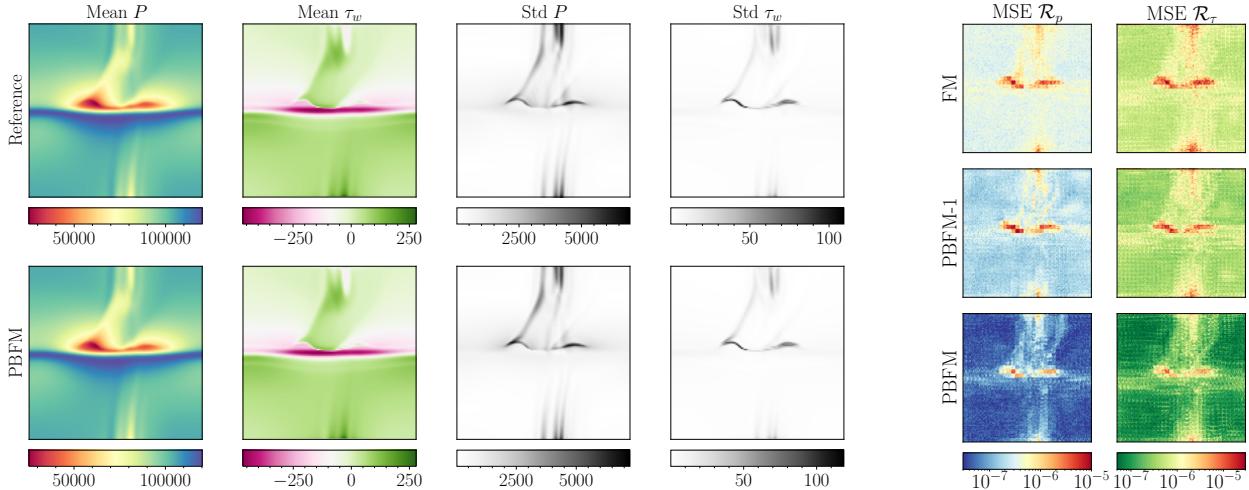


Figure 8: Dynamic stall example comparison of mean and standard deviation for PBFM framework computed over 128 samples using 20 FM steps. Examples of residual MSE for the three analyzed approaches.

filling in missing information using generative priors. Rixner and Koutsourelakis (2021) formulated a probabilistic generative model enforcing physical constraints through virtual observables.

More aligned with our work, Shu et al. (2023) proposed a framework that conditions the diffusion process on the residual gradient both at training and inference time, with a focus on super-resolution and reconstruction tasks based on random measures of turbulent flows. Despite being one of the first work in this direction, they do not directly enforce the residual. Similarly, Jacobsen et al. (2025) introduced CoCoGen, a method that employs the governing PDEs during inference, while leaving the training procedure unchanged. The approach improves the physical residual but slow down inference which is crucial in real applications. Giannone et al. (2023) demonstrated that aligning the sampling trajectory of diffusion models with trajectories obtained from physics-based iterative optimization methods can improve the quality of generated designs; however, their approach is limited to topology optimization problems rather than general PDEs. Another topology optimization work that enforces physical constraint is proposed by Mazé and Ahmed (2023). They included in the sampling process a surrogate neural network predicting and enforcing the compliance of the topology under given constraints. Finally, Basteck et al. (2025) proposed physics-based diffusion models (PIDM), which incorporate an additional term during diffusion training to minimize physical residuals. Despite their formal derivation of the objective, no concrete solution is proposed to address the inherent conflict between the generative and physical objectives.

6 Conclusions

In this paper, we introduce PBFM, a novel generative model designed to improve physical consistency while preserving the strengths of flow matching approach for high-dimensional data generation. Our framework provides a principled mechanism to minimize physical residuals, arising from PDEs or algebraic constraints, in a conflict-free manner. We conduct extensive benchmarks across three representative physical systems, demonstrating the versatility and robustness of our approach. Notably, PBFM achieves up to an $8\times$ reduction in physical residuals while accurately reproducing target data distributions. The incorporation of temporal unrolling plays a key role, enabling improved final state approximations and mitigating localized residual peaks in challenging regions of the solution space.

Additionally, our results highlight the benefits of stochastic sampling strategies, which outperform deterministic methods in cases involving complex target distributions. Beyond these specific benchmarks, the proposed framework generalizes to a wide class of PDE-constrained problems formulated via residual minimization. By combining the computational efficiency of flow matching with the interpretability and rigor of physics-based modeling, PBFM offers a powerful and flexible tool for surrogate modeling, uncertainty quantification, and simulation acceleration in physics and engineering applications.

References

- M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi:10.1016/j.inffus.2021.05.008. URL <https://www.sciencedirect.com/science/article/pii/S1566253521001081>.
- ANSYS. Fluent 2024R2, 2024. URL <https://www.ansys.com/products/fluids/ansys-fluent>.
- M. Baldan, G. Baldan, and B. Nacke. Solving 1D non-linear magneto quasi-static Maxwell's equations using neural networks. *IET Science, Measurement & Technology*, 15(2):204–217, 2021. doi:10.1049/smt2.12022.
- J.-H. Basteck. Physics-Informed Diffusion Models. Datasets and model checkpoints - CC BY 4.0. ETH Zurich, 2024. doi:10.3929/ethz-b-000674074.
- J.-H. Basteck, W. Sun, and D. Kochmann. Physics-Informed Diffusion Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tpYeermigp>.
- C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, J. Brandstetter, P. Garvan, M. Riechert, J. Weyn, H. Dong, A. Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024.
- J. Bose, T. Akhound-Sadegh, G. Huguet, K. FATRAS, J. Rector-Brooks, C.-H. Liu, A. C. Nica, M. Korablyov, M. M. Bronstein, and A. Tong. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=kJFIH23hXb>.
- S. L. Brunton and J. N. Kutz. Promising directions of machine learning for partial differential equations. *Nature Computational Science*, 4(7):483–494, Jul 2024. ISSN 2662-8457. doi:10.1038/s43588-024-00643-2.

- B. Chamberlain, J. Rowbottom, M. I. Gorinova, M. Bronstein, S. Webb, and E. Rossi. GRAND: Graph Neural Diffusion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1407–1418. PMLR, 18–24 Jul 2021.
- W. Chen, Q. Wang, J. S. Hesthaven, and C. Zhang. Physics-informed machine learning for reduced-order modeling of nonlinear problems. *Journal of Computational Physics*, 446:110666, 2021. ISSN 0021-9991. doi:10.1016/j.jcp.2021.110666.
- P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024. URL <https://arxiv.org/abs/2403.03206>.
- L. C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010. ISBN 9780821849743 0821849743.
- S. Fresca, L. Dede’, and A. Manzoni. A Comprehensive Deep Learning-Based Approach to Reduced Order Modeling of Nonlinear Time-Dependent Parametrized PDEs. *Journal of Scientific Computing*, 87(2):61, Apr 2021. ISSN 1573-7691. doi:10.1007/s10915-021-01462-7.
- R. Gao, E. Hoogeboom, J. Heek, V. D. Bortoli, K. P. Murphy, and T. Salimans. Diffusion models and gaussian flow matching: Two sides of the same coin. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL <https://openreview.net/forum?id=C8Yyg9wy0s>.
- G. Giannone, A. Srivastava, O. Winther, and F. Ahmed. Aligning Optimization Trajectories with Diffusion Models for Constrained Design Generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=KTR33hMnMX>.
- Z. Guo, J. Liu, Y. Wang, M. Chen, D. Wang, D. Xu, and J. Cheng. Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*, 2(2):136–154, Feb 2024. ISSN 2731-6092. doi:10.1038/s44222-023-00114-9.
- E. Haber, L. Ruthotto, E. Holtham, and S.-H. Jun. Learning across scales—multiscale methods for convolution neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- M. Herde, B. Raonic, T. Rohner, R. Käppeli, R. Molinaro, E. de Bezenac, and S. Mishra. Poseidon: Efficient Foundation Models for PDEs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=JC1VKK3UXk>.
- J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models, 2020. URL <https://arxiv.org/abs/2006.11239>.
- J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc., 2022.
- P. Holderrieth, M. Havasi, J. Yim, N. Shaul, I. Gat, T. Jaakkola, B. Karrer, R. T. Q. Chen, and Y. Lipman. Generator Matching: Generative modeling with arbitrary Markov processes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=RuP17cJtZo>.
- B. Holzschuh, S. Vegetti, and N. Thuerey. Solving inverse physics problems with score matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.
- B. Holzschuh, Q. Liu, G. Kohl, and N. Thuerey. PDE-Transformer: Efficient and Versatile Transformers for Physics Simulations. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=3BaJMRaPSx>.
- J. Huang, G. Yang, Z. Wang, and J. J. Park. DiffusionPDE: Generative PDE-solving under partial observation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=z0I2SbjNOR>.
- C. Jacobsen, Y. Zhuang, and K. Duraisamy. CoCoGen: Physically Consistent and Conditioned Score-Based Generative Models for Forward and Inverse Problems. *SIAM Journal on Scientific Computing*, 47(2):C399–C425, 2025. doi:10.1137/24M1636071.
- G. Karniadakis and S. Sherwin. *Spectral/hp Element Methods for Computational Fluid Dynamics*. Oxford University Press, 06 2005. ISBN 9780198528692. doi:10.1093/acprof:oso/9780198528692.001.0001.
- L. Kocić and W. J. Whiten. Computational investigations of low-discrepancy sequences. *ACM Trans. Math. Softw.*, 23(2):266–294, jun 1997. ISSN 0098-3500. doi:10.1145/264029.264064.
- Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=a-xFK8Ymz5J>.

- T. Li, L. Biferale, F. Bonaccorso, M. A. Scarpolini, and M. Buzzicotti. Synthetic lagrangian turbulence by generative diffusion models. *Nature Machine Intelligence*, 6(4):393–403, Apr 2024. ISSN 2522-5839. doi:10.1038/s42256-024-00810-0.
- Z. Li, K. Meidani, and A. B. Farimani. Transformer for partial differential equations’ operator learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=EPPqt3uERT>.
- Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=PqvMRDCJT9t>.
- Y. Lipman, M. Havasi, P. Holderrieth, N. Shaul, M. Le, B. Karrer, R. T. Q. Chen, D. Lopez-Paz, H. Ben-Hamu, and I. Gat. Flow matching guide and code, 2024. URL <https://arxiv.org/abs/2412.06264>.
- Q. Liu and N. Thuerey. Uncertainty-Aware Surrogate Models for Airfoil Flow Simulations with Denoising Diffusion Probabilistic Models. *AIAA Journal*, 62(8):2912–2933, 2024. doi:10.2514/1.J063440.
- Q. Liu, M. Chu, and N. Thuerey. ConFIG: Towards Conflict-free Training of Physics Informed Neural Networks. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://arxiv.org/abs/2408.11104>.
- Q. Liu, F. Köhler, and N. Thuerey. TorchFSM: Fourier Spectral Method with PyTorch, May 2025b. URL <https://doi.org/10.5281/zenodo.15350210>.
- F. Mazé and F. Ahmed. Diffusion models beat gans on topology optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9108–9116, 2023. doi:10.1609/aaai.v37i8.26093. URL <https://ojs.aaai.org/index.php/AAAI/article/view/26093>.
- V. Mikuni, B. Nachman, and M. Pettee. Fast point cloud generation with diffusion models in high energy physics. *Phys. Rev. D*, 108:036025, Aug 2023. doi:10.1103/PhysRevD.108.036025. URL <https://link.aps.org/doi/10.1103/PhysRevD.108.036025>.
- V. Monga, Y. Li, and Y. C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. doi:10.1109/MSP.2020.3016905.
- J. Morton, A. Jameson, M. J. Kochenderfer, and F. Witherden. Deep dynamical modeling and control of unsteady fluid flows. In *Advances in Neural Information Processing Systems*, 2018.
- A. Q. Nichol and P. Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021. URL <https://proceedings.mlr.press/v139/nichol21a.html>.
- V. M. Panaretos and Y. Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019. ISSN 2326-831X. doi:10.1146/annurev-statistics-030718-104938. URL <https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-030718-104938>.
- W. Peebles and S. Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023. doi:10.1109/ICCV51070.2023.00387.
- A. Pepe, M. Montanari, and J. Lasenby. Fengbo: a Clifford Neural Operator pipeline for 3D PDEs in Computational Fluid Dynamics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=VsxbWTDHjh>.
- M. Raissi, P. Perdikaris, and G. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi:10.1016/j.jcp.2018.10.045.
- M. Rixner and P.-S. Koutsourelakis. A probabilistic generative model for semi-supervised training of coarse-grained surrogates and enforcing physical constraints through virtual observables. *Journal of Computational Physics*, 434:110218, 2021. ISSN 0021-9991. doi:10.1016/j.jcp.2021.110218. URL <https://www.sciencedirect.com/science/article/pii/S0021999121001133>.
- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

- C. J. Roy and W. L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011. ISSN 0045-7825. doi:10.1016/j.cma.2011.03.016. URL <https://www.sciencedirect.com/science/article/pii/S0045782511001290>.
- A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. W. Battaglia. Learning to Simulate Complex Physics with Graph Networks, 2020. URL <https://arxiv.org/abs/2002.09405>.
- A. Schneuing, C. Harris, Y. Du, K. Didi, A. Jamasb, I. Igashov, W. Du, C. Gomes, T. L. Blundell, P. Lio, M. Welling, M. Bronstein, and B. Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, Dec 2024. ISSN 2662-8457. doi:10.1038/s43588-024-00737-x.
- SciPy. Wasserstein distance. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html#scipy.stats.wasserstein_distance.
- D. Shu, Z. Li, and A. Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023. ISSN 0021-9991. doi:10.1016/j.jcp.2023.111972. URL <https://www.sciencedirect.com/science/article/pii/S0021999123000670>.
- J. Song, C. Meng, and S. Ermon. Denoising Diffusion Implicit Models, 2022. URL <https://arxiv.org/abs/2010.02502>.
- N. Thuerey, K. Weißenow, L. Prantl, and X. Hu. Deep learning methods for reynolds-averaged navier–stokes simulations of airfoil flows. *AIAA Journal*, 58(1):25–36, 2020.
- A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=CD9Snc73AW>.
- M. L. Valencia, T. Pfaff, and N. Thuerey. Learning distributions of complex fluid simulations with diffusion graph networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=uKZdlihDDn>.
- S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-Attention with Linear Complexity, 2020. URL <https://arxiv.org/abs/2006.04768>.
- H. Wu, H. Luo, H. Wang, J. Wang, and M. Long. Transolver: A fast transformer solver for PDEs on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.
- Z. Ye, X. Huang, L. Chen, H. Liu, Z. Wang, and B. Dong. PDEformer: Towards a Foundation Model for One-Dimensional Partial Differential Equations. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024. URL <https://openreview.net/forum?id=GLDMCwdhTK>.
- H. Zhou, Y. Ma, H. Wu, H. Wang, and M. Long. Unisolver: PDE-Conditional Transformers Are Universal Neural PDE Solvers. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL <https://openreview.net/forum?id=6H1gUqkjmw>.
- Y. Zhu and N. Zabaras. Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, 366:415–447, 2018. ISSN 0021-9991. doi:<https://doi.org/10.1016/j.jcp.2018.04.018>. URL <https://www.sciencedirect.com/science/article/pii/S0021999118302341>.

A Flow Matching

Flow matching has recently gained attention as a compelling alternative to diffusion models, offering a more direct and computationally efficient framework for generative modeling (Lipman et al., 2023). It enables high-quality sample generation with significantly fewer function evaluations (Esser et al., 2024). Given a known source distribution p and an unknown target distribution q , flow matching learns a vector field u_t^θ , parameterized by a neural network, that generates a probability path p_t interpolating from $p_0 = p$ to $p_1 = q$ (Lipman et al., 2024). The learning objective is defined as:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1], x \sim p_t} \|u_t^\theta(x) - u_t(x)\|_2$$

where θ denotes the model parameters. While multiple formulations exist for the target velocity field u_t , a particularly simple and effective one leverages optimal transport (OT) (Tong et al., 2024). In this setting, samples from the base distribution $p_0 = \mathcal{N}(0, I)$ are linearly transported to p_1 via the conditional flow:

$$u_t(x | x_1) = \frac{x_1 - (1 - \sigma_{\min})x}{1 - (1 - \sigma_{\min})t},$$

with $\sigma_{\min} \sim 10^{-3}$, and the corresponding interpolant:

$$\psi_t(x) = (1 - (1 - \sigma_{\min})t)x + tx_1.$$

This yields a straight-line conditional flow with a time-independent vector field. Sampling from the trained model requires integrating the learned field over time:

$$x_1 = \int_0^1 u_t^\theta(x_t) dt,$$

typically using numerical ODE solvers such as Euler's method. Although the true OT vector field is constant, the learned approximation typically is not, and integration quality still depends on the time discretization.

B Dataset generation and residual computation

We provide a detailed description of the datasets used and generated, as well as the method employed to compute residuals during training.

B.1 Darcy flow

We use the dataset introduced by [Bastek \(2024\)](#) to enable direct comparison with their results. For completeness, we briefly summarize the key characteristics of the dataset. The underlying physical model is governed by the steady-state Darcy equations, which describe fluid flow through a porous medium:

$$\begin{aligned} \mathbf{u}(\mathbf{x}) &= -K(\mathbf{x})\nabla p(\mathbf{x}), \quad \mathbf{x} \in \Omega \\ \nabla \cdot \mathbf{u}(\mathbf{x}) &= f(\mathbf{x}), \quad \mathbf{x} \in \Omega \\ \mathbf{u}_{\hat{n}(\mathbf{x})} &= 0, \quad \mathbf{x} \in \partial\Omega \\ \int_{\Omega} p(\mathbf{x}) d\mathbf{x} &= 0 \end{aligned}$$

Here, $\hat{n}(\mathbf{x})$ denotes the unit outward normal on the boundary $\partial\Omega$. The source term $f(\mathbf{x})$ is defined as:

$$f(\mathbf{x}) = \begin{cases} r & \text{if } |x_i - \frac{1}{2}w| \leq \frac{1}{2}w \\ -r & \text{if } |x_i - 1 + \frac{1}{2}w| \leq \frac{1}{2}w \\ 0 & \text{otherwise} \end{cases}$$

with $r = 10$ and $w = 0.125$. The permeability field $K(\mathbf{x})$ is modeled as $K(\mathbf{x}) = \exp(G(\mathbf{x}))$, where $G(\mathbf{x})$ is a Gaussian random field. The pressure field is generated by solving a least-squares problem based on a 64-term truncated Karhunen–Loëve expansion, following the approach of [Jacobsen et al. \(2025\)](#).

To ensure consistency and avoid introducing numerical discrepancies, we adopt the same procedure for computing residuals during training as was used during dataset generation. Specifically, we employ identical finite difference stencils implemented via 2D convolutional layers, and apply the same reconstruction method for the forcing term f . The pressure field is also normalized by removing the integral contribution.

B.2 Kolmogorov flow

We generate two distinct datasets (training and validation) for the Kolmogorov flow problem with Reynolds numbers in the range [100, 500], using a spatial resolution of 128×128 . The simulation is based on the vorticity–stream function formulation. The velocity field is obtained from the computed vorticity through the stream function, while the pressure field is derived by solving the pressure Poisson equation. We employ TorchFSM ([Liu et al., 2025b](#)) to perform GPU-accelerated flow simulations using the Fourier spectral method. The training dataset includes 32 different flow conditions sampled via a Halton sequence ([Kocis and Whiten, 1997](#)), while the validation dataset contains 16 conditions. For each condition, 256 simulations are conducted with slightly perturbed initial states. Simulations are run for 10 000 time steps to reach a statistically steady state, followed by data sampling every 4 000 time steps. With a time step size of $dt = 1/\text{Re}$, this yields 1 024 snapshots per condition.

To ensure consistency, the divergence of the velocity fields is computed using the same numerical scheme as the spectral solver employed for dataset generation.

B.3 Dynamic stall

The dynamic stall datasets used for training and validation are generated by solving the unsteady, compressible, two-dimensional RANS equations around a sinusoidally pitching NACA0012 airfoil. An O-grid mesh is utilized, featuring 512 nodes along the airfoil surface and 128 nodes in the normal direction. Simulations are performed using ANSYS Fluent 2024R2 (ANSYS, 2024). The governing equations are discretized using a second-order upwind scheme for spatial accuracy and a second-order implicit scheme for time integration. Gradient reconstruction employs a least-squares cell-based method, while fluxes are computed using the Rhie-Chow momentum interpolation. Pressure-velocity coupling is handled through a coupled solver. Airfoil pitching is modeled by prescribing a rigid-body motion to the entire mesh, defined by the angular velocity function $\dot{\alpha}(t) = \omega \alpha_s \cos(\omega t)$. The mean angle of attack, α_0 , is applied at the start of each simulation by rotating the mesh accordingly. To close the RANS equations, the SST turbulence model with an intermittency transport equation is employed. Each oscillation cycle is resolved using 2048 time steps, and simulations are run until periodic convergence is achieved.

Table 1: Range of conditioning inputs that define the operating conditions of the pitching airfoil.

Variable	Min	Max
Mach	0.3	0.5
α_0	5°	10°
α_s	5°	10°
k	0.05	0.1

The design space is defined as a four-dimensional hypercube, with each axis corresponding to a conditioning input for the neural network. These inputs include the free-stream Mach number, the mean angle of attack α_0 , the pitching amplitude α_s , and the reduced frequency $k = \omega c/2V_\infty$. The ranges for each variable are provided in Table 1. Training and testing datasets are constructed using Halton sequences. The hypercube is sampled with 128 points for training and 16 points for testing. Each sampled point represents a nominal operating condition. Each nominal condition is perturbed as follows:

$$x_{\text{perturbed}} = (1 + \mathcal{N}(0, 0.02)) x_{\text{nominal}}$$

where $\mathcal{N}(0, 0.02)$ denotes a Gaussian noise term with zero mean and standard deviation 0.02. This results in 32 perturbed variations per nominal condition, yielding a total of 128×32 simulations for training and 16×32 for testing. To reduce computational costs, all simulation fields are downsampled to a resolution of 128×128 . The saved quantities include fields of absolute pressure, density, temperature, signed skin friction, and tangential velocity gradients across the airfoil surface over a full pitching cycle. Figure 9 illustrates a spatio-temporal representation of the wall shear stress (τ_w) across one complete pitching cycle, highlighting the pitch-up and pitch-down phases and demonstrating the structured mapping of surface data over time.

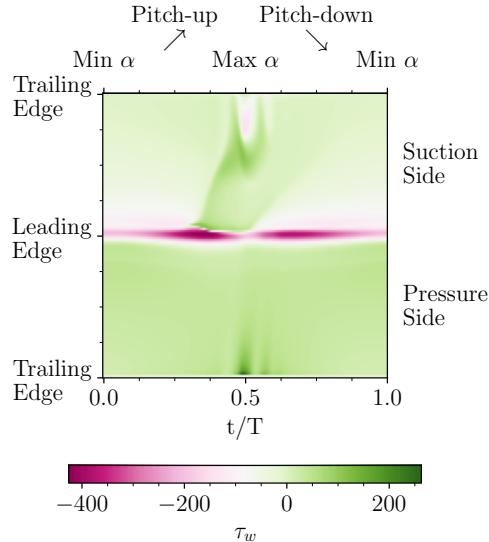


Figure 9: Example of a spatio-temporal contour of the post-processed τ_w distribution over an entire pitching cycle.

C Additional samples and analysis

We conclude the analysis of the proposed test cases by presenting additional comparisons and complete field prediction examples. We compute the 1D Wasserstein-1 distance to evaluate distributional differences. This approach treats each pixel independently rather than modeling the full 2D spatial structure.

C.1 Darcy flow

Figure 10 shows the pressure and permeability fields for the different methods under analysis. To ensure a fair comparison of the residuals, all fields have similar variable ranges, since higher magnitudes can lead to disproportionately large residual values. This choice is further justified by the residual plots, which exhibit higher absolute errors in regions with large permeability. Despite local differences, a clear trend emerges: starting from the FM baseline, where residual peaks reach values around 100, to the PBFM method, where peak residuals are reduced by approximately an order of magnitude.

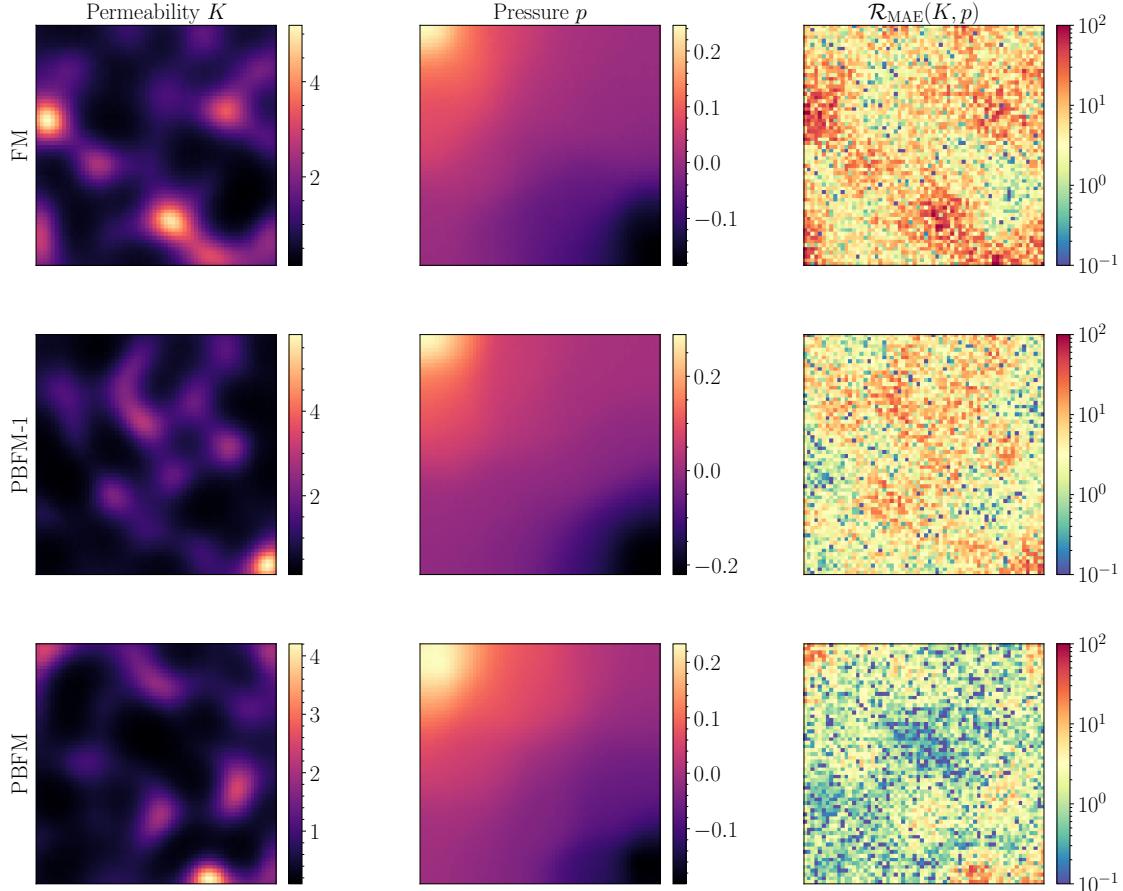


Figure 10: Darcy flow examples of pressure and permeability fields along with the physical residual for the proposed approaches.

To conclude our analysis of the Darcy flow, Figure 11 compares the DiT architecture with the UNet used by [Bastek et al. \(2025\)](#). This comparison corresponds to the FM setup without residual loss. In the first panel, which reports the FM loss during training, we observe that the UNet architecture reproduces the overfitting behavior noted by [Bastek et al. \(2025\)](#). In contrast, the DiT architecture exhibits a stable and monotonic reduction in training loss. The second panel shows the residual error, which remains comparable between the two models overall, with the exception of slight discrepancies at one function evaluation. Finally, in the third panel, the Wasserstein distance reveals a clear gap in performance: the UNet consistently incurs higher errors, with the distance being approximately twice as large for pressure and up to four times larger for permeability, underscoring the superior distributional accuracy of the DiT model. Additional samples produced by our method are shown in Figure 12.

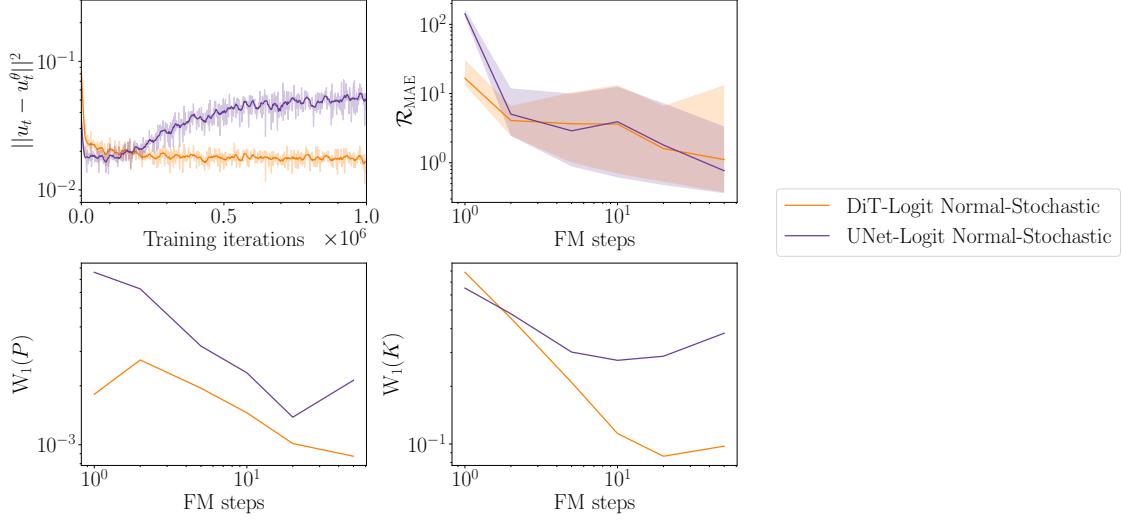


Figure 11: Comparison of UNet (Bastek et al., 2025) and DiT architectures for Darcy flow. The physical residual (error bars refer to min-max values within the validation dataset samples) and Wasserstein distance, as a function of FM steps, are computed over 1024 samples.

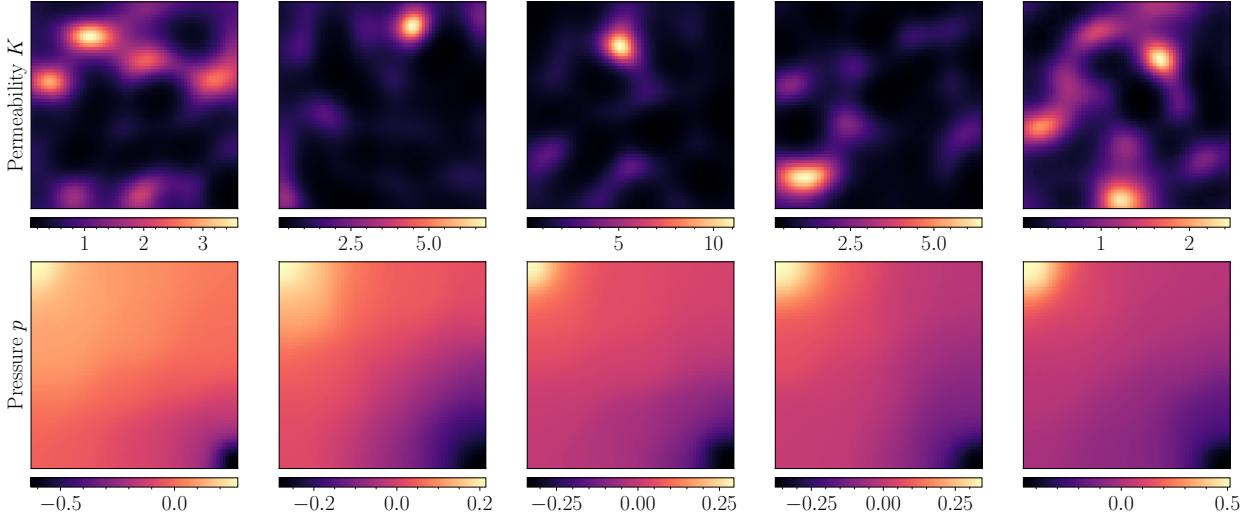


Figure 12: Representative Darcy flow field samples generated with PBPFM using 20 FM steps. The samples illustrate the diversity in flow behavior, particularly highlighting variability in the range extrema.

C.2 Kolmogorov flow

Figure 13 illustrates example predictions from the FM, PBPFM-1, and PBPFM models, along with their corresponding physical residuals, representing the divergence of the predicted fields. The baseline FM exhibits a residual MSE on the order of 10^{-1} across most of the domain, whereas the unrolled framework significantly reduces the error, dropping below 10^{-3} in large regions of the field, despite some localized areas with higher residuals.

Furthermore, Figure 14 shows the mean and standard deviation of the predicted fields for all models, alongside the reference data. All frameworks closely match the reference mean, with the unrolled version providing a smoother, less oscillatory solution compared to FM and PBPFM-1. Regarding the standard deviation, none of the models fully capture the reference distribution, particularly missing the peak value of approximately 0.95, although their performance remains broadly comparable.

Additional samples produced by PBPFM method are shown in Figure 15.

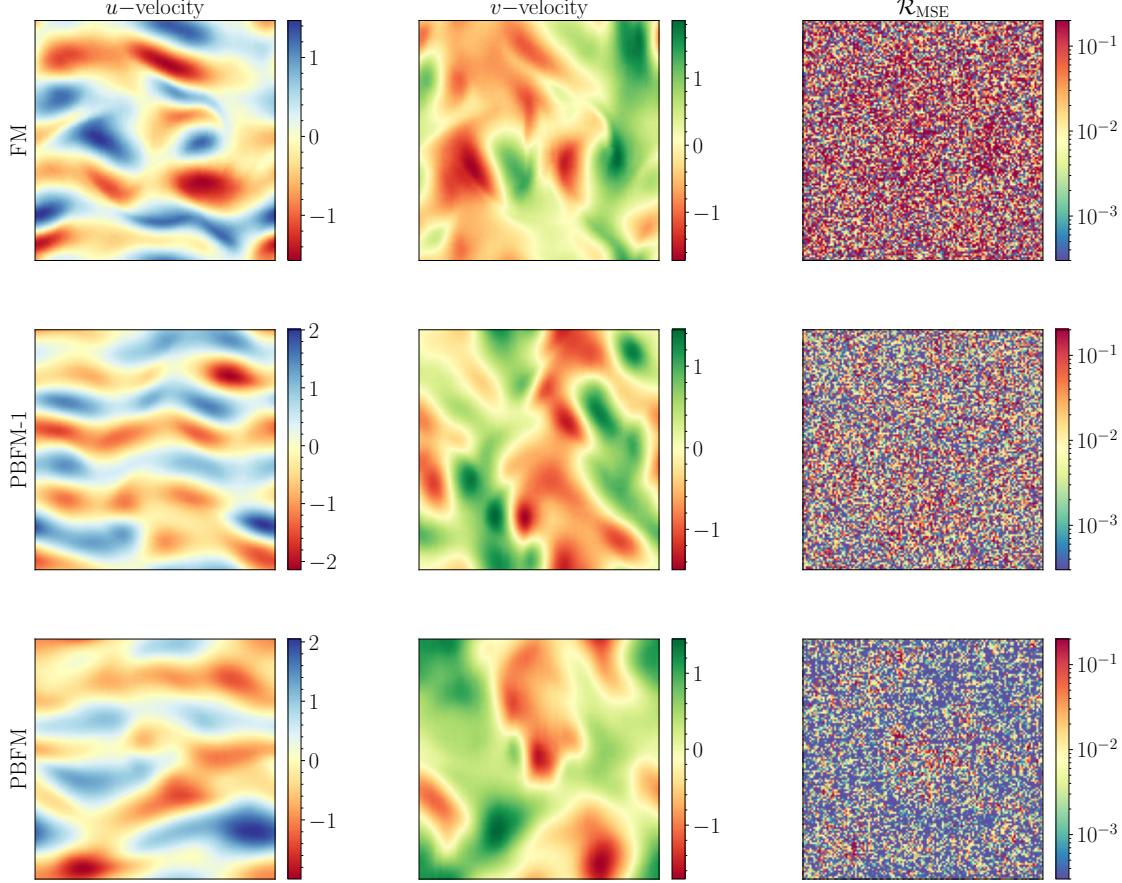


Figure 13: Kolmogorov flow example of u - and v -velocity fields and physical residual, divergence, MSE for the proposed approaches using 20 FM time steps.

C.3 Dynamic stall

For the dynamic stall case, Figure 16 presents the mean and standard deviation of the predicted fields obtained using PBFM. The predictions show strong agreement with the reference data, both in terms of overall distribution as well as capturing the extrema of each variable. The only notable deviation appears in the standard deviation of the skin friction, which slightly exceeds the corresponding reference values.

Figure 17 compares the performance of the proposed models as a function of the number of FM integration steps. The first panel focuses on the physical residual MSE, indicating that the best performance is achieved with 10 FM steps, with PBFM consistently delivering an 8 \times reduction compared to the baseline. The second panel reports the Wasserstein distance, showing that PBFM also narrows the gap with the reference data distribution, reaching its optimal value at 20 FM steps. Finally, the MSE of the mean and standard deviation reveals that the proposed methods preserve the already strong performance of the baseline, introducing only minor deviations.

Additional qualitative samples produced by our trained PBFM method are shown in Figure 18. They highlight the wide range of physical behavior modeled by this case: from small oscillation attached flow at the top, to deep dynamic stall cases with shock wave formation at the bottom.

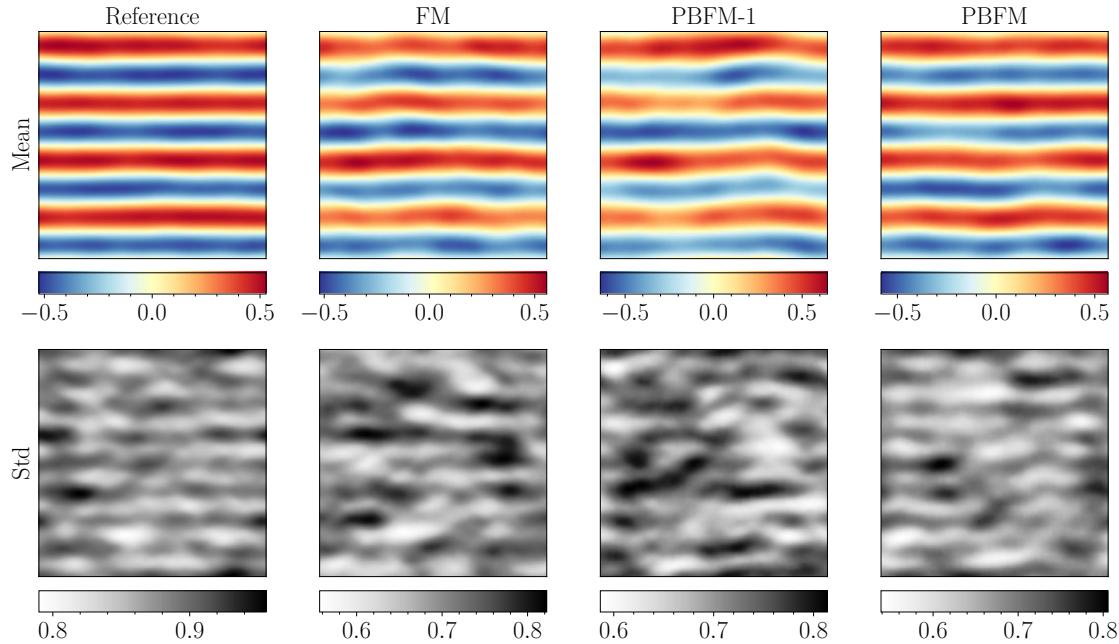


Figure 14: Example of mean and standard deviation of Kolmogorov flow computed over 20 FM steps with 128 samples for the proposed approaches.

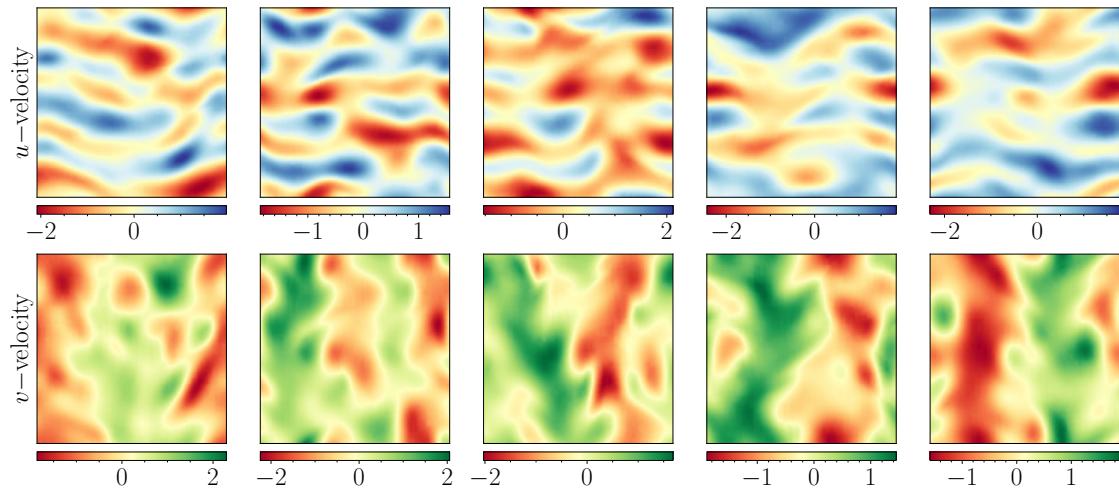


Figure 15: Example of Kolmogorov flow field samples generated with PBFM using 20 FM steps.

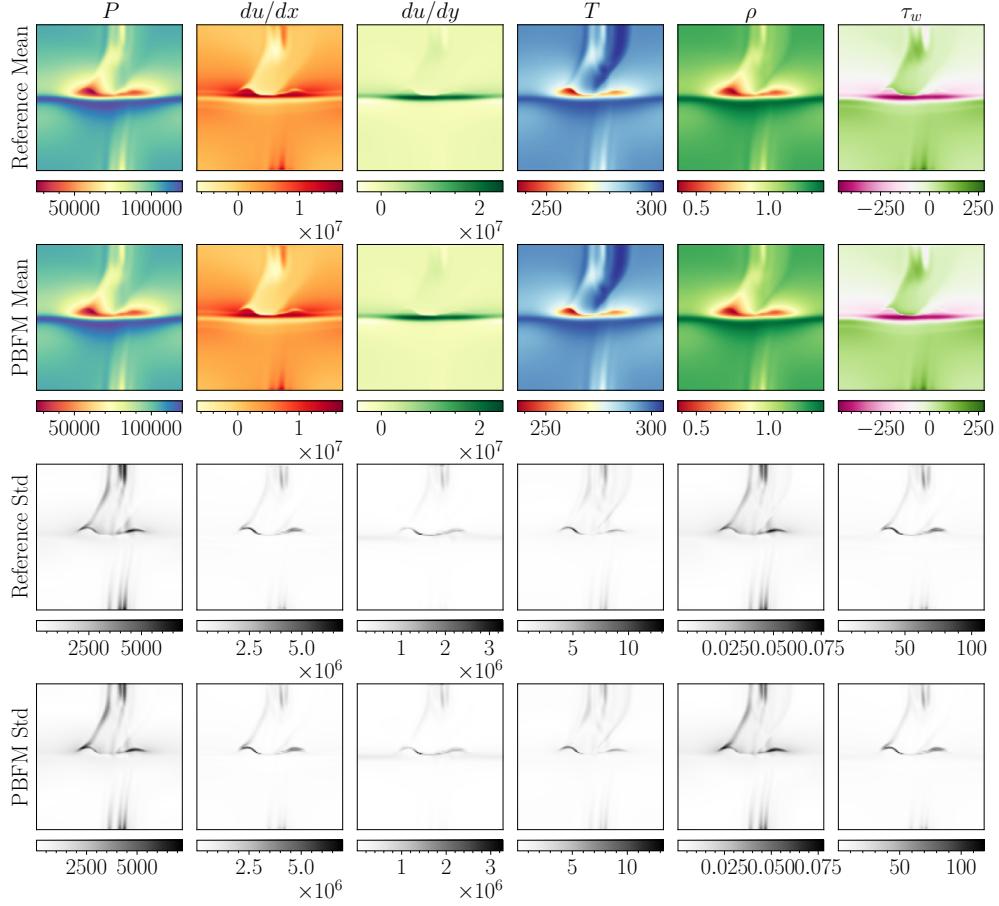


Figure 16: Example of mean and standard deviation of dynamic stall problem computed over 128 samples with 20 FM steps.

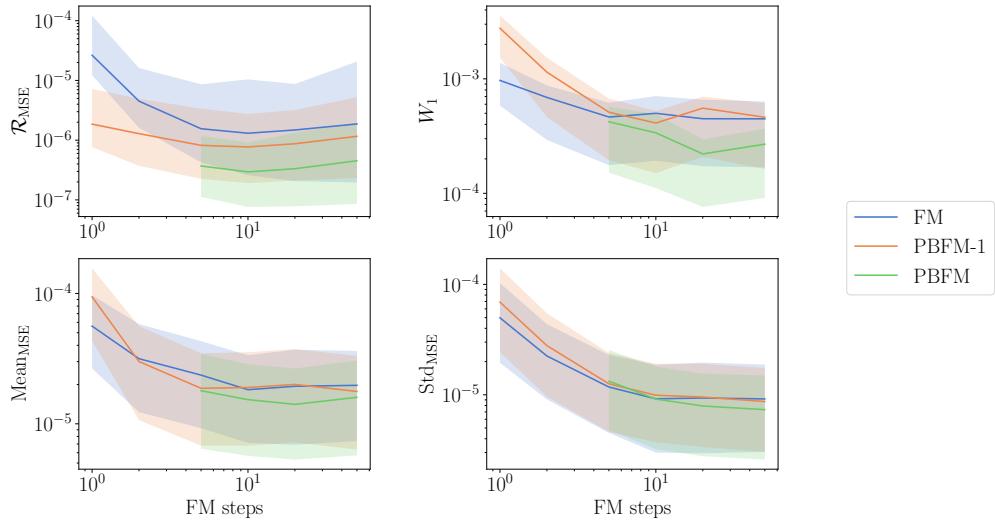


Figure 17: Comparison of the proposed approaches for physical residual and Wasserstein distance, mean and standard deviation MSE as a function of FM steps for dynamic stall case. Error bars refer to the different conditioning values within the validation dataset.

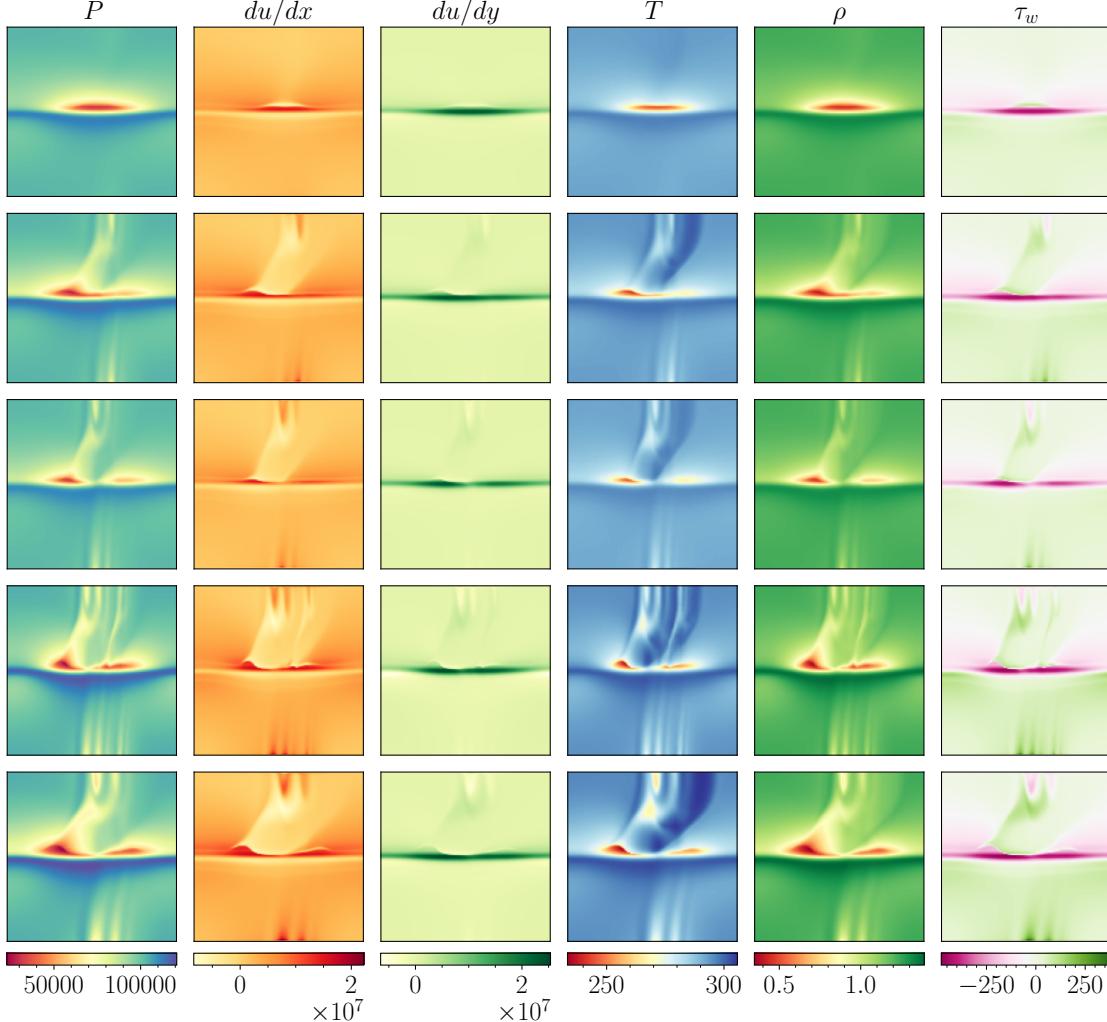


Figure 18: Example of dynamic stall flow field samples generated with PBFM using 20 FM steps. The outputs vary substantially depending on the conditioning inputs, illustrating the model’s sensitivity to different flow scenarios.

D Residual loss scaling laws

During training, the residual of each sample is computed starting from a given time t and evolved forward until $t = 1$. Trajectories initialized at times closer to $t = 0$ tend to exhibit larger errors, particularly when only a single integration step is used. To mitigate this effect, we introduce a weighting scheme based on a power law t^p , where the residual loss is scaled according to the starting time t . We investigate the impact of different power exponents p , focusing on the most challenging case of dynamic stall.

Figure 19 presents a comparison of the MSE for physical residuals, as well as the mean and standard deviation, for both the PBFM-1 and PBFM frameworks. The results show that unrolling helps regularize the error, producing a monotonic increase in error as a function of the power p , and also reduces sensitivity to the choice of p in the range $[1, 4]$. Notably, both frameworks achieve optimal performance when residuals are scaled linearly with time, $p = 1$. In contrast, using no scaling, $p = 0$, results in significantly higher errors, underscoring the importance of appropriately weighting residuals based on the starting time.

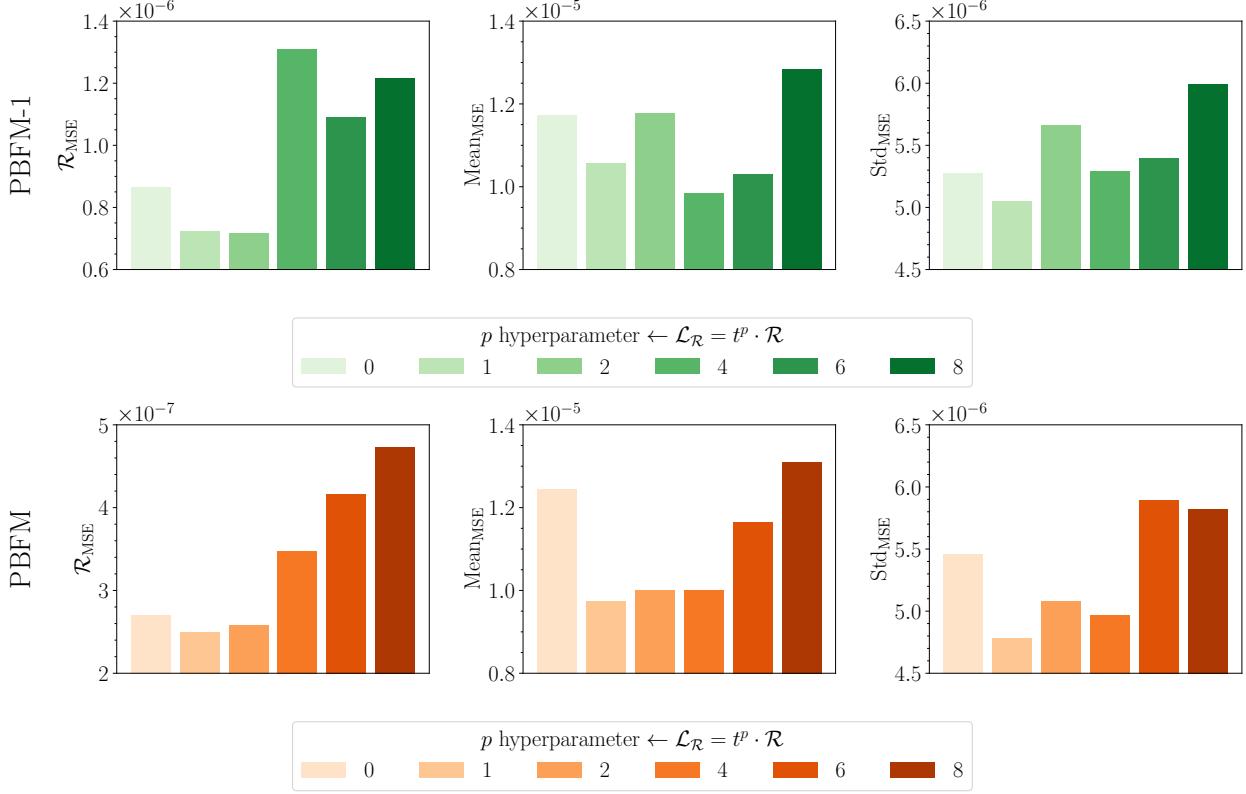


Figure 19: Comparison of physical residual, mean and standard deviation MSE for dynamic stall under various power-law scalings in the residual loss. Unrolling reduces sensitivity to the scaling exponent.

E Conflict-free updates and weighted loss terms

The introduction of the second loss term associated with the physical residual minimization transforms the framework into a multi-task learning problem. A seemingly attractive approach is to combine the two losses using a fixed weighting hyperparameter. However, this naive strategy requires manual tuning of the relative weight and often leads to suboptimal performance. In particular, optimization may get stuck in a local minimum of one loss due to conflicts between the tasks. To address this, we adopt the conflict-free updates (Liu et al., 2025a), which mitigate gradient conflicts by computing a non-conflicting optimization direction through the inverse of the loss-specific gradient covariance matrix. Furthermore, this approach has the potential to yield improved solutions.

We evaluate the proposed setup on the dynamic stall case, the most challenging scenario considered, comparing ConFIG to models trained with various fixed loss weights. Figure 20 reports the MSE for the physical residual, the predicted mean, and the standard deviation across all configurations. ConFIG consistently outperforms the fixed-weight approaches, achieving the lowest error in each individual metric and delivering the best overall performance.

F Architecture and training details

The framework is implemented in PyTorch v2.5.1, employing Distributed Data Parallel (DDP) for scalable training. All experiments are trained using the AdamW optimizer with weight decay set to 0, $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a fixed learning rate. To avoid learning rate adjustments across different hardware setups, we maintain a constant global batch size, independent of the number of GPUs. An Exponential Moving Average (EMA) of the model parameters is maintained throughout training, with a decay rate of 0.999, and is used during sampling.

We adopt the Diffusion Transformer (DiT) architecture proposed by Peebles and Xie (2023) as the backbone for our flow matching model, incorporating minor modifications. The model is conditioned via adaptive layer normalization (adaLN-Zero) blocks, which replace the standard normalization layers. The scale and shift parameters in these blocks are derived from the sum of the embedding vectors for the time step t , used in the flow matching process, and the

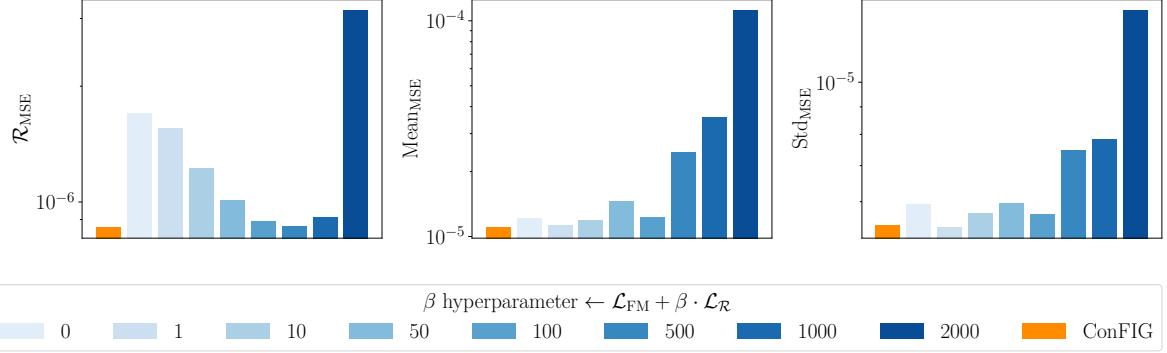


Figure 20: Comparison of MSE for physical residual, predicted mean, and standard deviation in the dynamic stall case, using ConFIG and fixed β hyperparameters for loss weighting. The optimal configuration minimizes the error across all three metrics simultaneously.

conditioning signal c . We introduce two modifications compared to the original DiT implementation. First, we incorporate linear attention (Wang et al., 2020) alongside the standard quadratic one. Second, we replace the original label embedder with a two-layer module: a linear layer followed by a SiLU activation function, and a second linear layer that produces the final conditioning embedding.

All the hyperparameters are summarized in Table 2.

Table 2: Architecture and training hyperparameters for proposed test cases.

	Darcy flow	Kolmogorov flow	Dynamic stall
Training iterations	1248000	512000	2048000
Learning rate	$3 \cdot 10^{-5}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
Batch size	64	64	64
Conditioning size	-	1	4
Output size	2	2	6
Patch size	4	4	4
Hidden size	256	256	128
DiT depth	8	8	4
Attention heads	8	8	4
Attention type	Quadratic	Linear	Linear
Parameters (M)	9.75	9.81	1.29
Gflops	2.72	13.00	1.68

G Performance evaluation

We provide an overview of the training performance, focusing on the impact of the residual loss along with ConFIG, as well as the effect of unrolling. Table 3 reports the wall-clock time, in seconds, for a single training iteration on an NVIDIA A100 64 GB GPU. Incorporating ConFIG (Liu et al., 2025a) increases the training time by approximately 80% compared to the baseline FM model, primarily due to the additional backward pass required to compute gradients associated with the residual loss. Similarly, unrolling negatively affects training performance: using four unroll steps roughly doubles the iteration time. Importantly, while training becomes more expensive, inference time remains unchanged, an essential requirement for deploying fast and reliable surrogate models.

Table 4 reports the inference wall-clock times for generating 32, 64, and 128 dynamic stall samples using 8 CPU cores of an Intel Xeon Platinum 8358 and an NVIDIA A100 64GB GPU. The same CPU was used to run the numerical simulations, ensuring a more consistent comparison of computational performance. On average, a single numerical simulation requires 76 minutes to complete. In contrast, the proposed generative model can produce 128 samples in just 2 and 4 minutes when using 10 and 20 FM steps on the CPU, respectively. When executed on a modern GPU, these inference times drop dramatically to 0.2 and 0.4 seconds, respectively. This substantial speedup highlights the model’s potential as a fast and reliable surrogate for dynamic stall prediction in helicopter and wind turbine applications.

Table 3: Comparison of wall-clock time in seconds for one training iteration on an NVIDIA A100 64GB GPU for the proposed approaches. Batch size is 64 for all cases. Inference time is unchanged.

	FM	PBFM-1	PBFM 2 steps	PBFM 3 steps	PBFM 4 steps
Darcy flow	$4.29 \cdot 10^{-2}$	$8.14 \cdot 10^{-2}$	$1.18 \cdot 10^{-1}$	$1.55 \cdot 10^{-1}$	$1.90 \cdot 10^{-1}$
Kolmogorov flow	$1.13 \cdot 10^{-1}$	$1.94 \cdot 10^{-1}$	$3.02 \cdot 10^{-1}$	$4.10 \cdot 10^{-1}$	$5.18 \cdot 10^{-1}$
Dynamic stall	$4.28 \cdot 10^{-2}$	$6.69 \cdot 10^{-2}$	$9.78 \cdot 10^{-2}$	$1.28 \cdot 10^{-1}$	$1.59 \cdot 10^{-1}$

Table 4: Wall-clock time in seconds to generate n samples using 8 cores of an Intel Xeon Platinum 8358 and an NVIDIA A100 64GB for dynamic stall case. Batch size is set equal to the number of samples. One simulation with the same cores takes on average 76 minutes ($4.56 \cdot 10^3$ s).

FM steps	Number of samples - GPU			Number of samples - CPU		
	32	64	128	32	64	128
1	$3.23 \cdot 10^{-3}$	$3.38 \cdot 10^{-3}$	$3.80 \cdot 10^{-3}$	$2.27 \cdot 10^0$	$6.16 \cdot 10^0$	$1.20 \cdot 10^1$
2	$9.76 \cdot 10^{-3}$	$1.50 \cdot 10^{-2}$	$2.53 \cdot 10^{-2}$	$4.48 \cdot 10^0$	$1.11 \cdot 10^1$	$2.40 \cdot 10^1$
5	$2.90 \cdot 10^{-2}$	$4.95 \cdot 10^{-2}$	$8.99 \cdot 10^{-2}$	$1.17 \cdot 10^1$	$2.96 \cdot 10^1$	$5.99 \cdot 10^1$
10	$6.15 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.98 \cdot 10^{-1}$	$2.25 \cdot 10^1$	$5.09 \cdot 10^1$	$1.19 \cdot 10^2$
20	$1.26 \cdot 10^{-1}$	$2.23 \cdot 10^{-1}$	$4.13 \cdot 10^{-1}$	$4.68 \cdot 10^1$	$1.15 \cdot 10^2$	$2.40 \cdot 10^2$
50	$3.21 \cdot 10^{-1}$	$5.70 \cdot 10^{-1}$	$1.06 \cdot 10^0$	$1.12 \cdot 10^2$	$3.08 \cdot 10^2$	$5.99 \cdot 10^2$

H Sampler implementation

Algorithm 2 outlines the proposed sampling procedure, which is implemented using the explicit Euler integration scheme with n equispaced time steps. The process begins by initializing x_t , which holds the sample values at time $t = 0$, with Gaussian noise. During the integration loop, each time step can be computed using either the standard deterministic FM sampler or the proposed stochastic variant. The choice between the two is governed by a user-defined boolean control parameter and is restricted to the initial segment of the trajectory, up to a threshold time t^* . In our experiments, we set $t^* = 0.2$, introducing additional stochasticity during the early phase of sampling while preserving high sample quality in later stages. In the stochastic sampler, the velocity u_t^θ is used to generate the final sample in a single forward step, followed by a backward update to time $t + dt$ using a new Gaussian noise sample.

Algorithm 2 Deterministic and Stochastic Samplers

```

 $dt \leftarrow 1/n$                                       $\triangleright n$  is the number of integration steps
 $x_t \leftarrow x_0 = \mathcal{N}(0, 1)$ 
for  $i = 0, i < n$  do
    if  $t < t^*$  and use stochastic sampler then
         $x_t \leftarrow x_t + (1 - t) \cdot u_t^\theta$             $\triangleright$  Integrate to  $t = 1$ 
         $t \leftarrow t + dt$ 
         $x_t \leftarrow (1 - t) \cdot \mathcal{N}(0, 1) + t \cdot x_t$     $\triangleright$  Return to  $t + dt$  using new generated normal noise
    else
         $x_t \leftarrow x_t + dt \cdot u_t^\theta$             $\triangleright$  Standard deterministic sampler, stochasticity is embedded in  $x_0$ 
         $t \leftarrow t + dt$ 
    end if
end for

```
