

Temporal self-attention improves field-based flow prediction under distribution shift and incomplete conditioning

Davide Dapelo

d.dapelo@liverpool.ac.uk

University of Liverpool

Shaun Lockett

University of Liverpool

John Bridgeman

University of Liverpool

Article

Keywords: Generative AI, Transformer, Computer Vision, CFD, Lattice-Boltzmann

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10458224/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Temporal self-attention improves field-based flow prediction under distribution shift and incomplete conditioning

Davide Dapelo^{1*}, Shaun Lockett¹ and John Bridgeman¹

^{1*}Department of Civil and Environmental Engineering, University of Liverpool,
Liverpool, United Kingdom.

*Corresponding author(s). E-mail(s): d.dapelo@liverpool.ac.uk;

Abstract

Data-driven fluid modelling now spans application-specific surrogates, multi-physics pre-training, and Earth-system foundation models. Nevertheless, many CFD predictors are evaluated within a single application family or on canonical configurations with fixed or simplified geometries and boundary conditions. Governing parameters are often held fixed or supplied explicitly, while substantial shifts in flow family or physical regime frequently require task-specific training or fine-tuning. We present Generating fLOws frOm Data (GLOD), a proof-of-concept autoregressive model combining a visual encoder-decoder with temporal self-attention over semantic snapshot embeddings. Using pressure and velocity fields alone, without explicit Reynolds number or geometric descriptors, GLOD is trained on synthetic laminar Lattice-Boltzmann obstacle flows and evaluated without retraining on circles, geometry-shifted squares, lid-driven cavities, and von Karman vortex shedding. Paired identity-backbone ablations isolate the contribution of temporal sequence modelling. Across aggregate losses, drag diagnostics, line profiles, fieldwise velocity decompositions, and wake-lag tests, temporal self-attention consistently improves prediction over direct previous-frame forwarding. Transfer remains non-trivial under moderate shifts but deteriorates under stronger changes in boundary-value problem and temporal organisation. The SWIN visual encoder-decoder is more robust than its UNet counterpart.

Keywords: Generative AI, Transformer, Computer Vision, CFD, Lattice-Boltzmann

1 Introduction

Artificial Intelligence (AI), and in particular neural-network-based modelling, is increasingly used in Computational Fluid Dynamics (CFD) to accelerate simulations and construct predictive surrogates. Existing work spans turbulence-model support, mesh generation, simulation monitoring, and direct field prediction [1–7]. Data-driven flow generation has progressed from application-specific convolutional, attention-based, and adversarial models [8–12] to high-fidelity turbulence generation [13–15], multi-physics pretraining [16], and foundation models spanning multiple PDEs or Earth-system applications [17, 18]. These developments demonstrate both the ambition of the field and the increasing breadth of learned physical prediction.

Important generalisation questions nevertheless remain. Many established surrogates are evaluated primarily within a single application family, while benchmark studies show substantial deterioration when geometries, boundary conditions, fluid properties, parameter ranges, or flow dynamics depart from those represented during training [19, 20]. Canonical neural-operator tests have also often used fixed forcing or periodic domains [21]. Geometry-aware methods can transfer across shapes and resolutions, but commonly encode the geometry and boundary data explicitly, for example through signed-distance functions or geometry-resolving meshes [22]. Likewise, simulation parameters are frequently held fixed within a task or supplied as conditioning variables. Emerging approaches are beginning to relax these restrictions: multiple-physics pretraining can infer the effect of unsupplied parameters from field history [16], and PDE foundation models can transfer learned representations to previously unseen physics [17]. However, substantial downstream changes still commonly involve task-specific fine-tuning, and robust zero-shot transfer across distinct flow families under incomplete conditioning remains an open problem.

Against this background, the present study asks a deliberately narrower question. Rather than pretraining across multiple physical systems or maximising performance on a closely matched benchmark, we train once on a restricted but heterogeneous family of obstacle flows and evaluate the same model, without retraining or fine-tuning, on progressively broader cases that differ in geometry, boundary-value problem, Reynolds-number regime, and temporal organisation. Reynolds number is not supplied explicitly, and internal boundaries are not provided through direct geometric descriptors, although their locations remain partially inferable from persistent pressure and velocity signatures. The objective is therefore not information-free geometry extrapolation or a production-ready general-purpose solver, but a controlled assessment of whether temporal self-attention over semantic field embeddings provides predictive value beyond the training family and beyond direct previous-frame forwarding.

1.1 Model overview

The central architectural idea of GLoOD is to apply self-attention along the temporal direction on semantic embeddings extracted from a time sequence of snapshots. The purpose is to let the model use multi-timestep latent context to predict temporal evolution while compensating for information that is not supplied explicitly, including hidden physical parameters and geometry available only indirectly through persistent field signatures. A high-level scheme of the model is shown in Figure 1. GLoOD combines a visual encoder-decoder with an autoregressive transformer backbone. Each flow snapshot is encoded into a latent representation, temporal

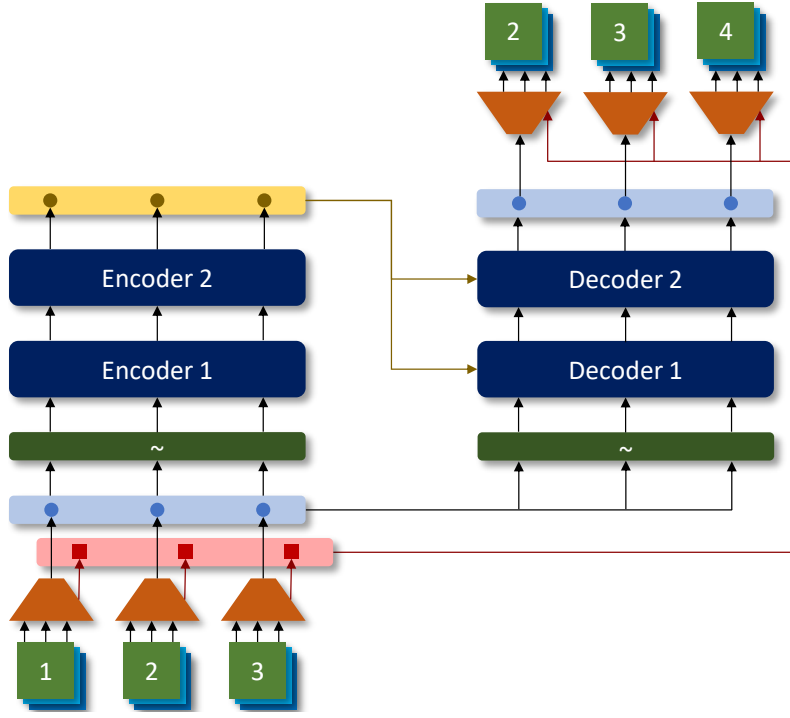


Figure 1: High-level representation of GLoOD’s architecture.

self-attention is applied over the latent sequence, and the updated latent state is decoded back into pressure and velocity fields. Two encoder-decoder variants are compared throughout: a CNN-based UNet design and a SWIN-based design. Unlike many previous approaches that rely on explicit geometric descriptors, masks, or pre-specified physical parameters, the present model takes only pressure and velocity fields as input. Obstacle interiors and boundary openings are therefore not provided explicitly, but remain implicitly encoded in the fields themselves through persistent spatial signatures, and physically relevant parameters such as Reynolds number are not supplied as separate conditioning variables. The encoder-decoder options and the transformer backbone are described in more detail in Methods.

The training set consists of synthetic 2D Lattice-Boltzmann simulations of laminar obstacle flows with circular obstacles and variable inlet/outlet layouts. Reynolds number is sampled from a lognormal distribution with physical-scale mean $\text{Re} = 90$ and coefficient of variation (CV) 1.31. This family is intentionally simpler than the broader inference campaign, but it is not trivial: it already contains moving flow around internal geometry, varying boundary layouts, and a randomly sampled parameter space. The trained models are then examined first on related obstacle-flow cases and subsequently on stronger shifts in geometry and flow organisation, including square-obstacle flows, lid-driven cavities, and von Karman vortex shedding. In all regimes, these tests are systematically repeated against identity-backbone ablations, in which the embedding obtained by encoding timestep $t - 1$ is passed directly to

the decoder for timestep t , while the same UNet or SWIN encoder-decoder is retained. This comparison is used to verify whether introducing self-attention along the temporal direction provides a real benefit beyond a pure previous-frame encode/decode route.

2 Results

2.1 Representative ground-truth sequences

Figure 2 collects six representative true-field sequences used to frame the subsequent inference tests: two in-distribution obstacle-flow examples, one geometry-shifted square-obstacle example, one cavity example, and two von Karman vortex-shedding examples. In each panel, pressure is shown on the top strip and velocity magnitude on the bottom strip, with timesteps arranged left to right from 0 to 18. Together, these examples illustrate the progressively broader temporal and physical regimes against which the same trained models are evaluated without retraining.

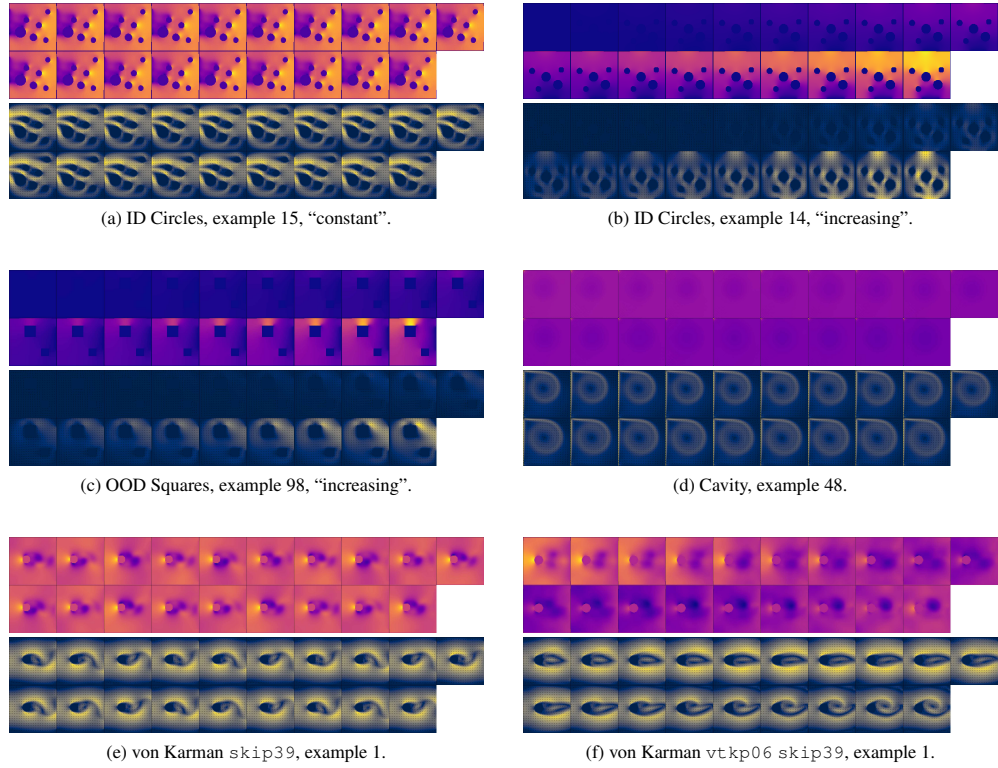


Figure 2: Representative true-field sequences spanning the main inference regimes. Each panel reports pressure on the top strip and velocity magnitude on the bottom strip.

2.2 Baseline performance on the training-family flows

A 10-fold cross-validation is used to assess whether the training dataset is sufficiently large and diverse for this proof-of-concept setting. Across the folds, the training and validation curves remain broadly consistent, with no obvious signs of severe overfitting. The SWIN-based encoder-decoder also performs better than UNet on the training-family flows, especially in velocity-magnitude error, while remaining comparable in directional metrics. Detailed foldwise curves and the percentile-stratified numerical summary are reported in Supplementary Figs. ??–?? and Supplementary Table ??, respectively.

2.3 Progressive inference stress tests

After training on a random 90/10 train/validation split, the models are evaluated without retraining on progressively broader inference datasets: related obstacle-flow circles, geometry-shifted square-obstacle cases, lid-driven cavities, and von Karman vortex-shedding cases with two temporal samplings. All these tests are paired with the respective identity-backbone ablations.

The Reynolds-number distributions are reported on the physical scale. The circle, square and cavity datasets have mean $Re = 166.8, 214.8$ and $12,252.8$, with coefficients of variation (CVs) of $0.61, 1.83$ and 0.91 , respectively. The original-timestep and reduced-timestep von Karman datasets have mean $Re = 749.0$ (CV 0.28) and 774.9 (CV 0.27), respectively. These statistics make explicit that the cavity and wake tests differ from the training family not only in geometry or boundary-value problem, but also in their Reynolds-number regimes.

The qualitative behaviour across these tests is illustrated in Fig. 3. The SWIN-based model generally preserves the large-scale organisation of the reference fields and produces the cleanest reconstructions. The UNet-based model recovers the main flow structures but is visibly noisier, particularly in spatially heterogeneous regions. Removing temporal self-attention has a much larger effect: the ID-SWIN predictions remain partly recognisable but show errors in magnitude and visible patch-wise artefacts, whereas the ID-UNet predictions are dominated by severe spatial noise and no longer provide a reliable reconstruction of the flow. The cavity pressure panels should be interpreted cautiously: the reference signal is very weak and offers limited visual insight, so this example is more useful for exposing the difficulty of the broader distribution shift than for discriminating fine pressure-reconstruction quality.

2.3.1 Inference loss and metrics

Figure 4 summarises the transformer-backbone losses and metrics across the inference ID (“Circles”), OOD (“Squares”), and cavity datasets, together with the K-fold distribution from the training-family runs. The complete numerical results, including the identity-backbone ablations, are reported in Supplementary Table ??. Overall, the transformer-backbone inference losses for the in-distribution circles and geometry-shifted squares remain close to the K-fold values, with the square cases still showing non-trivial transfer despite the unseen obstacle geometry. By contrast, the cavity cases are substantially harder, especially in pressure-dominated metrics, indicating a clearer distribution shift. Across these datasets, the identity-backbone ablations are consistently worse than the corresponding transformer-backbone models. On circles and squares, the deterioration is typically about one order of magnitude in aggregate loss, around one to almost three orders of magnitude in pressure MSE, and roughly one to

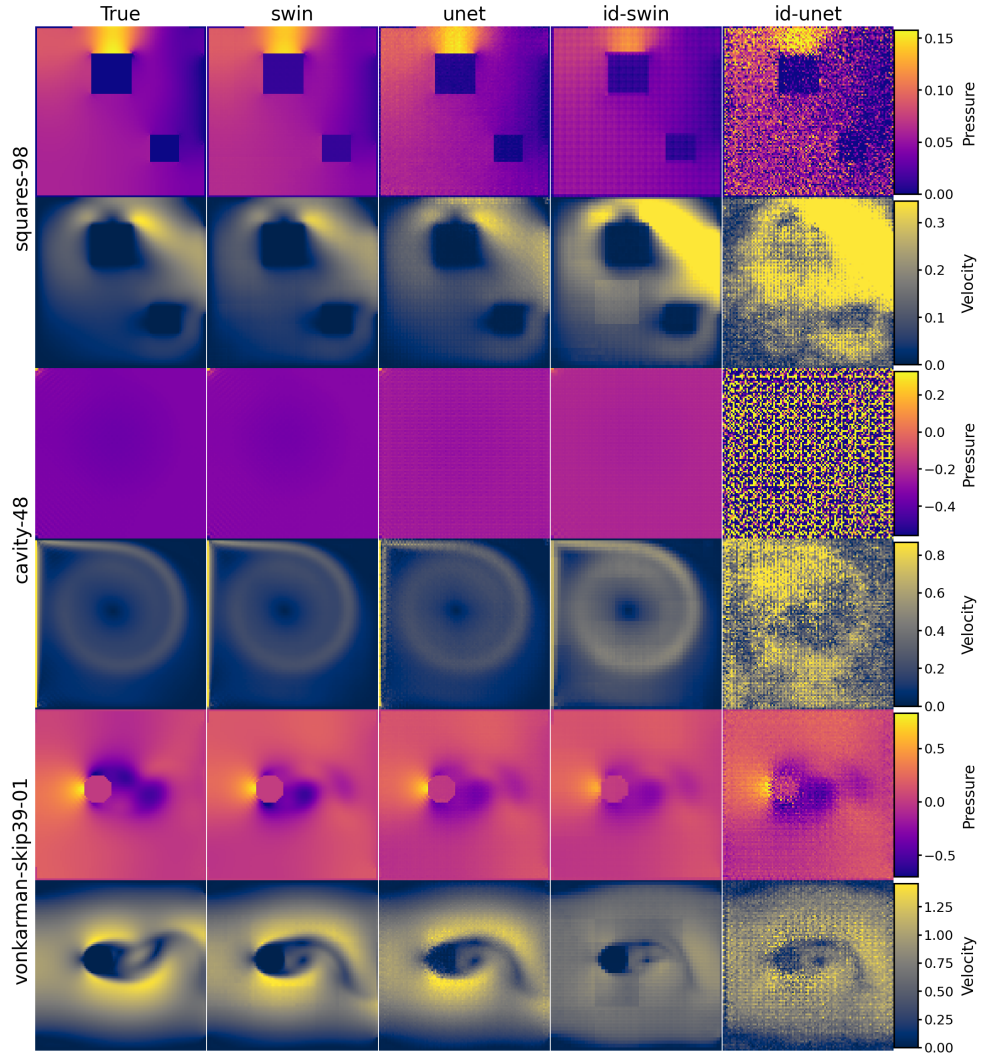


Figure 3: Qualitative comparison of representative ground-truth fields and predictions from the SWIN, UNet, ID-SWIN and ID-UNet models across selected square-obstacle, cavity and von Karman cases. The full SWIN model generally preserves the dominant spatial organisation, while UNet predictions are noisier. The identity-backbone ablations show substantially larger errors: ID-SWIN retains some recognisable structure but exhibits scale mismatch and patch-wise artefacts, whereas ID-UNet is dominated by spatially incoherent errors. The cavity pressure signal is very weak and should not be treated as a source of strong visual insight.

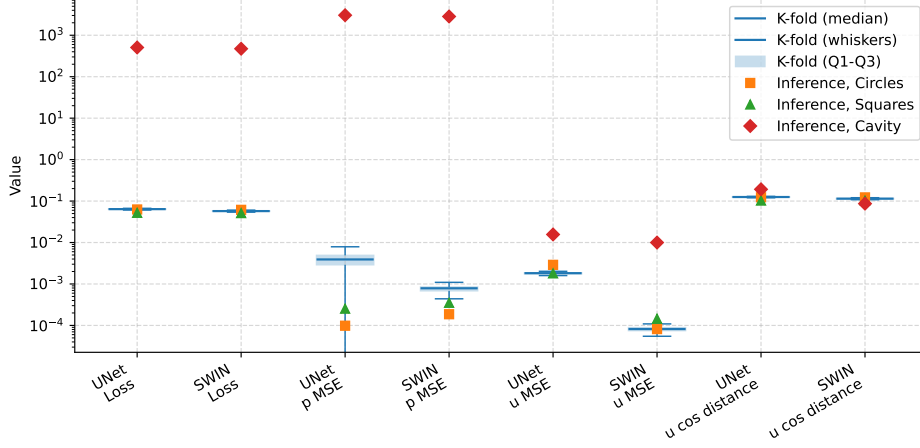


Figure 4: Summary plot of losses and metrics for the transformer-backbone runs. The corresponding numerical results, together with the identity-backbone ablations, are reported in Supplementary Table ??.

two orders of magnitude in velocity MSE; in the most extreme case, the cavity UNet identity run increases velocity MSE by nearly three orders of magnitude. This shows that temporal sequence modelling materially improves prediction beyond simple previous-frame forwarding.

2.3.2 Drag force

The complete total-drag, directional and pressure/viscous results are reported in Supplementary Tables ??, ?? and ?. Aggregate pointwise losses do not fully show whether physically meaningful derived quantities are preserved, so the obstacle-flow cases are also assessed through drag-force diagnostics on the obstacle boundaries. In the representative interquartile signal band, the total-drag magnitude error is 0.0524 for SWIN and 0.1059 for UNet on circles, and 0.0572 and 0.0780 on squares; the corresponding identity-backbone errors increase to 0.1735 and 0.4247 on circles, and 0.3287 and 0.8113 on squares for ID-SWIN and ID-UNet. Thus, the circle-to-square shift does not cause a uniform collapse of the transformer models, whereas removing temporal attention causes a marked deterioration. The total-drag direction gives the same qualitative ranking, with cosine-distance errors of 0.0002 versus 0.0009 on circles and 0.0031 versus 0.0036 on squares for SWIN and UNet. The decomposition further separates the models: in the same interquartile band, the viscous-drag magnitude error is 0.0624 for SWIN versus 0.1868 for UNet on circles, and 0.1092 versus 0.1355 on squares; the corresponding identity values reach 0.4524 and 0.8017 on circles, and 1.0134 and 2.0213 on squares. These values support the interpretation that the clearest SWIN advantage concerns near-wall velocity gradients and force direction, rather than total force magnitude alone. The complete total-drag table, pressure/viscous decompositions and sensitivity analyses are retained in the Supplementary Information.

2.3.3 Structure-aware velocity diagnostics

The complete pooled and section-wise profile results and the full fieldwise decompositions are reported in Supplementary Tables ??, ?? and ?. Two further diagnostics distinguish broad structural agreement from simpler amplitude or alignment effects. First, for each predicted velocity field, the streamwise component u_x is sampled along three fixed, representative vertical lines spanning the domain, at $x/L = 0.25, 0.50$, and 0.75 . The resulting one-dimensional profiles $u_x(y)$ therefore test whether the predicted velocity variation across the domain is correct at specific sections, rather than merely whether the global field average is close. Their median relative-profile errors are 0.0486 versus 0.2124 for SWIN and UNet on circles, 0.0657 versus 0.2290 on squares, and 0.2715 versus 0.5631 in the cavity. Second, the whole-field velocity prediction is compared after fitting a single global amplitude factor, an optimal spatial translation, or both. This asks whether the residual discrepancy can be explained by the prediction being globally too large/small or slightly displaced, rather than having the wrong internal structure. The raw speed-field relative L_2 errors are 0.0277 versus 0.1986, 0.0573 versus 0.1994, and 0.1939 versus 0.4359 for SWIN and UNet on circles, squares, and cavity, respectively. The identity ablations are substantially worse, with circle profile medians of 0.9098 for ID-SWIN and 2.2975 for ID-UNet. Shift-and-scale corrections reduce part of these discrepancies, especially for the identity models, but do not remove the ranking; the residual errors therefore reflect genuine differences in spatial structure rather than only global amplitude or minor alignment. The complete pooled profile table, section-wise results and full fieldwise decompositions remain in the Supplementary Information.

2.3.4 Von Karman vortex-shedding test

The complete lag, profile and aggregate-metric comparisons are reported in Supplementary Tables ??, ?? and ?. The von Karman datasets probe coherent time-periodic wake dynamics, for which same-timestep errors can be dominated by phase lag. Here, the “current” comparison means that a predicted field at timestep t is compared with the true field at the same t , whereas the “previous” comparison uses the true field at $t - 1$. The original-timestep `skip39` case starts from an already established alternating vortex street, so it is a particularly demanding test of phase-accurate rollout rather than of wake formation alone. In this case, the SWIN speed-field relative error falls from 0.3177 against the current frame to 0.0206 against the previous frame, while the corresponding vorticity errors fall from 1.0779 to 0.0940. This indicates that the wake is recognisable but temporally delayed. Reducing the output timestep by a factor of five largely alleviates this effect: in `vtkp06 skip39`, the same values are 0.0588 and 0.0219 for speed, and 0.1713 and 0.1103 for vorticity. The identity ablations do not show comparable recovery: in the original `skip39` test, ID-SWIN gives speed errors of 0.3705 and 0.3337 for current and previous frames, while ID-UNet gives 0.4435 and 0.3949. Thus, the lag is not equivalent to simple previous-frame copying. The complete comparison, including correlations and all variants, is reported in the Supplementary Information. The corresponding spatial evolution is shown in Fig. 5: the original-timestep SWIN sequence is visibly displaced relative to the reference wake, whereas the reduced-timestep sequence follows the successive vortex structures more closely.

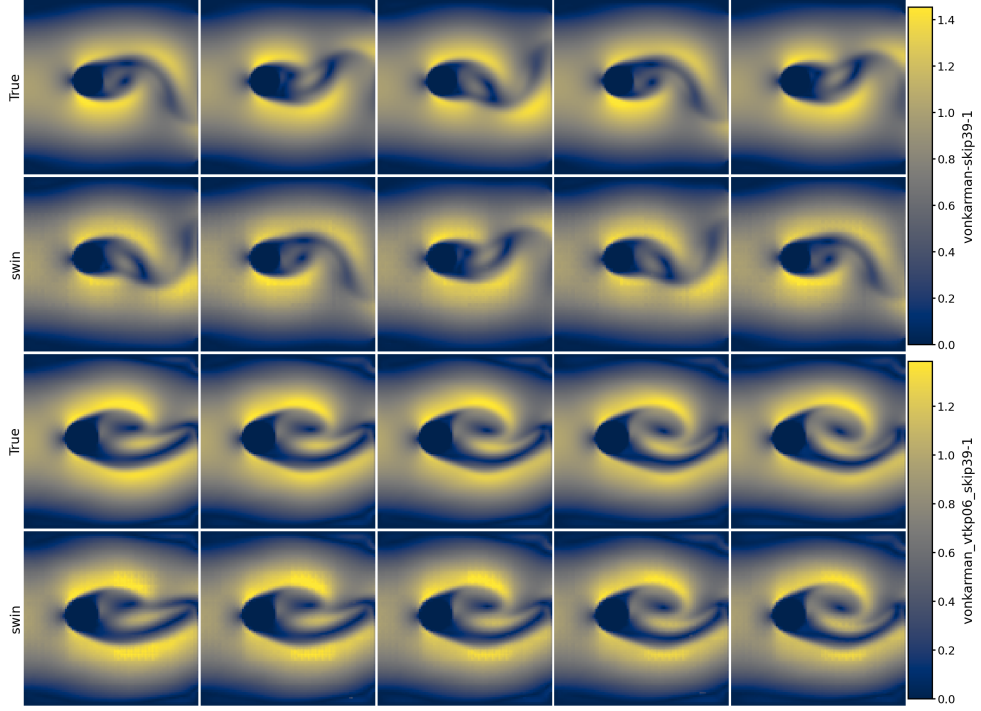


Figure 5: Representative von Karman velocity-magnitude snapshots for the original-timestep `skip39` case (top two rows) and the reduced-timestep `vtkp06 skip39` case (bottom two rows). Each pair compares the true sequence with the corresponding SWIN predictions at successive output timesteps. The reduced timestep produces smaller inter-frame changes and a closer temporal alignment of the predicted wake with the reference sequence.

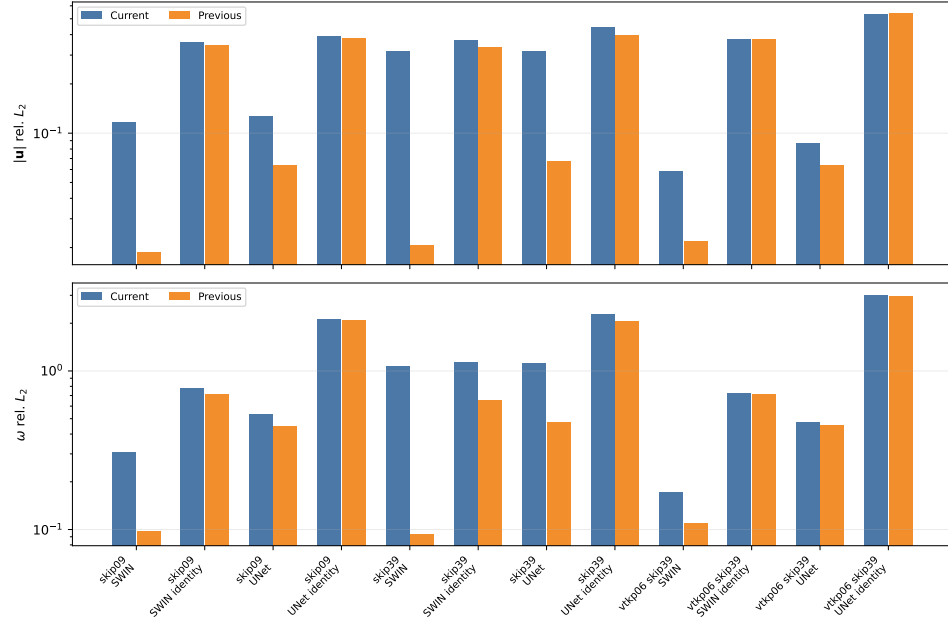


Figure 6: Von Karman lag diagnostic for the original-timestep skip09 and skip39 variants and for the reduced-timestep vtkp06 skip39 variant, including the identity-backbone ablations. Current compares predicted fields with the same true timestep, whereas previous compares them with the preceding true timestep.

3 Discussion

The results support four main conclusions. First, the present proof-of-concept architecture can reproduce the main structure of the training-family obstacle flows and can retain non-trivial predictive skill under moderate geometric shift. This is most clearly seen in the circle and square tests, where the aggregate losses remain comparatively small, the broad pressure and velocity organisation is preserved, and the physically derived diagnostics remain informative rather than collapsing completely. Second, in this proof-of-concept setting, the results show that a single field-based training regime can yield non-trivial, diagnostically measurable transfer without retraining across downstream flow families that differ substantially in boundary-value problem and temporal organisation. Third, the identity-backbone ablations are systematically worse across all regimes, indicating that temporal self-attention is not merely passing forward the previous latent frame but is contributing genuine sequence modelling, with a clear positive effect on temporal prediction under incomplete parametrisation. Fourth, the SWIN-based encoder-decoder is consistently more robust than the UNet-based one across aggregate losses, drag-derived quantities, line-profile errors, fieldwise velocity decompositions, and wake-sensitive diagnostics.

At the same time, the broader inference campaign makes the limitations of the approach visible. The cavity cases show that transfer deteriorates when the flow family differs substantially from that seen in training, particularly in situations where the pressure signal is low. The von Karman tests reveal a different limitation: the models can retain recognisable wake morphology, but phase-sensitive periodic dynamics remain difficult and can produce large same-timestep errors even when the qualitative wake structure is partially preserved. These two failure modes are informative in different ways. The cavity results indicate limits in extrapolating to cases where the input signal is weak, whereas the wake results indicate that temporal alignment and output sampling can become limiting even when the underlying spatial structures are broadly recognisable.

These findings are directly connected to the methodological choice made in this study. The training distribution was intentionally kept simpler than the final inference campaign in order to separate architectural effects from dataset tailoring while also moving away from the narrow-application or highly idealised setups that are common in the literature. In that sense, the contribution of the paper is not only a proof of feasibility for GLoOD, but also a proposal for evaluating generative CFD models under deliberately broader and less fully parameterised conditions. A model that performs well only when the target application closely matches the training set may still be useful in practice, but it is harder to interpret what such performance says about the architecture itself. The present design instead exposes both the regimes in which the model transfers and those in which it does not.

Taken together, this 2D, fixed-resolution study establishes a controlled baseline for assessing attention-based flow prediction under deliberately mismatched training and inference regimes. Within this scope, latent-sequence modelling provides a measurable benefit over a pure previous-frame encode/decode path, and this benefit remains visible across several progressively harder distribution shifts despite the absence of explicit parameter and geometry inputs. The results therefore motivate a next step: extending the same evaluation logic to richer flow families, broader Reynolds-number coverage, larger timesteps, three-dimensional settings, and partially observed flows. Such extensions will test how far the demonstrated field-history-based compensation can be carried beyond the present proof-of-concept setting.

4 Methods

4.1 Overview

GLOD combines a visual encoder-decoder with an autoregressive transformer backbone operating on 19-step sequences of semantic embeddings. Two encoder-decoder variants are considered throughout, namely a CNN-based UNet design and a SWIN-based design. The inputs are pressure and velocity fields alone; Reynolds number and explicit geometric masks or boundary descriptors are not supplied to the model.

The training data comprise 4,081 synthetic 2D Lattice-Boltzmann sequences on 101×101 domains with random circular obstacles, variable inlet/outlet layouts, and sampled Reynolds number. The inference campaign then evaluates the trained models on unseen circles, geometry-shifted squares, lid-driven cavities, and von Karman vortex-shedding cases, together with identity-backbone ablations.

Training uses teacher forcing, batches of 128 examples, Adam optimisation [23], a plateau-based learning-rate scheduler, and He initialisation [24]. The optimisation target combines pressure mean-squared error, velocity-magnitude mean-squared error, and cosine-distance information on velocity direction. Unless otherwise stated, the OOD results are paired comparisons on one fixed random 90/10 split and one trained instance per architecture/backbone configuration; they therefore characterise the controlled model instances studied here rather than seed-averaged population estimates.

Expanded layer-by-layer specifications, numerical setup, and the full procedures for the drag-force, profile, fieldwise, and lag-aware diagnostics are reported in the Supplementary Methods.

4.2 Architecture summary

Each snapshot is first encoded into a latent representation of size 768 together with intermediate feature maps used as skip information during decoding. The latent sequence is then processed autoregressively by a transformer backbone, and the updated embedding is decoded back into pressure and velocity fields.

In the UNet variant, pressure, u_x , and u_y are processed through separate convolutional pathways before being coupled at the embedding level. Each pathway downsamples the corresponding field through three strided 3×3 convolutions, produces a 256-dimensional pathway embedding, and contributes to the shared 768-dimensional latent state. The decoder mirrors this structure through three transposed-convolution upsampling pathways, with the intermediate feature maps forwarded as skip connections to recover local detail.

In the SWIN variant, the same three physical channels are again processed separately, but each pathway uses shifted-window self-attention instead of pure convolution. The fields are embedded into 2×2 patches, passed through three SWIN stages with patch merging, pooled to a 256-dimensional pathway embedding, and then coupled into the same 768-dimensional latent state. The decoder performs the reverse sequence of patch expansion and SWIN refinement, again reusing encoder-side intermediate feature maps.

The temporal backbone is a standard autoregressive transformer operating on 19-step sequences of these 768-dimensional embeddings. It uses sinusoidal positional encoding, two

encoder blocks and two decoder blocks, four attention heads per block, dropout of 0.1, and feed-forward layers with hidden width 1,536. In the full model, self-attention mixes information across the latent sequence before the next embedding is decoded. In the identity-backbone ablation, by contrast, the embedding encoded from timestep $t - 1$ is passed directly to the decoder for timestep t without this temporal mixing.

4.3 Metrics and loss function

Three quantitative diagnostics are used during training and evaluation. Let N denote the total number of evaluated space-time points. The pressure mean-squared error is

$$\text{MSE}_p = \frac{1}{N} \sum (p^{\text{pred}} - p^{\text{true}})^2. \quad (1)$$

The velocity mean-squared error is

$$\text{MSE}_u = \frac{1}{N} \sum \|\mathbf{u}^{\text{pred}} - \mathbf{u}^{\text{true}}\|_2^2. \quad (2)$$

To penalise directional disagreement in the velocity field, the cosine-distance metric is

$$\text{COSDIST} = \frac{1}{N} \sum \left(1 - \frac{\mathbf{u}^{\text{pred}} \cdot \mathbf{u}^{\text{true}}}{|\mathbf{u}^{\text{pred}}| |\mathbf{u}^{\text{true}}| + 10^{-8}} \right), \quad (3)$$

where the small additive term prevents division-by-zero in near-stagnant regions. The pressure and velocity mean-squared-error terms enforce magnitude accuracy in Galilean-invariant form, while the cosine distance prevents the model from matching only the speed and losing directional information.

The loss function is the weighted combination

$$\mathcal{L} = \alpha (\text{MSE}_p + \text{MSE}_u) + (1 - \alpha) \text{COSDIST}, \quad (4)$$

with $\alpha = 0.5$ throughout this study.

4.4 Diagnostic metrics

The diagnostics reported in the Results are computed after inference and are not included in the optimisation loss. They are intended to test whether low pointwise error corresponds to preservation of physically meaningful quantities, spatial structure, and temporal phase. All diagnostic errors are evaluated only on valid fluid nodes and are aggregated over the stated examples and predicted timesteps; the complete aggregation and sensitivity procedures are given in the Supplementary Information.

For obstacle-flow cases, the total force on the internal boundaries is decomposed into pressure and viscous contributions,

$$\mathbf{F} = \mathbf{F}_p + \mathbf{F}_v, \quad \mathbf{F}_p = - \int_{\Gamma} p \mathbf{n} \, ds, \quad \mathbf{F}_v = \frac{1}{\text{Re}} \int_{\Gamma} (\nabla \mathbf{u} + \nabla \mathbf{u}^T) \mathbf{n} \, ds, \quad (5)$$

where Γ is the union of obstacle boundaries. The model is not conditioned on Reynolds number and does not predict it as an additional output. In the drag computation, the same case-specific prefactor is therefore applied when comparing the predicted and true fields; for the viscous contribution, the $1/\text{Re}$ factor occurs in both quantities and cancels in their relative comparison. Equivalently, interpreting the predicted field as implying an effective Re^{pred} assumes $\text{Re}^{\text{pred}} \simeq \text{Re}^{\text{true}}$. This is not imposed as an input constraint, but is supported here by the absence of pathological field reconstructions in the transformer runs. Magnitude and directional errors are complementary measures of the same total force and are therefore defined together as

$$e_{\text{rel}}(q) = \frac{|q^{\text{pred}} - q^{\text{true}}|}{|q^{\text{true}}| + \varepsilon}, \quad d_{\text{cos}}(\mathbf{F}) = 1 - \frac{\mathbf{F}^{\text{pred}} \cdot \mathbf{F}^{\text{true}}}{|\mathbf{F}^{\text{pred}}| |\mathbf{F}^{\text{true}}| + \varepsilon}. \quad (6)$$

Because very small forces make relative errors unstable, results are reported both after pooling and within the interquartile signal band; a prefix-excluded view is also used to assess rollout-position effects.

To assess spatial structure independently of whole-field averages, vertical profiles of the streamwise velocity $u_x(y)$ are extracted at $x/L = 0.25, 0.50$, and 0.75 , using only valid fluid nodes. Their relative L_2 error is

$$e_{\text{prof}} = \frac{\|u_x^{\text{pred}}(y) - u_x^{\text{true}}(y)\|_2}{\|u_x^{\text{true}}(y)\|_2 + \varepsilon}. \quad (7)$$

Whole-field velocity errors are additionally recomputed after optimal global rescaling, optimal translation, and joint rescaling and translation. For a field $q \in \{u_x, u_y, |\mathbf{u}|\}$, the raw error is

$$e_{\text{raw}}(q) = \frac{\|q^{\text{pred}} - q^{\text{true}}\|_{2,\Omega}}{\|q^{\text{true}}\|_{2,\Omega} + \varepsilon}, \quad (8)$$

and the corrected errors are obtained by minimising the same numerator after applying the corresponding scale and/or translation. Their reduction, together with the fieldwise Pearson correlation, distinguishes amplitude or alignment errors from residual structural mismatch. These corrections are diagnostic only and are not applied to the reported predictions.

For von Karman wakes, temporal phase is assessed using both the speed field and vorticity, $\omega = \partial u_y / \partial x - \partial u_x / \partial y$. Each predicted field is compared with the true field at the same timestep (current) and with the preceding true field (previous):

$$e_{\text{current}}(g, t) = \frac{\|g_t^{\text{pred}} - g_t^{\text{true}}\|_{2,\Omega}}{\|g_t^{\text{true}}\|_{2,\Omega} + \varepsilon}, \quad e_{\text{previous}}(g, t) = \frac{\|g_t^{\text{pred}} - g_{t-1}^{\text{true}}\|_{2,\Omega}}{\|g_{t-1}^{\text{true}}\|_{2,\Omega} + \varepsilon}. \quad (9)$$

A lower previous-frame error is interpreted as evidence consistent with temporal lag, not as evidence that the model is intended to copy the previous frame. This distinction is tested directly through the identity-backbone ablations.

4.5 Use of AI tools in manuscript preparation

OpenAI ChatGPT was used during manuscript preparation for language editing, literature-search assistance, and code debugging assistance. All scientific claims, methodological choices, quantitative analyses, code changes, and final manuscript wording were reviewed by the authors.

Author contributions

DD conceived the research idea, carried out the research activities, implemented the code and wrote the first draft of the manuscript.

SL carried out the research activities, extended the code and edited the manuscript.

JB supervised the work, contributed to developing the research line, and edited the manuscript.

All authors reviewed the manuscript.

Data availability

The GLoOD source code is released under the GNU GPLv3 licence at <https://github.com/dav-dap/GLoOD/releases/tag/v1.0.0>. The synthetic-dataset generator is released at <https://github.com/dav-dap/GLoOD-dataset-generator/releases/tag/v1.0.0>. The datasets, trained model checkpoints and analysis artefacts supporting this study are available through the University of Liverpool DataCat repository at <https://doi.org/10.17638/datacat.liverpool.ac.uk/3101> (direct record: <https://datacat.liverpool.ac.uk/id/eprint/3101>).

Acknowledgements

The authors would like to thank the University of Liverpool's High Performance Computing (HPC) Service (Barkla), where the computational work was carried out.

Funding

This research did not receive funding.

Additional information

Competing interests The authors declare no competing interests.

References

- [1] Sun, X., Cao, W., Liu, Y., Zhu, L. & Zhang, W. High reynolds number airfoil turbulence modeling method based on machine learning technique. *Computers & Fluids* **236**, 105298 (2022). URL <https://www.sciencedirect.com/science/article/pii/S0045793021003868>.
- [2] Alhafiz, M. R., Palar, P. S. & Lavi, Z. An efficient data-driven neural network turbulence model for flow over periodic hills and parametric bump. AIAA SCITECH 2024 Forum

- (2024). URL <https://arc.aiaa.org/doi/abs/10.2514/6.2024-1985>. American Institute of Aeronautics and Astronautics, <https://arc.aiaa.org/doi/pdf/10.2514/6.2024-1985>.
- [3] Calzolari, G. & Liu, W. Deep learning to develop zero-equation based turbulence model for cfd simulations of the built environment. *Building Simulation* **17**, 399–414 (2024). URL <https://doi.org/10.1007/s12273-023-1083-4>.
 - [4] Lorsung, C. & Barati Farimani, A. Mesh deep q network: A deep reinforcement learning framework for improving meshes in computational fluid dynamics. *AIP Advances* **13**, 015026 (2023). URL <https://doi.org/10.1063/5.0138039>.
 - [5] Kochkov, D. *et al.* Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences* **118**, e2101784118 (2021). URL <https://www.pnas.org/doi/abs/10.1073/pnas.2101784118>.
 - [6] Tahmasebi, S. *et al.* Using diffusion models for reducing spatiotemporal errors of deep learning based urban microclimate predictions at post-processing stage. *Physics of Fluids* **37**, 035173 (2025). URL <https://doi.org/10.1063/5.0256658>.
 - [7] Jeon, J., Lee, J., Vinuesa, R. & Kim, S. J. Residual-based physics-informed transfer learning: A hybrid method for accelerating long-term cfd simulations via deep learning. *International Journal of Heat and Mass Transfer* **220**, 124900 (2024). URL <https://www.sciencedirect.com/science/article/pii/S0017931023010451>.
 - [8] Guo, X., Li, W. & Iorio, F. Convolutional neural networks for steady flow approximation. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16) (2016). URL <https://doi.org/10.1145/2939672.2939738>. Pp. 481–490. Association for Computing Machinery, New York, NY, USA.
 - [9] Bhatnagar, S., Afshar, Y., Pan, S., Duraisamy, K. & Kaushik, S. Prediction of aerodynamic flow fields using convolutional neural networks. *Computational Mechanics* **64**, 525–545 (2019). URL <https://doi.org/10.1007/s00466-019-01740-0>.
 - [10] Wu, M.-Y. *et al.* Fast prediction of flow field around airfoils based on deep convolutional neural network. *Applied Sciences* **12** (2022). URL <https://www.mdpi.com/2076-3417/12/23/12075>.
 - [11] Zuo, K., Ye, Z., Zhang, W., Yuan, X. & Zhu, L. Fast aerodynamics prediction of laminar airfoils based on deep attention network. *Physics of Fluids* **35**, 037127 (2023). URL <https://doi.org/10.1063/5.0140545>.
 - [12] Abaidi, R. & Adams, N. Gan-based generation of realistic compressible-flow samples from incomplete data. *Computers & Fluids* **269**, 106113 (2024). URL <https://www.sciencedirect.com/science/article/pii/S0045793023003389>.
 - [13] Du, P., Parikh, M. H., Fan, X., Liu, X.-Y. & Wang, J.-X. Conditional neural field latent diffusion model for generating spatiotemporal turbulence. *Nature Communications* **15**,

- 10416 (2024). URL <https://doi.org/10.1038/s41467-024-54712-1>.
- [14] Fan, X., Akhare, D. & Wang, J.-X. Neural differentiable modeling with diffusion-based super-resolution for two-dimensional spatiotemporal turbulence. *Computer Methods in Applied Mechanics and Engineering* **433**, 117478 (2025). URL <https://www.sciencedirect.com/science/article/pii/S0045782524007321>.
 - [15] Li, T., Biferale, L., Bonaccorso, F., Scarpolini, M. A. & Buzzicotti, M. Synthetic lagrangian turbulence by generative diffusion models. *Nature Machine Intelligence* **6**, 393–403 (2024). URL <https://doi.org/10.1038/s42256-024-00810-0>.
 - [16] McCabe, M. *et al.* Multiple physics pretraining for spatiotemporal surrogate models. *Advances in Neural Information Processing Systems* 37 (2024). URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d7cb9db5ade2db7814fbd01ee59f4c7b-Abstract-Conference.html. 38th Conference on Neural Information Processing Systems (NeurIPS 2024).
 - [17] Herde, M. *et al.* Poseidon: Efficient foundation models for pdes. *Advances in Neural Information Processing Systems* 37 (2024). URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/84e1b1ec17bb11c57234e96433022a9a-Abstract-Conference.html. 38th Conference on Neural Information Processing Systems (NeurIPS 2024).
 - [18] Bodnar, C. *et al.* A foundation model for the earth system. *Nature* **641**, 1180–1187 (2025). URL <https://doi.org/10.1038/s41586-025-09005-y>.
 - [19] Luo, Y., Chen, Y. & Zhang, Z. Cfdbench: A large-scale benchmark for machine learning methods in fluid dynamics. *arXiv preprint arXiv:2310.05963* (2023). URL <https://arxiv.org/abs/2310.05963>.
 - [20] Künzli, S., Grötschla, F., Mathys, J. & Wattenhofer, R. Surf: A generalization benchmark for gnn predicting fluid dynamics. *Proceedings of Machine Learning Research* **231**, 13:1–13:23 (2024). URL <https://proceedings.mlr.press/v231/kunzli24a.html>.
 - [21] Li, Z. *et al.* Fourier neural operator for parametric partial differential equations. 9th International Conference on Learning Representations (ICLR 2021) (2021). URL <https://openreview.net/forum?id=c8P9NQVtmnO>.
 - [22] Li, Z. *et al.* Geometry-informed neural operator for large-scale 3d pdes. *Advances in Neural Information Processing Systems* 36 (2023). URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/70518ea42831f02afc3a2828993935ad-Abstract-Conference.html. 37th Conference on Neural Information Processing Systems (NeurIPS 2023).
 - [23] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2015). URL <https://arxiv.org/abs/1412.6980>. Published as a conference paper at the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA.

- [24] He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015). URL https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html. Pp. 1026–1034. December 2015.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)