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

Davide Dapelo[1*], Shaun Lockett[1] and John Bridgeman[1]

[1*]Department of Civil and Environmental Engineering, University of Liverpool, Liverpool, United Kingdom.

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

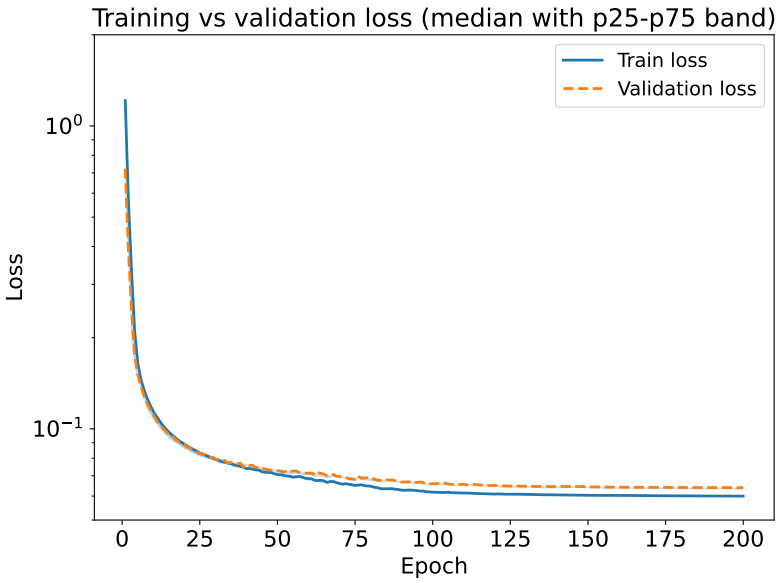
1 Supplementary Results

1.1 Baseline performance on the training-family flows

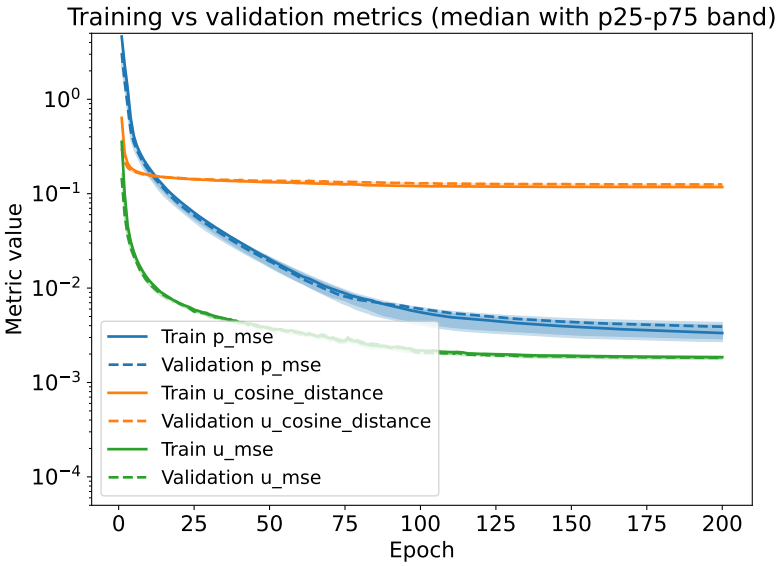
A K-fold cross-validation assesses whether the training dataset is sufficiently large and diverse. The training dataset is divided into $K = 10$ equally-sized folds. Then, $K = 10$ independent training runs are performed whereby one fold is used as the validation dataset, and the remaining nine are collated together and used as the training set. The values of loss and metrics are averaged across the ten runs epoch by epoch and depicted in Figures S1 (for the UNet architecture) and S2 (for the SWIN). The loss and metric values from the final epoch are stratified by percentile bands and presented in the summary table below, in Section 1.2.1. From the graphs in Figures S1 and S2, the training and validation curves are broadly consistent across the folds, with no obvious signs of severe overfitting. This suggests that, for this proof-of-concept model, the training dataset is sufficiently varied to learn a coarse representation of the flow patterns. The SWIN architecture yields better results than the UNet for pressure, clearly better for velocity magnitude MSE (around one order of magnitude lower), comparable cosine distance and a slightly better loss.

1.2 Progressive inference stress tests

The two architectures are trained again on a random train/validation split of 90/10%. After training, they are used to perform inference on examples not seen during training or validation in order to assess GLoOD's ability to generalize. Specifically, one in-distribution (ID) set of 86 examples and one out-of-distribution (OOD) set of 104 examples are produced in the



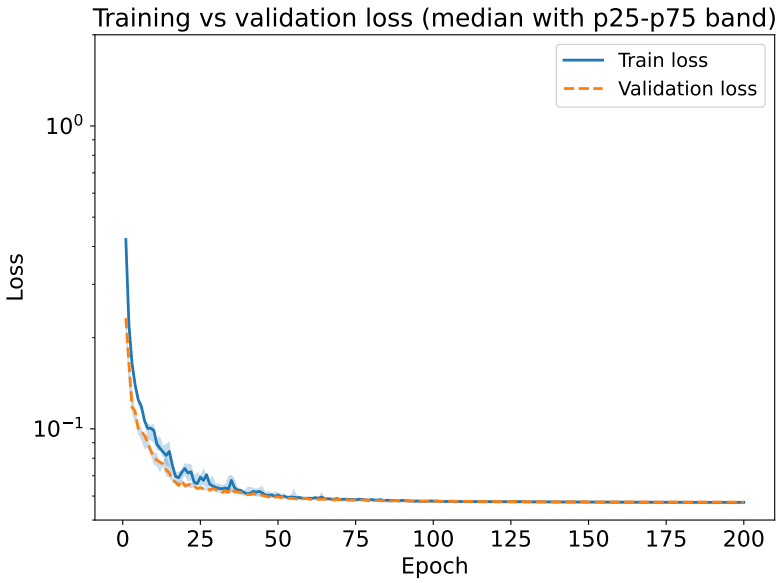
(a) UNet, loss.



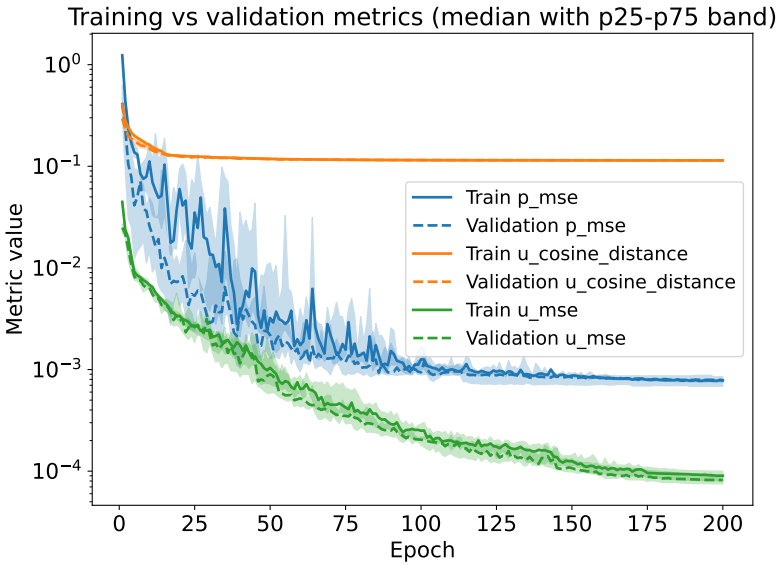
(b) UNet, metrics.

Supplementary Fig. S1: K-fold cross-validation for the UNet architecture.

same way as the training set. The OOD examples present square instead of circular obstacles. In addition, inference is performed on a further dataset of 200 examples, here denoted as



(a) SWIN, loss.



(b) SWIN, metrics.

Supplementary Fig. S2: K-fold cross-validation for the SWIN architecture.

"Cavity", which is fully distinct from the training distribution. The sampled Reynolds-number distributions are expressed on the physical scale: Circles has mean $\mathrm{Re} = 166.8$ and coefficient

of variation (CV) of 0.61, Squares has mean $\mathrm{Re} = 214.8$ and CV of 1.83, and Cavity has mean $\mathrm{Re} = 12{,}252.8$ and CV of 0.91. In this dataset, the flow takes place in a lid-driven cavity, and both the lid position and lid direction are sampled randomly. Two further von Karman vortex-shedding tests, with different output timesteps, are analysed separately in Section 1.2.8, where the main diagnostic is temporal lag rather than aggregate reconstruction loss alone. The standard-timestep von Karman set, which uses the same timestep as the other tests, contains 48 examples and has mean $\mathrm{Re} = 749.0$ and CV of 0.28, whereas the reduced-timestep `vtkp06` von Karman set, whose timestep is $1/5$ of the original value, contains 365 examples and has mean $\mathrm{Re} = 774.9$ and CV of 0.27. All inference results are also compared against an identity-ablation baseline. In this ablation, the temporal backbone is removed: the decoder at timestep t produces an embedding which is then passed directly to the decoder for timestep $t+1$, without the over-time self-attention block that in the full model mixes embeddings across the rollout. The identity baseline therefore tests whether simply forwarding the previous latent state is sufficient, or whether temporal self-attention between embeddings is materially contributing to the prediction.

1.2.1 Inference loss and metrics

The summary plot is reported in the main text, while the complete numerical table is reproduced here for reference. Overall, the transformer-backbone inference losses for the "Circles" and "Squares" datasets remain close to the K-fold values, with the OOD "Squares" set even yielding slightly lower aggregate losses than the K-fold medians. This suggests that the model retains a non-trivial ability to generalise to obstacle geometries not seen during training. By contrast, the "Cavity" case shows a much larger loss, dominated by the pressure MSE, indicating that this configuration lies substantially farther from the training distribution. The identity-backbone runs, for which no K-fold experiment was performed, are consistently worse than the corresponding transformer-backbone runs on all three inference datasets. On the circle and square tests, the deterioration is typically about one order of magnitude in aggregate loss, one to almost three orders of magnitude in pressure MSE, and one to two orders of magnitude in velocity MSE; in the most extreme case, the cavity UNet-identity run increases the velocity MSE by nearly three orders of magnitude.

A more detailed view is provided by the individual metrics. For the "Circles" and "Squares" datasets, the pressure error remains small, while the velocity-magnitude error, measured through u MSE, is consistently lower for SWIN than for UNet. The cavity case instead produces very large pressure errors for both architectures, although SWIN still reduces the velocity-direction and velocity-magnitude errors. The cosine-distance metric should be interpreted cautiously: in regions where the velocity magnitude is close to zero, and in particular inside the obstacles, the velocity direction is effectively undefined; consequently, small perturbations can produce locally arbitrary directions and hence substantial cosine-distance noise. Taken together, the numerical results suggest that the present proof-of-concept model preserves the large-scale flow structure on the circle and square test sets, while the cavity configuration exposes a clear out-of-distribution limitation. The systematic gap with respect to the identity ablation also indicates that the temporal self-attention strategy over the latent embeddings is clearly useful: the backbone is not merely forwarding information, but is materially improving the next-state prediction.

Supplementary Table S1: Summary of losses and metrics for the K-fold and inference runs. Identity backbones are reported only for inference, because no K-fold runs were performed for them.

Set	Run	n	Loss	p MSE	u MSE	u cos distance
K-fold, median	UNet	–	0.063934	0.0039052	0.0018225	0.12524
	SWIN	–	0.057338	0.00078758	8.1868e-05	0.11437
K-fold, Q1	UNet	–	0.063069	0.0029075	0.0017548	0.1238
	SWIN	–	0.056452	0.00068478	7.5077e-05	0.11262
K-fold, Q3	UNet	–	0.064566	0.0049028	0.0018613	0.12623
	SWIN	–	0.058057	0.00084874	8.86e-05	0.11578
K-fold, lower whisker	UNet	–	0.060824	0	0.0015951	0.12016
	SWIN	–	0.054045	0.00043885	5.4793e-05	0.10787
K-fold, upper whisker	UNet	–	0.066811	0.0078958	0.0020211	0.12988
	SWIN	–	0.060465	0.0010947	0.00010888	0.12052
Circles	SWIN	86	0.061387	0.00018631	8.1673e-05	0.12266
Circles	SWIN identity	86	0.079335	0.011782	0.015264	0.14457
Circles	UNet	86	0.062709	9.8438e-05	0.0028966	0.12345
Circles	UNet identity	86	0.14723	0.056728	0.040067	0.24883
Squares	SWIN	104	0.051254	0.00035126	0.00014487	0.10229
Squares	SWIN identity	104	0.066597	0.0070873	0.010213	0.12402
Squares	UNet	104	0.052053	0.00025364	0.0018019	0.10282
Squares	UNet identity	104	0.13279	0.041325	0.032803	0.22993
Cavity	SWIN	200	472.75	2836.2	0.0099868	0.086619
Cavity	SWIN identity	200	504.39	3025.8	0.030906	0.15497
Cavity	UNet	200	506.94	3041	0.015642	0.19216
Cavity	UNet identity	200	820.92	4896.7	12.884	1.0289

1.2.2 Inspection of the produced output

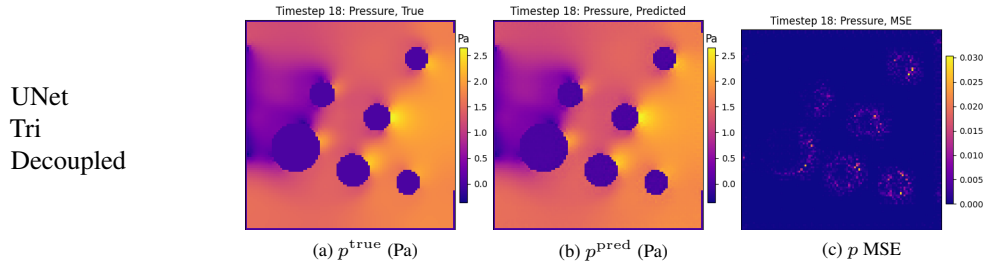
Two samples, one "constant" and one "increasing", from the ID dataset with circular obstacles, an "increasing" from the OOD dataset with square obstacles, and one example from the cavity dataset are selected and passed through both models. The predicted fields are then compared visually against the corresponding true fields.

Representative true sequences for these regimes are collected in the main text. The supplementary material therefore focuses directly on predicted-versus-true comparisons and error maps.

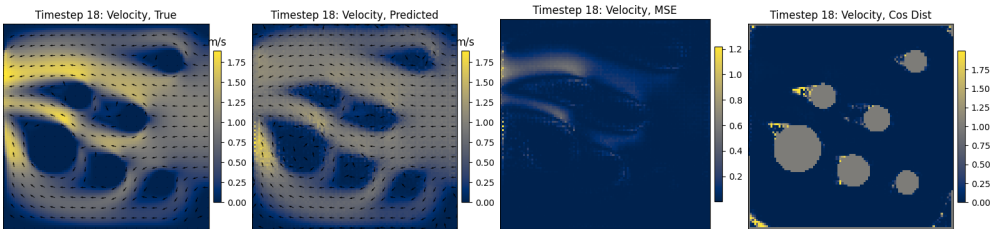
1.2.3 In-distribution: circular obstacles

The "constant" ID example 15 is shown in Figure S3. The predicted fields qualitatively reproduce the true pressure and velocity fields, with discrepancies occurring predominantly in the UNet architecture. These discrepancies mainly consist of noise at the borders of the obstacles and around the borders of the inlets and outlets, as well as overestimation of the velocity patterns. In particular, the models correctly recover regions of near-zero velocity inside obstacles and the approximate location of inlet and outlet openings, even though these

UNet
Tri
Decoupled

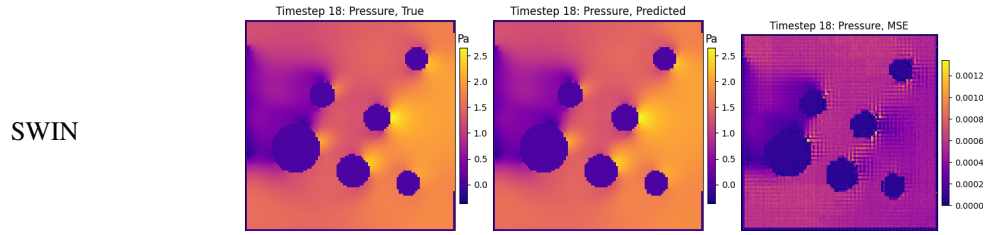


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

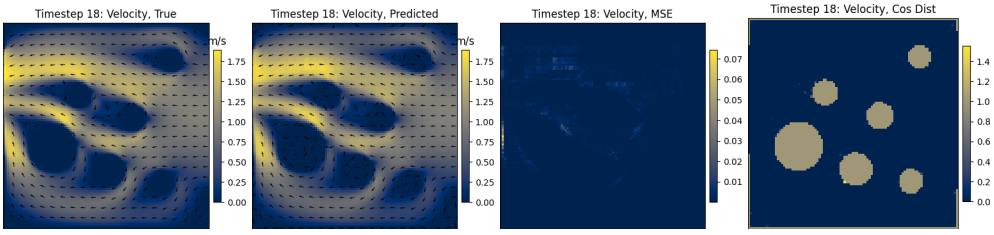


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



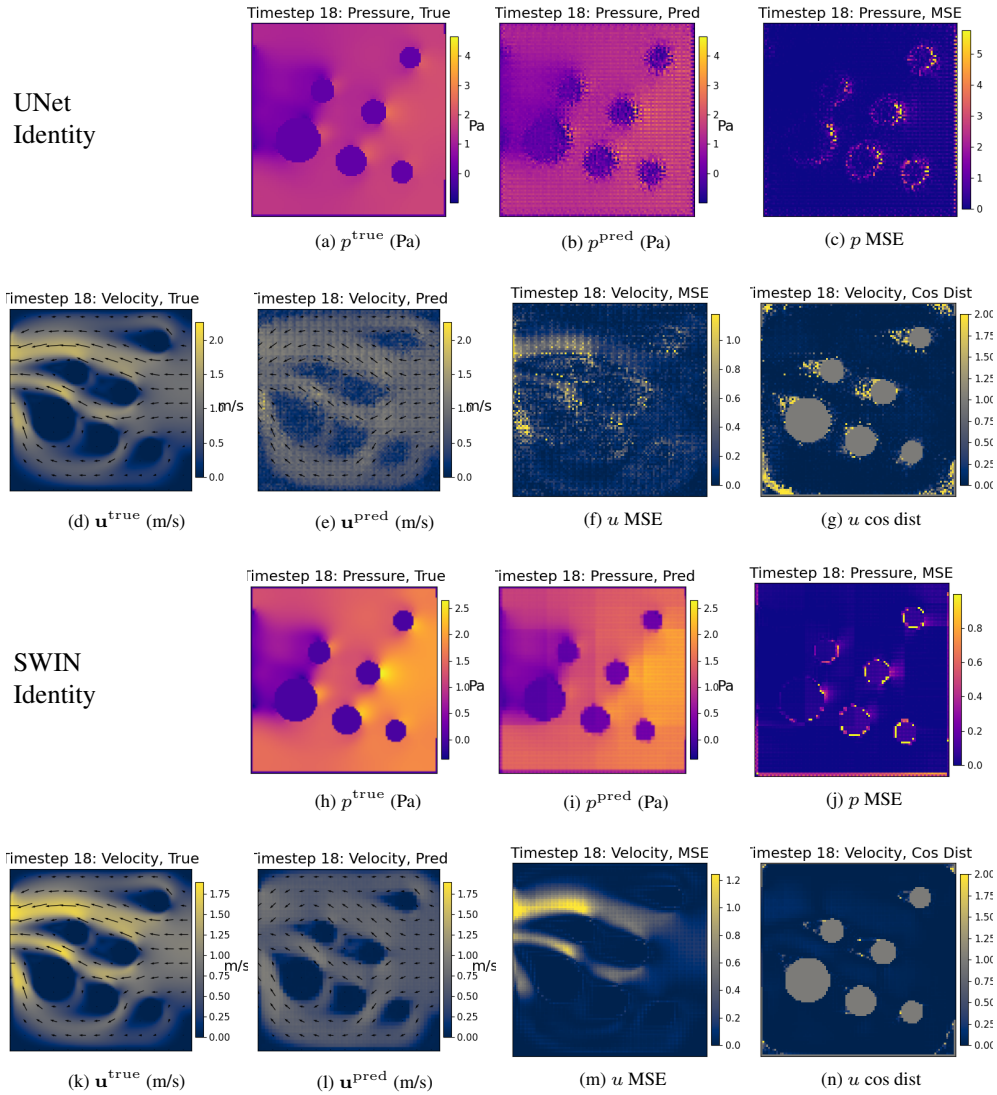
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE



(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S3: ID Circles example 15, timestep 18, “constant”.

geometrical features are not provided through explicit masks or coordinate inputs, but are only available through their implicit signatures in the input pressure fields. The magnitude of the predicted velocity field is zero inside the obstacles, but as mentioned above in Section 1.2.1, the lack of directional information in those zones leads to higher cosine distance and contributes to noisier, harder-to-minimise loss functions. The corresponding identity-ablation snapshots in Figure S4 are visibly worse: the broad structure is still partially recognisable, but both pressure



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

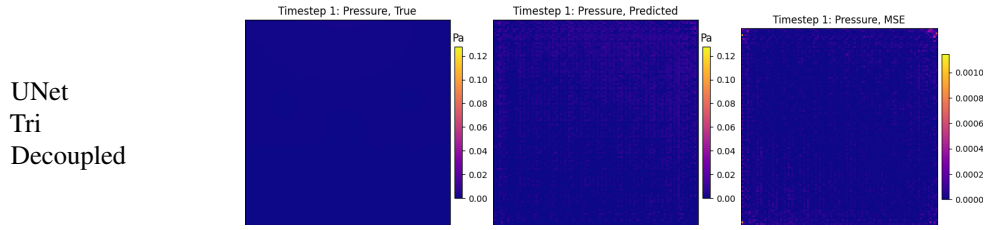
(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S4: ID Circles identity-ablation example 15, timestep 18, "constant".

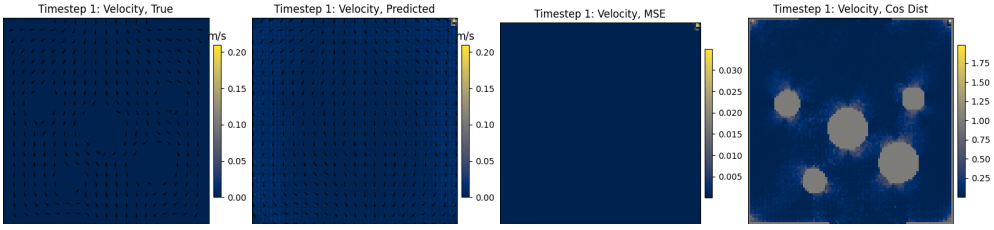
and velocity are less coherent, and the errors are no longer confined as cleanly to narrow boundary regions.

The "increasing" ID example 14 is shown in Figures S5, S6 and S7. Again, the main flow patterns are qualitatively reproduced and the same considerations as in the "constant" case apply, including visible noise near boundaries. The temporal increase in flow intensity appears to be recovered across the three timesteps. As before, the models still recover the approximate layout of obstacles and boundary openings from the pressure and velocity fields

UNet
Tri
Decoupled

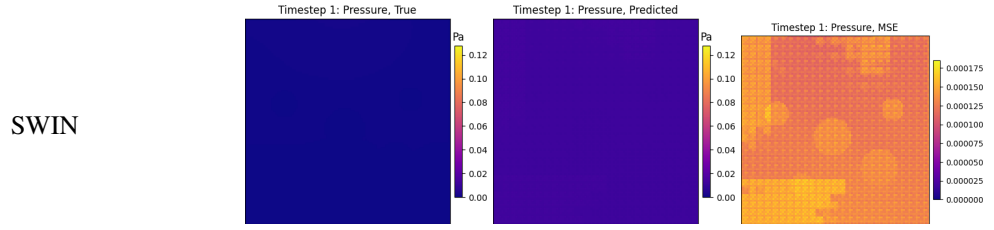


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

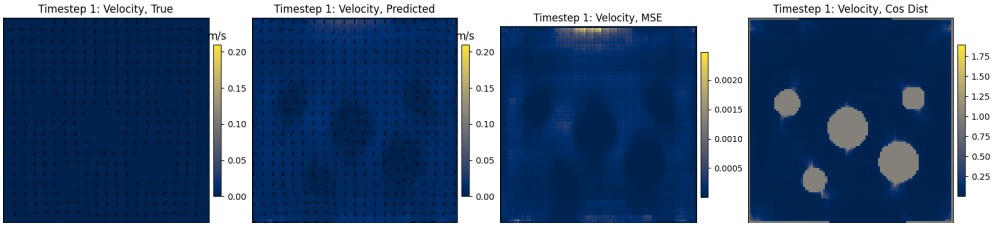


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

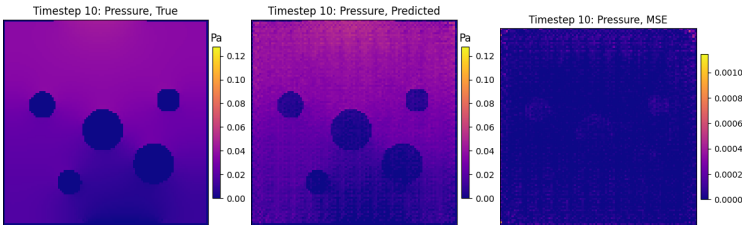


(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

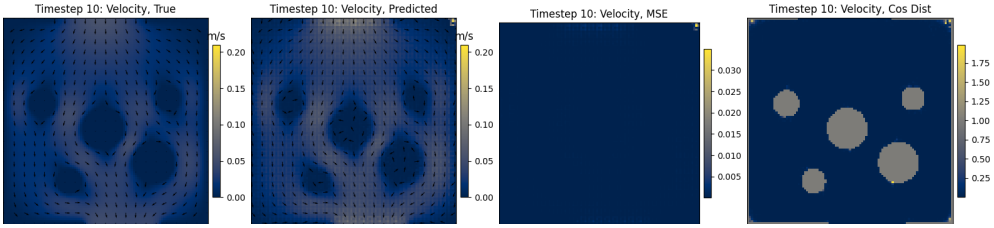
Supplementary Fig. S5: ID Circles example 14, timestep 1, "increasing".

alone, without separate geometric input channels. A clear qualitative difference between the two architectures is that the UNet produces periodic tile-like noise, whereas this artefact is not apparent in the SWIN predictions. This effect is visible in both pressure and velocity, but it is more marked in pressure. Consistently with this visual impression, the colourbar ranges of the p MSE and u MSE panels for SWIN are roughly one-half to one-third of those for UNet, while the cosine-distance maps appear broadly similar between the two models. The

UNet
Tri
Decoupled

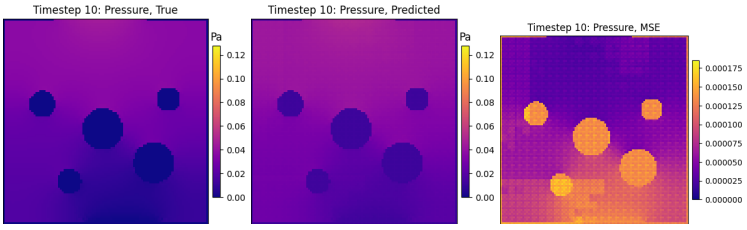


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

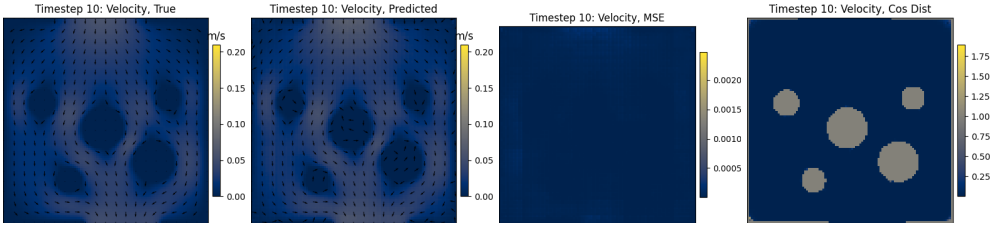


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE



(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

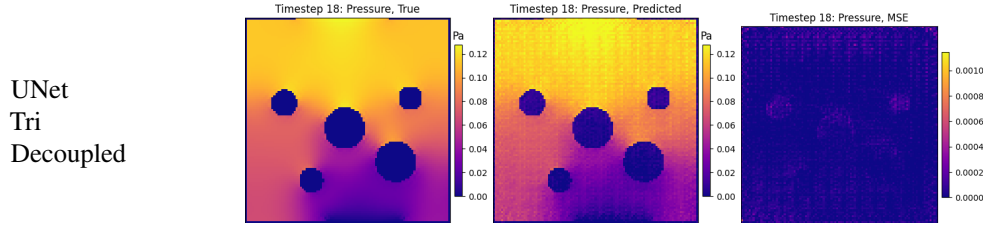
Supplementary Fig. S6: ID Circles example 14, timestep 10, “increasing”.

identity-ablation counterparts in Figures S8–S10 remain recognisable, but the fields are visibly less sharp and the error maps are substantially broader, again supporting the conclusion that direct latent forwarding is not competitive with the temporal-attention backbone.

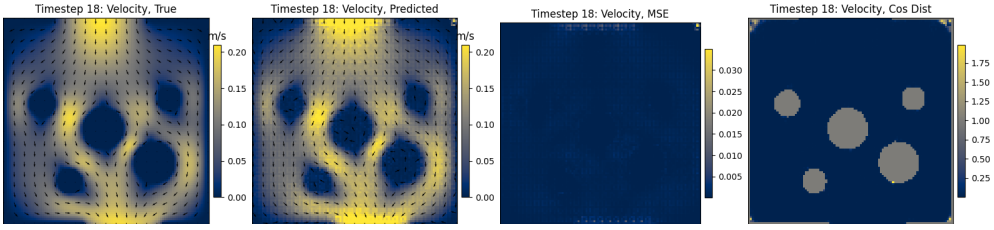
1.2.4 Out-of-distribution: square obstacles

The “increasing” OOD example 98 is shown in Figures S11, S12 and S13. The qualitative

UNet
Tri
Decoupled

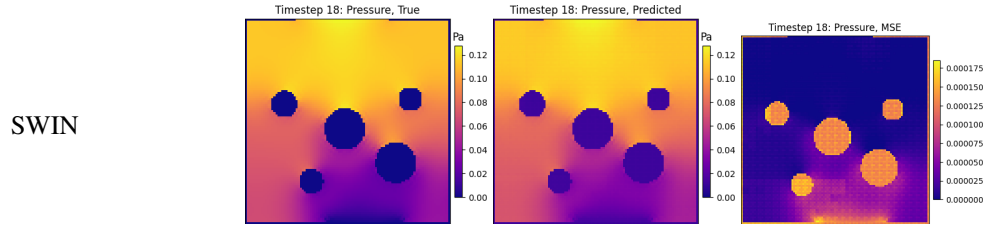


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

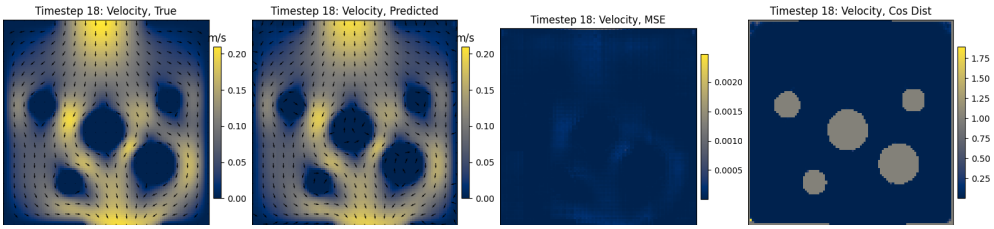


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



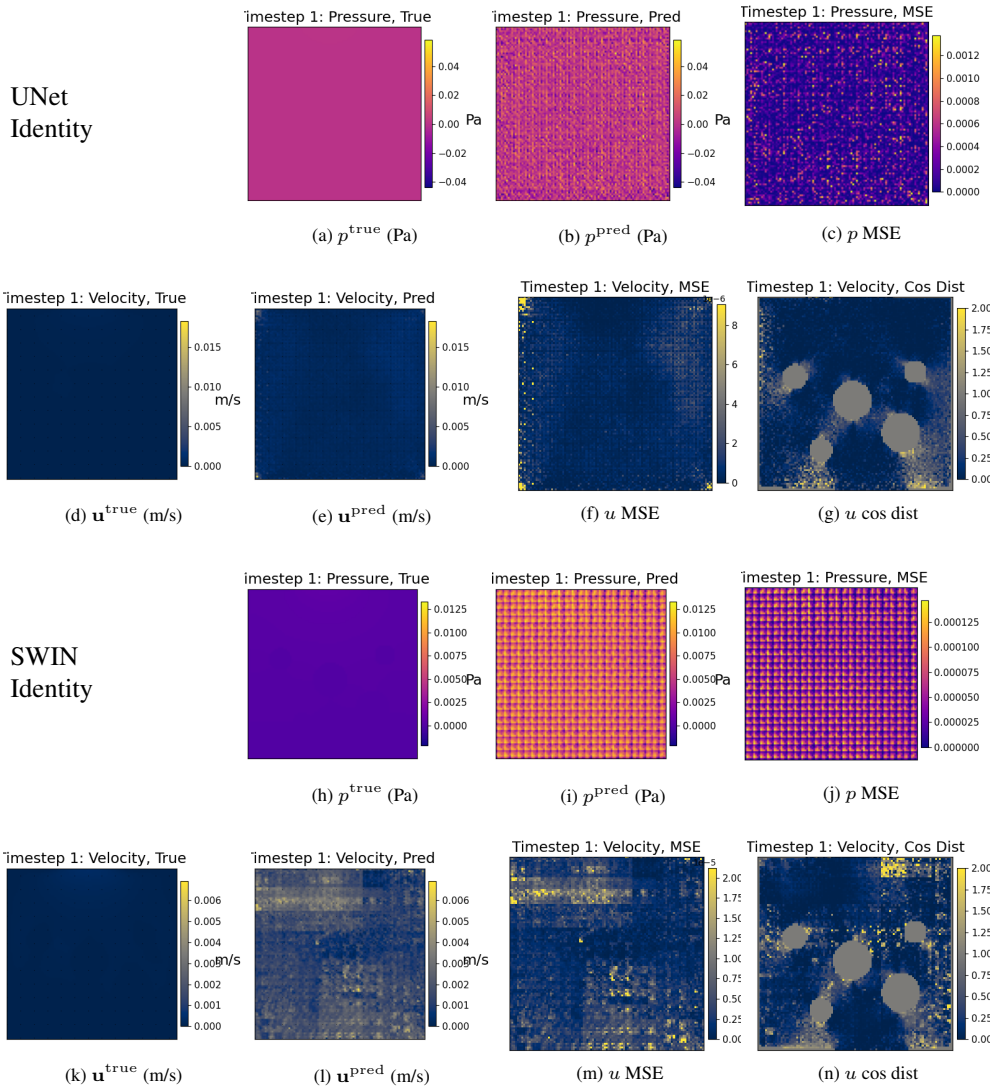
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE



(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S7: ID Circles example 14, timestep 18, "increasing".

behaviour is similar to that of the "increasing" ID example: the main flow structures are reproduced, the approximate position and extent of the square obstacles are recovered, and the tile-like artefact remains visible only in the UNet predictions. As in the ID case, this artefact affects both pressure and velocity, but is more evident in pressure. Consistently with the numerical summary and the main-text summary figure, the qualitative advantage of SWIN over UNet is clearer in the pressure field than in the velocity magnitude: the p MSE panels



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

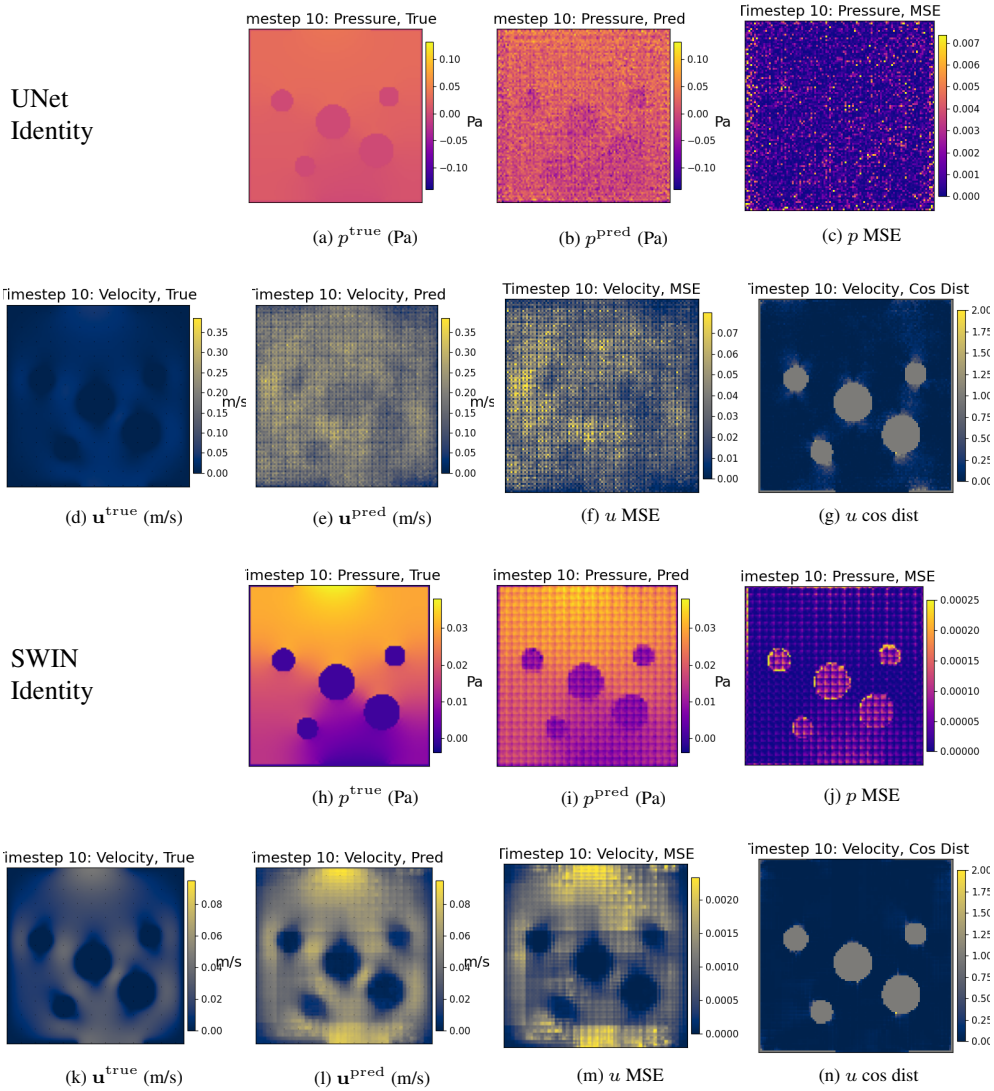
(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S8: ID Circles identity-ablation example 14, timestep 1, "increasing".

indicate a marked reduction of the error scale for SWIN, whereas the corresponding difference is less evident in the u MSE panels. The cosine-distance maps, by contrast, remain broadly comparable between the two models. The identity-ablation square examples in Figures S14–S16 are again clearly worse, with less stable pressure structure and broader velocity errors, so the temporal backbone remains important even in the OOD geometry case.



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

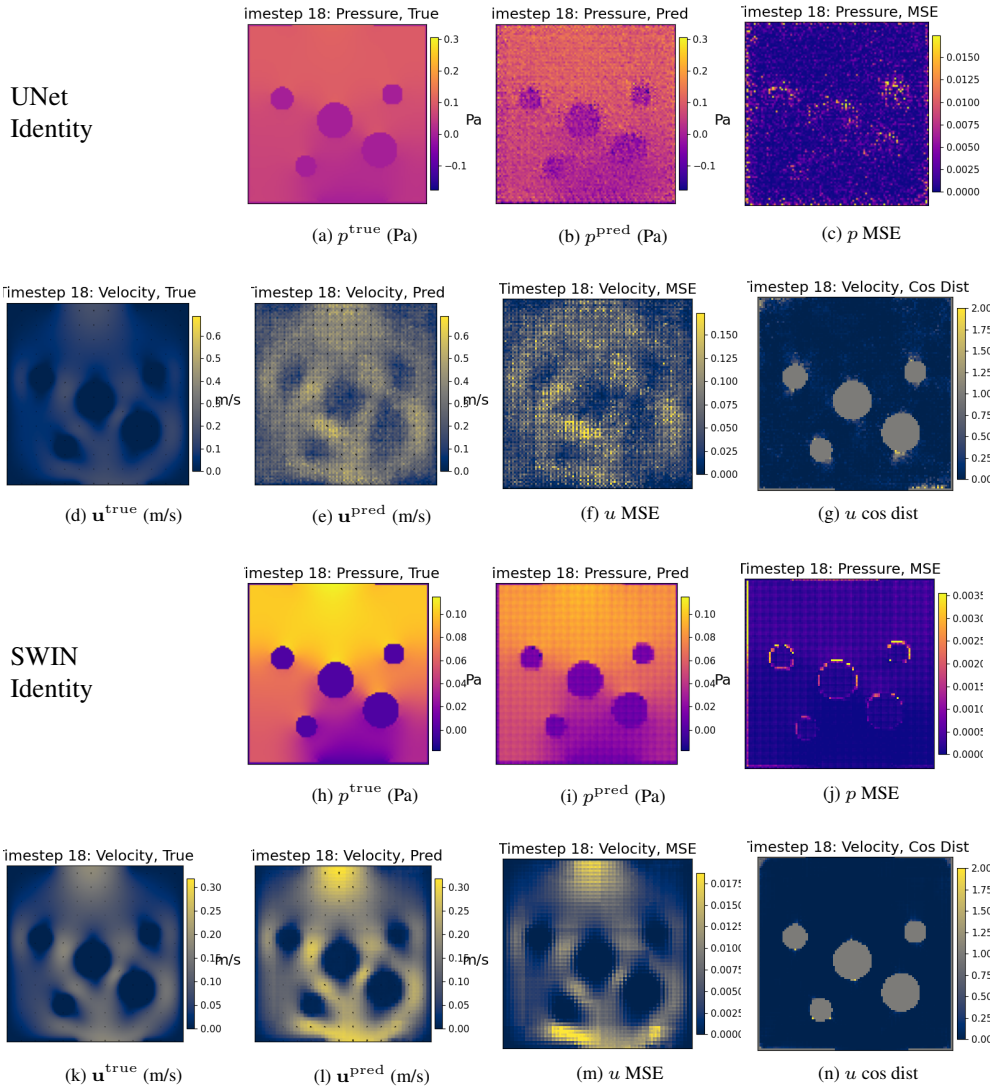
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S9: ID Circles identity-ablation example 14, timestep 10, “increasing”.

1.2.5 Out-of-distribution: lid-driven cavity

The cavity example 48 is shown in Figures S17, S18 and S19. A further final-timestep cavity example, example 49, is shown in Figure S23. The cavity case is visually less aligned with the training distribution than the circles and squares examples. This is consistent with the aggregate metrics in Section 1.2.1, where the cavity dataset produces substantially larger pressure errors. The visual panels also clarify why the pressure MSE is especially severe in



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

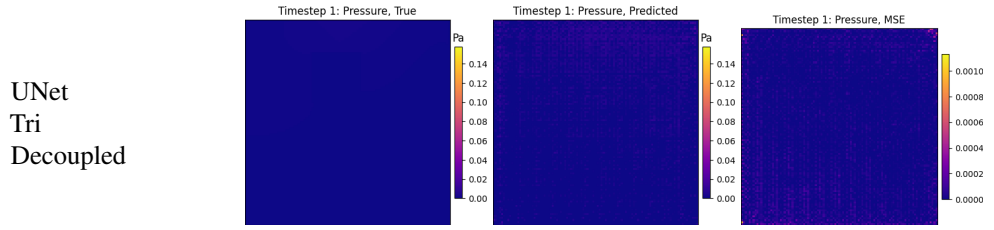
(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S10: ID Circles identity-ablation example 14, timestep 18, "increasing".

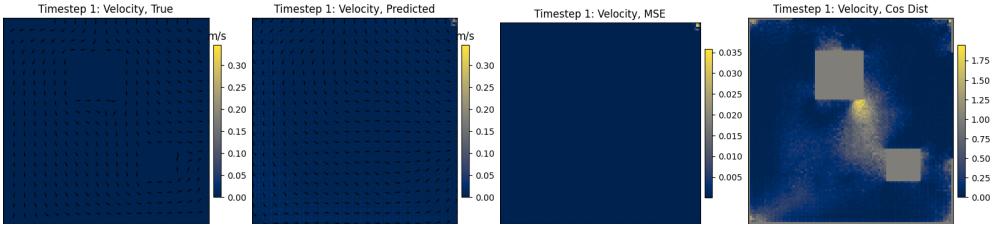
this case. In the cavity examples, the true pressure field has a much lower dynamic range than in the obstacle-flow datasets and is visibly affected by structured numerical noise. In this regime, the pressure MSE should not be interpreted only as a failure to recover the broad cavity pattern: small offsets, smoothing of weak structures and residual local artefacts can dominate the squared error when the coherent pressure signal itself is weak.

The same issue is present to a lesser extent in the velocity field. Example 48 retains a more readable cavity circulation: across the three timesteps, both architectures correctly

UNet
Tri
Decoupled

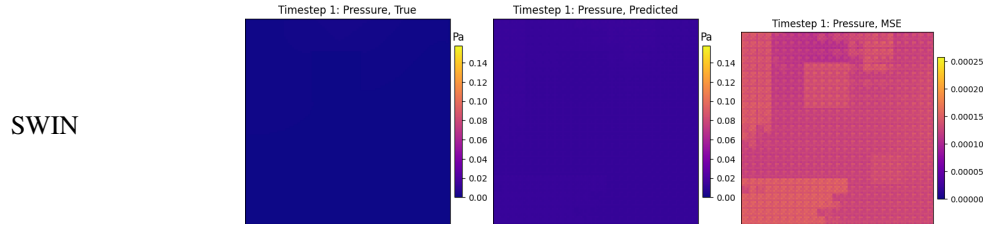


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

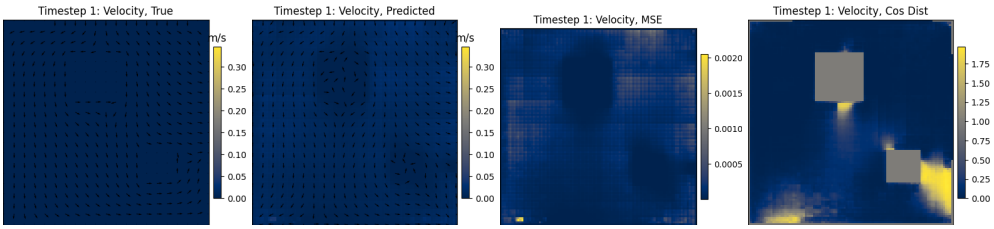


(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

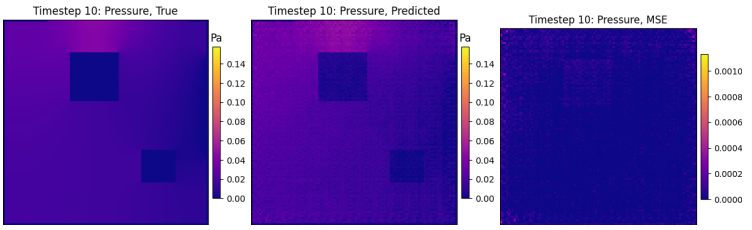


(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

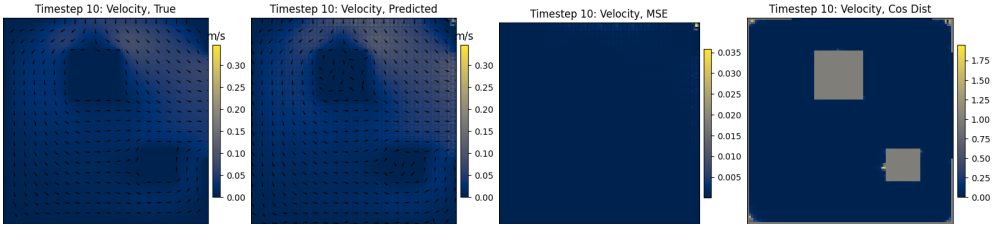
Supplementary Fig. S11: OOD Squares example 98, timestep 1, “increasing”.

reproduce the expansion of the main vortex, while local differences in field intensity and fine-scale structure remain visible. Example 49 at timestep 18 provides a clearer low-field case for velocity, with most of the domain carrying only weak motion and the strongest signal concentrated near the moving lid. Both architectures still recover the qualitative organisation of the flow, but weak velocity magnitudes make local direction and magnitude errors less stable. The cavity identity-ablation snapshots in Figures S20–S22 are clearly less stable still, with

UNet
Tri
Decoupled

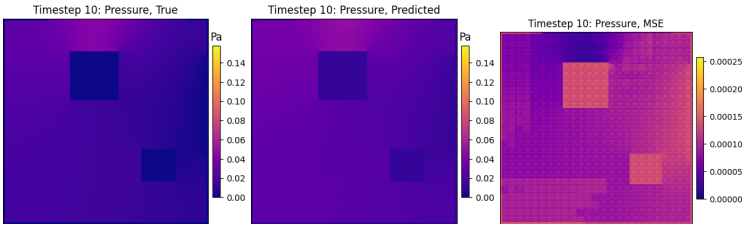


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

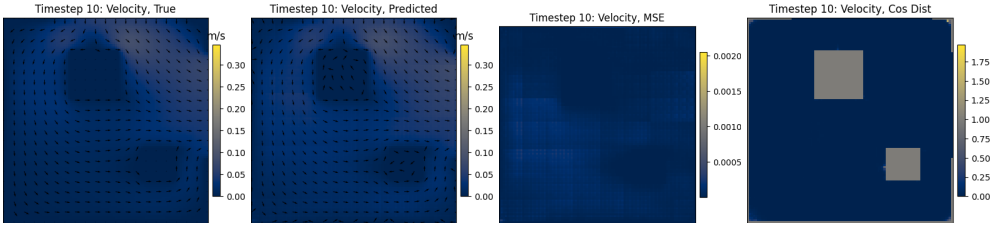


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE



(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

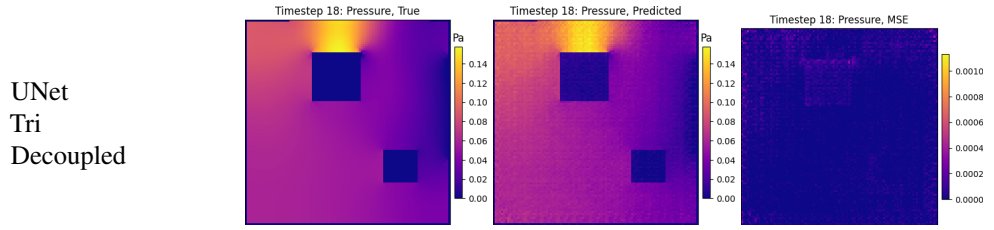
Supplementary Fig. S12: OOD Squares example 98, timestep 10, “increasing”.

weaker structural coherence in both pressure and velocity, consistent with the large quantitative deterioration seen in the numerical summary table above.

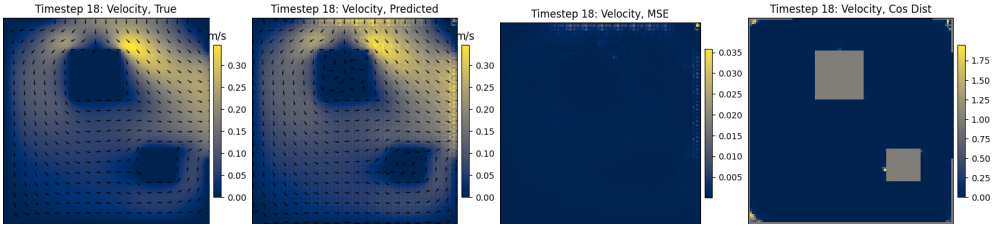
1.2.6 Physical diagnostics: drag force

Pointwise metrics and visual inspection indicate that both architectures reproduce the broad flow structure, but they do not fully resolve which model better preserves physically relevant

UNet
Tri
Decoupled

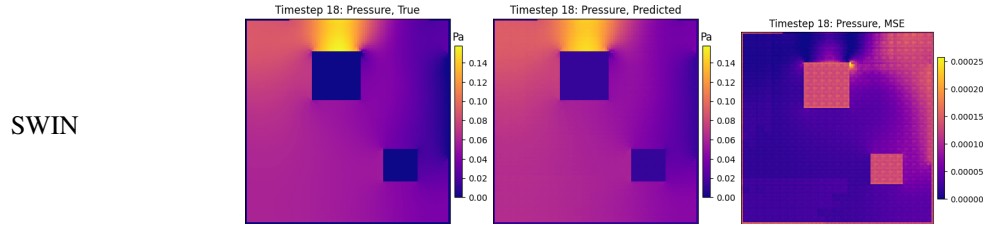


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

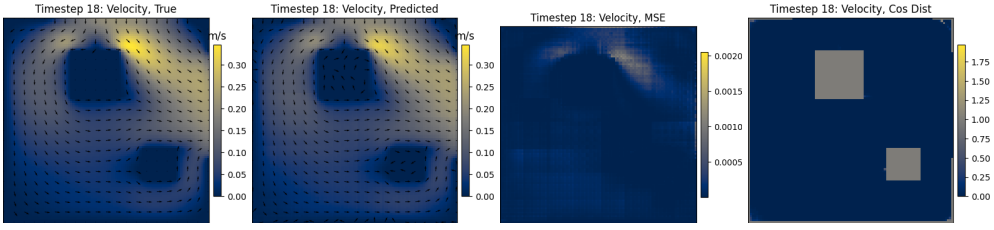


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

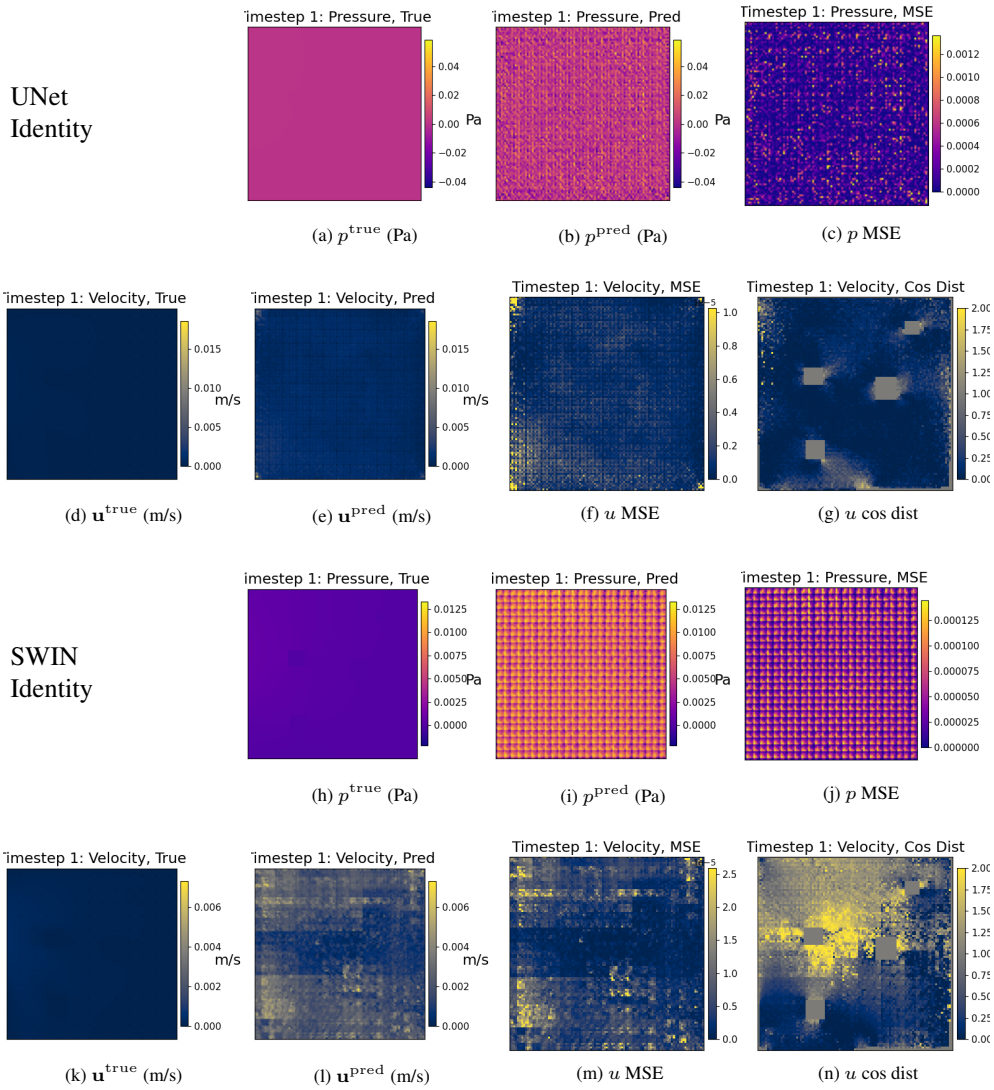


(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S13: OOD Squares example 98, timestep 18, “increasing”.

derived quantities. In the present setting, fieldwise pressure and velocity errors alone are therefore not sufficient to assess the models comprehensively. For this reason, a complementary analysis is carried out on the drag force, evaluated on the ID (circles) and OOD (squares) inference datasets.

Let Γ denote the union of the boundaries of all obstacles present in a given example, and let $\mathbf{n}$ be the outward unit normal to these boundaries. The total drag-force vector computed



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

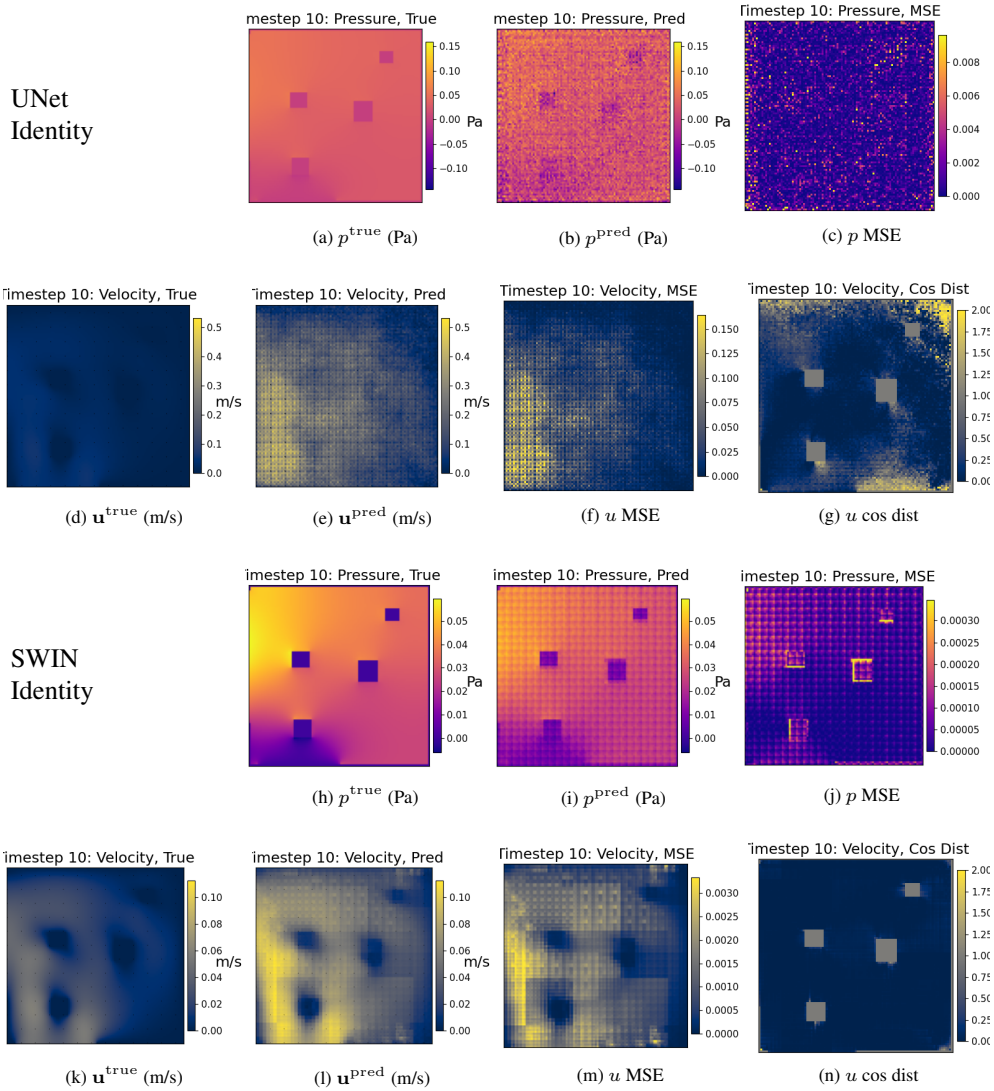
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S14: OOD Squares identity-ablation example 18, timestep 1, "increasing".

here is therefore the aggregate drag over all obstacles in the domain, and it is decomposed into pressure and viscous contributions as

$$\mathbf{F} = \mathbf{F}_p + \mathbf{F}_v, \tag{1}$$



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

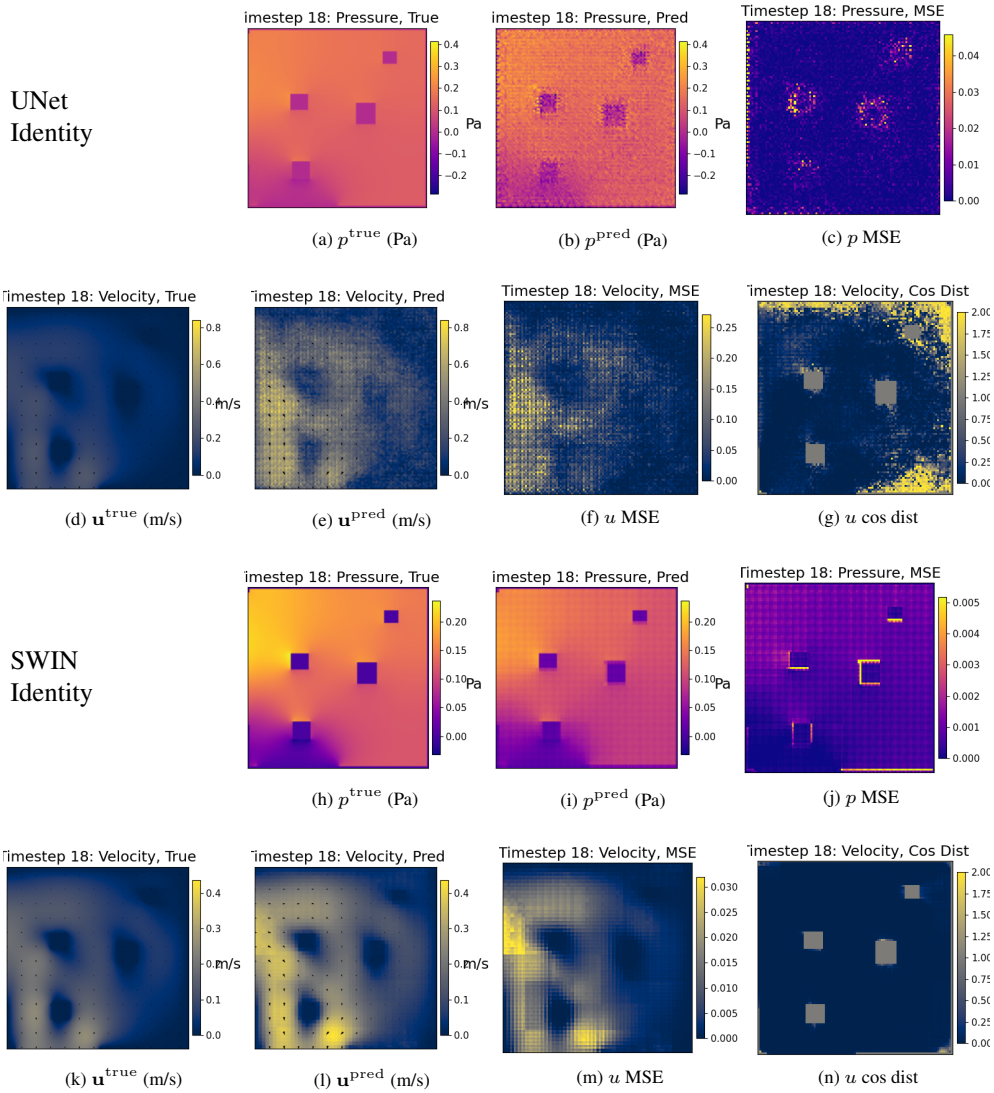
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S15: OOD Squares identity-ablation example 18, timestep 10, "increasing".

with, in non-dimensional form:

$$\mathbf{F}_p = -\int_\Gamma p\,\mathbf{n}\,\mathrm{d}s, \qquad \mathbf{F}_v = \frac{1}{\mathrm{Re}}\int_\Gamma \left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)\mathbf{n}\,\mathrm{d}s. \tag{2}$$



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

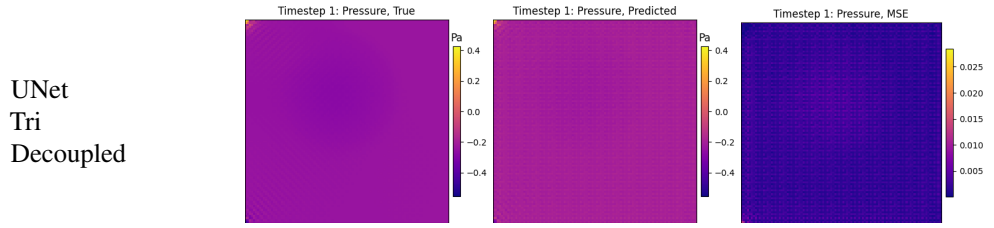
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

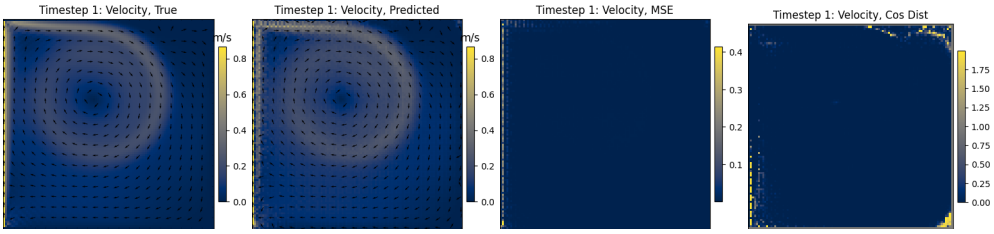
Supplementary Fig. S16: OOD Squares identity-ablation example 18, timestep 18, "increasing".

In this formulation, the viscous term carries a Reynolds-number-dependent prefactor. However, the drag assessment is based on relative errors between predicted and true quantities for the same inferred case, so this prefactor is expected to cancel to leading order provided that the effective prefactors in the predicted and true fields do not differ substantially. This assumption is supported by the results of Section 1.2.2, which show that the main flow patterns are recovered with reasonable fidelity. The drag errors can therefore be interpreted primarily as

UNet
Tri
Decoupled

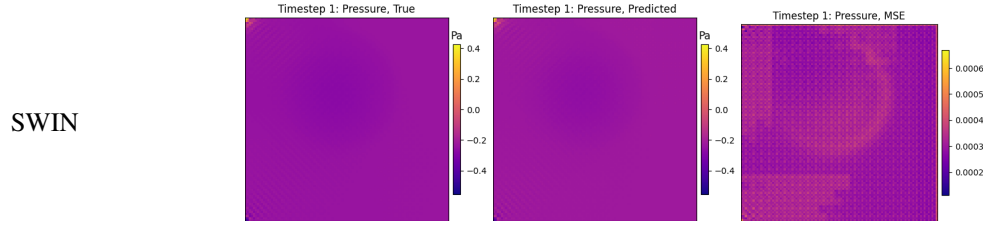


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

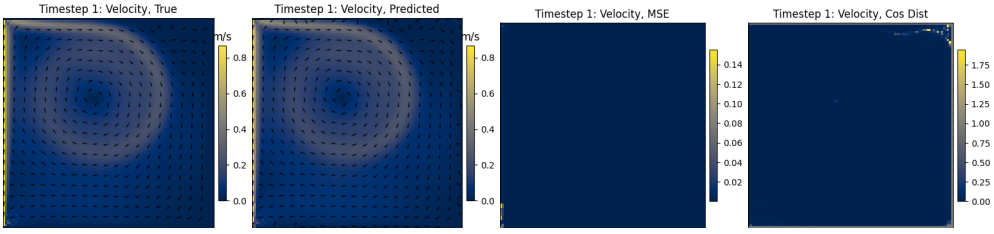


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE



(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

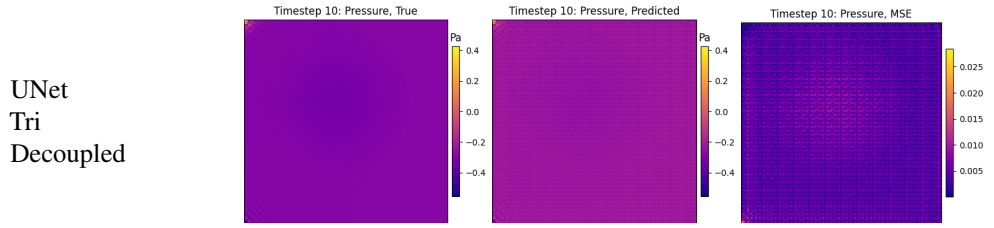
Supplementary Fig. S17: Cavity example 48, timestep 1.

reflecting differences between p^{pred}, $\mathbf{u}^{\mathrm{pred}}$ and their true counterparts, rather than differences in the global Reynolds scaling.

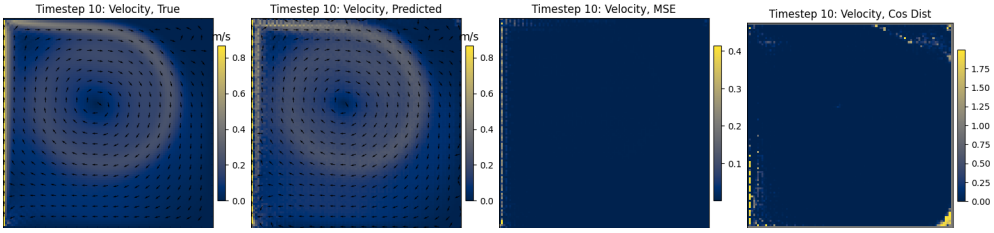
The analysis reported in this section is performed on the absolute relative error of a given quantity, namely:

$$e_{rel}(q) = \frac{\left|q^{\mathrm{pred}} - q^{\mathrm{true}}\right|}{\left|q^{\mathrm{true}}\right| + \varepsilon}, \tag{3}$$

UNet
Tri
Decoupled

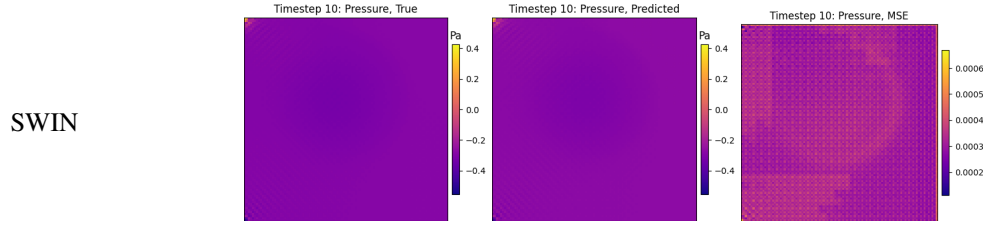


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

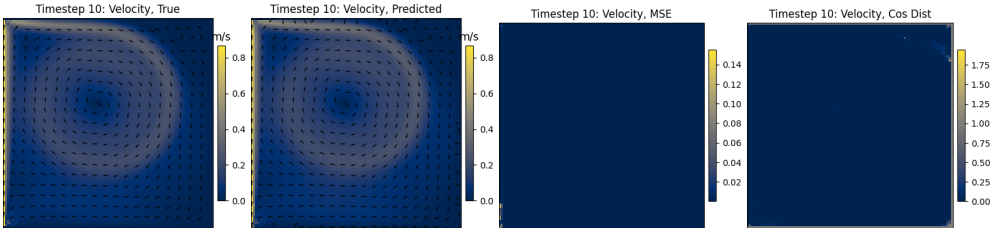


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

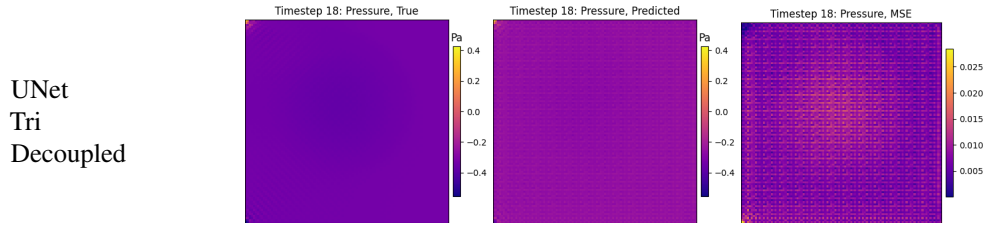


(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

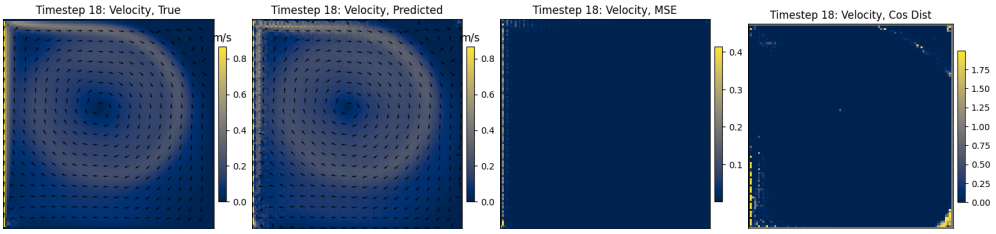
Supplementary Fig. S18: Cavity example 48, timestep 10.

where q^{pred} and q^{true} denote respectively the predicted and true values of the quantity of interest, and ε is a small positive constant introduced to avoid division-by-zero in the vanishing-signal limit.

UNet
Tri
Decoupled

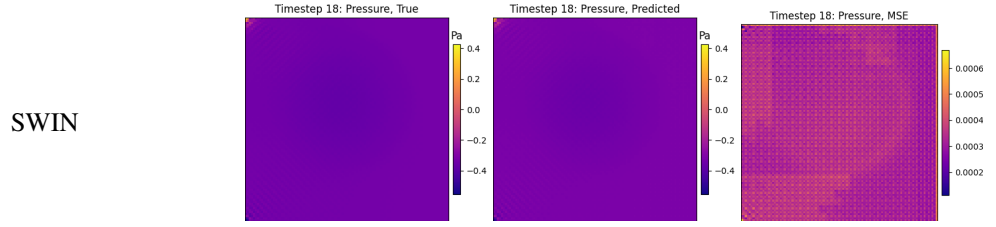


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

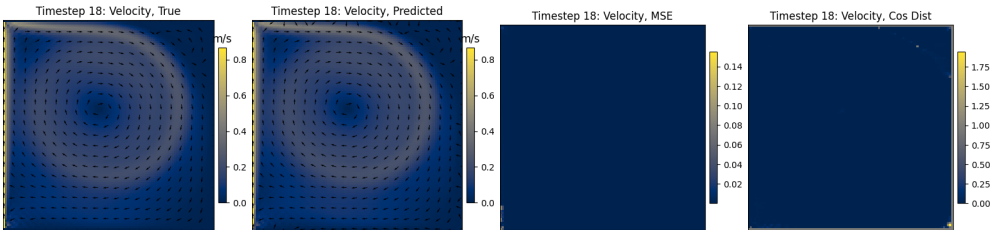


(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

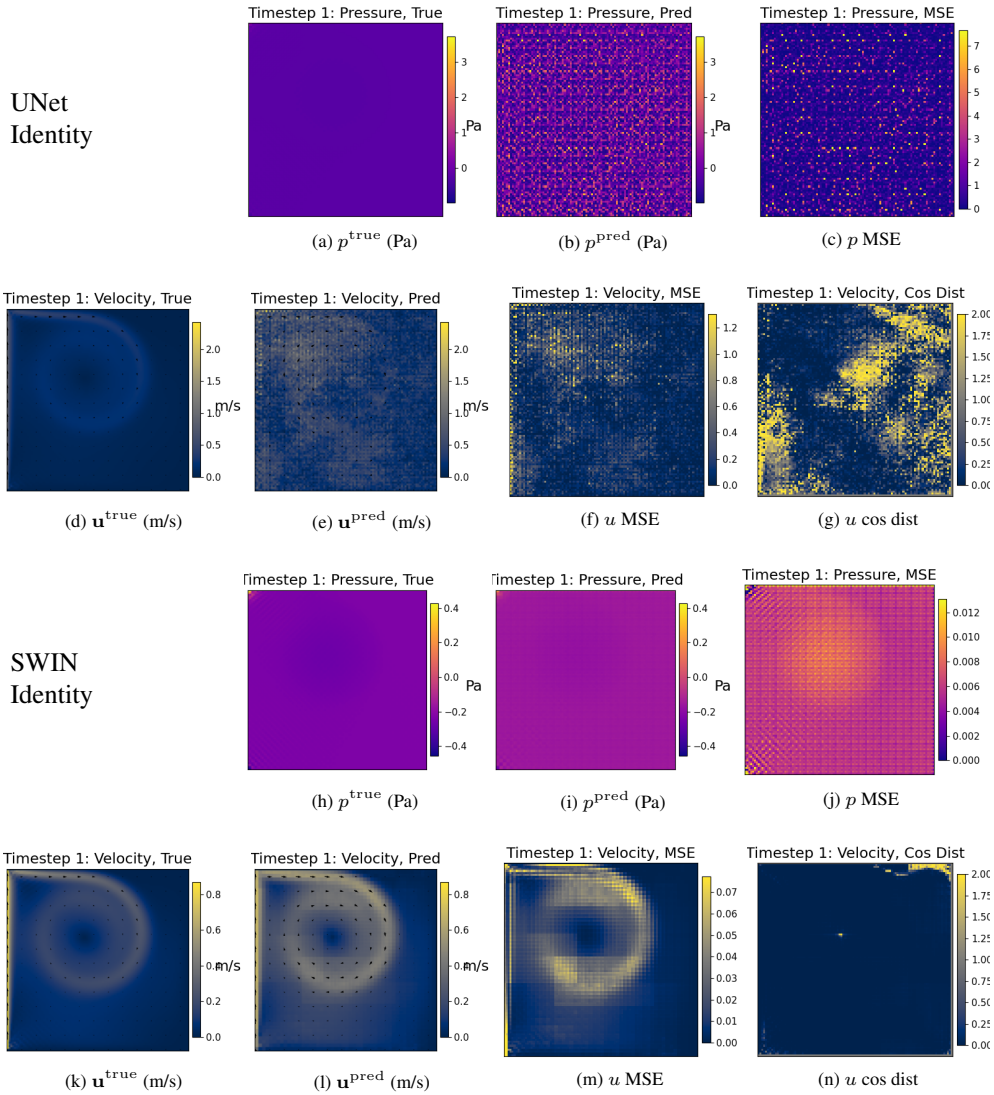


(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S19: Cavity example 48, timestep 18.

Sensitivity analysis

A sensitivity analysis is performed to assess the impact of two effects on the drag predictions. The first is the role of signal intensity, *viz.,* whether very weak or very strong signals are associated to less stable errors. To assess this, Table S2 reports the absolute relative error on the total drag magnitude, stratified by percentile band. Compared to the interquartile range (25th–75th percentile), the second band (5th–25th percentile) exhibits higher errors



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

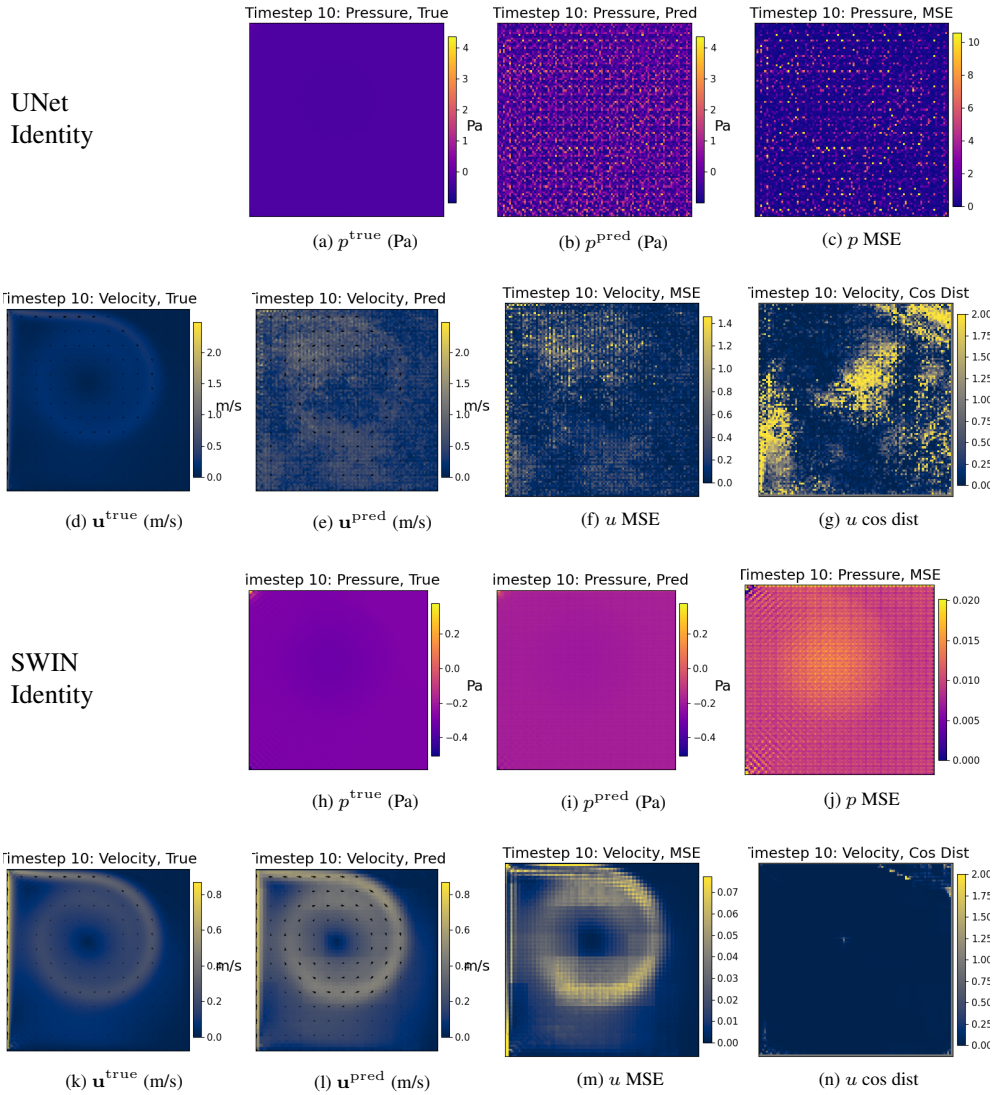
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S20: Cavity identity-ablation example 48, timestep 1.

and therefore performs worse, while the first band (<5th percentile) performs substantially worse still. This suggests that the models have difficulty predicting low- and very-low-intensity signals. By contrast, the upper two bands (75th–95th percentile and >95th percentile) do not show a comparable deterioration: their errors are either close to the interquartile band or lower, especially for SWIN.

Second, the impact of rollout position is assessed. In other words, the first predicted timesteps are expected to suffer from reduced temporal context compared to the central and



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

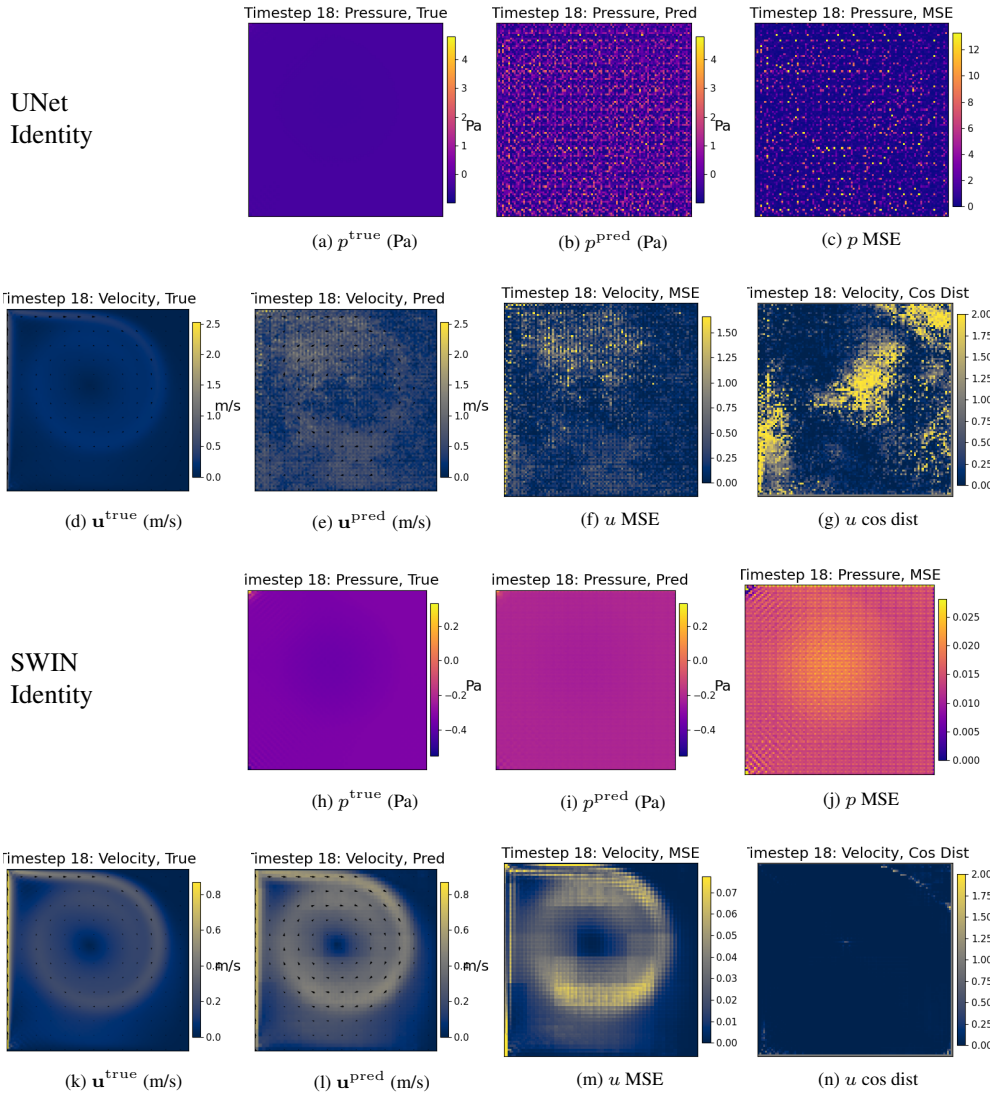
(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S21: Cavity identity-ablation example 48, timestep 10.

final ones, and are therefore expected to perform worse. Figures S24, S25, S26, and S27 show how the component-wise absolute relative error on the pressure- and viscous-force vectors, $\mathbf{F}_p$ and $\mathbf{F}_v$, changes when the first k timesteps are removed from the aggregate calculation.

These figures show that removing the first timesteps reduces the upper-percentile errors across most force components, especially around the 75th and 95th percentiles. The lower-percentile curves are instead comparatively flat. This suggests a rollout-related deterioration in predictive capability at the beginning of the sequence, affecting mainly the less stable tail of



(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

(d) $\mathbf{u}^{\text{true}}$ (m/s) (e) $\mathbf{u}^{\text{pred}}$ (m/s) (f) u MSE (g) u cos dist

(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE

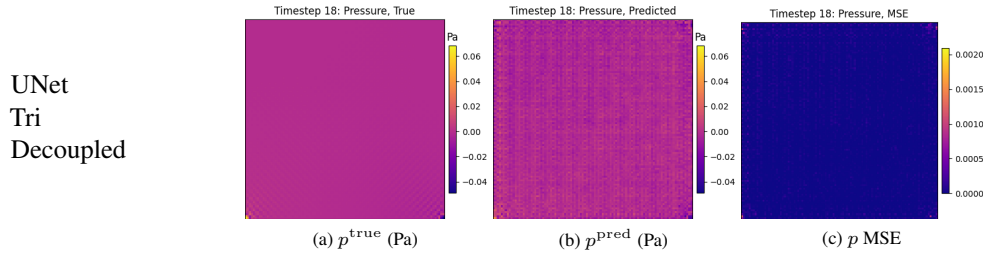
(k) $\mathbf{u}^{\text{true}}$ (m/s) (l) $\mathbf{u}^{\text{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S22: Cavity identity-ablation example 48, timestep 18.

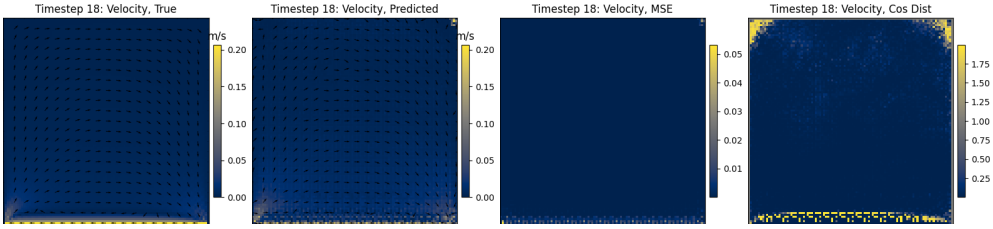
the error distribution. The identity-ablation curves are also decisively outside the transformer-backbone range in all four panels, which indicates that the temporal-backbone contribution is not marginal in this derived-force diagnostic.

It should be noted that the two effects cannot be fully separated in the present dataset. Because of the "increasing" setup, the lowest-intensity signals are more likely to occur during the initial timesteps, so low-signal and low-context effects are partially confounded. Nevertheless, both effects are considered plausible contributors to the observed deterioration.

UNet
Tri
Decoupled

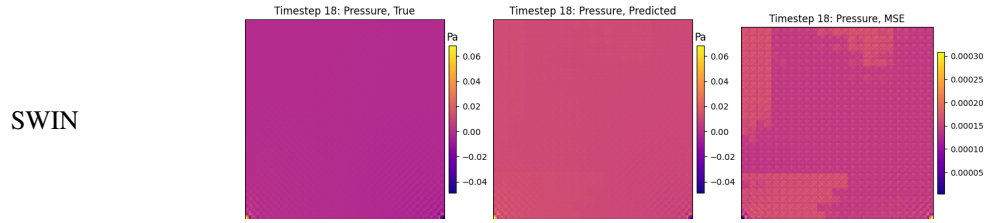


(a) p^{true} (Pa) (b) p^{pred} (Pa) (c) p MSE

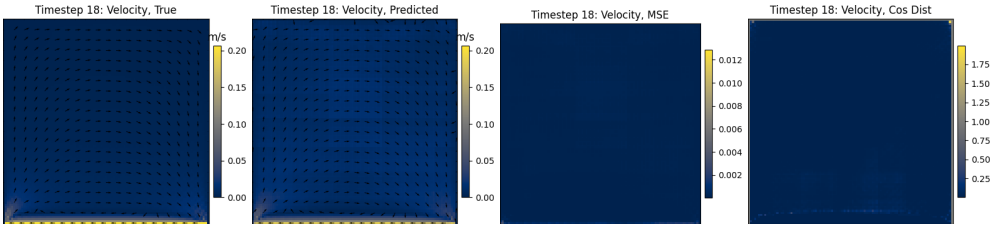


(d) $\mathbf{u}^{\mathrm{true}}$ (m/s) (e) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (f) u MSE (g) u cos dist

SWIN



(h) p^{true} (Pa) (i) p^{pred} (Pa) (j) p MSE



(k) $\mathbf{u}^{\mathrm{true}}$ (m/s) (l) $\mathbf{u}^{\mathrm{pred}}$ (m/s) (m) u MSE (n) u cos dist

Supplementary Fig. S23: Cavity example 49, timestep 18.

In particular, the low-signal contribution is consistent with the independent cavity inspection in Section 1.2.5, where low-intensity pressure fields, and to a lesser extent low-intensity velocity fields, were associated with less stable fieldwise errors. For this reason, the analysis below focuses conservatively on three measures of the absolute relative error: fully pooled; restricted to the 25th–75th percentile range; and pooled after excluding the first 10 timesteps.

Supplementary Table S2: Absolute relative error of the total drag magnitude stratified by percentile band.

Run	$<$ 5th	5–25th	25–75th	75–95th	$>$ 95th
SWIN, Circles	8.2990	0.8649	0.0524	0.0237	0.0186
SWIN identity, Circles	2.8903	0.5352	0.1735	0.2923	0.3389
UNet, Circles	7.6167	0.1728	0.1059	0.1415	0.1416
UNet identity, Circles	12.7138	1.3837	0.4247	0.4588	0.5843
SWIN, Squares	14.8614	1.3602	0.0572	0.0096	0.0106
SWIN identity, Squares	9.4476	1.0722	0.3287	0.2644	0.3162
UNet, Squares	17.8037	0.3379	0.0780	0.0705	0.0816
UNet identity, Squares	22.9169	3.3035	0.8113	0.3408	0.3883

Total drag

Table S3 reports the error on the physically most direct scalar quantity, namely the magnitude of the total drag force. The interquartile band is taken as the most representative signal regime,

Supplementary Table S3: Absolute relative error of the total drag-force magnitude $|\mathbf{F}|$ across the inference datasets.

Run	Pooled $\overline{\lvert e_{rel}\rvert}$	25th–75th $\overline{\lvert e_{rel}\rvert}$	Prefix-10 $\overline{\lvert e_{rel}\rvert}$
SWIN, Circles	0.6214	0.0524	0.0566
SWIN identity, Circles	0.4143	0.1735	0.2219
UNet, Circles	0.5052	0.1059	0.1026
UNet identity, Circles	1.2482	0.4247	0.4728
SWIN, Squares	1.0475	0.0572	0.0557
SWIN identity, Squares	0.9207	0.3287	0.3520
UNet, Squares	1.0167	0.0780	0.0770
UNet identity, Squares	2.3017	0.8113	0.7926

because it excludes both the very small-force cases, where relative errors are unstable, and the extreme upper tail, which contains comparatively few samples. In this representative band, and also in the prefix-excluded view, SWIN yields lower total-drag magnitude errors than UNet on both datasets. The identity-backbone ablations are markedly worse than the corresponding transformer-backbone models, indicating that the previous-frame visual encoding alone is not sufficient to preserve this integrated force quantity. The circle-to-square shift does not produce a uniform increase in total-magnitude error for the transformer-backbone models: SWIN remains nearly unchanged, while UNet is lower on squares than on circles in these two robust views. This indicates non-trivial generalisation to the OOD geometry, but the pooled and total-magnitude views should still be interpreted cautiously, because pressure and viscous contributions can compensate and make the total drag appear comparatively accurate even when the underlying decomposition is less accurate.

To complement the observations above, the directional error between predicted and true total drag is introduced through the cosine distance:

$$d_{\cos} = 1 - \frac{\mathbf{F}^{\text{pred}} \cdot \mathbf{F}^{\text{true}}}{|\mathbf{F}^{\text{pred}}|\,|\mathbf{F}^{\text{true}}|}, \tag{4}$$

clipped to the interval $[0, 2]$. Lower values indicate better directional agreement, with $d_{\cos} = 0$ corresponding to perfect alignment. The results are reported on Table S4. The total-force

Supplementary Table S4: Directional error of the total drag force, measured by cosine distance.

Run	Pooled $\overline{d_{\cos}}$	25th–75th $\overline{d_{\cos}}$	Prefix-10 $\overline{d_{\cos}}$
SWIN, Circles	0.0071	0.0002	0.0002
SWIN identity, Circles	0.0529	0.0152	0.0147
UNet, Circles	0.0441	0.0009	0.0010
UNet identity, Circles	0.3175	0.1715	0.1839
SWIN, Squares	0.0137	0.0031	0.0021
SWIN identity, Squares	0.1143	0.0596	0.0572
UNet, Squares	0.0795	0.0036	0.0030
UNet identity, Squares	0.3835	0.3251	0.2897

direction analysis provides a more discriminating view of the OOD shift. Directional errors increase from circles to squares for both architectures, but SWIN remains more accurate than UNet in all three views. Here too the identity-ablation rows sit well outside the transformer-backbone range, so the temporal backbone is clearly necessary to preserve the force direction.

Pressure and viscous contributions

The errors in the magnitudes of the separate pressure and viscous drags are assessed. Table S5 reports the corresponding absolute relative errors in the pooled, interquartile and prefix-excluded views. The pressure-drag magnitude does not uniformly favour SWIN: in the interquartile and prefix-excluded views, UNet gives lower pressure-magnitude errors, especially on the square dataset. By contrast, viscous-drag magnitude errors show a clearer and more physically relevant separation in the robust views, where SWIN is substantially better than UNet on both geometry families. This indicates that the main advantage of SWIN lies in reconstructing the near-boundary velocity field and its gradients, rather than in the pressure contribution alone. The identity-ablation runs are again far outside the competitive range, especially in the viscous contribution, reinforcing the conclusion that latent self-attention is needed for robust near-wall reconstruction.

Direction of the total drag

Taken together, the total-drag magnitude, its pressure/viscous decomposition, and the directional analysis show that physically derived quantities provide discrimination between the

Supplementary Table S5: Absolute relative errors for the magnitudes of pressure drag and viscous drag across the inference datasets. Lower is better.

Run	Type	Pooled $\overline{\lvert e_{rel}\rvert}$	25th–75th $\overline{\lvert e_{rel}\rvert}$	Prefix-10 $\overline{\lvert e_{rel}\rvert}$
SWIN, Circles	Pressure	0.3409	0.0764	0.0707
SWIN, Circles	Viscous	1.8543	0.0624	0.0701
SWIN identity, Circles	Pressure	0.5708	0.3978	0.4029
SWIN identity, Circles	Viscous	1.0009	0.4524	0.4717
UNet, Circles	Pressure	0.4477	0.0396	0.0357
UNet, Circles	Viscous	1.0710	0.1868	0.1977
UNet identity, Circles	Pressure	1.3300	0.3406	0.3901
UNet identity, Circles	Viscous	1.5212	0.8017	0.7799
SWIN, Squares	Pressure	0.8399	0.1009	0.0924
SWIN, Squares	Viscous	2.4380	0.1092	0.0917
SWIN identity, Squares	Pressure	1.0978	0.4131	0.4148
SWIN identity, Squares	Viscous	1.5179	1.0134	0.8666
UNet, Squares	Pressure	0.7262	0.0575	0.0529
UNet, Squares	Viscous	2.6945	0.1355	0.1500
UNet identity, Squares	Pressure	2.2617	0.4626	0.5091
UNet identity, Squares	Viscous	2.9830	2.0213	1.6656

models that is less evident from local pressure and velocity errors alone. The OOD square geometry does not lead to a uniform deterioration of total-drag magnitude, which suggests a non-trivial degree of generalisation. However, the decomposed and directional drag metrics remain more diagnostic and indicate that SWIN is the more robust architecture, especially for total-force direction and the viscous contribution. Across all these drag views, the identity-ablation results are decisively outside the transformer-backbone range and therefore support the same qualitative conclusion.

1.2.7 Structure-aware velocity diagnostics

The drag-force analysis assesses a physically derived integrated quantity. The following two diagnostics instead focus directly on the structure of the predicted velocity fields, first through one-dimensional profiles and then through fieldwise decompositions.

Line-profile errors

A first intermediate diagnostic is obtained by comparing one-dimensional velocity profiles extracted from the predicted and true fields. In particular, vertical profiles of the streamwise velocity component $u_x(y)$ are extracted at three fixed sections, $x/L = 0.25$, 0.50, and 0.75, using only valid fluid nodes. For a given section, the relative profile error is defined as

$$e_{\mathrm{prof}} = \frac{\left\| {u_x}^{\mathrm{pred}}(y) - {u_x}^{\mathrm{true}}(y) \right\|_2}{\left\| {u_x}^{\mathrm{true}}(y) \right\|_2 + \varepsilon}, \tag{5}$$

where ${u_x}^{\mathrm{pred}}$ is the predicted profile and ${u_x}^{\mathrm{true}}$ is the true profile.

Table S6 reports the resulting profile errors pooled over the three sections, all examples, and all predicted timesteps, including the identity-backbone ablations. The gated columns repeat the calculation after retaining only the strongest 5% of true-profile norms within each run. The profile analysis confirms the ranking observed in the fieldwise metrics: SWIN gives

Supplementary Table S6: Fixed-section $u_x(y)$ profile relative L_2 errors, including the identity-backbone ablations. Gated-95 retains the strongest 5% of true-profile norms within each run.

Geometry	Run	n	Median	Q25–Q75	Gated-95 median	Gated-95 Q75
Circles	SWIN	86	0.0486	0.0217–0.4633	0.0242	0.0308
Circles	SWIN identity	86	0.9098	0.4056–2.2090	0.4870	0.5333
Circles	UNet	86	0.2124	0.1066–0.4518	0.1181	0.2114
Circles	UNet identity	86	2.2975	0.6420–10.1167	0.4686	0.5768
Squares	SWIN	104	0.0657	0.0196–0.6781	0.0229	0.0251
Squares	SWIN identity	104	1.0616	0.3981–2.7811	0.3536	0.3801
Squares	UNet	104	0.2290	0.1042–0.5078	0.0976	0.1208
Squares	UNet identity	104	2.8521	0.7411–10.2886	0.4300	0.5132
Cavity	SWIN	200	0.2715	0.0661–7.4693	0.1640	0.2202
Cavity	SWIN identity	200	1.3697	0.7507–8.4755	0.5506	0.6337
Cavity	UNet	200	0.5631	0.2506–1.8383	0.3341	0.4181
Cavity	UNet identity	200	8.3752	2.8974–17.8583	4.8159	9.9376

substantially lower median errors than UNet on both the ID circles and the OOD squares. The difference persists for cavity, although the cavity profile distribution is much broader. The identity-backbone variants are consistently worse than their transformer-backbone counterparts, even after strong-signal gating. This supports the interpretation that temporal self-attention contributes to the next-state prediction, and that part of the cavity deterioration is driven by low-signal profiles rather than by geometric novelty alone.

The dependence on the position of the vertical section is reported in Table S7. The absolute values vary across sections, but the architecture ranking is stable: SWIN remains lower than UNet for all three sections and all three geometries. The same median profile errors are shown

Supplementary Table S7: Median vertical $u_x(y)$ line-profile relative L2 error on the three sections, including the identity-backbone ablations.

Geometry	Run	$x/L = 0.25$	$x/L = 0.50$	$x/L = 0.75$
Circles	SWIN	0.0447	0.0373	0.0656
Circles	SWIN identity	0.9597	0.8653	0.9449
Circles	UNet	0.2607	0.1695	0.2452
Circles	UNet identity	2.2495	2.1138	2.9036
Squares	SWIN	0.0572	0.0659	0.0836
Squares	SWIN identity	1.0563	0.9842	1.1324
Squares	UNet	0.2292	0.1936	0.2944
Squares	UNet identity	2.8521	2.4190	3.6289
Cavity	SWIN	0.3016	0.2498	0.2718
Cavity	SWIN identity	1.1773	1.3812	1.4161
Cavity	UNet	0.6050	0.4039	0.7311
Cavity	UNet identity	8.6627	5.1993	9.2940

in Figure S28. Figure S29 shows the effect of removing the first k predicted timesteps for the transformer-backbone runs. The prefix effect is visible but comparatively mild in the median profile error, especially relative to the low-signal effect captured by the gated view in Table S6.

Fieldwise velocity-error decomposition

A further diagnostic assesses whether velocity-field errors are mainly due to simple global effects, such as amplitude mismatch or small spatial misalignment, or whether they reflect residual structural differences. For each predicted velocity field, three quantities are considered: u_x, u_y, and the speed magnitude $|\mathbf{u}|$. For a generic field $q \in \{u_x, u_y, |\mathbf{u}|\}$ over the valid-fluid domain Ω, the raw relative L_2 error is defined as

$$e_{\mathrm{raw}}(q) = \frac{\left\| q^{\mathrm{pred}} - q^{\mathrm{true}} \right\|_{2,\Omega}}{\left\| q^{\mathrm{true}} \right\|_{2,\Omega} + \varepsilon}. \tag{6}$$

This value is compared with three corrected views. In the scale-only view, the optimal scalar

$$\alpha^\star = \arg\min_{\alpha} \left\| \alpha q^{\mathrm{pred}} - q^{\mathrm{true}} \right\|_{2,\Omega} \tag{7}$$

is applied to the prediction. In the shift-only view, the prediction is translated by the optimal displacement

$$\boldsymbol{\delta}^\star = \arg\min_{\boldsymbol{\delta} \in \mathcal{D}} \left\| \mathcal{T}_{\boldsymbol{\delta}} q^{\mathrm{pred}} - q^{\mathrm{true}} \right\|_{2,\Omega}, \tag{8}$$

where $\mathcal{T}_{\boldsymbol{\delta}}$ is the translation operator and $\mathcal{D}$ is the bounded search window used in the analysis. Finally, the shift+scale view solves

$$(\alpha^\star, \boldsymbol{\delta}^\star) = \arg\min_{\alpha,\, \boldsymbol{\delta} \in \mathcal{D}} \left\| \alpha \mathcal{T}_{\boldsymbol{\delta}} q^{\mathrm{pred}} - q^{\mathrm{true}} \right\|_{2,\Omega}. \tag{9}$$

The same denominator $\|q^{\mathrm{true}}\|_{2,\Omega} + \varepsilon$ is used for all corrected relative errors. The reduction reported in Table S8 is

$$R_{\mathrm{shift+scale}} = 1 - \frac{e_{\mathrm{shift+scale}}(q)}{e_{\mathrm{raw}}(q)}. \tag{10}$$

Pearson correlation is also computed fieldwise as

$$r(q) = \frac{\sum_\Omega \left(q^{\mathrm{pred}} - \overline{q^{\mathrm{pred}}}\right)\left(q^{\mathrm{true}} - \overline{q^{\mathrm{true}}}\right)}{\sqrt{\sum_\Omega \left(q^{\mathrm{pred}} - \overline{q^{\mathrm{pred}}}\right)^2}\sqrt{\sum_\Omega \left(q^{\mathrm{true}} - \overline{q^{\mathrm{true}}}\right)^2} + \varepsilon}. \tag{11}$$

The corrections are not used as a post-processing proposal, but as diagnostics: a large improvement after correction indicates that a substantial part of the raw error can be explained by amplitude or alignment effects.

Table S8 reports the speed-field diagnostics, including the identity-backbone ablations. The residual column is the fraction of raw error left after shift+scale correction. SWIN has lower raw speed-field relative errors than UNet on all three geometries. The circle-to-square shift increases the SWIN error but keeps it well below the corresponding UNet value, while cavity remains the most difficult dataset for both models. The identity-backbone ablations are substantially worse, especially for cavity. Shift+scale correction reduces part of their error, indicating a strong amplitude or intensity component, but the corrected errors remain above those of the transformer-backbone models. This reinforces the conclusion that temporal self-attention does more than preserve a static spatial representation.

Supplementary Table S8: Median whole-field speed diagnostics, including the identity-backbone ablations. Lower relative L_2 values are better; higher Pearson correlations are better.

Geometry	Run	Raw L2	Scale L2	Corr. L2	Pearson	α	Residual
Circles	SWIN	0.0277	0.0267	0.0267	0.9989	0.9939	0.9423
Circles	SWIN identity	0.4341	0.2805	0.2790	0.9135	0.7744	0.6799
Circles	UNet	0.1986	0.1307	0.1302	0.9798	0.9601	0.7929
Circles	UNet identity	0.9509	0.4035	0.3960	0.7499	0.5244	0.4411
Squares	SWIN	0.0573	0.0508	0.0508	0.9978	0.9936	0.9155
Squares	SWIN identity	0.8610	0.2739	0.2738	0.9185	0.5383	0.2120
Squares	UNet	0.1994	0.1245	0.1216	0.9831	0.9157	0.7485
Squares	UNet identity	1.9806	0.4008	0.3882	0.7771	0.3268	0.1810
Cavity	SWIN	0.1939	0.1761	0.1761	0.9769	0.9152	0.7380
Cavity	SWIN identity	1.1081	0.4294	0.4191	0.8397	0.4657	0.2396
Cavity	UNet	0.4359	0.3660	0.2845	0.8342	0.7926	0.6288
Cavity	UNet identity	5.3207	0.8575	0.6865	0.1171	0.0537	0.1422

Figure S30 reports the same raw, scale-only, and shift+scale relative errors for the speed field. The cavity case again stands out, especially for UNet identity, whose raw Pearson correlation is also substantially lower than in the circle and square cases; the shift+scale

correction improves the relative error but does not remove the gap. Taken together with Sections 1.2.5, 1.2.6, and 1.2.7, this supports the view that low-signal regimes and field-intensity mismatch are recurrent contributors to the residual errors.

1.2.8 Von Karman vortex-shedding test

The preceding diagnostics focus on obstacle-flow and cavity examples. The final tests are carried out on von Karman vortex-shedding datasets. These cases introduce a different type of out-of-distribution behaviour: coherent, time-periodic wake structures. They are assessed separately because aggregate pointwise errors can be dominated by phase lag. A prediction that captures the wake morphology but is slightly delayed in time can therefore appear poor under same-timestep relative-error metrics.

Three variants are considered. Two use the original output timestep and contain 48 examples with mean $\mathrm{Re} = 749.0$ and CV of 0.28 on the physical scale: `skip09`, which samples the early onset of alternating vortices, and `skip39`, which starts from an already established vortex street. The third variant, `vtkp06 skip39`, uses the same established-wake setup but reduces the output timestep by a factor of five; it contains 365 examples with mean $\mathrm{Re} = 774.9$ and CV of 0.27.

Representative true sequences for the von Karman regimes are reported in the main text.

To probe phase lag directly, the predicted speed field $\left|\mathbf{u}^{\mathrm{pred}}\right|$ and vorticity field $\omega^{\mathrm{pred}} = \partial {u_y}^{\mathrm{pred}}/\partial x - \partial {u_x}^{\mathrm{pred}}/\partial y$ are compared with two true fields: the corresponding true timestep, denoted as current, and the immediately preceding true timestep, denoted as previous. For $g \in \{|\mathbf{u}|, \omega\}$, the two relative errors are

$$
\begin{aligned}
e_{\mathrm{current}}(g,t) &= \frac{\left\| g_t^{\mathrm{pred}} - g_t^{\mathrm{true}} \right\|_{2,\Omega}}{\left\| g_t^{\mathrm{true}} \right\|_{2,\Omega} + \varepsilon}, \\
e_{\mathrm{previous}}(g,t) &= \frac{\left\| g_t^{\mathrm{pred}} - g_{t-1}^{\mathrm{true}} \right\|_{2,\Omega}}{\left\| g_{t-1}^{\mathrm{true}} \right\|_{2,\Omega} + \varepsilon}.
\end{aligned}
\tag{12}
$$

The corresponding Pearson correlations are computed by replacing g_t^{true} with either g_t^{true} or g_{t-1}^{true} in the fieldwise expression above. The previous-frame comparison is only a diagnostic baseline; it is not used as a deployable predictor. If the previous true frame matches the prediction better than the current true frame, the error is consistent with a temporal delay in the predicted wake.

Table S9 shows that the previous-frame comparison is better than the current-frame comparison for the transformer-backbone models in all three variants, but the size of the effect is not constant. At the original timestep, the lag is already visible in `skip09` and becomes much more severe in `skip39`, especially for vorticity. This indicates that the difficulty is larger once the alternating wake is fully established. When the same established-wake setup is sampled more finely in time, as in `vtkp06 skip39`, the current-frame errors decrease substantially. The identity-backbone ablations, by contrast, are worse in a broader sense: their correlations are lower and their previous-frame comparison does not recover transformer-like accuracy. Thus, the lag depends both on the temporal sampling of the input data and on the

Supplementary Table S9: Lag diagnostics for the von Karman variants, including the identity-backbone ablations. Current compares predicted fields with the same true timestep; previous compares them with the preceding true timestep.

Variant	Run	Field	Current rel. L2	Previous rel. L2	Current Pearson	Previous Pearson
skip09	SWIN	$\|\mathbf{u}\|$	0.1161	0.0186	0.9785	0.9993
skip09	SWIN	ω	0.3095	0.0978	0.9546	0.9953
skip09	SWIN identity	$\|\mathbf{u}\|$	0.3601	0.3452	0.7895	0.8178
skip09	SWIN identity	ω	0.7799	0.7173	0.6456	0.7048
skip09	UNet	$\|\mathbf{u}\|$	0.1269	0.0635	0.9726	0.9936
skip09	UNet	ω	0.5359	0.4515	0.8490	0.8939
skip09	UNet identity	$\|\mathbf{u}\|$	0.3907	0.3793	0.6868	0.7176
skip09	UNet identity	ω	2.1170	2.1011	0.2863	0.3171
skip39	SWIN	$\|\mathbf{u}\|$	0.3177	0.0206	0.8015	0.9993
skip39	SWIN	ω	1.0779	0.0940	0.4097	0.9957
skip39	SWIN identity	$\|\mathbf{u}\|$	0.3705	0.3337	0.7336	0.8271
skip39	SWIN identity	ω	1.1446	0.6594	0.1816	0.7519
skip39	UNet	$\|\mathbf{u}\|$	0.3182	0.0675	0.8002	0.9923
skip39	UNet	ω	1.1308	0.4775	0.3140	0.8802
skip39	UNet identity	$\|\mathbf{u}\|$	0.4435	0.3949	0.6149	0.7033
skip39	UNet identity	ω	2.2924	2.0757	0.0882	0.2953
vtkp06 skip39	SWIN	$\|\mathbf{u}\|$	0.0588	0.0219	0.9947	0.9991
vtkp06 skip39	SWIN	ω	0.1713	0.1103	0.9854	0.9940
vtkp06 skip39	SWIN identity	$\|\mathbf{u}\|$	0.3766	0.3754	0.7920	0.7937
vtkp06 skip39	SWIN identity	ω	0.7288	0.7180	0.6921	0.7016
vtkp06 skip39	UNet	$\|\mathbf{u}\|$	0.0867	0.0634	0.9887	0.9929
vtkp06 skip39	UNet	ω	0.4774	0.4589	0.8815	0.8907
vtkp06 skip39	UNet identity	$\|\mathbf{u}\|$	0.5342	0.5354	0.3418	0.3367
vtkp06 skip39	UNet identity	ω	3.0166	2.9752	0.2157	0.2190

phase of the wake, while the identity ablation confirms that the transformer-backbone models are not merely producing a fixed delayed copy.

Figure S31 summarises the same lag diagnostic for the transformer and identity backbones. The previous-frame baseline remains lower for the transformer-backbone models, reinforcing the interpretation that a phase delay contributes to the error. The identity-backbone variants remain comparatively inaccurate under both current- and previous-frame comparisons, which separates phase lag from the loss of temporal sequence modelling.

The same pattern is visible in the velocity-profile diagnostics. Table S10 reports the current-target median profile relative L2 error. The established-wake `skip39` case has much larger current-frame errors than `skip09`; after reducing the timestep in `vtkp06 skip39`, these errors drop markedly. SWIN remains lower than UNet in the current-frame view, and both identity-backbone variants remain worse than their corresponding transformer-backbone models.

The aggregate inference losses for the reduced-timestep `vtkp06 skip39` set are reported in Table S11. In this dataset, SWIN has lower loss, pressure MSE, velocity-magnitude

Supplementary Table S10: Current-target fixed-section $u_x(y)$ profile errors for the von Karman variants, including the identity-backbone ablations.

Variant	Run	Profile median	Q25	Q75
skip09	SWIN	0.0786	0.0416	0.1525
skip09	SWIN identity	0.4337	0.3124	0.4558
skip09	UNet	0.1081	0.0862	0.1602
skip09	UNet identity	0.4187	0.3407	0.5080
skip39	SWIN	0.3445	0.1006	0.5109
skip39	SWIN identity	0.4071	0.3358	0.5038
skip39	UNet	0.3464	0.1458	0.5093
skip39	UNet identity	0.5191	0.4506	0.5751
`vtkp06` skip39	SWIN	0.0511	0.0373	0.0684
`vtkp06` skip39	SWIN identity	0.4313	0.3377	0.4883
`vtkp06` skip39	UNet	0.0913	0.0730	0.1104
`vtkp06` skip39	UNet identity	0.7948	0.4347	1.3596

MSE, and velocity cosine distance than UNet. These aggregate values are consistent with the lag-aware diagnostics above, but are less informative about the temporal origin of the error.

Supplementary Table S11: Aggregate inference losses and metrics on the von Karman wake test set, including the identity-backbone ablations.

Run	n examples	Loss	p MSE	u MSE	u cos distance
SWIN	365	0.0358	0.0128	0.0020	0.0660
SWIN identity	365	0.0697	0.0253	0.0611	0.0901
UNet	365	0.0400	0.0196	0.0038	0.0709
UNet identity	365	0.4084	0.2152	0.2883	0.5529

Figure S32 shows two consecutive predicted timesteps for the three variants, using the available inference summary panels for example 1. The standard-timestep `skip09` and `skip39` rows show that the qualitative phase relation between prediction and truth changes as the wake develops. The reduced-timestep `vtkp06` `skip39` row preserves a similar established-wake morphology but with smaller inter-frame changes, consistent with the reduced lag observed quantitatively.

Figure S33 reports the corresponding identity-backbone snapshots immediately afterwards, so that the qualitative wake structure can be compared without changing example or timestep.

The identity-backbone panels in Figure S33 are qualitatively worse across all three variants. The alternating wakes are still hinted at, but the phase structure is blurrier and less stable, which matches the lag-aware metrics and reinforces the role of the temporal backbone.

1.2.9 Cross-regime synthesis of identity-backbone effects

The analyses above indicate that the models recover meaningful flow evolution, but they do not by themselves isolate the role of temporal self-attention in the backbone. To probe this point, two additional ablations are considered. For each encoder-decoder family, an identity-backbone variant is evaluated in which the encoded representation of timestep $t-1$ is passed directly to the decoder for timestep t, with no self-attention over the latent sequence. These ablations therefore preserve the same visual encoder-decoder route, while removing the explicit temporal model.

For convenience, the main obstacle/cavity and von Karman identity-backbone diagnostics are also collected in Tables S12 and S13.

Supplementary Table S12: Identity-backbone ablation on obstacle-flow and cavity diagnostics. Drag reports the representative 25–75% true-force band and is not defined for cavity. Profile errors are fixed-section $u_x(y)$ relative L_2 errors; Gated-95 retains the strongest 5% of true-profile norms.

Geometry	Run	n	Drag band	Profile median	Profile Gated-95	Speed raw L2	Speed corr. L2	Speed Pearson
Circles	SWIN	86	0.0524	0.0486	0.0242	0.0277	0.0267	0.9989
Circles	SWIN identity	86	0.1735	0.9098	0.4870	0.4341	0.2790	0.9135
Circles	UNet	86	0.1059	0.2124	0.1181	0.1986	0.1302	0.9798
Circles	UNet identity	86	0.4247	2.2975	0.4686	0.9509	0.3960	0.7499
Squares	SWIN	104	0.0572	0.0657	0.0229	0.0573	0.0508	0.9978
Squares	SWIN identity	104	0.3287	1.0616	0.3536	0.8610	0.2738	0.9185
Squares	UNet	104	0.0780	0.2290	0.0976	0.1994	0.1216	0.9831
Squares	UNet identity	104	0.8113	2.8521	0.4300	1.9806	0.3882	0.7771
Cavity	SWIN	200	–	0.2715	0.1640	0.1939	0.1761	0.9769
Cavity	SWIN identity	200	–	1.3697	0.5506	1.1081	0.4191	0.8397
Cavity	UNet	200	–	0.5631	0.3341	0.4359	0.2845	0.8342
Cavity	UNet identity	200	–	8.3752	4.8159	5.3207	0.6865	0.1171

The ablation results are integrated into Tables S3, S6, S8, S9, and S10, and the compact summaries above show the same ranking in a more direct side-by-side form. Across drag, profile, fieldwise, and von Karman diagnostics, the identity-backbone variants are consistently worse than their corresponding transformer-backbone models. Scale and shift+scale corrections reduce part of the identity-model error, especially on squares and cavity, but the corrected errors remain above those of the transformer-backbone models. The Von Karman results also separate the identity effect from phase lag: transformer-backbone predictions can be closer to the previous true frame than to the current one, but identity-backbone predictions remain comparatively inaccurate under both comparisons. Taken together, these results indicate that temporal self-attention materially improves the predicted field evolution, rather than merely preserving a static previous-frame representation.

Supplementary Table S13: Identity-backbone ablation on von Karman lag diagnostics. Current compares g_t^{pred} with g_t^{true}, whereas previous compares g_t^{pred} with g_{t-1}^{true}.

Variant	Run	Field	Current rel. L2	Previous rel. L2	Current Pearson	Previous Pearson
skip09	SWIN	$\lvert\mathbf{u}\rvert$	0.1161	0.0186	0.9785	0.9993
skip09	SWIN	ω	0.3095	0.0978	0.9546	0.9953
skip09	SWIN identity	$\lvert\mathbf{u}\rvert$	0.3601	0.3452	0.7895	0.8178
skip09	SWIN identity	ω	0.7799	0.7173	0.6456	0.7048
skip09	UNet	$\lvert\mathbf{u}\rvert$	0.1269	0.0635	0.9726	0.9936
skip09	UNet	ω	0.5359	0.4515	0.8490	0.8939
skip09	UNet identity	$\lvert\mathbf{u}\rvert$	0.3907	0.3793	0.6868	0.7176
skip09	UNet identity	ω	2.1170	2.1011	0.2863	0.3171
skip39	SWIN	$\lvert\mathbf{u}\rvert$	0.3177	0.0206	0.8015	0.9993
skip39	SWIN	ω	1.0779	0.0940	0.4097	0.9957
skip39	SWIN identity	$\lvert\mathbf{u}\rvert$	0.3705	0.3337	0.7336	0.8271
skip39	SWIN identity	ω	1.1446	0.6594	0.1816	0.7519
skip39	UNet	$\lvert\mathbf{u}\rvert$	0.3182	0.0675	0.8002	0.9923
skip39	UNet	ω	1.1308	0.4775	0.3140	0.8802
skip39	UNet identity	$\lvert\mathbf{u}\rvert$	0.4435	0.3949	0.6149	0.7033
skip39	UNet identity	ω	2.2924	2.0757	0.0882	0.2953
`vtkp06` skip39	SWIN	$\lvert\mathbf{u}\rvert$	0.0588	0.0219	0.9947	0.9991
`vtkp06` skip39	SWIN	ω	0.1713	0.1103	0.9854	0.9940
`vtkp06` skip39	SWIN identity	$\lvert\mathbf{u}\rvert$	0.3766	0.3754	0.7920	0.7937
`vtkp06` skip39	SWIN identity	ω	0.7288	0.7180	0.6921	0.7016
`vtkp06` skip39	UNet	$\lvert\mathbf{u}\rvert$	0.0867	0.0634	0.9887	0.9929
`vtkp06` skip39	UNet	ω	0.4774	0.4589	0.8815	0.8907
`vtkp06` skip39	UNet identity	$\lvert\mathbf{u}\rvert$	0.5342	0.5354	0.3418	0.3367
`vtkp06` skip39	UNet identity	ω	3.0166	2.9752	0.2157	0.2190

2 Supplementary Methods

2.1 Architecture details

The model combines two main components:

- A visual encoder-decoder (architecture-specific parameters). For the UNet-based option, the encoder-decoder follows a CNN [1] UNet [2]-style design, whereas the SWIN-based follows [3, 4]. The encoder converts each snapshot into a tuple consisting of a high-level semantic embedding vector and a set of intermediate feature maps. The decoder performs the reverse operation.
- A Transformer backbone (66,735,360 parameters). The Transformer applies self-attention to a sequence of embeddings produced by the encoder, and generates the next embedding of the sequence.

The model uses ELU activation functions [5]. The snapshot sequence is processed as follows:

1. Each snapshot is converted into a sequence of semantic embeddings by the UNet/SWIN's encoder.

2. The sequence of embeddings is processed by the Transformer's encoder into a context-aware latent representation of the sequence.
3. The latent representation is then passed to the Transformer's decoder, which predicts the next embedding of the sequence.
4. The updated sequence of embeddings is then converted into a sequence of snapshots by the UNet/SWIN's encoder. The intermediate feature maps are passed directly to the decoder at the corresponding position in the sequence, shifted forward by one timestep.

2.1.1 UNet encoder-decoder

This autoencoder employs a UNet-based architecture to encode a snapshot of pressure and velocity patterns (x and y component) into a semantic embedding vector, and to decode an embedding back into a flow and pressure pattern snapshot. No explicit binary masks or coordinate fields for obstacles, inlets or outlets are provided; however, their presence remains implicitly encoded in the pressure and velocity patterns. In particular, cells inside obstacles are outside the CFD computational domain and therefore retain characteristic values inherited from the solver, including persistent zero pressure. A set of intermediate feature maps are generated as well, in order to facilitate the reconstruction of low-level local details. Unlike the original U-Net architecture, pressure and each component of the velocity field are processed separately through dedicated pathways. This design choice arises because these fields differ significantly—unlike the RGB channels in the original UNet application—so combining them in a single pathway would lead to destructive interference. Following the encoder's downsampling, the pathway embeddings are concatenated and passed through a fully connected layer, enabling cross-pathway feature exchange. Symmetrically, in the decoder, the combined embedding is again processed by a fully connected layer before being split and upsampled across the dedicated pathways.

The architecture of the UNet's encoder is as follows:

1. The encoder accepts an input tensor of shape (batch_size, 101, 101, 3), where the last dimension corresponds to the three channels: p, u_x, and u_y.
2. The input is split across the last dimension into three tensors of shape (batch_size, 101, 101, 1), thereby separating p, u_x, and u_y from one another, and each of them is fed into a dedicated downsampling pathway.
3. Each downsampling pathway consists of three convolutional layers with 3×3 kernels, strides of 2, and respectively 8, 16, and 32 filters. This arrangement creates intermediate feature maps of shapes (batch_size, 52, 52, 8), (batch_size, 26, 26, 16) and (batch_size, 13, 13, 32).
4. The deepest feature map is then flattened into a vector of shape (batch_size, 5408) and passed to a fully-connected layer. The final pathway embedding has shape (batch_size, 256).
5. The three pathway embeddings are concatenated and passed to a fully-connected layer that maintains the same shape. The final embedding has shape (batch_size, 768).

The architecture of the UNet's decoder mirrors the encoder and is as follows:

1. The embedding is passed to a fully-connected layer that maintains the same shape, and then split into three pathway embeddings, each one of shape (batch_size, 256).
2. Each pathway embedding is passed to a fully-connected layer which outputs a tensor of shape (batch_size, 5408), and then reshaped into a tensor of shape (batch_size, 13, 13, 32).
3. Each tensor shaped in this way is processed through a dedicated upsampling pathway consisting of three transposed convolutional layers with 3×3 kernels, strides of 2, respectively 8, 16, and 32 filters. This arrangement creates intermediate tensors of shapes (batch_size, 26, 26, 16), (batch_size, 52, 52, 8) and (batch_size, 104, 104, 1). The final intermediate tensors are cropped on the right to a final shape of (batch_size, 101, 101, 1). Each intermediate tensor is concatenated with the corresponding feature map before being passed to the transposed convolutional layer.
4. The tensors produced by the upsampling pathways are finally concatenated into the final output of shape (batch_size, 101, 101, 3).

2.1.2 SWIN encoder-decoder

This autoencoder architecture is based on a Shifted Window Transformer. This autoencoder utilises a type of self-attention called "Shifted Window Attention" that aims to exploit self-attention's capability to propagate relevant information throughout the flow field based on the context of surrounding flow, while also reducing the required computational cost by dividing the (current stage) flow snapshot representation into 'windows' and only performing Self-Attention (W-MSA) within each window (block 1 of the SWIN layer). The window borders then "shift" by half a window length and Shifted Window Self Attention (SW-MSA) is performed (block 2 of the layer) to allow information and context to propagate between windows without having to perform full global attention. This occurs at each SWIN layer.

In the configuration used within this work, the window size is 13x13 patches, so the window will shift by 6 patches. This means that at the last SWIN layer in the encoder, W-MSA is performed twice as there is no need to shift as one window comprises the entire snapshot representation at this layer. A separate encoder and decoder are trained for each channel (p, u_x, u_y). Each decoder layer (except the lowest-dimension first layer) sees skip connections (the output) from the same layer of the encoder, acting as an intermediate feature map to reconstruct low-level local details. The SWIN Transformer Autoencoder has 30,324,240 parameters.

The architecture of the SWIN UNet Encoder is as follows:

1. The encoder accepts an input tensor of shape (batch_size, 101, 101, 3) where the last dimension corresponds to the three channels: (p, u_x, u_y).
2. The input is split across the last dimension into three tensors of shape (batch_size, 101, 101, 1), thereby separating the (p, u_x, u_y) channels from one another, and each of them is fed into a dedicated down-sampling pathway.
3. Each pathway starts by permuting to channel-first and padding the image to size (batch_size, 1, 104, 104), and then embedding it into 2×2 patches (batch_size, 48, 52, 52).
4. There are three consecutive SWIN Transformer layers (each with 2 SWIN blocks), with each layer but the last followed by a 2×2 patch-merging step. This results in feature maps of size (batch_size, 48, 52, 52) and (batch_size, 96, 26, 26), as the deepest encoder/decoder stage does not utilise a feature map.

5. The output of the deepest layer is passed to an adaptive average pooling with output shape (batch_size, 192, 8, 8) and then flattened to (batch_size, 12288) and passed to a fully-connected bottleneck with size (batch_size, 256).
6. The three separate pathways are concatenated and passed to another fully-connected layer with the same input/output shape to facilitate coupling between the channels, leaving a shape (batch_size, 768).

The architecture of the SWIN UNet Decoder is as follows:

1. The embedding is passed to a fully-connected layer that has the same input/output shape and is split into three pathways each with shape (batch_size, 256).
2. Each pathway is passed to a fully-connected layer with output shape (batch_size, 12288), reshaped to (batch_size, 192, 8, 8), and up-sampled by bilinear interpolation to (batch_size, 192, 13,13) before being passed to the up-sampling pathway.
3. Each up-sampling pathway consists of an initial patch expansion layer (no attention), and 2 SWIN layers (2 blocks each), each followed by a patch expansion. For all up-sampling layers after the initial one, the corresponding feature map from the encoder is concatenated to the tensor, and the output is linearly projected back to the pre-concatenation size before patch expansion.
4. The final up-sample is followed by a 1×1 convolutional layer to reduce channel size back to 1.
5. The final output of shape (batch_size, 104, 104, 1) is cropped to (batch_size, 1, 101, 101), concatenated with all pathway outputs, and permuted to channel-last format to shape (batch_size, 101, 101, 3).

2.1.3 Transformer backbone

The backbone is a standard autoregressive transformer that applies self-attention to a sequence of embeddings produced by UNet/SWIN's encoder, in order to predict the next embedding of the sequence.

The Transformer comprises an encoder and a decoder. The encoder is structured as follows:

1. The encoder accepts a sequence of n_seq embeddings (with n_seq set to 19), arranged in a tensor of shape (batch_size, n_seq, 768).
2. A standard sinusoidal positional encoding [6] is added to the input.
3. The input is passed through two identical standard encoder blocks, each featuring four attention heads. A dropout rate of 0.1 is applied to their dropout layers. The first fully connected layer contains 1,536 perceptrons, while the second reduces the embedding dimension back to 768 and does not use an activation function.
4. The final output of the encoder is a context-aware latent representation of the sequence.

The decoder is structured as follows:

1. The decoder accepts an input sequence with the same characteristics as the encoder's input.

2. A standard sinusoidal positional encoding is added to the input.
3. The input is passed through two identical standard decoder blocks, again each featuring four attention heads. A dropout rate of 0.1 is applied to their dropout layers. The first fully connected layer contains 1,536 perceptrons, while the second reduces the embedding dimension back to 768 and does not use an activation function. In each block, the context-aware latent produced by the encoder is passed to the second multi-head attention layer as key/value.
4. The final output of the decoder is a sequence of embeddings of the same type and size as the input, but shifted by one timestep forward for autoregressive prediction. The first $\mathrm{n\_seq} - 1$ embeddings align with the last $\mathrm{n\_seq} - 1$ embeddings of the input in order to allow a teacher-forcing training strategy, whereas the last embedding of the last sequence is generated ex novo.

2.2 Training dataset generation

The dataset consists of 4,081 sequences of 20 2D snapshots each, artificially generated through a standard Smagorinsky-LES BGK Lattice-Boltzmann model implemented through OpenLB, an open-source library for Lattice-Boltzmann modelling. Under the assumption of small Mach and Knudsen numbers, the Lattice-Boltzmann Method reproduces a fluid obeying the Navier-Stokes equations with a compressibility error proportional to Ma^2, where Ma is the Mach number, and the equation of state $p = \rho\, c_s^2$ where $c_s \equiv \frac{1}{\sqrt{3}} \frac{\delta x}{\delta t}$ is the tuneable speed of sound, which depends on the lattice size δx, and the timestep δt. It has been shown that, under the assumption of low Mach number, the speed of sound (and therefore, the Mach number itself) can be altered and used as a tuneable parameter in order to improve the balance between stability, accuracy and numerical expense, without altering the system significantly [7].

The computational domain is a 2D square grid of 101×101 lattice sites, containing randomly-generated inlet, outlet, and multiple impenetrable circles within its interior. The inlet is placed on one of the four sides of the domain sampled uniformly, while the outlet is placed on one of the remaining three sides, also sampled uniformly. For both the inlet and outlet, the size of the opening is sampled uniformly between 0.3 and 1.0 times the domain side length. The lower bound is selected to prevent excessively narrow inlets/outlets, which can lead to spurious acoustic artifacts, a well-known issue in Lattice-Boltzmann simulations [7]. The number of impenetrable internal circles is sampled uniformly between 1 and 6. For each of them, position and radius are randomly generated in such a way that the minimum radius and the minimum distance from the borders (including the other circles) are 0.05 times the domain side length. The inlet is defined as a fixed-velocity boundary following a Poiseuille profile; the outlet as a fixed-pressure boundary; the external boundaries and the boundaries of the circles are respectively defined as bounce-back and Bouzidi [8] boundaries (both equivalent to second-order no-slip). An example of a generated configuration is reported in Figure S34.

The Smagorinsky constant and the lattice velocity are set to 0.14 and 0.1 respectively. The Reynolds number is evaluated using domain side length and average inlet velocity, and is sampled from a lognormal distribution specified by its physical-scale mean, 90, and coefficient of variation, 1.31. The physical quantities of domain side length and reference average inlet velocity are conventionally set to 1 m and 0.5 m/s respectively. All the other parameters are tuned accordingly, following standard definitions and laws of similarity as per [7], Chapter 7.

Pressure and flow patterns are stored every 80 timesteps, and a total of 20 snapshots are stored. Each run is uniformly randomly chosen to follow either an "increasing" or a "constant" simulation. In the "increasing" simulation, the inlet average velocity is ramped linearly from zero to the reference value throughout the simulation. In the "constant" simulation, the average inlet velocity is first ramped up to the reference value and then kept constant; the timesteps are stored after the initial plateau.

2.3 Training and evaluation protocol

Training follows a teacher-forcing approach. Each example is a sequence of $\mathrm{n\_seq} + 1 = 20$ snapshots: the first n_seq are used as inputs, and the last n_seq provide the true sequence, denoted below with the superscript "true". The dataset is pipelined onto the model through batches of 128 examples. An Adam optimizer [9] is used. A learning rate scheduler starts from a value of 0.001, and then reduces it by a factor of 0.5 after five consecutive epochs of no improvement (on plateau) of the validation loss, down to a minimum of 10^{-6}. The maximum number of epochs is set to 200. All model weights are initialized using the He normal [10] scheme.

2.3.1 Metrics and loss function

Three metric functions are used. In the notation below, the superscripts "pred" and "true" respectively denote the model prediction and the corresponding CFD field. The metrics are: (*i*) mean squared error over the pressure alone:

$$\begin{aligned} \mathrm{MSE}_p &= \frac{1}{N} \sum \left(p^{\mathrm{pred}} - p^{\mathrm{true}} \right)^2 \\ &\equiv \frac{1}{BLN_xN_y} \sum_{b=1}^{B} \sum_{\ell=1}^{L} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[y^{\mathrm{pred}}\left(b,\ell,i,j,0\right) - y^{\mathrm{true}}\left(b,\ell,i,j,0\right) \right]^2 \ , \end{aligned} \tag{13}$$

where B, L, N_x and N_y are respectively batch size, sequence length and number of points along the x and y direction; (*ii*) mean square error over the velocity magnitude:

$$\begin{aligned} \mathrm{MSE}_u &= \frac{1}{N} \sum \left(\mathbf{u}^{\mathrm{pred}} - \mathbf{u}^{\mathrm{true}} \right)^2 \\ &\equiv \frac{1}{2BLN_xN_y} \sum_{b=1}^{B} \sum_{\ell=1}^{L} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k=1}^{2} \left[y^{\mathrm{pred}}\left(b,\ell,i,j,k\right) - y^{\mathrm{true}}\left(b,\ell,i,j,k\right) \right]^2 \ ; \end{aligned} \tag{14}$$

and (*iii*) cosine distance between predicted and true velocity directions:

$$\begin{aligned}\mathrm{COSDIST} &= \frac{1}{N}\sum\left(1-\frac{\mathbf{u}^{\mathrm{pred}}\cdot\mathbf{u}^{\mathrm{true}}}{|\mathbf{u}^{\mathrm{pred}}|\,|\mathbf{u}^{\mathrm{true}}|}\right)\\ &\equiv\frac{1}{2BLN_xN_y}\sum_{b=1}^{B}\sum_{\ell=1}^{L}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\\ &\quad\left[1-\frac{\sum_{k=1}^{2}x^{\mathrm{pred}}(b,\ell,i,j,k)\,x^{\mathrm{true}}(b,\ell,i,j,k)}{\sum_{k=1}^{2}x^{\mathrm{pred}}(b,\ell,i,j,k)\sum_{k=1}^{2}x^{\mathrm{true}}(b,\ell,i,j,k)+10^{-8}}\right]\end{aligned} \tag{15}$$

where the addendum 10^{-8} in the denominator prevents divide-by-zero errors at points with no flow motion. The reason for introducing both mean-squared-error metrics (Equations 13 and 14) is that they are Galilean-invariant, therefore forcing the predictions to retain physical meaning. The reason for introducing the cosine distance (Equation 15) is that the mean squared error alone would allow the reconstruction of the magnitude of the velocity field while losing all information about its direction. It can also be noted that the cosine distance is Galilean-invariant as well.

The loss function is a linear combination of the mean squared error over all the output components, and the cosine distance (defined as 1 minus cosine similarity) between predicted and true velocity directions, each weighted by $\alpha = 0.5$:

$$\mathcal{L} = \alpha\left(\mathrm{MSE}_p + \mathrm{MSE}_u\right) + (1-\alpha)\,\mathrm{COSDIST}\ . \tag{16}$$

2.3.2 Numerical setup

The numerical jobs are performed on NVIDIA Volta V100 GPU cards for around five hours.

2.3.3 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.

References

[1] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[2] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (2015). URL https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer International Publishing, Cham.

[3] Cao, H. *et al.* Swin-unet: Unet-like pure transformer for medical image segmentation (2021). arXiv:2105.05537.

[4] Liu, Z. *et al.* Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883* (2021).

[5] Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2016). URL https://arxiv.org/abs/1511.07289. Published as a conference paper at the International Conference on Learning Representations (ICLR 2016).

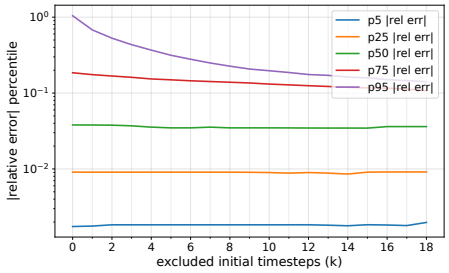
[6] Vaswani, A. *et al.* Attention is all you need. Advances in Neural Information Processing Systems 30 (2017). URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Pp. 5998–6008. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[7] Krueger, T. *et al. The Lattice Boltzmann Method: Principles and Practice* Graduate Texts in Physics (Springer, 2016).

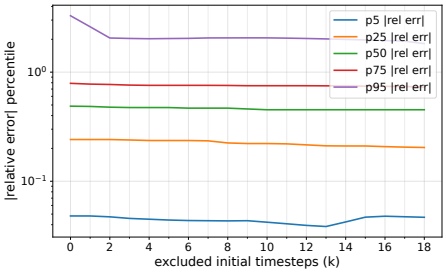
[8] Bouzidi, M., Firdaouss, M. & Lallemand, P. Momentum transfer of a boltzmann-lattice fluid with boundaries. *Physics of Fluids* **13**, 3452–3459 (2001).

[9] 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.

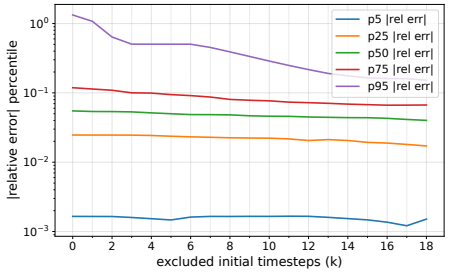
[10] He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015). URL https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html.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



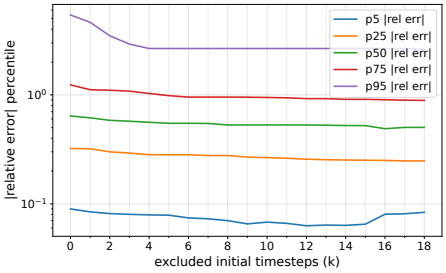
(a) SWIN C



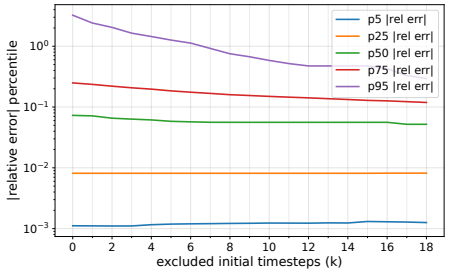
(b) SWIN id C



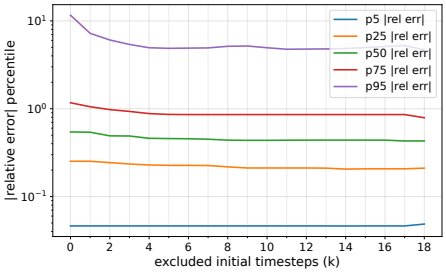
(c) UNet C



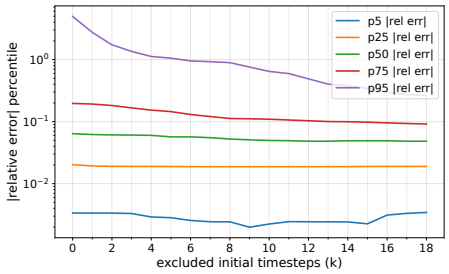
(d) UNet id C



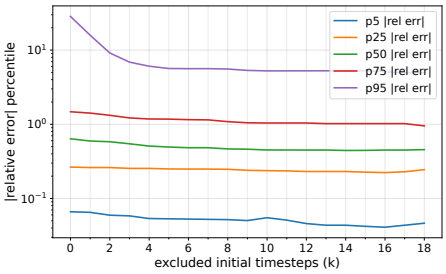
(e) SWIN S



(f) SWIN id S

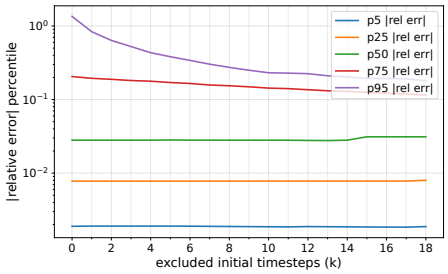


(g) UNet S

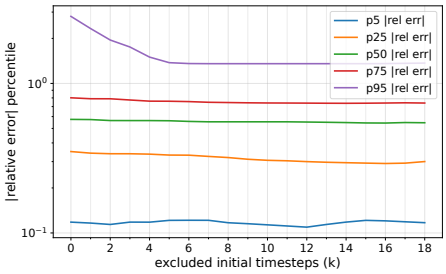


(h) UNet id S

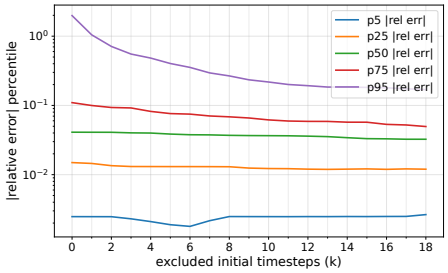
Supplementary Fig. S24: Effect of removing the first k timesteps on pressure force $F_{x,p}$.



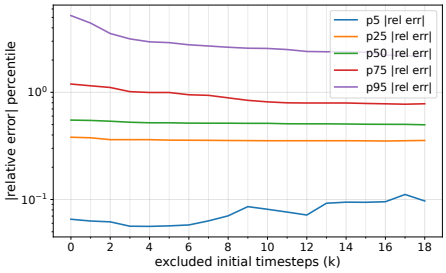
(a) SWIN C



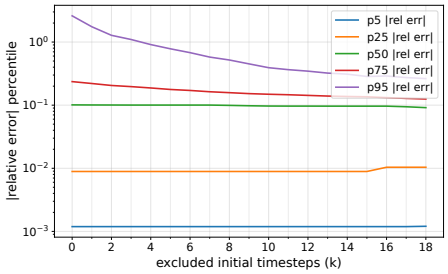
(b) SWIN id C



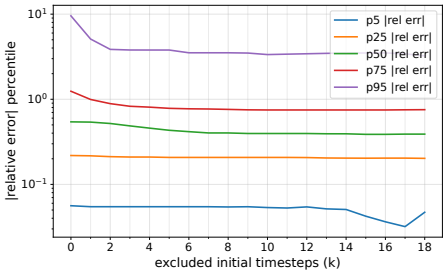
(c) UNet C



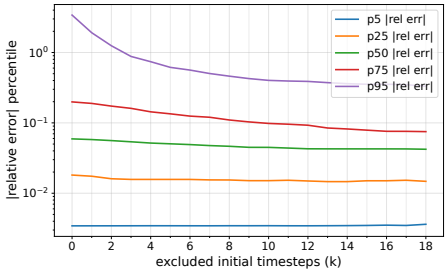
(d) UNet id C



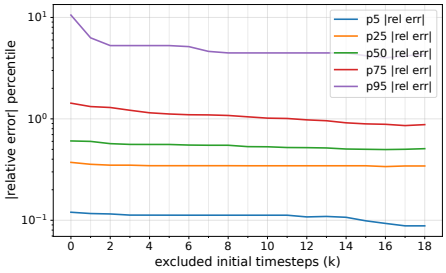
(e) SWIN S



(f) SWIN id S

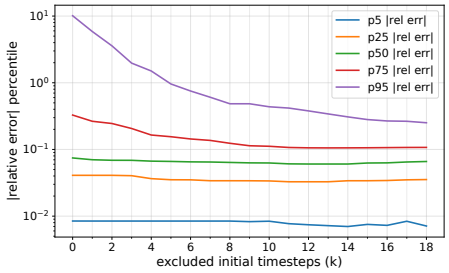


(g) UNet S

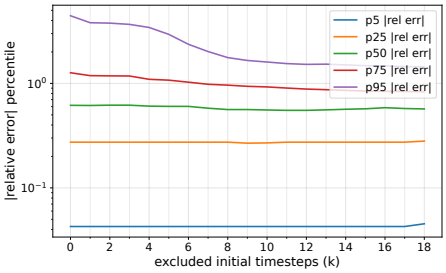


(h) UNet id S

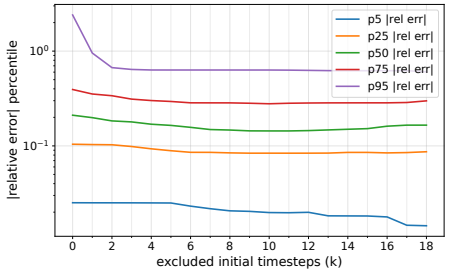
Supplementary Fig. S25: Effect of removing the first k timesteps on transverse pressure drag $F_{y,p}$.



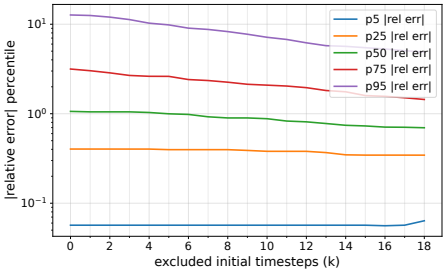
(a) SWIN C



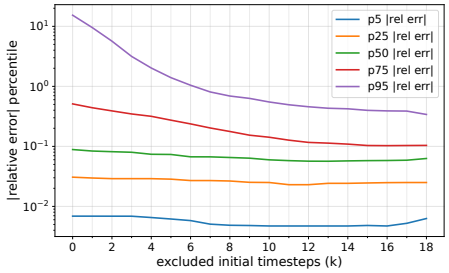
(b) SWIN id C



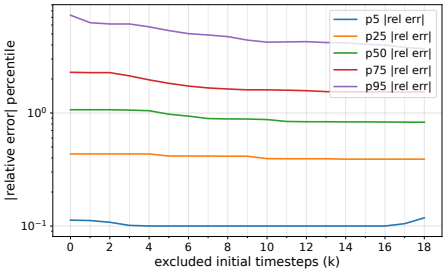
(c) UNet C



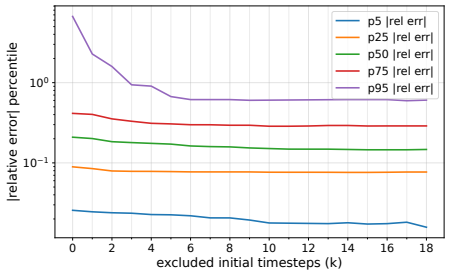
(d) UNet id C



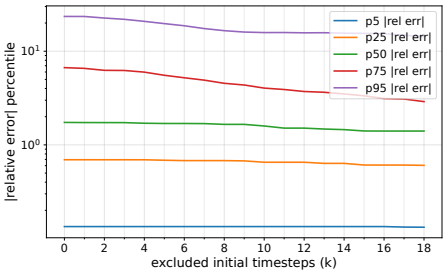
(e) SWIN S



(f) SWIN id S

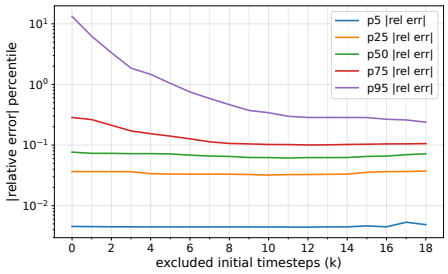


(g) UNet S

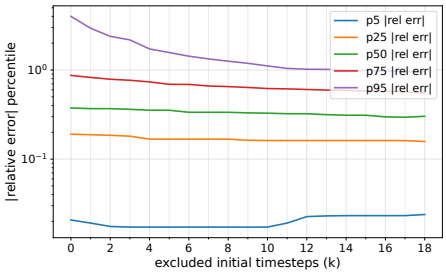


(h) UNet id S

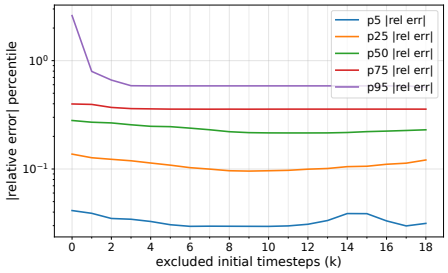
Supplementary Fig. S26: Effect of removing the first k timesteps on viscous force $F_{x,v}$.



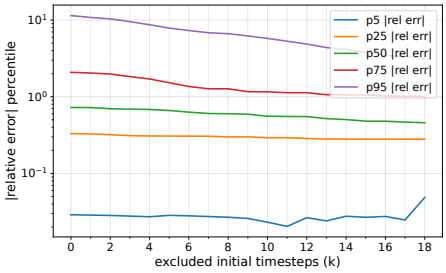
(a) SWIN C



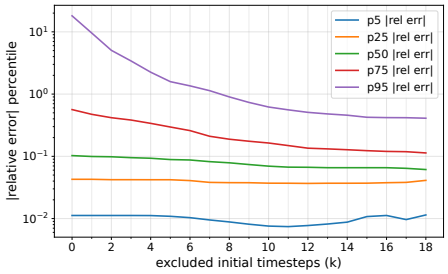
(b) SWIN id C



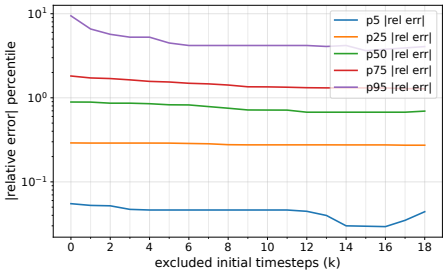
(c) UNet C



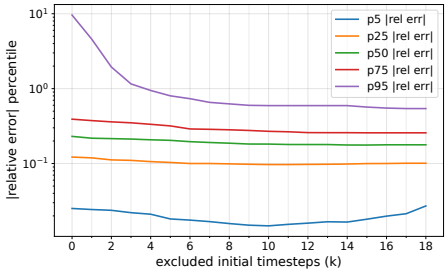
(d) UNet id C



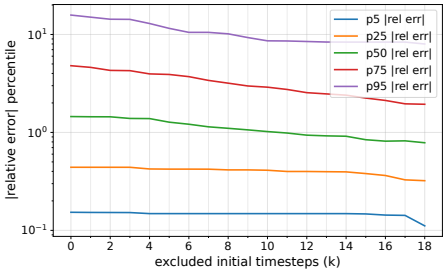
(e) SWIN S



(f) SWIN id S

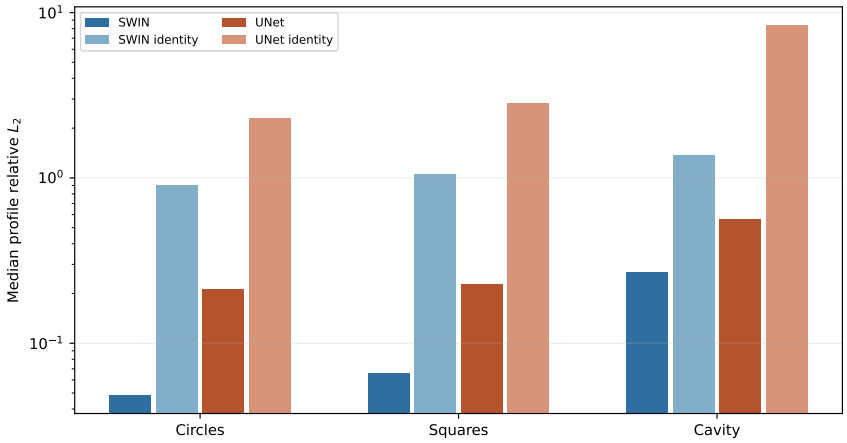


(g) UNet S

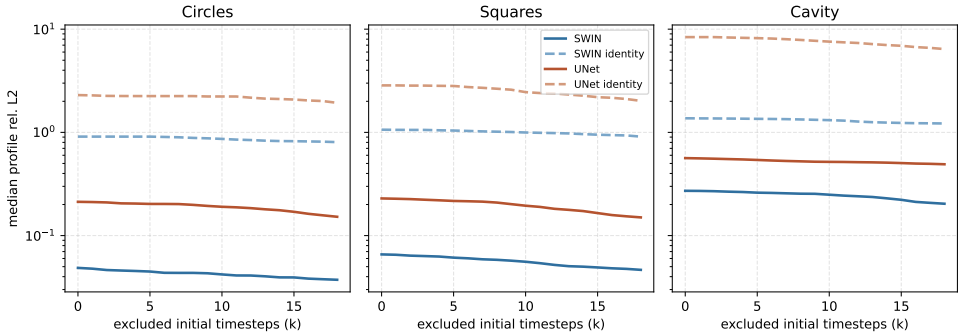


(h) UNet id S

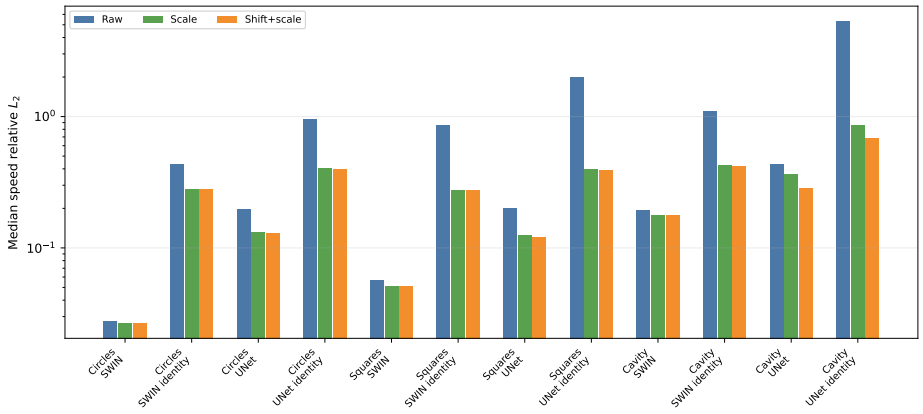
Supplementary Fig. S27: Effect of removing the first k timesteps on transverse viscous drag $F_{y,v}$.



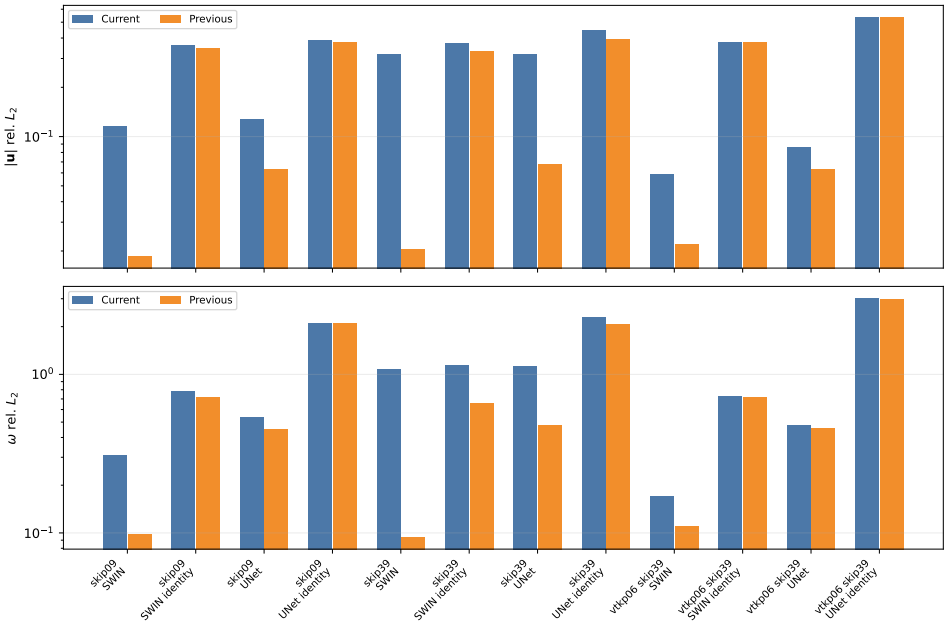
Supplementary Fig. S28: Median vertical $u_x(y)$ profile relative L2 error, including the identity-backbone ablations.



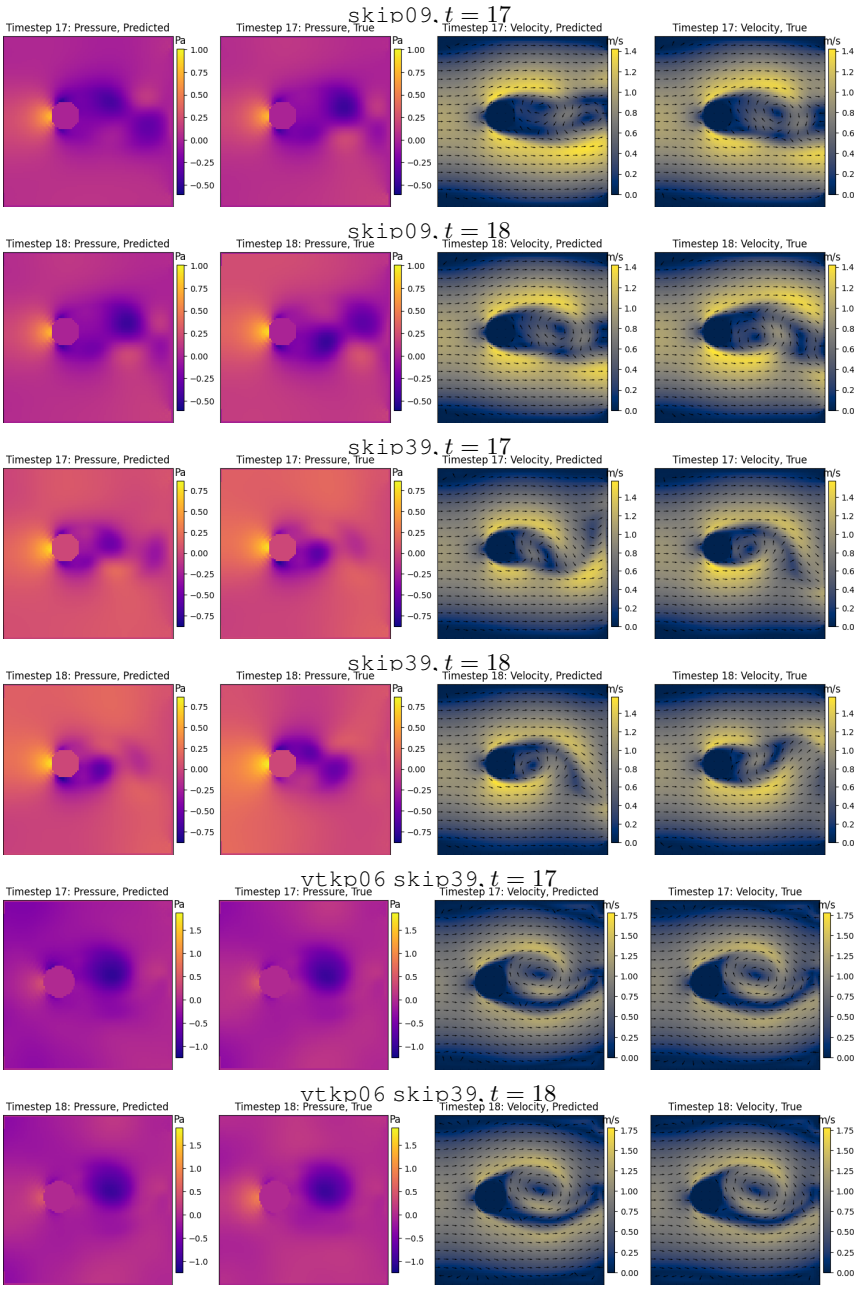
Supplementary Fig. S29: Effect of removing the first k predicted timesteps on the median vertical $u_x(y)$ profile relative L2 error.



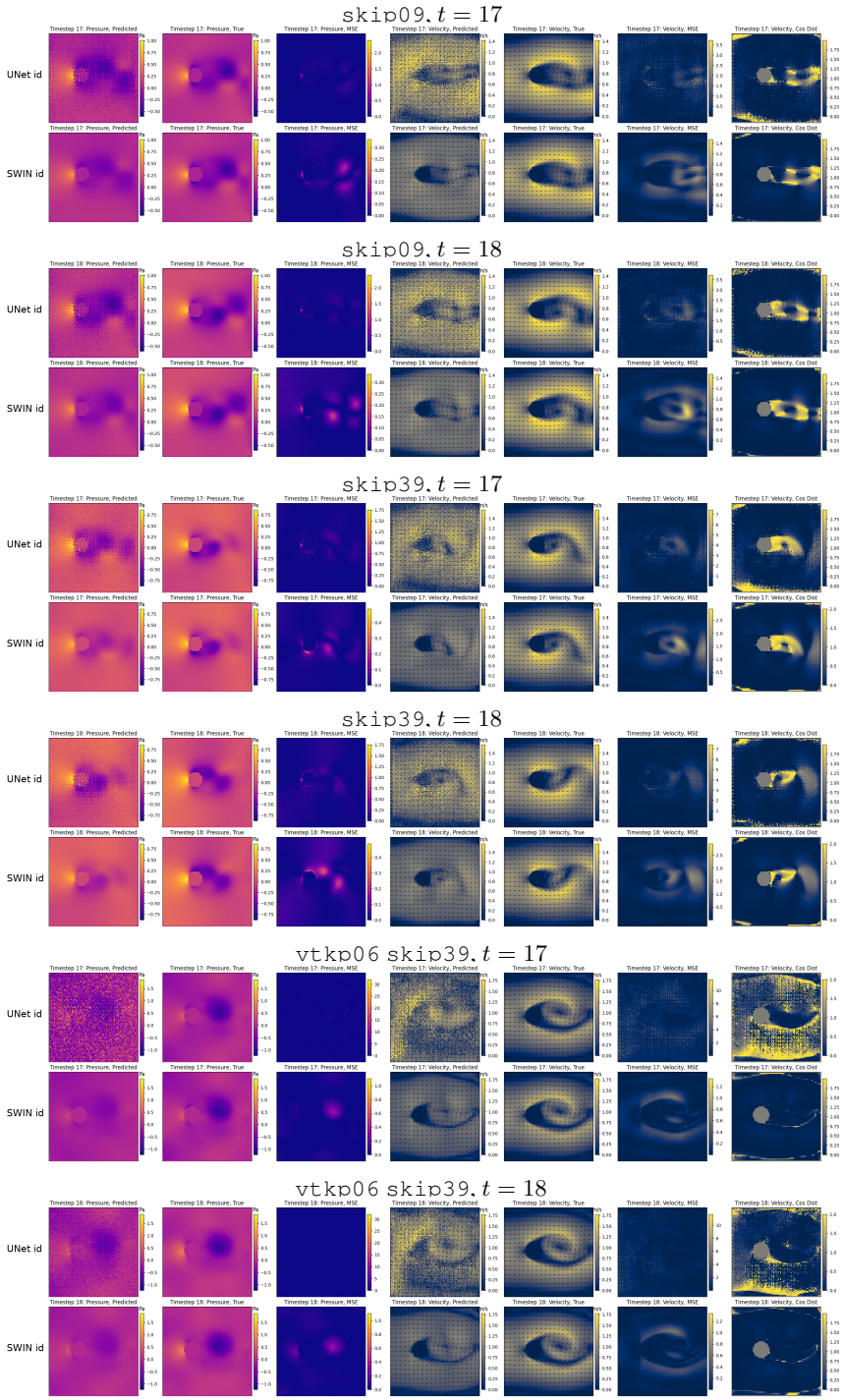
Supplementary Fig. S30: Median speed-field relative L2 error before and after scale-only and shift+scale corrections, including the identity-backbone ablations.



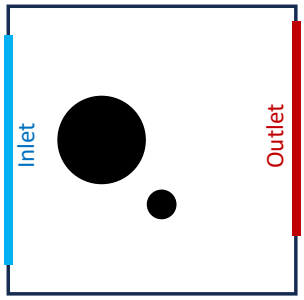
Supplementary Fig. S31: Von Karman lag diagnostic for the original-timestep `skip09` and `skip39` variants and for the reduced-timestep `vtkp06 skip39` variant, including the identity-backbone ablations. Current compares predicted fields with the same true timestep, whereas previous compares them with the preceding true timestep.



Supplementary Fig. S32: Two consecutive von Karman snapshots for example 1. Each panel reports pressure and velocity prediction/ground truth for the corresponding variant and timestep.



Supplementary Fig. S33: Identity-backbone von Karman snapshots for example 1. In each panel, UNet identity is shown on the top row and SWIN identity on the bottom row. Relative to Figure S32, the wake structures remain only partially recognisable and are visibly less coherent.



Supplementary Fig. S34: Example of a generated computational domain.