

Supplementary Information

When does physics supervision help a transformer? A significance-tested analysis of the loss channel

Sehaj Randhir Singh
Correspondence: sehajrsinghs@gmail.com

SI 1. Per-seed held-out error

Table 1 lists the five per-seed held-out rel-MAE values for each domain $\times$ architecture (from `significance.json`).

Table 1: Held-out rel-MAE per seed (5 seeds).

Domain	arch	s0	s1	s2	s3	s4
beam	phys	0.1407	0.1536	0.1479	0.1540	0.1380
beam	nophys	0.2327	0.1359	0.2950	0.1716	0.1287
beam	mlp	0.1338	0.1351	0.1705	0.1867	0.1549
cantilever	phys	0.1320	0.1562	0.1828	0.1458	0.1231
projectile	phys	0.1701	0.0607	0.1000	0.0468	0.1603
burgers	phys	0.0088	0.0082	0.0067	0.0083	0.0129
heat2d	phys	0.0119	0.0124	0.0107	0.0257	0.0124

(Full per-seed tables for every cell, including residual values, are committed in `significance.json`; only a representative subset is printed here to keep the SI compact.)

SI 2. Pooled statistics

Pooled over 25 paired (phys, nophys) runs, 5 seeds $\times$ 5 domains:

Table 2: Pooled paired statistics.

Metric	Wilcoxon p	Cliff's δ	median ratio
phys_resid (consistency)	1.79×10^{-7}	-0.92	$19.33 \times$
phys_resid vs. mlp	2.21×10^{-5}	-0.69	$6.52 \times$
val_rel_mae (accuracy)	0.647	+0.04	1.00 (n.s.)

SI 3. Normalization

Of the 75 matrix runs, 60 use shape normalization (each trajectory normalized by its own peak) and 15 use global normalization. The shape-norm protocol is the default for all trajectory domains because deflections span orders of magnitude; the ablation in the main text shows normalization effects of the same order as the physics-loss effect. All normalization code is in `physx/dataset.py`.

SI 4. Reproducibility

```
python3 physx/run_matrix.py          # 75-run matrix (< 1 hour, 20-core CPU)
python3 physx/lca_significance.py --matrix  # pooled + per-domain stats
python3 -m unittest physx.test_physx      # 49 tests
```

Every statistic in the main text is reproduced by these scripts from the committed per-run logs.