This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and the watermark; the final published version is available on SpringerLink.

Supplementary Material

ICUWorld-QR: A Residual Latent World Model for Hourly ICU Physiology

MICCAI 2026 Workshop on Medical World Models — Camera-Ready

This supplement provides per-seed numerical detail and supporting figures for results summarised in the main text, covering interval calibration, latent-norm abstention, the multi-seed ablation breakdown, bootstrap confidence intervals, full-cohort transfer, and multi-step rollouts.

Contents

	Content
Figure S1	Reliability of nominal prediction intervals
Figure S2	Interval width by volatility quintile against the empirical reference
Figure S3	Risk–coverage curves for latent-norm abstention
Figure S4	A→B seed spread at two cohort sizes
Figure S5	eICU versus Challenge marginal distributions
Table S1	Multi-seed ablation breakdown, per seed and aggregate
Table S2	Paired bootstrap confidence intervals
Table S3	Full-cohort transfer results and one-step A→B seed values
Table S4	Multi-step rollout error at h = 1, 3, 6
Table S5	Per-quintile interval coverage and width

Protocol

All results use seeds {42, 43, 44}. Unless stated otherwise, data are PhysioNet Challenge 2019 vitals, error is masked mean absolute error over genuinely observed entries, and the frozen configuration sets the JEPA weight to zero with the variance hinge active. Throughout, WM denotes ICUWorld-QR, LV denotes last-value persistence, and negative WM−baseline differences indicate that the world model is more accurate.

Where interval widths are compared against an empirical reference, that reference is the constant-within-quintile width required for 80% coverage on the seven dense vitals (HR, O_2Sat, Temp, SBP, MAP, DBP, Resp). It is computed from realised residuals rather than produced by any model.

Supplementary Figures

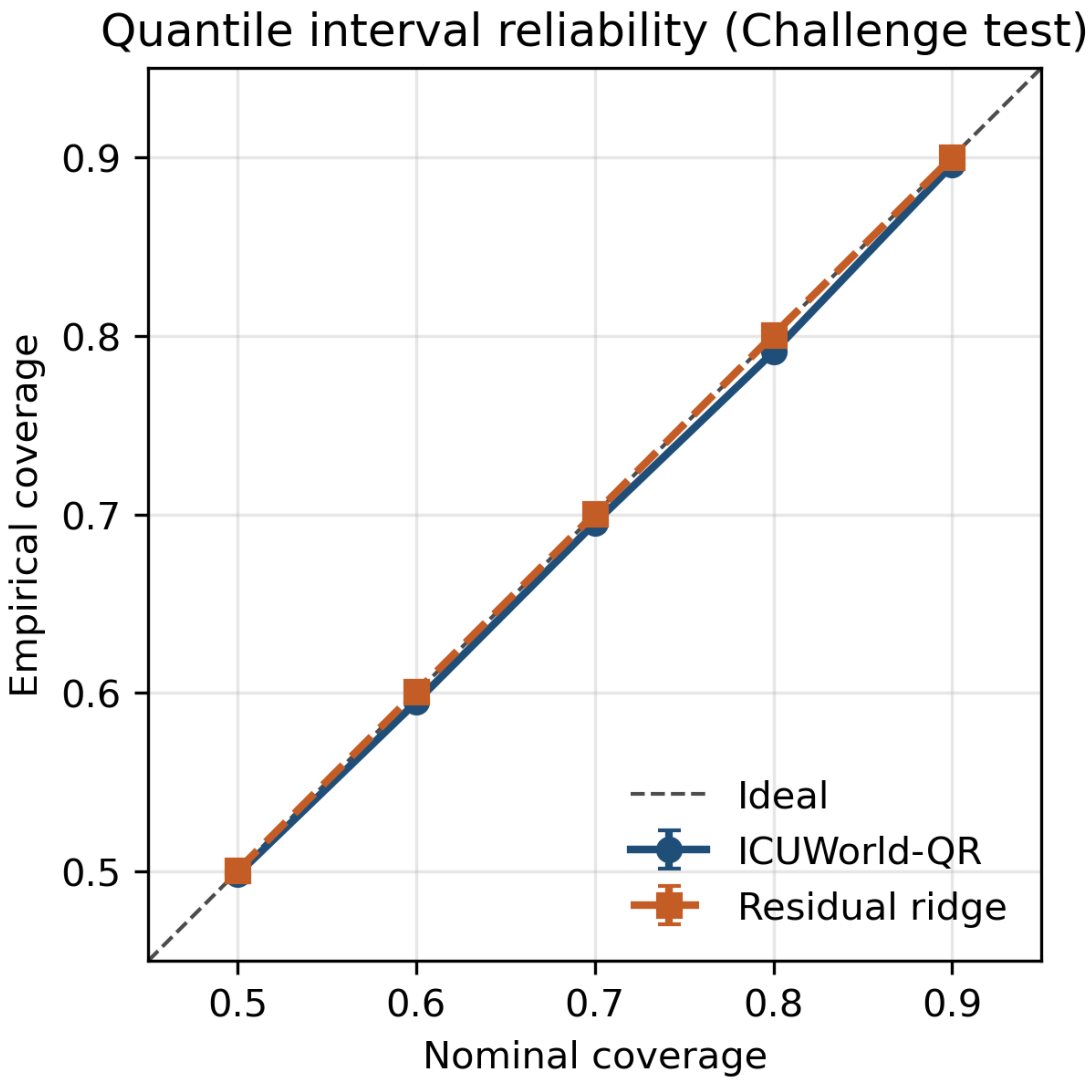

Figure S1. Reliability of nominal prediction intervals from pinball $q_{0.1}$ / $q_{0.9}$ heads on the frozen trunk, against a flat ridge residual interval. Curves give empirical coverage versus nominal coverage; the dashed diagonal is perfect calibration.

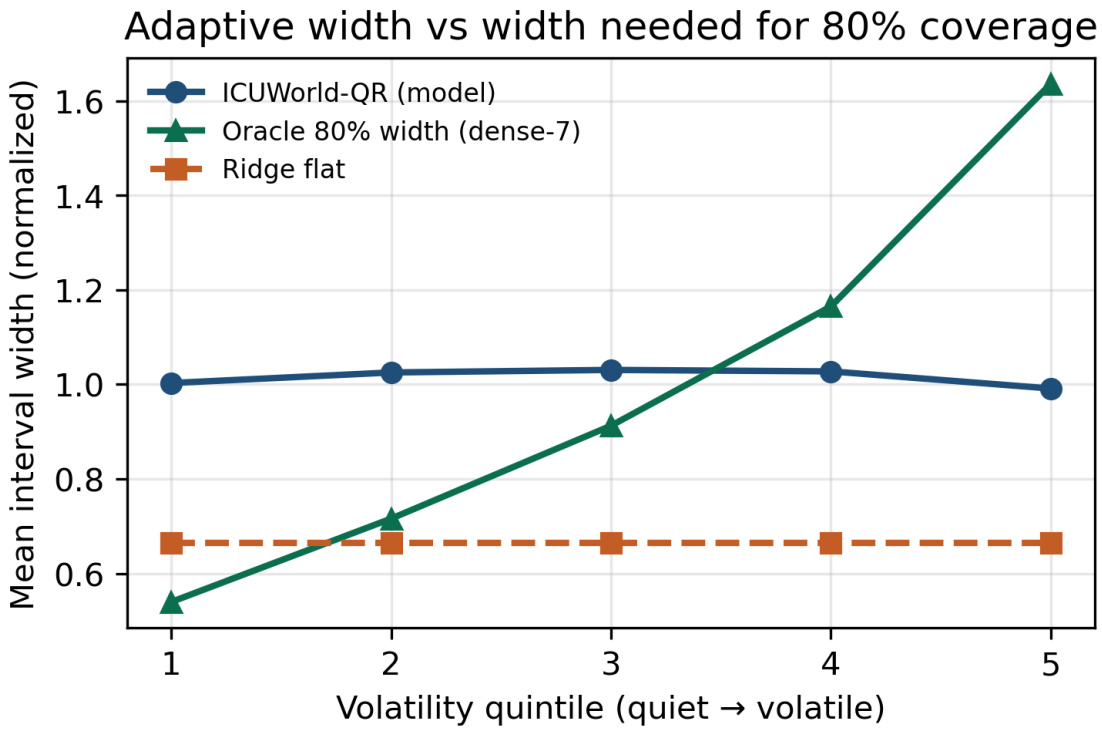

Figure S2. Mean interval width by volatility quintile: model pinball widths against the empirical width required for 80% coverage within each quintile, with the flat ridge width for reference. The model widens from quiet to non-quiet hours but under-adapts relative to the reference, by a factor of roughly 1.5 against 3.0 from Q1 to Q5.

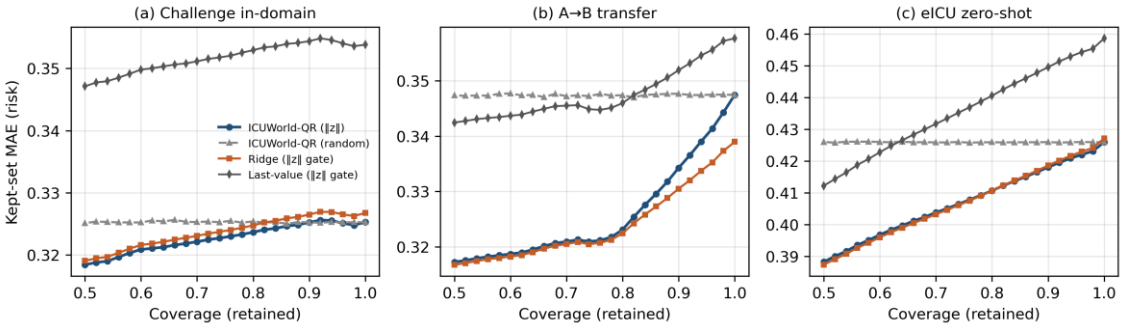

Figure S3. Risk–coverage curves for latent-norm abstention in-domain and under transfer. In-domain gating is indistinguishable from random; under shift it improves on random but is matched by the ridge gated on the same latent norm, indicating detection of generally difficult hours rather than model-specific failure.

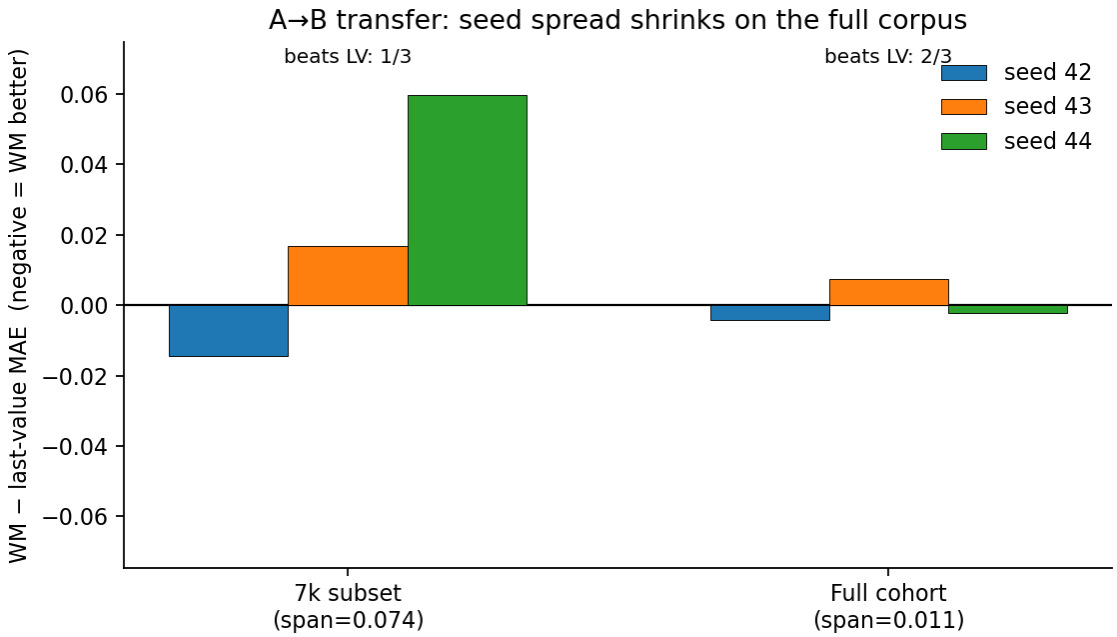

Figure S4. A→B seed spread of one-step WM−LV error on the 7,000-patient subset (left) and the full corpus (right), seeds 42–44. The subset span of 0.074 shrinks to 0.011 on the full cohort, where the model matches rather than beats persistence on average.

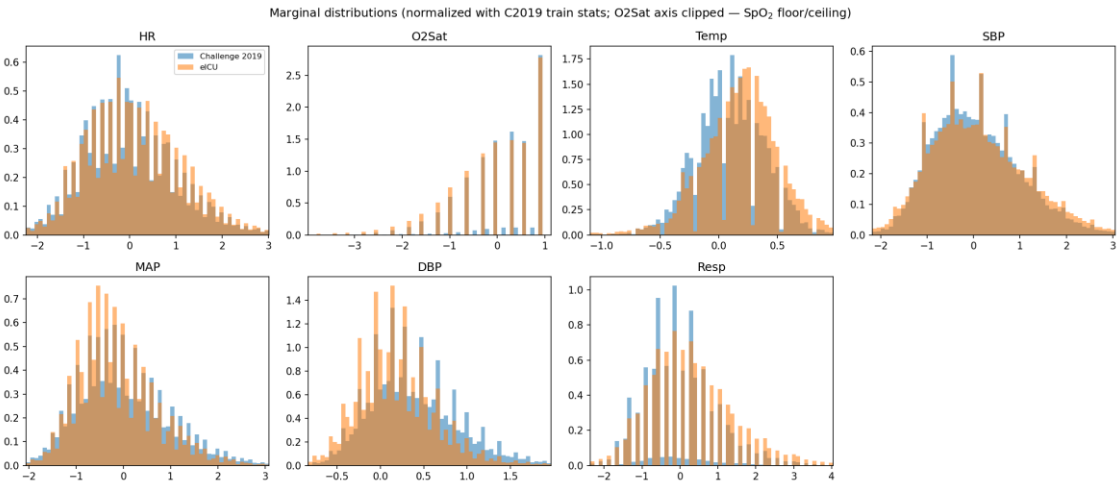

Figure S5. Marginal distributions of the seven dense vitals in eICU against PhysioNet Challenge 2019, both standardized with Challenge training statistics. All channel shifts fall below 0.5 SD. The SpO_2 axis is clipped; the right-hand spike is the 100% ceiling and the left tail reflects the 30% physiological floor.

Supplementary Tables

Table S1. Multi-seed ablation breakdown

Rows follow the Table 1 protocol: 7,000 training and 1,500 test patients, patient-level split, checkpoint selection by validation masked MAE. A latent is counted as alive when its median per-dimension standard deviation exceeds 0.1 and sepsis AUROC exceeds 0.70. The no-hinge ablation collapses on all three seeds.

Variant	Seed	MAE	MAE hard	AUROC	z SD	cos	Alive
Freeze (hinge)	42	0.3329	0.4564	0.773	1.704	0.648	Y
	43	0.3330	0.4567	0.784	1.871	0.722	Y
	44	0.3325	0.4555	0.789	1.767	0.602	Y
Without hinge	42	0.3336	0.4648	0.660	0.000	1.000	N
	43	0.3388	0.4730	0.650	0.000	1.000	N
	44	0.3239	0.4552	0.646	0.000	1.000	N
Without ridge	42	0.3635	0.5380	0.742	3.161	0.919	Y
	43	0.3662	0.5441	0.736	3.802	0.932	Y
	44	0.3518	0.5238	0.716	4.144	0.925	Y
Without history	42	0.3713	0.4955	0.751	1.268	0.486	Y
	43	0.3781	0.5047	0.753	1.251	0.537	Y
	44	0.3652	0.4858	0.775	1.238	0.440	Y
Without context	42	0.3345	0.4654	0.705	2.725	0.905	Y
	43	0.3401	0.4737	0.718	2.695	0.890	Y
	44	0.3249	0.4559	0.692	2.994	0.911	N

Table S1b. Aggregate across seeds.

Variant	MAE	AUROC	z SD	Alive
Freeze (hinge)	0.3328 ± 0.0002	0.782 ± 0.006	1.781 ± 0.069	3/3
Without hinge	0.3321 ± 0.0062	0.652 ± 0.006	0.000 ± 0.000	0/3
Without ridge	0.3605 ± 0.0063	0.731 ± 0.011	3.702 ± 0.407	3/3
Without history	0.3715 ± 0.0053	0.760 ± 0.011	1.252 ± 0.012	3/3
Without context	0.3331 ± 0.0062	0.705 ± 0.011	2.805 ± 0.135	2/3

Table S2. Paired bootstrap confidence intervals

Paired patient-level bootstrap with 2,000 resamples. The difference is MAE(WM) − MAE(baseline), so negative values favour the world model. The final column marks whether the 95% interval excludes zero.

Setting	Seed	Versus	Difference	95% CI	Excl. 0
In-domain	42	ridge	-0.0012	[-0.0017, -0.0007]	Y
		LV	-0.0271	[-0.0286, -0.0257]	Y
	43	ridge	-0.0015	[-0.0020, -0.0009]	Y
		LV	-0.0267	[-0.0283, -0.0251]	Y
	44	ridge	-0.0017	[-0.0022, -0.0012]	Y
		LV	-0.0263	[-0.0279, -0.0247]	Y
A→B (7k)	42	ridge	+0.0046	[0.0033, 0.0063]	Y
		LV	-0.0144	[-0.0165, -0.0122]	Y
	43	ridge	+0.0353	[0.0290, 0.0422]	Y
		LV	+0.0168	[0.0100, 0.0239]	Y
	44	ridge	+0.0767	[0.0623, 0.0916]	Y
		LV	+0.0597	[0.0452, 0.0754]	Y
B→A (7k)	42	ridge	+0.0030	[0.0023, 0.0039]	Y

		LV	-0.0182	[-0.0199, -0.0165]	Y
	43	ridge	+0.0046	[0.0038, 0.0055]	Y
		LV	-0.0175	[-0.0193, -0.0158]	Y
	44	ridge	+0.0026	[0.0015, 0.0039]	Y
		LV	-0.0167	[-0.0186, -0.0145]	Y

Table S3. Full-cohort results (40,336 stays)

Frozen configuration, batch size 256, eight epochs, checkpoint by validation masked MAE. In-domain uses all Challenge stays with a patient-level split and roughly 20% held out. A→B trains on all 20,336 site-A stays and tests on all 20,000 site-B stays; B→A reverses this.

Setting	Seed	WM	LV	Ridge	WM−LV	WM−Ridge	AUROC
In-domain	42	0.3277	0.3571	0.3317	-0.0294	-0.0040	0.778
	43	0.3267	0.3560	0.3311	-0.0292	-0.0044	0.804
	44	0.3282	0.3561	0.3312	-0.0279	-0.0029	0.801
A→B	42	0.3803	0.3844	0.3663	-0.0041	+0.0140	0.732
	43	0.3888	0.3815	0.3631	+0.0073	+0.0256	0.623
	44	0.3818	0.3841	0.3655	-0.0023	+0.0162	0.642
B→A	42	0.3729	0.3880	0.3665	-0.0151	+0.0064	0.717
	43	0.3679	0.3874	0.3662	-0.0195	+0.0018	0.641
	44	0.3683	0.3870	0.3654	-0.0187	+0.0029	0.713

Table S3b. Aggregate across seeds.

Setting	WM MAE	WM−LV	WM−Ridge	AUROC
In-domain	0.3276 ± 0.0006	-0.0288 ± 0.0007	-0.0038 ± 0.0006	0.794 ± 0.011
A→B	0.3836 ± 0.0037	+0.0003 ± 0.0050	+0.0186 ± 0.0050	0.665 ± 0.048
B→A	0.3697 ± 0.0023	-0.0177 ± 0.0019	+0.0037 ± 0.0020	0.690 ± 0.035

Table S3c. One-step A→B error by seed. These are the values reported in the main text; they come from the transfer experiment rather than the rollout evaluation, and differ from Table S4 at h = 1 for the reason given there.

Cohort	Seed	WM	LV	WM−LV
7,000 patients	42	0.3698	0.3842	-0.0144
	43	0.4025	0.3858	+0.0168
	44	0.4381	0.3784	+0.0597
Full corpus	42	0.3803	0.3844	-0.0041
	43	0.3888	0.3815	+0.0073
	44	0.3818	0.3841	-0.0023

The seed span narrows from 0.074 on the subset to 0.011 on the full corpus.

Table S4. Multi-step rollout

Open-loop rollout error, mean ± std over seeds. Rollouts are evaluated on a subsample of up to 2,000 stays using history-conditioned open-loop prediction, so the h = 1 values are not numerically identical to the one-step test error reported elsewhere even for the same trained weights.

Setting	h	WM	LV	Ridge	WM−LV	WM−Ridge
In-domain, 7k	1	0.3346 ± 0.0002	0.3617	0.3367	-0.0271	-0.0021
	3	0.3636 ± 0.0007	0.3960	0.3626	-0.0324	+0.0010
	6	0.4054 ± 0.0029	0.4302	0.3918	-0.0248	+0.0136
In-domain, full	1	0.3433 ± 0.0068	0.3764	0.3480	-0.0331	-0.0048

	3	0.3731 ± 0.0108	0.4102	0.3746	-0.0371	-0.0015
	6	0.4108 ± 0.0154	0.4387	0.3999	-0.0278	+0.0110
A→B, full	1	0.3688 ± 0.0092	0.3787	0.3582	-0.0099	+0.0105
	3	0.4081 ± 0.0138	0.4187	0.3944	-0.0106	+0.0137
	6	0.4550 ± 0.0187	0.4590	0.4335	-0.0040	+0.0215
B→A, full	1	0.3788 ± 0.0112	0.3988	0.3746	-0.0201	+0.0042
	3	0.4190 ± 0.0100	0.4301	0.4039	-0.0111	+0.0151
	6	0.4763 ± 0.0122	0.4580	0.4334	+0.0182	+0.0429
A→B, 7k	1	0.3867 ± 0.0278	0.3778	0.3586	+0.0090	+0.0281
	3	0.4384 ± 0.0389	0.4177	0.3953	+0.0208	+0.0432
	6	0.5001 ± 0.0547	0.4579	0.4353	+0.0422	+0.0648
B→A, 7k	1	0.3796 ± 0.0086	0.3982	0.3755	-0.0186	+0.0041
	3	0.4162 ± 0.0051	0.4295	0.4042	-0.0133	+0.0120
	6	0.4672 ± 0.0015	0.4573	0.4334	+0.0099	+0.0338

Table S5. Per-quintile interval coverage and width

Volatility quintiles are formed on mean absolute next-hour change. The nominal interval is $[q_{0.1}, q_{0.9}]$, giving 80% target coverage. Overall empirical coverage is 0.791 ± 0.003 for ICUWorld-QR and 0.801 ± 0.001 for the ridge.

Quintile	WM coverage	WM width	Empirical width for 80%	Ridge width
Q1 (quiet)	0.948 ± 0.000	0.697 ± 0.024	0.540 ± 0.013	0.664
Q2	0.898 ± 0.001	1.056 ± 0.028	0.716 ± 0.017	0.664
Q3	0.831 ± 0.001	1.064 ± 0.027	0.913 ± 0.018	0.664
Q4	0.750 ± 0.005	1.072 ± 0.024	1.165 ± 0.026	0.664
Q5 (volatile)	0.657 ± 0.005	1.075 ± 0.025	1.636 ± 0.031	0.664

Table S5b. Coverage by seed.

Seed	Overall	Q1	Q2	Q3	Q4	Q5
42	0.795	0.947	0.898	0.832	0.756	0.664
43	0.789	0.948	0.900	0.830	0.746	0.653
44	0.789	0.948	0.897	0.830	0.749	0.654