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

Arti Taneja[1], Hasmila A. Omer[2], Shantaram Vasikarla[3], and Priya Ranjan[4]

[1] Independent Researcher tanejaai@mail.com, arti.taneja@gmail.com
[2] Coventry University, Coventry, UK Hasmilaakmar@gmail.com
[3] California State University, Northridge, CA, USA shantaram@computer.org
[4] Shinkai-Labs, Chennai, India pranjan@ieee.org

Abstract. Static ICU risk models fail to track rapid hourly physiological dynamics in critically ill patients. We introduce ICUWorld-QR, a residual latent world model for ICU trajectory forecasting. The architecture combines a history-conditioned encoder, a transition module driven by weak context (FiO_2 and ICU hour), and a decoder predicting residual corrections relative to a linear baseline. We identify a core vulnerability in residual architectures: a strong linear prior removes gradient incentive for representation learning, collapsing the latent state into a constant. Enforcing a batch-wise variance hinge constraint prevents collapse and restores sepsis ranking, whereas a JEPA-style objective does neither. On PhysioNet Challenge 2019, ICUWorld-QR outperforms last-value persistence by 7.5% overall (MAE 0.333 vs. 0.360) and 15% on hard hours, while its latent state ranks sepsis onset (AUROC 0.782 ± 0.006). Cross-site transfer exhibits seed fragility at small cohort sizes; training on the full corpus stabilizes forecasting but not sepsis ranking. Predictive intervals under-adapt to volatility, leaving a fixed-width ridge better calibrated. Zero-shot transfer across 204 eICU hospitals matches a source-fitted ridge, while target fine-tuning provides modest gains that offer no advantage over the frozen initialization when target hospitals are scarce.

Keywords: World models · ICU physiology · Time-series forecasting · Sepsis · Representation collapse

1 Introduction

Intensive care patients are monitored continuously, yet most machine learning on these data reduces the resulting time series to a static prediction over a fixed window: mortality, or whether sepsis will occur. Such single-horizon classifiers cannot forecast how physiology will evolve over the next several hours, quantify uncertainty across future time steps, or be queried about sensitivity to context such as oxygen delivery. Clinicians reason in terms of trajectories, whereas models predicting isolated events lack an internal representation of temporal dynamics.

World models address this by learning how a patient's latent state evolves under varying observations and context [4,9], with recent clinical applications spanning treatment-conditioned disease progression [15], EHR latent models for

treatment policy learning [13], and ICU trajectory forecasting [8]. Two gaps remain. First, existing ICU forecasting frameworks [7,5,14,8] evaluate models strictly on predictive accuracy; because their latent representations are never tested on downstream clinical tasks, it is unclear whether they capture physiological structure or merely minimize next-step error. We do not re-benchmark them, as their horizons, feature sets, and splits differ substantially from ours. Second, while clinical world models are typically developed on MIMIC and eICU, the PhysioNet Challenge 2019 corpus [11], comprising 40,336 open ICU stays with hourly physiology and sepsis labels, is used almost exclusively for static classification, leaving its temporal dynamics unexamined.

Forecasting ICU vitals one hour ahead appears simple yet is difficult to improve upon. Because vitals change slowly, last-value carry-forward is already strong and a ridge on recent values and trends is stronger still. A learned model can either predict from scratch against that baseline or predict a residual correction on top of it; the residual guarantees baseline performance but leaves less signal to learn. We adopt the residual approach and study its trade-offs.

We present ICUWorld-QR, a residual latent world model for hourly ICU physiology. An encoder maps a short window of observations to a latent state, a transition module advances it under weak observational context such as inspired oxygen fraction (FiO_2) and time since admission, and a decoder predicts a correction on top of a frozen linear prior. The same state supports a sepsis readout [12], quantile prediction intervals [6], and an abstention signal from the latent norm, so it must encode information useful beyond next-hour prediction alone.

Contributions. We introduce a variance-regularized residual latent world model combining a ridge residual, a latent transition module, and a batch standard deviation hinge [2], and assess whether a JEPA-style objective [1] is needed alongside it. Through ablation we identify when the encoder fails to learn patient representations and the constraint needed to restore them. We evaluate the latent on three tasks rather than forecast error alone: next-hour prediction, sepsis ranking [12], and prediction intervals that respond to volatility [6]. Finally, we evaluate transfer across corpora, covering cross-site within Challenge, zero-shot across 204 eICU hospitals [10], and target fine-tuning against from-scratch and refit-ridge baselines.

2 Methods

2.1 State, Context, and History

Let $s_t \in \mathbb{R}^{d_s}$ denote standardized vitals and lab values at hour t, with binary observation mask $m_t \in \{0,1\}^{d_s}$. Values are standardized using training statistics from the PhysioNet Challenge 2019 split. The context vector a_t captures weak observational features (FiO_2, ICU length of stay, unit indicators, demographics) without explicit treatment policy modeling. History h_t concatenates values, missingness indicators, context, and first-difference velocities over a lookback window of $H = 4$ hours.

2.2 Residual Latent Dynamics

$$z_t = e(h_t), \quad \hat{z}_{t+1} = f(z_t, a_t), \quad \hat{s}_{t+1} = s_t + \Delta_{\text{ridge}}(h_t) + d(\hat{z}_{t+1}, a_t) \tag{1}$$

where e encodes history h_t, f advances latent state z_t, and d decodes residuals. The frozen prior Δ_{ridge} drives predictions until the near-zero-initialized decoder learns residual capacity.

2.3 Variance Hinge

A strong ridge prior allows a constant encoder to yield near-optimal dynamic predictions, providing little gradient incentive to learn an informative latent state z and leading to latent collapse. To prevent this, we apply a VICReg-style [2] batch standard deviation hinge loss across each latent dimension j:

$$\mathcal{L}_{\text{hinge}} = \sum_j \max\bigl(0,\, 1 - \sqrt{\text{Var}_j(z) + \epsilon}\bigr), \quad \epsilon = 10^{-4}, \tag{2}$$

weighted by $\lambda_{\text{hinge}} = 1$. We also evaluate a JEPA-style latent prediction loss [1] aligning predicted state $\hat{z}_{t+1}$ with a stop-gradient encoding of h_{t+1}. Its impact is analyzed in Sec. 3; our frozen configuration sets $\lambda_{\text{jepa}} = 0$.

2.4 Sepsis Ranking Head

A shallow prediction head $r(z_t)$ estimates the patient's sepsis status (`SepsisLabel`) in the subsequent hour [11,12] using class-balanced binary cross-entropy, providing a queryable clinical risk ranking directly from the latent state. Because r reads exclusively from z_t, its discriminative performance serves as a direct probe of whether the latent state retains patient-specific clinical information. We do not target the official Challenge utility score.

2.5 Quantile Heads

After freezing the dynamics trunk, auxiliary heads trained with pinball loss [6] predict residual quantiles ($q_{0.1}$ and $q_{0.9}$) while leaving the median forecast path unchanged. This yields a prediction interval that adapts to patient state via z_t, unlike residual ridge baselines that report fixed quantiles.

2.6 Training Objective

The model is trained end-to-end using a composite objective:

$$\mathcal{L} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{mae}}\mathcal{L}_{\text{mae}} + \lambda_{\text{hinge}}\mathcal{L}_{\text{hinge}} + \lambda_{\text{risk}}\mathcal{L}_{\text{risk}}. \tag{3}$$

Here $\mathcal{L}_{\text{mse}}$ and $\mathcal{L}_{\text{mae}}$ are the masked squared and absolute errors of $\hat{s}_{t+1}$, and $\mathcal{L}_{\text{risk}}$ is the class-balanced cross-entropy of the sepsis head. Dynamics terms are masked by observation indicators ($m_{t+1} = 1$), so gradients are computed only

over observed clinical values rather than missing entries. Channels are reweighted by physiological volatility, with higher-volatility channels receiving larger weights, clipped to $[0.25, 4]$. The frozen configuration uses λ_{mse}=0.45, λ_{mae}=0.25, λ_{hinge}=1, and λ_{risk}=0.08, and applies no hard-hour sample reweighting; hard-hour performance is evaluated post-hoc only.

2.7 Architecture and Optimization

The encoder and transition module are multilayer perceptrons with two hidden layers of width 256, latent dimension 64, and ReLU activations (linear outputs). The residual decoder maps $[\hat{z}_{t+1}, a_t]$ through a single width-256 ReLU hidden layer to d_s outputs, and the sepsis head is a 128-unit multilayer perceptron on z_t. Training uses Adam with initial learning rate 5×10^{-4} and multiplicative per-epoch decay $\text{lr}_e = 5 \times 10^{-4} \times 0.85^{e-1}$, batch size 512, and 8 epochs, with checkpoint selection by validation masked MAE. The model has 325,794 trainable parameters. All experiments use a reproducible NumPy/CPU implementation with seed 42 unless otherwise stated.

2.8 Transfer Protocols

The architecture remains unchanged across three transfer protocols. *Site transfer:* train on Challenge Set A (or B) and test on the other without retraining. *Zero-shot:* evaluate the Challenge-trained model on eICU dense vitals [10], standardizing target features with Challenge statistics on shared channels to measure covariate shift directly rather than absorbing it. *Adaptation:* fine-tune the transition model and decoder on eICU hospital splits while keeping the ridge prior, encoder, and sepsis head frozen.

3 Experiments

Dataset and Splitting Protocol. Experiments use PhysioNet Challenge 2019 [11,3] Sets A and B with patient-level splits. Validation data is partitioned from the training set to select model checkpoints via validation masked MAE. The frozen setup uses 7,000 training and 1,500 test patients (seed 42). All MAE metrics use standardized units (z-scores from the Challenge training split) computed strictly on observed entries ($m = 1$).

Evaluation Metrics. We evaluate next-step (one-hour-ahead) masked MAE, *hard-hour* MAE (the top quartile of absolute next-hour vital change), and sepsis AUROC using official Challenge labels. We also evaluate performance across volatility quintiles.

Baselines. We compare against last-value persistence and a history-and-velocity residual ridge model, which is Challenge-trained when used in transfer experiments.

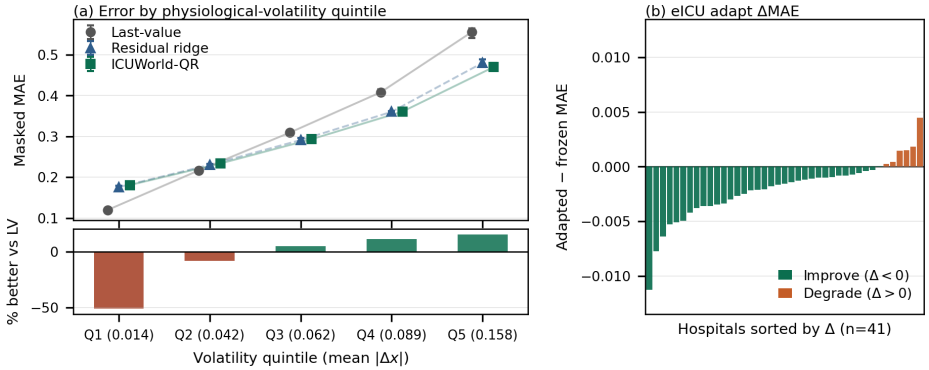

Fig. 1: **(a)** Masked MAE by physiological-volatility quintile; tick labels give mean absolute next-hour change, with approximately 10,400 hours per bin. The lower panel shows relative change against last-value. Hard-hour MAE in Table 1 uses the top quartile rather than Q5. **(b)** Per-hospital adapted minus frozen MAE across 41 held-out eICU hospitals, sorted by effect size; negative values indicate improvement (seed 42).

Table 1: In-domain results and latent diagnostics, mean over seeds 42–44 (per-seed values in the supplement). Baselines predict physiology only. Lower rows ablate individual components of the frozen configuration. z SD is the median per-dimension standard deviation of the latent and $\cos(z, \bar{z})$ its mean cosine to the batch mean; "alive" counts seeds with z SD > 0.1 and AUROC > 0.70.

Method	Next-step	Hard-hour	AUROC	z SD	$\cos(z, \bar{z})$	Alive
Last-value	0.360	0.537	—	—	—	—
Residual ridge	0.334	0.465	—	—	—	—
Frozen configuration	**0.333**	**0.456**	**0.782**	1.78	0.658	3/3
JEPA + hinge	0.333	0.465	0.745	3.74	0.937	3/3
Without hinge	0.332	0.464	0.652	$\sim 10^{-6}$	1.000	**0/3**
Without ridge prior	0.361	0.535	0.731	3.70	0.925	3/3
Without history	0.372	0.495	0.760	1.25	0.488	3/3
Without context	0.333	0.465	0.705	2.81	0.902	2/3

3.1 Main Results and Ablations

ICUWorld-QR reduces next-step MAE by ∼7.5% and hard-hour MAE by ∼15% versus last-value. Next-step MAE matches the ridge in aggregate, though a paired patient-level bootstrap gives MAE differences of −0.0015 vs. ridge and −0.027 vs. last-value, both with 95% CIs excluding zero. Stratifying error by volatility quintile (Fig. 1a) shows the model slightly underperforms last-value on quiet hours but improves as volatility increases.

Without the hinge, the encoder outputs a constant vector across patients: median z SD collapses to 10^{-6} and mean cosine to the batch mean reaches 1.00. Because forecast error is nearly unchanged (MAE ≈ 0.332, at ridge level),

collapse is invisible to prediction loss and requires latent statistics or sepsis AUROC (which drops to 0.652) to detect. Under the frozen configuration the latent retains full numerical rank across all 64 dimensions, and omitting the JEPA term improves both forecast MAE and sepsis AUROC over joint JEPA–hinge training. Omitting the ridge prior or history window degrades forecasting substantially, while omitting context leaves aggregate MAE unchanged but lowers sepsis AUROC below our liveness threshold in one of three seeds.

3.2 Site Transfer (A↔B)

Table 2: Cross-site transfer performance between Challenge hospital systems (mean ± std across seeds 42–44). ICUWorld-QR AUROC is a range (min–max) given non-normal seed variance. All runs converged with live latents and comparable source validation error, ruling out optimization failure or latent collapse.

Direction	Last-value	Ridge	ICUWorld-QR	AUROC	Mean $\|z\|$
A→B (test B)	0.383 ± 0.004	0.364 ± 0.003	0.404 ± 0.034	0.617–0.723	∼38
B→A (test A)	0.379 ± 0.002	0.358 ± 0.002	0.361 ± 0.002	0.550–0.687	∼22

Cross-site transfer exhibits marked, asymmetric seed instability. While B→A transfer is consistent across seeds, outperforming persistence by 0.017 ± 0.002 MAE, A→B forecasting varies widely across seeds 42–44 (delta vs. last-value: −0.014, +0.017, +0.060), beating persistence in only one run of three. Sepsis ranking is similarly seed-dependent (0.617–0.723 for A→B; 0.550–0.687 for B→A). Forecasting and sepsis instabilities are uncorrelated: the seed with the strongest A→B sepsis transfer yielded the worst forecasting MAE. On the full 40,336-stay cohort the A→B seed span narrows from 0.074 to 0.011 (delta vs. last-value: −0.004, +0.007, −0.002), with the model matching rather than beating persistence; sepsis transfer remains seed-dependent (0.665 ± 0.048). Ranking hours by latent norm improves kept-set error over random abstention under transfer (AURC 0.163 vs. 0.174 on A→B; 0.204 vs. 0.213 on eICU) but not in-domain (0.157 vs. 0.158), and the same ranking benefits the ridge equally. A threshold fixed at the source validation P90 yields coverage of 0.89, 0.05, and 0.98 across the three settings.

3.3 Sensitivity Probe and External Transfer

Sweeping input FiO_2 across its physiological range monotonically increases predicted next-hour SpO_2. Modifying FiO_2 by ± 0.10 raises predicted SpO_2 in nearly all test hours (median increase +0.12 percentage points), reflecting observational model sensitivity rather than a causal effect.

Under zero-shot transfer to eICU dense vitals [10] (∼25,000 stays, 1.18M transitions, 204 hospitals), ICUWorld-QR achieves 0.428 ± 0.001 MAE across seeds 42–44 versus 0.459 for last-value. Performance closely matches the ridge baseline (difference -0.0010 ± 0.0012; per-seed −0.0017, +0.0003, −0.0017).

Among hospitals with ≥ 200 recorded hours, hospital-level MAE is 0.425 (SD = 0.054) for ICUWorld-QR, 0.427 (SD = 0.055) for ridge, and 0.458 for last-value. Sepsis evaluation is omitted due to missing eICU label endpoints.

3.4 Quantile Heads

Pinball [6] $q_{0.1}/q_{0.9}$ heads on the frozen trunk leave median MAE unchanged at ≈ 0.33. Mean interval width is 0.71 at Q1, then 1.07–1.09 across Q2 to Q5, while the ridge is flat at 0.66. The change is a step from quiet to non-quiet hours rather than a smooth monotone rise. Width adaptation is not calibration: empirical coverage of the nominal 80% interval is 0.791 ± 0.003 overall, against 0.801 ± 0.001 for the ridge, and drifts from 0.95 at Q1 to 0.66 at Q5. The empirical width required for 80% coverage rises from 0.54 to 1.64, a factor of 3.0, whereas the model widens by 1.5.

3.5 Multi-Step Rollout

Autoregressive rollout to three and six hours, evaluated on a 2,000-stay subsample, preserves the margin over last-value but not over the ridge. In-domain, ICUWorld-QR reduces MAE against last-value by 0.027, 0.032, and 0.025 at one, three, and six hours, while the gap against the ridge moves from -0.002 at one hour to $+0.014$ at six as errors compound. The same ordering holds on the full cohort, where B$\rightarrow$A at six hours falls behind last-value ($+0.018$).

3.6 eICU Dynamics Adaptation

To adapt to target dynamics across seeds 42–44, we fine-tune the transition model and decoder on eICU hospital splits (144 train, 21 validation, 41 test) with the encoder, sepsis head, and ridge prior fixed. Baselines are a target-refitted vital-only ridge and the architecture trained from scratch for 12 epochs.

At full budget, adaptation consistently improves over the frozen model by 0.0022 ± 0.0001 MAE and the vital-only ridge by 0.0044 ± 0.0002, improving 35 of 41 test hospitals on seed 42 (Fig. 1b). Recomputing normalization statistics without gradient updates degrades performance, while unfreezing the encoder at $0.1\times$ learning rate yields no gain. Mild source-domain forgetting occurs, shifting Challenge test MAE from 0.333 to 0.336.

Across training budgets, adapted and frozen models perform identically at 3 and 8 hospitals because validation selection retains the initialization. Training from scratch performs worse at small budgets, matches adaptation at 20 hospitals, and surpasses it at full scale.

4 Discussion

Across all settings, the residual ridge prior remains a strong baseline. It is marginally behind ICUWorld-QR in-domain (bootstrap CI excludes zero), marginally

ahead on B→A, and substantially ahead on A→B. The ridge is also more stable at our primary cohort size: it is unaffected by seed choice, whereas ICUWorld-QR's A→B error varies across seeds by an order of magnitude more than the margin separating them, a gap that closes on the full corpus. Zero-shot across 204 eICU hospitals, the two match. Margins of this size are typical in short-horizon ICU forecasting, where recent architectures yield sub-percent gains over strong baselines [8].

Crucially, aggregate error masks representation failures. Our ablations show that a collapsed latent reproduces ridge-level forecasting while sepsis classification degrades sharply. Evaluating ICU models solely on prediction loss, the standard practice in prior work [7,5,14], fails to detect state loss, just as source validation error fails to signal cross-site failure on A→B. Though the latent space supports a working sepsis readout, its auxiliary capabilities are limited: predictive intervals under-adapt to volatility by roughly half, while the latent-norm signal flags generally hard hours rather than model-specific failures.

This behavior reflects a structural trade-off. Because a strong linear prior absorbs most predictable next-hour variance, the encoder receives little gradient incentive to build a rich representation. Removing JEPA does not solve the issue, while removing the prior severely harms forecasting accuracy. An explicit variance constraint is thus required to prevent latent collapse, a problem likely to arise whenever neural networks are trained as residuals on top of strong linear priors.

Limitations. Our evaluation uses observational data, so we cannot test real interventions, and the FiO_2 probe measures model sensitivity rather than causal effect because clinicians raise oxygen when saturation drops. External transfer applies strictly to vital signs, as eICU lacks a matching sepsis endpoint, and masked evaluation covers only hours clinicians chose to measure. Single-seed transfer results are unreliable at our primary cohort size, sepsis transfer stays seed-dependent even on the full corpus, the latent norm gate requires target-side recalibration, and prediction intervals are miscalibrated across volatility strata. We omit bedside scores like NEWS2, which require consciousness assessments absent from this corpus and a supplemental oxygen indicator that FiO_2 charting, present in only 8% of hours, cannot reliably provide.

5 Conclusion

We present ICUWorld-QR, a variance-regularized residual latent world model for intensive care physiology, in which pairing a strong linear prior with a variance hinge preserves an informative latent space and its sepsis readout. The model beats last-value in-distribution and transfers zero-shot to 204 eICU hospitals while matching linear baselines on aggregate error. Cross-site transfer is seed-sensitive at small cohort sizes, and both the prediction intervals and the latent norm gate are weaker than their construction suggests.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: CVPR, pp. 15619–15629 (2023)
2. Bardes, A., Ponce, J., LeCun, Y.: VICReg: variance-invariance-covariance regularization for self-supervised learning. In: ICLR (2022)
3. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.Ch., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
4. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: NeurIPS, pp. 2455–2467 (2018)
5. He, R.Y., Chiang, J.N.: TFT-multi: simultaneous forecasting of vital sign trajectories in the ICU. Scientific Reports (2025). arXiv:2409.15586
6. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)
7. Lim, B., Arık, S.Ö., Loeff, N., Pfister, T.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting **37**(4), 1748–1764 (2021)
8. Liu, W., Xiong, J., Ni, C., Zhu, Y., Lin, X., Malin, B.A., Yin, Z.: DRIFT: direct–recursive intervention-conditioned forecasting of ICU physiological trajectories. arXiv:2607.25864 (2026)
9. Liu, K., Li, M., Bao, Y., Zhang, T., Chu, C., Bu, J., Wang, H.: Medical world models: representing medical states, modelling clinical dynamics and guiding intervention policies. arXiv:2606.16721 (2026)
10. Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data **5**, 180178 (2018)
11. Reyna, M.A., Josef, C.S., Jeter, R., Shashikumar, S.P., Westover, M.B., Nemati, S., Clifford, G.D., Sharma, A.: Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine **48**(2), 210–217 (2020)
12. Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.-D., Coopersmith, C.M., Hotchkiss, R.S., Levy, M.M., Marshall, J.C., Martin, G.S., Opal, S.M., Rubenfeld, G.D., van der Poll, T., Vincent, J.-L., Angus, D.C.: The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA **315**(8), 801–810 (2016)
13. Xu, Q., Habib, G., Perera, D., Feng, M.: MedDreamer: model-based reinforcement learning with latent imagination on complex EHRs for clinical decision support. arXiv:2505.19785 (2025)
14. Xu, W., Dai, Y., Yang, Y., Loza, M., Zhang, W., Cui, Y., Zeng, X., Park, S.J., Nakai, K.: OmniTFT: omni target forecasting for vital signs and laboratory result trajectories in multi-center ICU data. arXiv:2511.19485 (2025)
15. Yang, Y., Wang, Z.-Y., Liu, Q., Sun, S., Wang, K., Chellappa, R., Zhou, Z., Yuille, A., Zhu, L., Zhang, Y.-D., Chen, J.: Medical world model: generative simulation of tumor evolution for treatment planning. In: ICCV, pp. 8319–8329 (2025)