CardioWorld-GQ: An Observation-Conditioned Cardiac Imaging World Model with a Differentiable Clinical Readout and Calibrated Endpoint Uncertainty

Jing Yuan[1][0000−0002−3548−5875], Shengming Wang[1][0009−0000−8704−6535], and Ningtao Liu[2][0000−0002−5590−3229]

[1] Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
{jyuan,wangshengming}@zjnu.edu.cn
[2] Luoyang Institute of Science and Technology, Luoyang, Henan, 471023, China
nt_liu@lit.edu.cn

Abstract. Cardiac magnetic resonance analysis requires consistent reasoning across anatomy, cine dynamics, scar, and clinical endpoints. We present *CardioWorld-GQ*, an observation-conditioned cardiac imaging world model whose differentiable readout computes volumes, mass, scar burden, and all-frame LVEF from the segmentation probability field. On the official validation split, the full model attains mean Dice 0.780, mean IoU 0.661, ASSD 2.49 mm, soft LVEF MAE 0.47%, and 91.2% marginal coverage at a nominal 90% target. The differentiable-readout ablation reduces soft LVEF MAE from 0.70% to 0.44%; other modules have metric-specific effects. The framework targets L1-L2 imaging functionality without claiming causal intervention or control.

Keywords: Cardiac MRI · Imaging world models · Joint segmentation-quantification · Differentiable clinical readout · Evidential deep learning · Conformal prediction · Trustworthy uncertainty.

1 Introduction

CMR couples cine imaging of cardiac function with LGE imaging of myocardial scar. CMR-MULTI provides both sequences across short-axis (SAX), two-chamber (2CH), four-chamber (4CH), and an additional right-atrium LGE (RAS) view, with annotations spanning cardiac chambers, myocardium, and scar.

Conventional pipelines segment each view independently and compute endpoints with separate, non-differentiable scripts. Consequently, masks and reported measurements are not optimized together, views do not share a cardiac representation, LVEF depends on two selected frames, and endpoint uncertainty is uncalibrated.

World models organize prediction around latent state, evolution, and observation [4,5]; medical imaging applications include radiograph representation and ultrasound guidance [14,13,7]. CardioWorld-GQ treats sequence, view, and

phase as observation queries over a shared cardiac state. Its probability field supports both segmentation and differentiable quantification, its phase-indexed states produce the LVEF volume curve, and evidential plus conformal components quantify uncertainty.

Contributions.

1. **Observation-conditioned cardiac state.** Multi-view, multi-sequence CMR shares a latent state with sequence-, view-, and phase-conditioned emission.
2. **Differentiable clinical readout.** Exact functionals of the probability field reproduce official voxel-count endpoints in the hard limit and compute LVEF from the complete volume curve.
3. **Calibrated uncertainty and controlled evaluation.** Evidential scores and split-conformal intervals accompany endpoint predictions; component-wise ablations retain the official ResUNet++ interfaces and distinguish marginal guarantees from vendor-shift stress tests.

2 Related Work

Well-configured U-Nets remain strong medical segmentation baselines [6]. Unlike index regression or mask post-processing, our probability-field functionals preserve mask–measurement consistency; calibrated summation is unbiased, whereas near-binary soft-Dice solutions can bias volume [3]. Evidential segmentation supplies single-pass dense scores [9,16], but does not cleanly separate uncertainty types [2]. Split conformal prediction supplies marginal endpoint coverage under exchangeability [1]; vendor shift is therefore evaluated as an empirical stress test [10].

3 Method

3.1 Cardiac analysis as an observation-conditioned world model

For case i, $\mathcal{O}_i = \{(I_{s,v}, A_{s,v}, m_{s,v})\}$ contains images indexed by sequence s and view v, optional pose $A_{s,v} \in SE(3)$, and endpoint/vendor metadata.

Definition 1 (Cardiac world model). *A cardiac world model is defined as a tuple* $(\mathcal{S}, \Phi_\theta, \mathcal{R}, F, \mathcal{Q})$*,where* $\mathcal{S}$ *is a latent state space; the encoder* $E : \mathcal{O} \to \mathcal{S}$ *infers a state* $z \in \mathcal{S}$*, materialized as a probabilistic segmentation field* $\Phi_\theta(\cdot\,; z, a) \in \Delta^K$ *over* K *structures on the probability simplex* Δ^K*; the emission operator* $\mathcal{R}_a : \mathcal{S} \to \Delta^K$ *renders the state to the plane selected by an action* $a = (s, v, t)$ *(sequence, view, cardiac phase* $t \in [0, 1]$*); the dynamics* F *advances the state through the cardiac cycle,* $\dot{z}(t) = F(z(t))$*; and the readout* $\mathcal{Q} : \mathcal{S} \to \mathbb{R}^d$ *maps the state to* d *clinical endpoints. An* action *is an imaging-observation request.*

The official multi-head ResUNet++ instantiates this tuple (Fig. 1). Its four encoder stages use $(64, 128, 256, 512)$ channels, and the bottleneck $z \in \mathbb{R}^{B\times 512\times H_4\times W_4}$ is the latent state. Per-view decoders implement $\mathcal{R}_a$ for action $a = (s, v, t)$ and emit $\Phi_\theta(\cdot; z, a)$, which is passed directly to $\mathcal{Q}$. Views share encoder parameters; cross-view consistency is used only when case-level correspondence exists.

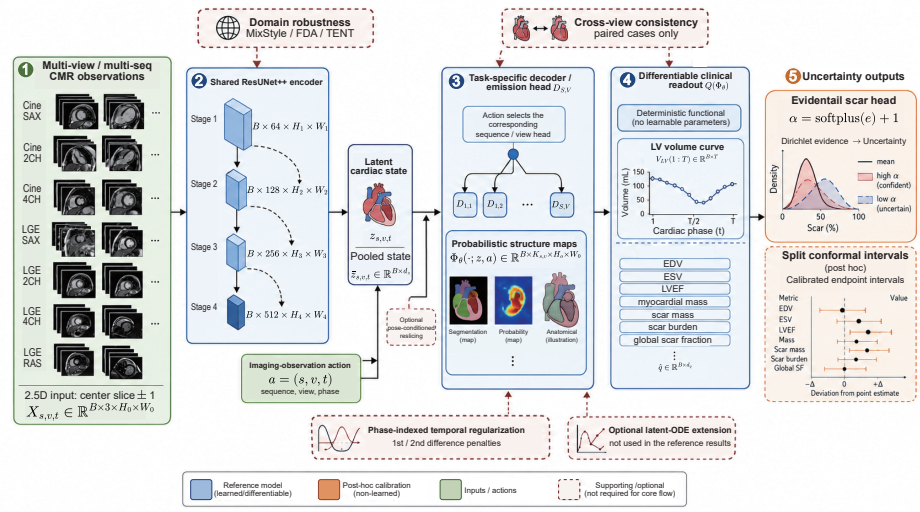

Fig. 1: **CardioWorld-GQ architecture.** A shared four-stage ResUNet++ encoder maps each 2.5D observation to a 512-channel latent state. Sequence-, view-, and phase-conditioned heads emit probability fields; the differentiable readout produces the LV volume trajectory and clinical endpoints. The scar head predicts Dirichlet evidence, and conformal calibration is fitted post hoc. Dashed blocks are supporting or optional modules.

3.2 Differentiable clinical readout

Let $p_c(\mathbf{x})$ be the softmax probability of structure c at voxel $\mathbf{x}$ for a given view, and let $\Delta_V = s_x s_y s_z$ be the voxel volume from the NIfTI header (in mm^3). We define the soft volume, mass, scar burden, and a global scar fraction as

$$V_c = \frac{\Delta_V}{1000} \sum_{\mathbf{x}} p_c(\mathbf{x}) \quad [\mathrm{mL}], \qquad M_c = \rho\, V_c \quad [\mathrm{g}], \quad \rho = 1.05 \ \mathrm{g/mL}, \tag{1}$$

$$B_{\mathrm{scar}} = \frac{M_{\mathrm{scar}}}{M_{\mathrm{scar}} + M_{\mathrm{myo}} + \varepsilon}, \qquad G_{\mathrm{scar}} = \frac{\sum_{\mathbf{x}} p_{\mathrm{scar}}(\mathbf{x})}{\sum_{\mathbf{x}} \left(p_{\mathrm{scar}}(\mathbf{x}) + p_{\mathrm{myo}}(\mathbf{x})\right) + \varepsilon}. \tag{2}$$

In the hard limit $p_c(\mathbf{x}) \in \{0, 1\}$, Eq. (1) exactly recovers official voxel-count definitions. Masks and endpoints are therefore internally consistent, although correctness still depends on calibration, spacing, and labels. We call G_{scar} a global scar fraction, not transmurality, because the latter requires wall-normal or sector-wise geometry.

Proposition 1 (Unbiasedness and bias of the soft volume). *Let the per-voxel label be Bernoulli with true posterior probability* $q_c(\mathbf{x}) = \mathbb{P}(\mathit{label}(\mathbf{x})=c \mid I)$*, and let the network output* $p_c(\mathbf{x})$ *estimate* $q_c(\mathbf{x})$*. The soft volume* $\widehat{V}_c = \frac{\Delta_V}{1000} \sum_{\mathbf{x}} p_c(\mathbf{x})$ *satisfies* $\mathbb{E}[\widehat{V}_c] = \frac{\Delta_V}{1000} \sum_{\mathbf{x}} \mathbb{E}[p_c(\mathbf{x})]$*, so if* p_c *is calibrated,* $\mathbb{E}[p_c(\mathbf{x})] =$

$q_c(\mathbf{x})$, then $\widehat{V}_c$ is an unbiased estimator of the expected true volume $\frac{\Delta_V}{1000}\sum_{\mathbf{x}} q_c(\mathbf{x})$. If instead the training objective drives p_c toward a thresholded map $\mathbf{1}[q_c(\mathbf{x}) > \frac{1}{2}]$, the volume bias is $\frac{\Delta_V}{1000}\sum_{\mathbf{x}}\big(\mathbf{1}[q_c(\mathbf{x}) > \frac{1}{2}] - q_c(\mathbf{x})\big)$, whose magnitude grows with the aleatoric uncertainty concentrated near $q_c(\mathbf{x}) \approx \frac{1}{2}$.

Proof. Apply linearity of expectation; substituting the thresholded map and subtracting the expected true volume gives the stated bias.

Accordingly, cross-entropy is retained for volume-bearing classes and post-hoc temperature scaling is fitted on validation data; neither guarantees calibration under finite samples or shift.

Ejection fraction from the whole curve. Let $V_{\mathrm{LV}}(t)$ be the LV cavity volume at cardiac phase t, computed by Eq. (1) per phase. Rather than picking two frames, we define end-diastolic and end-systolic volumes through temperature-weighted soft extrema over *all* phases,

$$\mathrm{EDV}_\tau = \sum_t w_t^+ V_{\mathrm{LV}}(t), \quad \mathrm{ESV}_\tau = \sum_t w_t^- V_{\mathrm{LV}}(t), \quad w_t^\pm = \mathrm{softmax}\big(\pm V_{\mathrm{LV}}(t)/\tau\big), \tag{3}$$

and set $\mathrm{LVEF}_\tau = (\mathrm{EDV}_\tau - \mathrm{ESV}_\tau)/(\mathrm{EDV}_\tau + \varepsilon)\cdot 100$. As $\tau \to 0^+$, Eq. (3) recovers the official hard extrema; for $\tau > 0$ it is differentiable and uses all phases. We anneal $\tau(t) = \max(\tau_{\min}, \tau_0 \gamma^t)$ from smooth early gradients toward the hard definition and report hard and soft EF separately.

Agreement loss and its gradient. For a scalar endpoint q with prediction $\hat{q}$ and reference q over a batch, we minimize $1 - \mathrm{CCC}$, where Lin's concordance correlation [8] is

$$\mathrm{CCC}(\hat{q}, q) = \frac{2\, s_{\hat{q}q}}{s_{\hat{q}}^2 + s_q^2 + (\bar{\hat{q}} - \bar{q})^2}, \mathcal{L}_{\mathrm{quant}} = \sum_q \big(1 - \mathrm{CCC}(\hat{q}, q)\big) + \lambda_{\mathrm{mse}} \sum_q \|\hat{q} - q\|^2, \tag{4}$$

with sample covariance $s_{\hat{q}q}$, variances $s_\cdot^2$, and means $\bar{\cdot}$. CCC penalizes mean and scale disagreement; the MSE term stabilizes constant batches. The loss is applied to LVEF, scar mass, and scar burden when references are available.

3.3 Supporting consistency, dynamics, and uncertainty

Views share an encoder and are aligned by a latent consistency loss; when poses are available, differentiable grid sampling can additionally reslice a shared field. A latent transition predicts cardiac phase, while first- and second-difference penalties regularize the resulting volume curve. These terms encourage, but do not guarantee, geometric or physiological consistency.

Uncertainty. The scar head predicts Dirichlet evidence $\alpha_c = \mathrm{softplus}(e_c) + 1$, with total-evidence uncertainty $u = K/\sum_c \alpha_c$. For endpoint q, split conformal calibration uses validation residuals $r_i = |q_i - \hat{q}_i|$ and returns $[\hat{q} - q_{1-\alpha}, \hat{q} + q_{1-\alpha}]$, where $q_{1-\alpha}$ is the finite-sample residual quantile. Coverage is marginal under exchangeability; vendor-held-out coverage is therefore reported only empirically.

3.4 Supporting robustness modules and the full objective

Supporting robustness modules comprise shallow-stage MixStyle [15], Fourier low-frequency amplitude transfer [12], and optional one-step TENT adaptation [11]; TENT is disabled for the scar head by default. The complete objective is

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}}_{\text{segmentation}} + \lambda_q \mathcal{L}_{\text{quant}} + \lambda_e \mathcal{L}_{\text{evi}} + \lambda_t \mathcal{L}_{\text{temp}} + \lambda_c \mathcal{L}_{\text{cons}} \; [+ \lambda_J \, \|\partial_z f_\phi\|_2^2], \quad (5)$$

where the bracketed Jacobian-spectral term is used only in the optional latent-ODE variant and is absent from the reference model. We select fixed loss weights on a held-out split; the readout, evidential, conformal, and consistency components add negligible parameters.

4 Experimental Protocol and Pipeline Verification

4.1 Dataset and faithful baseline

CMR-MULTI provides multi-center cine and LGE NIfTI data across SAX, 2CH, 4CH, and an LGE RAS subset. Available case pairings are read from the dataset spreadsheet rather than assumed. Reference LVEF comes from SAX cine metadata; scar mass follows the official SAX-LGE voxel-count rule. All reported values are estimates on the official validation split, not an independent challenge test set.

As mapped explicitly to the world-model operators in Sec. 3 and Fig. 1, we extend the official multi-head ResUNet++ baseline (shared encoder, per-view heads, and 2.5D center-slice ± 1 pseudo-RGB inputs) with the readout, evidential and conformal uncertainty, latent consistency, volume-curve regularization, and domain-generalization modules without changing the backbone interfaces. Disabling them recovers the traditional baselines.

4.2 Implementation

We use Adam and round-robin sampling over available modalities/views. Loss weights are selected on a held-out split; evidential KL and the soft-extrema temperature are annealed. Split-conformal calibration uses a validation subset disjoint from training. Readout and uncertainty layers add negligible parameters.

4.3 Tasks, metrics, and ablations

Our T1–T5 analysis groups are not official challenge subtasks. T1 reports Dice, IoU, HD95, and ASSD; T2 hard/soft LVEF and volume MAE, CCC, and scar errors; T3 cross-view endpoint agreement; T4 curve morphology and 5%–15% temporal-index perturbations; and T5 ECE, marginal/LOVO coverage, width, and failure-detection AUROC. LOVO is an empirical vendor-shift stress test.

Each "+" row independently adds the named component to *Ours (Baseline)*; the Full Model combines them. Conformal calibration is post hoc and does not change point predictions. Components are judged on their target metrics rather than expected to improve every column.

The differentiable readout reduces soft LVEF MAE from 0.70% to 0.44% and raises CCC from 0.932 to 0.978. Evidential learning primarily affects calibration/soft volumes; dynamics and consistency yield the lowest soft LVEF MAE (0.42%). Split conformal prediction provides 91.2% marginal coverage at nominal 90% with 26.26 mL width without changing point predictions.

(T1) Segmentation. Table 1 shows joint-highest mean Dice (0.780), highest mean IoU (0.661), and lowest ASSD (2.49 mm) for the full model. Relative to the 3D baseline, these correspond to changes of +0.014 Dice, +0.023 IoU, and −0.24 mm ASSD. The 2D baseline nevertheless retains the lowest HD95 (8.87 vs. 9.56 mm for the full model). Thus the integrated objectives improve average overlap and surface agreement, but do not eliminate the localized boundary deviations emphasized by HD95.

Table 1: Comprehensive segmentation performance evaluation (Task 1). We report spatial overlap (Mean Dice, Mean IoU) and boundary adherence accuracy (Mean HD95, Mean ASSD) across all CMR tasks. Best performance per column is highlighted in bold.

Method	Mean Dice ↑	Mean IoU ↑	Mean HD95 (mm) ↓	Mean ASSD (mm) ↓
2D Baseline	0.77	0.65	**8.87**	2.56
3D Baseline	0.77	0.64	9.69	2.73
Ours (Baseline)	**0.78**	0.66	9.39	2.69
+ diff. readout	0.77	0.65	9.55	2.64
+ evidential	**0.78**	0.65	9.33	2.57
+ dynamics & cons.	**0.78**	0.65	8.97	2.72
+ domain gen.	**0.78**	**0.66**	9.94	2.97
Ours (Full Model)	**0.78**	**0.66**	9.56	**2.49**

(T2) Clinical quantification. For the full model, Table 2 reports soft LVEF MAE 0.47%, CCC 0.963, soft EDV/ESV MAEs 0.50/0.49 mL, scar MAE 10.99 g, and scar RVE 47.4%. Relative to Ours (Baseline), the differentiable-readout variant increases CCC from 0.932 to 0.978 and reduces soft LVEF MAE from 0.70% to 0.44%. The evidential variant has the lowest soft EDV/ESV MAEs (0.45/0.45 mL), the domain-generalization variant the lowest scar MAE (10.88 g), and the full model the lowest hard EDV/ESV MAEs (15.46/15.91 mL) and scar RVE. Effects are therefore metric-dependent.

Supporting analyses. For the full model, soft versus hard temporal-index shift MAEs were 0.33/0.33, 0.67/0.65, and 1.06/1.01 mL at 5%, 10%, and 15%, showing no robustness improvement. At nominal 90%, it achieved 91.2% marginal and 87.9% LOVO coverage, 26.26 mL width, and AUROC 0.868; other config-

Table 2: Clinical quantification (Task 2). Hard denotes discrete frame/mask extraction and Soft the differentiable full-curve readout. Across configurations, soft EDV/ESV MAEs are approximately 0.45-0.62 mL; the highest LVEF CCC is 0.978. Best results are highlighted in bold.

Method	LVEF MAE (%)		CCC	EDV MAE (mL)		ESV MAE (mL)		Scar MAE	RVE
	Hard	**Soft**	↑	Hard	**Soft**	Hard	**Soft**	(g)	(%)
2D Baseline	1.15	0.55	0.94	16.87	0.59	17.28	0.56	12.19	52.2
3D Baseline	2.33	0.96	0.71	16.97	0.57	17.27	0.51	16.83	74.5
Ours (Baseline)	1.51	0.70	0.93	16.31	0.50	16.74	0.49	11.35	60.1
+ diff. readout	1.07	0.44	**0.98**	16.47	0.62	16.87	0.60	14.59	74.5
+ evidential	1.46	0.69	0.89	18.63	**0.45**	19.08	**0.45**	12.42	118.5
+ dynamics & cons.	**0.73**	**0.42**	0.97	17.51	0.56	17.96	0.56	12.25	115.9
+ domain gen.	1.07	0.48	0.97	18.54	0.62	19.03	0.59	**10.88**	72.2
Ours (Full Model)	0.99	0.47	0.96	**15.46**	0.50	**15.91**	0.49	10.99	**47.4**

Table 3: Trustworthiness results. LOVO coverage is empirical under vendor shift and does not inherit the split-conformal guarantee.

Method	ECE (%) ↓	Marginal Cov	LOVO Cov ↑	Width (mL) ↓	Failure AUC ↑
2D Baseline	**10.94**	91.2%	87.9%	28.14	**0.909**
3D Baseline	15.35	91.5%	84.8%	32.13	0.809
Ours (Baseline)	14.99	94.2%	84.8%	31.70	0.883
+ diff. readout	15.34	91.1%	**90.9%**	**25.18**	0.833
+ evidential	14.71	90.3%	84.8%	26.95	0.828
+ dynamics & cons.	14.88	89.5%	84.8%	26.36	0.904
+ domain gen.	14.23	92.0%	**90.9%**	27.76	0.859
Ours (Full Model)	14.23	**91.2%**	87.9%	26.26	0.868

urations reached LOVO 90.9%, width 25.18 mL, and AUROC 0.909. Thus the full model is balanced rather than column-best throughout.

Interpreting hard and soft quantification. Hard-frame values apply the conventional discrete pipeline to selected ED/ES masks, whereas soft values evaluate differentiable functionals of the full probability curve. For the full model, LVEF MAE changes from 0.99% (hard) to 0.47% (soft), EDV MAE from 15.46 to 0.50 mL, and ESV MAE from 15.91 to 0.49 mL. The same direction holds for both external baselines, so the comparison supports the readout strategy rather than proving that every added module improves segmentation. Conversely, the best soft LVEF MAE is 0.42% for dynamics and consistency, and the best CCC is 0.978 for the readout ablation; the full model is close but not column best. Low endpoint error can therefore coexist with a non-optimal boundary metric.

5 Discussion

Retaining ResUNet++ isolates the readout, regularization, and uncertainty layers from backbone capacity. The readout is the clearest contributor to functional

Table 4: Temporal curve morphology and robustness to temporal-index perturbations.

Method	MASD ↓	Trough ↑	5% shift		10% shift		15% shift	
			Hard	Soft	Hard	Soft	Hard	Soft
2D Baseline	**1.19**	26.7%	0.30	0.31	0.68	0.72	0.98	1.03
3D Baseline	1.32	26.7%	**0.29**	**0.30**	**0.62**	**0.64**	1.01	1.05
Ours (Baseline)	1.32	**53.3%**	0.34	0.35	0.66	0.70	1.07	1.12
+ diff. readout	1.23	26.7%	0.31	0.32	**0.62**	**0.64**	**0.96**	**1.01**
+ evidential	1.23	26.7%	0.33	0.34	0.65	0.67	1.00	1.04
+ dynamics & cons.	1.28	26.7%	0.34	0.35	0.63	0.67	1.02	1.07
+ domain gen.	1.33	46.7%	0.33	0.35	0.70	0.73	0.98	1.03
Ours (Full Model)	1.30	33.3%	0.33	0.33	0.65	0.67	1.01	1.06

quantification; other modules affect different metric families. The full model has joint-highest mean Dice (0.780), highest mean IoU (0.661), lowest ASSD (2.49 mm), and lowest scar RVE (47.4%), while other rows remain preferable for HD95, LVEF MAE, LOVO coverage, width, or failure detection.

These metric families answer different questions: Dice/IoU and surface distances assess anatomy, endpoint MAE/CCC assess its clinical consequences, and coverage, width, and AUROC assess uncertainty. Their optima need not coincide: the full model's 91.2% marginal coverage is close to the 90% target, but its 87.9% LOVO coverage and 0.868 AUROC are not column best. Reporting these trade-offs prevents a favorable aggregate result from obscuring a clinically relevant failure mode; pooled calibration may still conceal domain dependence.

Limitations are specific. Latent consistency without reliable poses does not establish view-invariant 3D geometry. Soft volumes inherit calibration, spacing, and annotation errors; global scar fraction is not transmurality. Conformal coverage is marginal under exchangeability, as reflected by 91.2% pooled versus 87.9% LOVO coverage. Evaluation is limited to one challenge dataset, requiring multicenter and arrhythmia-stratified validation before clinical use. CardioWorld-GQ therefore represents L1–L2 observation-conditioned imaging, not causal intervention, treatment response, or control.

6 Conclusion

CardioWorld-GQ formulates multi-view, multi-sequence CMR analysis as an observation-conditioned imaging world model with differentiable clinical endpoints. On the official validation split, the full model attained mean Dice 0.78, mean IoU 0.66, ASSD 2.49 mm, soft LVEF MAE 0.47%, and 91.2% marginal coverage at the nominal 90% target. These values represent a balanced configuration rather than the best result for every metric. The readout recovers the official voxel-count definitions in the hard-mask limit, ensuring internal consistency between masks and endpoints. The submitted system supports L1-L2 observation-conditioned imaging functionality with a differentiable clinical readout, without causal counterfactual simulation, treatment modelling, or control.

References

1. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023). https://doi.org/10.1561/2200000101
2. Bengs, V., Hüllermeier, E., Waegeman, W.: On second-order scoring rules for epistemic uncertainty quantification. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 2078–2091. PMLR (2023)
3. Bertels, J., Robben, D., Vandermeulen, D., Suetens, P.: Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty. Medical Image Analysis **67**, 101833 (2021). https://doi.org/10.1016/j.media.2020.101833
4. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Advances in Neural Information Processing Systems. vol. 31, pp. 2450–2462 (2018)
5. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations (2020)
6. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jäger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15009, pp. 488–498. Springer Nature Switzerland (2024)
7. Jiang, H., Sun, Z., Jia, N., Li, M., Sun, Y., Luo, S., Song, S., Huang, G.: Cardiac copilot: Automatic probe guidance for echocardiography with world model. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15001, pp. 190–199. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72378-0_18
8. Lin, L.I.K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**(1), 255–268 (1989). https://doi.org/10.2307/2532051
9. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems. vol. 31, pp. 3179–3189 (2018)
10. Tibshirani, R.J., Foygel Barber, R., Candès, E.J., Ramdas, A.: Conformal prediction under covariate shift. In: Advances in Neural Information Processing Systems. vol. 32, pp. 3690–3700 (2019)
11. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations. pp. 104–105 (2021)
12. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4085–4095 (2020)
13. Yue, Y., Wang, Y., Jiang, H., Liu, P., Song, S., Huang, G.: EchoWorld: Learning motion-aware world models for echocardiography probe guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25993–26003 (2025). https://doi.org/10.1109/CVPR52734.2025.02421
14. Yue, Y., Wang, Y., Tao, C., Liu, P., Song, S., Huang, G.: CheXWorld: Exploring image world modeling for radiograph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20778–20788 (2025). https://doi.org/10.1109/CVPR52734.2025.01935

15. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with MixStyle. In: International Conference on Learning Representations. pp. 102–103 (2021)
16. Zou, K., Chen, Y., Huang, L., Zhou, N., Yuan, X., Shen, X., Wang, M., Goh, R.S.M., Liu, Y., Tham, Y.C., Fu, H.: Toward reliable medical image segmentation by modeling evidential calibrated uncertainty. IEEE Transactions on Cybernetics **55**(12), 5975–5988 (2025). https://doi.org/10.1109/TCYB.2025.3604432