



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 this watermark; the final published version is available on SpringerLink.

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

Jing Yuan¹[0000–0002–3548–5875], Shengming Wang¹[0009–0000–8704–6535], and Ningtao Liu²[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

A Additional Experimental Results

This appendix collects the result figures and complete supporting tables included in the full version but omitted from, or replaced by compact summaries in, the short version. The main-text conclusions are unchanged; these materials provide qualitative examples, agreement plots, and the complete T3–T5 ablation results.

A.1 Qualitative segmentation comparison

Figure 1 provides representative qualitative comparisons across the seven CMR tasks.

A.2 Bland–Altman analysis of the clinical readout

Figure 2 compares the hard-frame and soft full-curve readouts for EDV, ESV, and EF.

A.3 Complete supporting tables

Tables 1–3 give the full T3–T5 results summarized in the main text. They are grouped on one full-width appendix page to avoid one-table float pages.

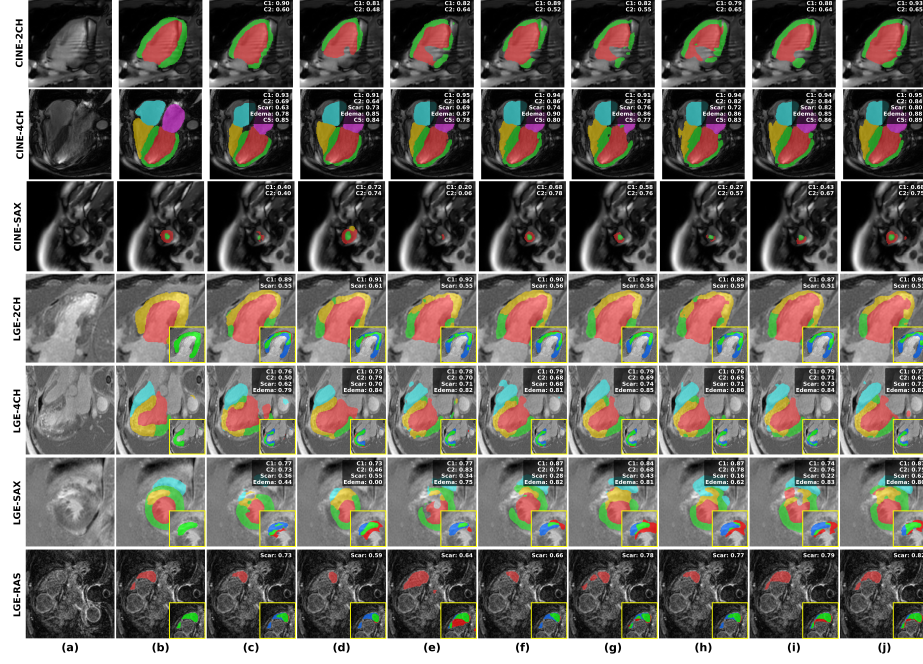


Fig. 1: Qualitative comparison of different segmentation methods across seven cardiac magnetic resonance (CMR) imaging tasks. From top to bottom, the rows correspond to CINE-2CH, CINE-4CH, CINE-SAX, LGE-2CH, LGE-4CH, LGE-SAX, and LGE-RAS views, respectively. Panels from left to right represent: (a) Raw Image, (b) Ground Truth, (c) 2D baseline, (d) 3D baseline, (e) Baseline, (f) +Readout, (g) +Evidential, (h) +Dyn&Cons, (i) +MixStyle, and (j) Ours (Full). The localized Dice scores for each anatomical structure are annotated at the top-right corner of the respective panels. For LGE tasks containing focal scar tissues, the bottom-right insets provide seamless zoomed-in error maps (green: true positives, red: false positives, blue: false negatives).

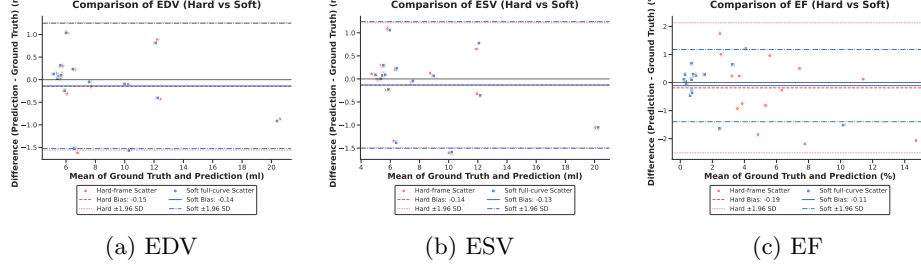


Fig. 2: Bland-Altman analysis comparing the traditional discrete hard-frame selection (red circles) and the proposed temperature-weighted soft full-curve continuous integration (blue squares) on the CardioWorld-GQ (full) model. Solid and dashed lines denote the mean biases, while dash-dotted and dotted lines represent the 95% limits of agreement (± 1.96 SD) for the soft and hard methods, respectively. Compared to the hard-frame baseline, our soft full-curve approach demonstrates tighter agreement limits and reduced bias against the ground truth. EDV: End-Diastolic Volume; ESV: End-Systolic Volume; EF: Ejection Fraction.

Table 1: Complete Task 3 cross-view consistency results (MAE; lower is better).

Method	SAX vs. 2CH (MAE ↓)						SAX vs. 4CH (MAE ↓)					
	EDV (mL)		ESV (mL)		EF (%)		EDV (mL)		ESV (mL)		EF (%)	
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft
2D Baseline	16.87	17.00	17.28	17.12	5.31	1.42	9.77	9.88	10.09	9.97	5.31	1.42
3D Baseline	16.97	17.07	17.27	17.14	3.88	0.82	11.08	11.17	11.33	11.23	3.88	0.82
Ours (Baseline)	16.31	16.46	16.74	16.58	5.19	1.39	11.12	11.24	11.48	11.35	5.19	1.39
+ diff. readout	16.47	16.58	16.87	16.70	5.46	1.60	10.57	10.67	10.91	10.78	5.46	1.60
+ evidential	18.63	18.76	19.08	18.89	5.59	1.49	11.33	11.44	11.68	11.54	5.59	1.49
+ dynamics	17.51	17.66	17.96	17.79	5.65	1.51	9.30	9.43	9.66	9.53	5.65	1.51
+ domain gen.	18.54	18.71	19.03	18.85	6.11	1.68	11.08	11.22	11.49	11.35	6.11	1.68
Ours (Full Model)	15.46	15.60	15.91	15.73	5.69	1.53	10.78	10.89	11.18	11.01	5.69	1.53

Table 2: Complete Task 4 temporal dynamics and index-perturbation results. Lower MASD/shift MAE and higher single-trough rate are better.

Method	MASD Single-Trough		$\sigma = 5\%$ Shift		$\sigma = 10\%$ Shift		$\sigma = 15\%$ Shift	
	($\downarrow$)	($\uparrow$)	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	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

Table 3: Complete Task 5 calibration and failure-detection results. The coverage target is 90%; LOVO coverage is empirical under vendor shift.

Method	ECE (%)	Conformal Target $\approx 90\%$		Width (mL)	Failure AUC
	($\downarrow$)	Marginal Cov	LOVO Cov $\uparrow$	($\downarrow$)	($\uparrow$)
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