An Auditable Geometry Contract for Multi-View Cardiac Magnetic Resonance Imaging Analysis

Ekaterina Antipushina*[3,4,5], Daniil Sukhorukov*[1,2], Elizaveta Cherkasskaya[3,4], Ruslan Kalimullin[3,4], Ilya Makarov[1,6], and Yury Koush[3,4]

[1] AIRI, Moscow, Russia
[2] ITMO University, Saint Petersburg, Russia
[3] Center for Bio- and Medical Technologies, Moscow, Russia
[4] Skolkovo Institute of Science and Technology, Moscow, Russia
[5] Applied AI Institute, Moscow, Russia
[6] Trusted Artificial Intelligence Center, Russian Academy of Sciences, Moscow, Russia

Abstract. Automated cardiac magnetic resonance (CMR) segmentation assumes that image metadata preserve the anatomical meaning of array axes. In the CMR-MULTI challenge, cine volumes concatenate spatial slices and cardiac phases along one axis, but their reported spacing causes nnU-Net to sample this mixed frame axis in-plane. We identify this axis-semantic failure and introduce an auditable geometry contract that changes only metadata, exposing coherent anatomical frames without resampling voxel values or altering output geometry. The failure affects all 450 released cine cases but none of the 299 late gadolinium enhancement (LGE) cases. In a controlled experiment using fold 0 (the first of five cross-validation splits), the correction increases the Task 1 Dice score by 0.029–0.084 across cine views and improves all ten structures; paired 95% bootstrap confidence intervals exclude zero, and exact sign tests yield $p \leq 9.8{\times}10^{-4}$. As a negative control, all 400 LGE training tensors remain unchanged, and Dice on public validation differs by at most 1.1×10^{-5}. The final view-specific five-fold ensemble achieves an overall public-validation score of 0.740 (cine/LGE Dice: 0.906/0.764; left ventricular ejection fraction correlation: 0.920). These results demonstrate that explicit axis semantics can eliminate a systematic preprocessing error without changing voxel data or network architecture.

Keywords: Cardiac magnetic resonance imaging · Late gadolinium enhancement · Multi-view segmentation · nnU-Net · Ejection fraction · Scar quantification

1 Introduction

Cardiac magnetic resonance (CMR) imaging provides quantitative assessment of cardiac anatomy and function. Automated delineation of chambers, myocardium

* E. Antipushina and D. Sukhorukov contributed equally.

and scar can reproduce clinically relevant measurements at scale [3,6]; on curated cohorts, modern segmentation systems approach expert performance [4]. Their robustness, however, remains sensitive to acquisition and data provenance. The Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation challenge showed that a model successful on one vendor or protocol need not generalise to another [5]. The gap is larger for late gadolinium enhancement (LGE), where small, heterogeneous lesions remain much harder to delineate than cardiac chambers [9]. Multi-sequence benchmarks exploit complementary CMR contrasts for myocardial pathology segmentation [16,10], while newer LGE resources extend the task to atrial anatomy [15,2,1]. Joint evaluation across sequences, views, structures and derived biomarkers nevertheless remains uncommon.

The CMR-MULTI challenge [12] brings these settings together: cine and LGE images span short-axis (SAX), two-chamber (2CH) and four-chamber (4CH) views, with an additional right-atrium-segmentation (RAS) LGE view. The targets include ventricles, myocardium, atria and scar, and the evaluation links segmentation to left-ventricular ejection fraction (LVEF) and scar mass. For a medical world model, these segmentations form part of the observable cardiac state. Their validity therefore depends not only on the network, but also on whether an input tensor retains the anatomical meaning implied by its metadata.

This dependence is easy to overlook. Before training, nnU-Net runs an automated planning stage that analyses image sizes and voxel spacings to choose the 2D/3D configuration, slice axis, target spacing and patch size [7]; we refer to this stage as the *planner*. Conversion between Digital Imaging and Communications in Medicine (DICOM) and Neuroimaging Informatics Technology Initiative (NIfTI) formats is known to be sensitive to vendor- and sequence-specific metadata [11]. In CMR-MULTI, the third cine axis does not describe a coherent spatial direction: it concatenates slices and cardiac phases. Its spacing nevertheless makes that axis appear in-plane to the planner, so the resulting two-dimensional (2D) samples mix anatomy and time (Fig. 1). This is a representation error upstream of learning, not a conventional appearance-domain shift.

We test the hypothesis that this mismatch is a measurable source of segmentation error. Our geometry contract preserves the voxel array and the submitted physical space while presenting the planner with the intended 2D sampling plane. A controlled cine intervention isolates its effect under a fixed model and training budget; an LGE falsification control tests whether the same mechanism remains inert when the stored geometry is already valid. The contract changes the selected plane in all released cine cases and improves every cine view and structure, whereas LGE tensors remain exactly invariant. These controls, together with end-to-end biomarker validation, make input geometry an explicit experimental variable rather than an undocumented preprocessing choice.

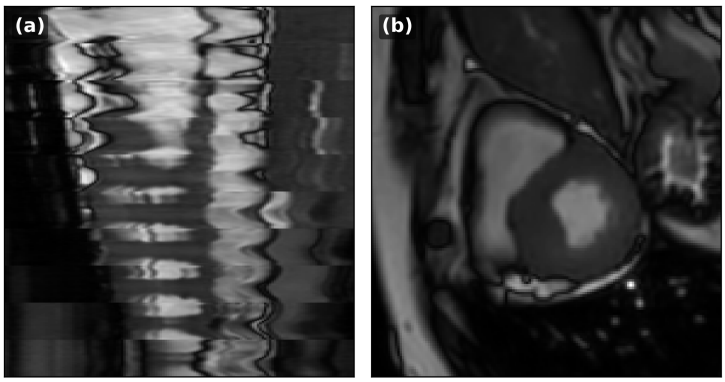

Fig. 1. Axis-semantic failure in pseudo-volumetric cine CMR. The third array axis concatenates $D = n_s n_t$ frames (spatial slices $\times$ cardiac phases). **(a)** Its reported spacing leads the planner to treat the frame axis as in-plane, producing samples that mix anatomy and time. **(b)** The geometry contract changes only the planning metadata, exposes coherent SAX frames for learning, and restores predictions to the original physical geometry.

2 Methodology

2.1 Problem formulation and dataset

The dataset contains paired image and annotation volumes for each view. Cine provides 105/15/30 training/validation/test cases for each of SAX, 2CH and 4CH; LGE provides 40/7/12–13 cases for these views and 80/14/26 for RAS. Because labels are defined per view, the same stored value may identify different structures and some label sets are non-contiguous (Table 1).

Each cine array has shape (H, W, D), but D indexes frames rather than a physical depth. In SAX, the metadata give $D = n_s \cdot n_t$; in the long-axis views, the same axis is likewise a frame stack rather than coherent three-dimensional anatomy. Its stored spacing therefore cannot be interpreted as inter-slice distance. The planner selects a different axis in all 450 cine cases (150 per view),

Table 1. Per-view label conventions used by the challenge. LV_c: LV cavity; Myo: myocardium; RV_c: RV cavity; RA/LA: right/left atrium.

Task / View	1	2	3	4	5
Cine SAX	Myo	LV_c	RV_c	–	–
Cine 2CH	LV_c	Myo	–	–	–
Cine 4CH	LV_c	Myo	RV_c	RA	LA
LGE SAX	LV_c	Myo	Scar	RV_c	–
LGE 2CH	LV_c	Myo	Scar	–	–
LGE 4CH	LV_c	Myo	Scar	RV_c	–
LGE RAS	RA	–	–	–	–

while it correctly identifies axis 2 in all 299 LGE cases. The supplement reports the complete audit and spacing distributions.

2.2 Geometry as a model input

Let A, y and M denote an image array, its label array and the image header, and let $P(M)$ be the slice axis selected by the nnU-Net planner. For cine, the intended learning representation must satisfy

$$A' = A, \qquad y' = y, \qquad P(M') = 2, \qquad M_{\text{out}} = M. \tag{1}$$

The first two equalities exclude interpolation or voxel reordering; the third exposes the frame axis as through-plane; the last returns predictions to the original physical space. The transformation is defined only when metadata identify axis 2 as non-spatial, $P(M) \neq 2$, and the untouched header is available. Volumes that do not satisfy these conditions retain their original representation.

The implementation uses a diagonal M'. For cine in-plane spacings (s_x, s_y), its entries are $(s_x, s_y, 999, 1)$. The last spatial value is a sentinel rather than a physical spacing; it enforces the ordering on which the 2D planner depends. Any value above $\max(s_x, s_y)$ produces the same plan. For SAX, the median training plane changes from the incoherent 275×144 plane to 144×149. Image and label receive the same header, with `sform` code 2 and `qform` code 0, while their arrays remain bitwise unchanged. After prediction and inverse label mapping, the mask is written with the stored input header M. This restoration is necessary because boundary distances are evaluated in physical coordinates.

2.3 View-specific segmentation

Seven nnU-Net models [13,7] represent the three cine and four LGE views. This view-specific design accommodates their different target sets and avoids treating the cine frame stack as three-dimensional anatomy. We use the `2d` configuration for both cine and the thin LGE stacks (3–10 slices). Prior cine experiments found spatial neighbours more informative than temporal neighbours [14]; here, independent frames provide the cleaner test of whether the sampling plane itself is valid. nnU-Net is a competitive, well-characterised baseline, allowing the study to focus on representation rather than architectural novelty [8].

Each view's native labels are mapped to contiguous training indices $0, \ldots, K$ and mapped back after inference. The mapping is estimated from the union of training cases so that classes absent in an individual volume, such as scar in a non-infarct case, are retained. Five fold-specific softmax outputs are averaged. Mirroring-based test-time augmentation (TTA) is used for cine and LGE RAS, but not for LGE 2CH, 4CH or SAX. In SAX, a predicted scar voxel is retained only when $p(\text{scar}) \geq 0.9996$; otherwise it receives the most probable non-scar label. These inference choices were selected on public validation and are analysed as such rather than as independently validated improvements.

2.4 Biomarkers from segmentation

Cine metadata factor the SAX frame axis as $D = n_s n_t$. After reshaping predictions to (H, W, n_s, n_t), the left-ventricular cavity volume at phase ϕ is proportional to

$$V_\phi = \sum_{x,y,s} \mathbf{1}[\hat{y}_{xys\phi} = \mathrm{LV}_c]. \tag{2}$$

A cyclic seven-frame mean stabilises the volume curve, and ejection fraction (EF) is $100(V_{\max} - V_{\min})/V_{\max}$. Scar mass follows directly from the number of SAX scar voxels and the original voxel volume, $m = N_{\mathrm{scar}}(s_x s_y s_z/1000) \times 1.05\,\mathrm{g}$, using the myocardial density specified by the challenge baseline. Applied to reference masks, both estimators reproduce the supplied biomarkers to numerical precision (Section 3.4); subsequent biomarker error therefore reflects the predicted masks rather than an inconsistent reshaping or unit conversion.

3 Experiments

3.1 Dataset and protocol

All experiments use nnU-Net v2.8.1 and its automatically generated `2d` plans. Five fold-specific models are trained per view on NVIDIA A100 accelerators, using the default random split with seed 12345. The selected schedule is 250 epochs, except for LGE 4CH and RAS (1000 epochs). Evaluation uses the public validation split; consequently, results used for training or inference selection are development estimates rather than held-out performance. The challenge data are available from the organisers [12], and the supplement records complete configurations and artifact provenance.

The primary experiment compares native and corrected cine geometry with the architecture, fold and wall-clock budget fixed. LGE provides a negative control because its stored through-plane axis is already consistent with the planner. Additional matched comparisons separate the effects of TTA, scar post-processing and five-fold ensembling. Segmentation is measured by the Dice similarity coefficient (DSC), maximum Hausdorff distance (HD) and average symmetric surface distance (ASD). Empty structures receive DSC 1 only when both masks are empty; their surface distances are omitted. Our implementation reproduces the official scorer (cine HD 9.13 versus 9.16 mm).

3.2 Geometry intervention and negative control

Table 2 isolates representation from model capacity. Corrected geometry improves every challenge component in all three cine views, with Task 1 gains from 0.029 to 0.084. In 20 000 paired patient-bootstrap replicates, every mean-DSC 95% interval remains above zero. Mean DSC improves in 15/15 2CH cases, 15/15 4CH cases and 14/15 SAX cases; exact two-sided sign tests give $p \leq 9.8 \times 10^{-4}$ and remain significant after Bonferroni correction. The largest change occurs

Table 2. Matched cine geometry experiment (fold 0; original physical-space scoring). Models differ only in the learning representation and share a wall-clock budget. ΔDSC is reported with a 95% percentile paired patient-bootstrap interval (20,000 replicates; seed 20260731).

View	Geometry	DSC	HD (mm)	ASD (mm)	Task-1	ΔDSC [95% CI]
Cine SAX	original	0.884	12.03	0.190	0.631	
	rewrite	0.911	11.60	0.119	0.660	+0.027 [0.017, 0.038]
Cine 2CH	original	0.853	11.58	0.142	0.632	
	rewrite	0.891	7.57	0.082	0.671	+0.038 [0.027, 0.051]
Cine 4CH	original	0.858	23.22	0.414	0.595	
	rewrite	0.907	8.22	0.066	0.679	+0.049 [0.033, 0.068]

in 4CH, where maximum HD decreases from 23.2 to 8.2 mm. The effect is also structure-consistent: DSC rises for all ten annotations (range +0.008 to +0.084), with the largest gain for the right atrium in 4CH. The complete per-structure values are reported in the supplement.

The LGE control tests the converse condition. Its genuine through-plane axis is already the largest-spacing axis in all 299 released cases. Native and contract paths consequently produce bitwise-identical values for all 400 preprocessed training image/label tensors and the same four 2D configurations. Public DSC is identical in 2CH, 4CH and RAS; the SAX difference is 1.1×10^{-5} in favour of the native header. The positive cine result and negative LGE result jointly localise the effect to the planner–metadata mismatch.

3.3 End-to-end segmentation

The selected system reaches mean DSC 0.906 for cine, with maximum HD 7.99 mm and ASD 0.293 mm. Mean LGE DSC is 0.764, ranging from 0.711 in 2CH to 0.887 in RAS (Table 3). Chambers and myocardium are consistently delineated more accurately than scar; scar DSC is 0.559, 0.441 and 0.549 in 2CH, 4CH and SAX, respectively. This pattern agrees with the persistent lesion-level difficulty reported by prior LGE benchmarks [9,16,10]. Across fold-specific models, cine DSC varies by ±0.002–0.008, compared with ±0.03–0.04 on the seven-case LGE validation subsets. Thus the apparent LGE differences are both anatomically concentrated and statistically less stable.

3.4 Biomarker validity and accuracy

The reference-mask experiment first isolates the biomarker implementation from segmentation. Across all 15 LVEF and seven scar-mass cases, computed and supplied values differ by less than 0.01 (for example, 10.582 versus 10.582 g and 54.988 versus 54.988). On predicted masks, smoothed LVEF has Pearson correlation coefficient (PCC) 0.920 (Fisher-z 95% confidence interval (CI)

Table 3. Per-class validation Dice (five-fold ensemble), by view. LV_c: LV cavity; Myo: myocardium; RV_c: RV cavity; RA/LA: right/left atrium. Chambers and myocardium are segmented well throughout; myocardial scar is consistently the hardest structure.

View	LV_c	Myo	RV_c	RA	LA	Scar
Cine 2CH	0.921	0.864	–	–	–	–
Cine 4CH	0.944	0.857	0.909	0.907	0.933	–
Cine SAX	0.939	0.912	0.893	–	–	–
LGE 2CH	0.899	0.676	–	–	–	0.559
LGE 4CH	0.890	0.772	0.860	–	–	0.441
LGE SAX	0.814	0.786	0.723	–	–	0.549
LGE RAS	–	–	–	0.887	–	–

$[0.772, 0.974]$). Its mean absolute error (MAE) is 5.9 EF points, with bias -4.8 and 95% limits of agreement (LoA) $[-16.0, +6.4]$. Smoothing therefore improves the challenge correlation from 0.891 while worsening agreement relative to the raw curve (MAE 4.4, bias +0.2).

Scar mass remains uncertain. Its PCC (vPCC in the challenge notation) is 0.851 (95% CI $[0.272, 0.978]$), and relative absolute error (RAE) is 0.479 (patient-bootstrap 95% CI $[0.160, 0.784]$). MAE is 15.2 g, bias is -15.2 g and LoA are $[-61.4, +31.0]$ g. Leave-one-case-out threshold selection retains SAX/scar DSC 0.715/0.536, but vPCC/RAE decrease to 0.804/0.668. We therefore treat scar-mass accuracy as preliminary and report all seven pairs in the supplement.

With predicted biomarkers (EF PCC 0.920; scar-mass vPCC 0.851, RAE 0.479), the selected system obtains 0.740 on public validation (Task 1 0.716; Task 2 0.764). These values summarise the selected development configuration and are not external-test estimates; the official score definition is reproduced in the supplement.

3.5 Ablation studies

The ablations distinguish robust contributions from challenge-specific choices. For cine, test-time augmentation (TTA) reduces maximum HD from 9.13 to 8.31 mm while leaving DSC nearly unchanged, indicating improved boundary robustness rather than better overlap. For LGE, disabling TTA increases 2CH DSC from 0.706 to 0.711 and has negligible effect in 4CH. We therefore disable it for LGE 2CH, 4CH and SAX, but retain it for RAS.

The SAX scar threshold produces the largest post-processing gain: requiring $p(\text{scar}) \geq 0.9996$ increases mean/scar DSC from 0.685/0.392 to 0.718/0.549 and yields mass vPCC/RAE of 0.851/0.479. In contrast, retaining only the largest connected component reduces LGE DSC from 0.740 to 0.702 by removing multifocal scar. Neither extending training to 1000 epochs nor foreground oversampling improves the 0.685 baseline (DSC 0.676 and 0.678, respectively). Dice plus cross-entropy also performs best among the tested losses, with mean/scar DSC

of 0.718/0.550, compared with 0.681/0.408 for Tversky loss and 0.714/0.542 for focal loss.

Five-fold ensembling provides smaller, task-dependent gains. Cine DSC increases by 0.0012 (95% CI $[-0.0006, 0.0029]$), while maximum HD decreases from 8.31 to 7.99 mm; LGE DSC increases from 0.737 to 0.751. We therefore retain ensembling in the final system. Full per-view, per-class and out-of-fold (OOF) controls are reported in the supplement.

4 Discussion and Conclusion

The principal finding is that a data representation assumption can dominate an otherwise unchanged segmentation pipeline. Cardiac benchmarks have established the importance of appearance diversity across pathologies, centres and vendors [4,5]; our results identify axis semantics as a complementary source of shift. This distinction matters for self-configuring systems: rigorous architectural validation [8] still presupposes that spacing describes anatomy. Here, positive and negative controls show that performance changes only when that presupposition is violated.

The geometry contract is an input validity condition, not a resampling algorithm. DICOM-to-NIfTI conversion can assemble temporal frames or 2D instances into multidimensional arrays and varies by vendor and sequence [11]. The result does not justify blind spacing inflation: applicability requires independent evidence of a non-spatial axis, unchanged voxels, verified plane selection and restoration of physical space.

The biomarker results also qualify what good segmentation means. Temporal smoothing raises LVEF correlation but introduces negative bias, demonstrating that a challenge metric can reward ranking while agreement worsens. Scar is the limiting anatomy, consistent with EMIDEC, MS-CMR and MyoPS findings [9,16,10]; with only seven cases, the wide mass interval and leave-one-out degradation preclude a clinical claim. Similarly, the score's weighted HD term, $h(H) = 0.21/(1 + H)$, has derivative $h'(H) = -0.21/(1 + H)^2$. Near 9 mm, a 1-mm improvement changes the total score by only ≈ 0.002, even though one distant voxel can change maximum HD sharply. The 95th-percentile Hausdorff distance (HD95) or surface Dice would better characterise typical boundaries.

Limitations include evaluation within a single challenge, testing the matched original-versus-corrected geometry comparison with only one of the five training folds, choosing inference settings on public validation, and estimating scar mass from seven cases; external generalisation remains untested. Still, the cine gain and LGE invariance show that axis semantics are part of the model: learning cannot recover meaning discarded by its input representation.

References

1. Bai, J., Qiu, R., Chen, J., et al.: A two-stage method with a shared 3D U-Net for left atrial segmentation of late gadolinium-enhanced MRI images. Cardiovascu-

lar Innovations and Applications **8**(1) (2023). https://doi.org/10.15212/CVIA.2023.0039
2. Bai, J., Zhu, J., Chen, Z., et al.: A benchmark framework for the right atrium cavity segmentation from LGE-MRIs. IEEE Transactions on Medical Imaging **44**(12), 5290–5305 (2025). https://doi.org/10.1109/TMI.2025.3590694
3. Bai, W., Sinclair, M., Tarroni, G., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance **20**(1), 65 (2018). https://doi.org/10.1186/s12968-018-0471-x
4. Bernard, O., Lalande, A., Zotti, C., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502
5. Campello, V.M., Gkontra, P., Izquierdo, C., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021). https://doi.org/10.1109/TMI.2021.3090082
6. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine **7**, 25 (2020). https://doi.org/10.3389/fcvm.2020.00025
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
8. 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. pp. 488–498. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72114-4_47
9. Lalande, A., Chen, Z., Pommier, T., et al.: Deep learning methods for automatic evaluation of delayed enhancement-MRI: The results of the EMIDEC challenge. Medical Image Analysis **79**, 102428 (2022). https://doi.org/10.1016/j.media.2022.102428
10. Li, L., Wu, F., Wang, S., Luo, X., et al.: MyoPS: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. Medical Image Analysis **87**, 102808 (2023). https://doi.org/10.1016/j.media.2023.102808
11. Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C.: The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods **264**, 47–56 (2016). https://doi.org/10.1016/j.jneumeth.2016.03.001
12. Qu, T., Su, Z., Zhang, H., et al.: Universal multi-sequence, multi-center and multi-view CMR segmentation challenge (2026). https://doi.org/10.5281/zenodo.19728181
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Sun, X., Garg, P., Plein, S., van der Geest, R.J.: SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging. Medical Physics **48**(4), 1750–1763 (2021). https://doi.org/10.1002/mp.14752

15. Zhu, J., Bai, J., Zhou, Z., et al.: RAS Dataset: A 3D cardiac LGE-MRI dataset for segmentation of right atrial cavity. Scientific Data **11**(1), 401 (2024). https://doi.org/10.1038/s41597-024-03253-9
16. Zhuang, X., Xu, J., Luo, X., et al.: Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge. Medical Image Analysis **81**, 102528 (2022). https://doi.org/10.1016/j.media.2022.102528