Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Metadata Supervised Imaging Representations for Modelling and Controlling Acquisition Variability

Mehmet Yigit Avci

yigit.avci@kcl.ac.uk

King's College London https://orcid.org/0009-0008-6228-5933

Pedro Borges

King's College London

Virginia Fernandez

King's College London

Natalia Glazman

King's College London https://orcid.org/0009-0003-2551-4198

Paul Wright

University College London

Mehmet Yigitsoy

AstraZeneca https://orcid.org/0000-0001-6598-0933

Sebastien Ourselin

School of Biomedical Engineering & Imaging Sciences, King's College London https://orcid.org/0000-0002-5694-5340

Jorge Cardoso

Kings College London https://orcid.org/0000-0003-1284-2558

Article

Keywords: Magnetic resonance imaging, DICOM metadata, Representation learning, Domain shift, Image harmonisation, Contrastive learning, Quality control

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10380756/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. M.Y. Avci's PhD is partially funded by deepc GmbH. The other authors declare no competing interests.

Metadata Supervised Imaging Representations for Modelling and Controlling Acquisition Variability

Mehmet Yigit Avci[1*], Pedro Borges[1], Virginia Fernandez[1], Natalia Glazman[1], Paul Wright[2], Mehmet Yigitsoy[3], Sebastien Ourselin[1], M. Jorge Cardoso[1]

[1]School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK.
[2]University College London, London, UK.
[3]deepc GMBH, Munich, Germany.

*Corresponding author(s). E-mail(s): yigit.avci@kcl.ac.uk;
Contributing authors: pedro.borges@kcl.ac.uk;
virginia.fernandez@kcl.ac.uk; natalia.glazman@kcl.ac.uk;
paul.w@ucl.ac.uk; mehmet.yigitsoy@gmail.com;
sebastien.ourselin@kcl.ac.uk; m.jorge.cardoso@kcl.ac.uk;

Abstract

Biomedical imaging data exhibit substantial acquisition variability, where identical biological structures can appear markedly different due to differences in imaging devices, acquisition protocols, sites, and reconstruction settings. Consequently, learned representations often entangle underlying biological information with acquisition-dependent appearance, limiting interpretability, generalisation, and clinical deployment. We show that these sources of variation can be separated by jointly modelling medical images and acquisition metadata. Using large-scale clinical brain MRI data as a case study, we learn representations that disentangle anatomical structure from contrast-dependent appearance. The resulting acquisition-aware representations organise heterogeneous imaging protocols, support sequence understanding, and detect image–metadata inconsistencies, whereas anatomical representations suppress acquisition-specific variation while preserving biologically relevant information. Building on these disentangled representations, we introduce a unified anatomy-preserving harmonisation model for cross-modality and cross-site adaptation, conditioned on image content or acquisition metadata. Our findings suggest that acquisition variability is a structured component of the imaging process that can be modelled, audited, and

controlled, providing a foundation for acquisition-aware representation learning in large-scale biomedical imaging.

Keywords: Magnetic resonance imaging, DICOM metadata, Representation learning, Domain shift, Image harmonisation, Contrastive learning, Quality control

1 Introduction

Biomedical imaging is a cornerstone of modern healthcare and life sciences, enabling non-invasive visualisation and quantitative assessment of biological structure, function, and pathology across diverse clinical and research applications. However, despite rapid progress in machine learning for medical image analysis, robust deployment of automated imaging models across institutions remains challenging. A key limitation is acquisition shift, where differences in imaging devices, acquisition protocols, and reconstruction settings induce substantial changes in image appearance without corresponding changes in the underlying biology [1]. Unlike natural images, biomedical images are not solely determined by the object being imaged; instead, image appearance reflects a complex interaction between biological structure and the acquisition process. As a result, the same images can exhibit markedly different intensity distributions, textures, and contrast under different acquisition conditions, while visually similar scans may originate from distinct protocols rather than shared biology [2–4]. When learned representations entangle acquisition-specific factors with anatomical content, downstream models become sensitive to domain shifts, limiting their generalisability and hindering reliable deployment across scanners, institutions, and patient populations [5, 6].

This challenge reflects a broader limitation of current MRI representation learning. Existing approaches address acquisition variability through two main strategies. The first is domain generalisation, which trains models to learn representations that are invariant to scanner or site identity, typically through adversarial objectives that suppress domain-discriminative features [7, 8]. The second is image harmonisation, which maps images from one acquisition setting into the appearance of another, either through statistical correction of batch effects [9] or through learned image-to-image translation [10–13]. Both strategies rest on a shared assumption that acquisition-related variation is a nuisance factor to be removed or ignored. However, in MRI, this variability is a structured and physically meaningful consequence of the imaging process itself. By ignoring this underlying physical process, current methods rarely provide representations that explicitly separate anatomical structure from acquisition-dependent appearance, limiting controlled image synthesis, principled cross-site comparison and systematic quality control.

Crucially, much of the information governing acquisition-dependent appearance is already recorded during routine clinical imaging. Every MRI scan is accompanied by DICOM metadata that describes the physical and technical parameters underlying image formation, including echo time, repetition time, inversion time, pulse sequence type, scanner vendor, field strength and reconstruction settings [14]. These parameters

are not merely descriptive labels; they are causal determinants of image contrast and appearance.

Prior work has leveraged metadata for sequence classification [15–18], metadata-conditioned reconstruction [19] and scanner harmonisation [20], while multimodal contrastive learning has demonstrated the broader utility of aligning medical images with auxiliary information sources [21]. However, metadata itself is often incomplete, inconsistent or manually specified in large retrospective archives [22], creating a dual challenge: metadata provides a rich description of acquisition variability, yet its reliability cannot be assumed. Understanding the relationship between image appearance and acquisition metadata is therefore important not only for harmonisation and representation learning, but also for auditing data quality at scale.

Here we introduce **MaRaI** (**M**ultimodal **A**cquisition-aware **R**adiology **AI**), a framework that treats DICOM metadata not as auxiliary labels but as structured supervision for learning representations in which anatomy and acquisition are explicitly separable (Fig. 1). MaRaI has two components that share a common representation space. **MR-CLIP** [23, 24] aligns MRI volumes with natural-language prompts constructed from DICOM acquisition tags, using a supervised contrastive objective [25] over discretised metadata groups to pull together volumes that share the same acquisition protocol regardless of anatomy or subject identity. The resulting embeddings capture contrast as a structured variable, supporting downstream tasks, including cross-modal retrieval, few-shot sequence classification and archive-level quality control. **DIST-CLIP** [26] uses these embeddings to factorise each scan into an acquisition-invariant anatomical map β and a contrast representation θ, extending and unifying the line of disentangled harmonisation methods [10, 11, 27]. To enable controllable image synthesis, we introduce the Style Fusion Decoder, a new style transfer architecture that fuses anatomical and contrast information via cross-attention and Adaptive Instance Normalisation (AdaIN) [28]. It modulates anatomical features using contrast-specific representations to achieve structure-preserving appearance transfer. A single model supports all major brain MRI contrasts, where conditioning is obtained either from a reference image or directly from acquisition metadata, enabling synthesis without requiring a target scan. Our main contributions are:

- **Anatomy–contrast disentanglement with clinical relevance.** We show that anatomical maps derived from DIST-CLIP reduce acquisition-driven variability in tissue morphometry across protocols and improve cross-vendor Alzheimer's disease classification on ADNI relative to raw MRI, establishing that explicit disentanglement of anatomy from contrast delivers measurable clinical benefit beyond visual consistency.
- **Unified harmonisation across contrasts and sites.** We propose a novel style transfer architecture to inject contrast information into disentangled anatomical representations, enabling precise, layer-wise appearance control without altering the underlying anatomy. Image-guided and metadata-conditioned modes share a single set of trained weights, yielding a unified model that performs high-fidelity harmonisation across all major brain MRI contrasts and generalises to scanners unseen during training, a combination not achieved by prior harmonisation frameworks.

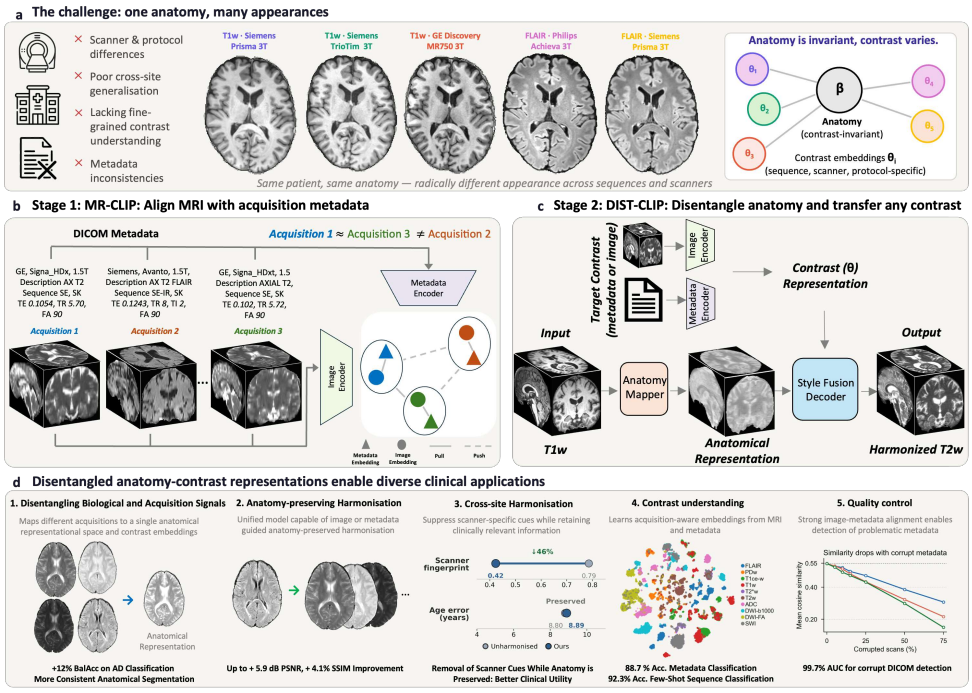


Fig. 1 Disentangled MRI representation learning under acquisition variability. a, MRI appearance varies substantially across scanners, protocols and contrasts despite shared underlying anatomy, motivating separation of anatomy from acquisition-dependent appearance. **b,** Image–metadata contrastive alignment pairs each MRI volume with a natural-language prompt built from its DICOM tags in a shared embedding space. **c,** A disentanglement and harmonisation model built on pretrained contrast embeddings factorises each volume into an anatomical map (β) and a contrast code (θ). Synthesis holds β fixed from a source scan and generates target appearance from a reference image or target metadata alone. **d,** Factorised anatomical and contrast representations enable diverse downstream applications, including anatomically consistent harmonisation, cross-site alignment, sequence understanding and quality control.

- **Metadata-driven contrast-aware representation learning.** We introduce a metadata discretisation strategy that converts heterogeneous DICOM acquisition parameters into structured supervision signals, enabling large-scale supervised contrastive learning. The resulting MR-CLIP embeddings support retrieval, few-shot classification, and quality control, outperforming supervised baselines and demonstrating that acquisition metadata can effectively structure large-scale clinical archives.

Together, we show that routine acquisition metadata, already present in every clinical DICOM archive, can anchor representations in which anatomy and contrast are separable, target appearance is controllable without always requiring a reference scan, and large imaging collections can be screened for metadata integrity before analysis. The framework opens a path toward clinical AI systems that are robust to the acquisition heterogeneity that characterises real-world hospital data.

2 Results

2.1 Anatomical representations suppress acquisition-dependent variability

DIST-CLIP decomposes each scan into two complementary representations: an anatomical map β that captures tissue structure independently of how that structure was imaged, and a contrast code θ that encodes acquisition-dependent appearance. For this factorisation to be meaningful, β must preserve structural information like sulcal geometry, ventricular shape, tissue boundaries that is consistent across protocols while relegating sequence-specific intensity characteristics to θ. We evaluated this claim through two experiments: assessment of morphometric stability across registered multi-contrast acquisitions, and cross-vendor Alzheimer's disease (AD) classification on ADNI [29].

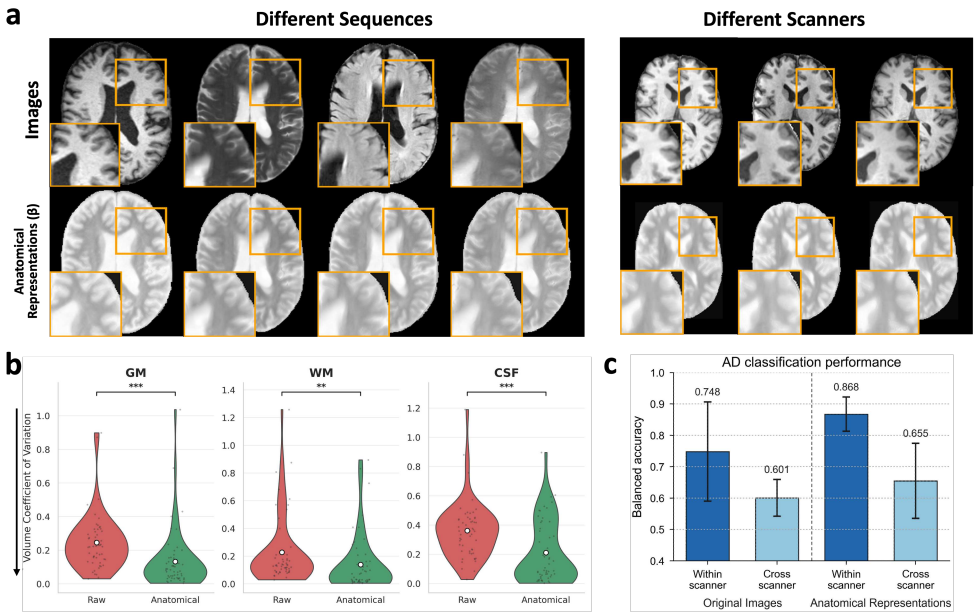


Fig. 2 Anatomical representations suppress acquisition-dependent variability. a, Axial slices from raw MRI acquisitions from different sequences and scanners exhibit substantial appearance variability despite shared underlying anatomy. The proposed framework maps heterogeneous acquisitions into anatomically consistent representations while preserving fine-grained structural boundaries. **b,** Coefficient of variation analysis of tissue-specific segmentation volumes, evaluated across different sequences with at least three acquisitions, demonstrates reduced variability in anatomical representations compared with raw MRI images, indicating suppression of acquisition-dependent variation while preserving shared anatomical structure. **c,** On ADNI, Alzheimer's disease classifiers trained on one scanner generalise better to data from another scanner with anatomical representations than with raw MRI, indicating reduced acquisition variability without loss of diagnostic signal.

Qualitatively, raw slices from the same subject varied substantially in image contrast across sequences and scanners, whereas our anatomical representations showed aligned sulcal and ventricular anatomy (Fig. 2a). Importantly, we do not expect anatomical representations from different sequences to become identical: each acquisition still carries sequence-specific signal. The objective is to suppress non-biological acquisition variability while preserving meaningful anatomical content, which appears as patchwise structural consistency rather than exact intensity matching. For quantitative analysis, we applied SPM12 unified segmentation [30] to raw images and matched β-maps across sessions with > 3 acquisitions (Fig. 2b). Mean Dice overlap averaged across gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) was similar between raw and anatomical images (0.608 versus 0.582; all tissue-wise Dice comparisons not significant), whereas mean inter-sequence volume coefficient of variation (CV) was significantly lower for anatomical maps (0.161 versus 0.279; $p < 0.05$ for GM, WM and CSF). These results indicate that anatomical outputs remain tissue-consistent while reducing protocol-driven variability.

We next evaluated whether this stability improves downstream prediction under scanner shift. On ADNI, AD classifiers (AD vs Cognitively Normal) trained on Siemens scans and tested without fine-tuning on GE scans resulted in higher balanced accuracy (results are reported as mean of 5-fold cross-validation) using β-inputs than raw MRI (0.65 vs. 0.60, $p > 0.05$). Within-scanner evaluation also showed that β-inputs consistently outperformed raw images, yielding an approximately 12% improvement ($p < 0.05$) (Fig. 2c). Overall, the anatomy pathway improves cross-vendor robustness while preserving clinically relevant signal.

2.2 Anatomy-preserved unified harmonisation across contrasts and scanners

Harmonisation directly follows from anatomy-contrast disentanglement: once anatomical structure is separated from acquisition-dependent contrast, new image appearances can be generated by manipulating the contrast component. In this setting, a source scan provides the anatomical content, while the target contrast is specified either through acquisition metadata or a reference image, enabling synthesis of cross-contrast outputs. We trained on registered multi-contrast pairs and compared against image-based HACA3 [11] and metadata-based TUMSyn [20] under matched preprocessing and evaluation. To assess image-based conditioning (DIST-CLIP/I and HACA3) fairly, we use a reference image from a different subject acquired under the same target protocol to extract the target style, while evaluation is performed against the corresponding registered ground-truth target scan of the source subject. Pixel-level fidelity was quantified with SSIM and PSNR on held-out source→target pairs (758 pairs across the four main contrasts; Supplementary Table S1; all directions in Supplementary Fig. S2). For the representative directions in Fig. 3a,b (T2w→PDw, PDw→FLAIR, FLAIR→T2w), DIST-CLIP/I and DIST-CLIP/T achieved the strongest SSIM/PSNR, reaching 30.3 dB and 0.969 SSIM on PDw→T2w. Qualitatively, HACA3 produced locally sharp results but occasionally generated implausible tissue intensities, particularly in FLAIR where ventricular signal was not faithfully preserved. TUMSyn reduced high-frequency noise but often caused oversmoothing and grid-like artefacts,

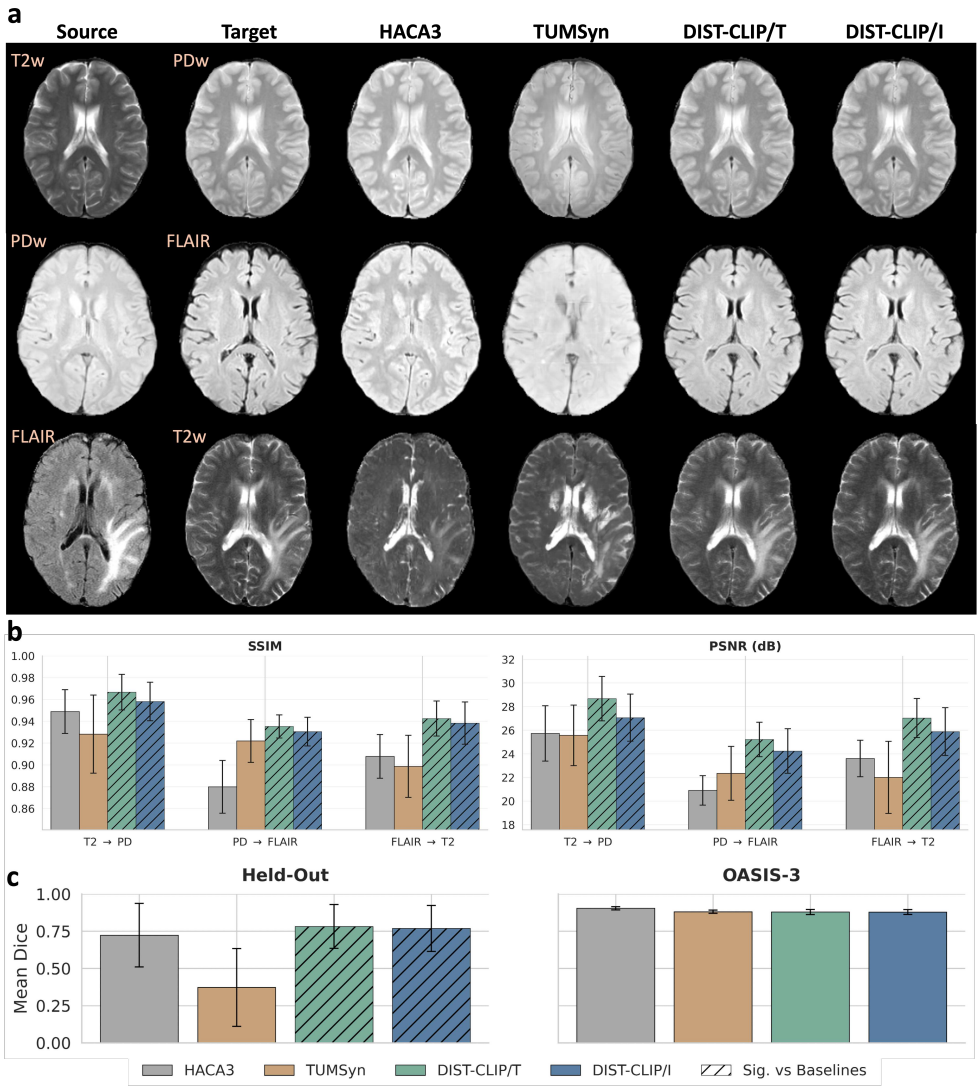


Fig. 3 Cross-contrast harmonisation and segmentation consistency. a, Qualitative T2w→PDw, PDw→FLAIR and FLAIR→T2w examples (source, target, HACA3, TUMSyn, DIST-CLIP/T, DIST-CLIP/I). **b,** Mean SSIM and PSNR (±s.d.) on the same three translation directions. **c,** Mean Dice (±s.d.) between SynthSeg [31] parcellations of reference T2w volumes and T1w→T2w harmonised outputs on held-out clinical test data ($n = 20$) and OASIS-3 ($n = 14$). Hatched bars denote methods that achieve statistically significant improvements over both HACA3 and TUMSyn ($p < 0.05$).

with inconsistent target-domain appearance across volumes. In some cases, pathology signals were attenuated or partially removed; similar suppression effects were also observed for HACA3.

For volumetric consistency, we further trained a DIST-CLIP-2.5D variant without the Anatomy Mapper due to memory constraints, feeding the source image directly into the Style Fusion Decoder. A single 3D MR-CLIP contrast embedding was shared across all slices, while a shared 2D decoder performed slice-wise harmonisation. This enforces volumetric consistency, improving sagittal and coronal coherence (Supplementary Fig. S4), with a downside of oversmoothing cortical texture.

We next asked whether improved pixel fidelity translates into better preservation of anatomically meaningful structure. Using SynthSeg [31], we compared segmentations of T1w→T2w harmonised outputs against the corresponding reference T2w scans (mean Dice across structures; Fig. 3c). On held-out clinical data ($n = 20$), DIST-CLIP/T and DIST-CLIP/I achieved the highest Dice (0.78), performing significantly better than HACA3 (0.70; $p < 0.01$) and substantially outperforming TUMSyn (0.38; $p < 0.001$). On OASIS-3 ($n = 14$), a more homogeneous research dataset on which HACA3 and TUMSyn were trained but DIST-CLIP was not, all methods achieved similarly high Dice scores (0.87–0.89). Dice scores also exhibited less variance than on held-out test, reflecting the substantially higher heterogeneity of routine clinical acquisitions.

We further evaluated clinical utility and robustness to domain shift using cross-site harmonisation on the external ON-Harmony travelling-heads dataset [32]. This dataset comprises ten healthy participants, each scanned at six 3T sites using GE, Philips, and Siemens systems, with no overlap between these scans and the harmonisation training data (Fig. 4). For each method, we synthesised target-site appearances for each subject from a given source scan and assessed performance using both image-level fidelity and downstream task behaviour. For image-based approaches, the target style was extracted by randomly selecting a reference scan from another subject of the desired target site.

To quantify whether harmonisation preserves clinically meaningful information while removing acquisition-specific variability, we trained two downstream 3D convolutional neural networks (CNNs) on the harmonised outputs of the ON-Harmony dataset. Specifically, we used a 4-layer 3D CNN for the brain-age regression model (evaluated via mean absolute error, MAE). For the site classification task, we employed a 3D ResNet-18 to predict the acquisition site. All models were trained and evaluated using a leave-one-subject-out cross-validation. Lower site classification accuracy indicates reduced scanner- and protocol-specific information, while lower brain-age MAE indicates better preservation of anatomical structure relevant to normative ageing. An effective harmonisation method should therefore reduce site identifiability without degrading downstream predictive performance.

DIST-CLIP/I and DIST-CLIP/T achieved the highest SSIM and PSNR ($p < 0.001$ vs. all baselines), indicating superior reconstruction quality under cross-site translation. On downstream tasks, unharmonised data yielded a brain-age MAE of 8.80 years and site classification accuracy of 0.79. DIST-CLIP/I reduced site classification accuracy to 0.42 while preserving brain-age performance (MAE 8.89 years). TUMSyn showed weaker scanner fingerprint removal (accuracy 0.45) and slightly higher MAE (8.96 years), consistent with its lower SSIM/PSNR and contrast discrepancies (Fig. 4). HACA3 retained substantial site-specific signal (accuracy 0.78) and

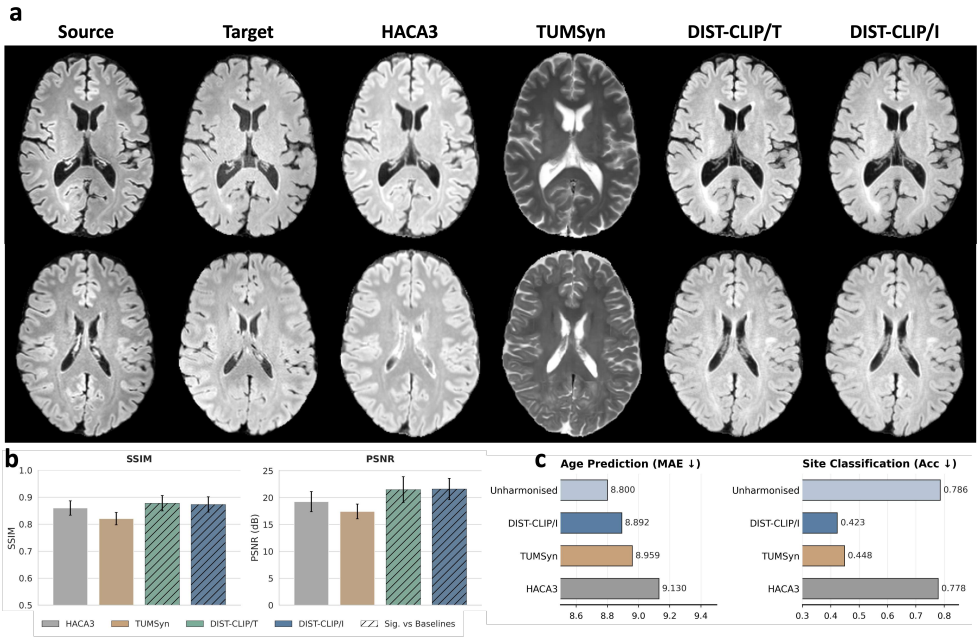


Fig. 4 Cross-scanner harmonisation on the external ON-Harmony travelling-head dataset ($n = 10$ participants). a, Representative examples of FLAIR images harmonised across scanners while preserving anatomical structure. **b,** Quantitative evaluation of harmonisation quality using structural similarity index (SSIM↑) and peak signal-to-noise ratio (PSNR↑), measured against scanner-matched reference images. **c,** Downstream evaluation on the same cohort using brain-age prediction (MAE↓) and scanner/site classification (Acc↓). Effective harmonisation reduces scanner-specific information, reflected by lower site-classification accuracy, while preserving biologically relevant features required for accurate brain-age estimation.

produced the highest brain-age error (9.13 years). Overall, DIST-CLIP/I provides a more favourable balance, successfully reducing scanner-related confounding while maintaining the utility of harmonised images for downstream clinical prediction.

Finally, zero-shot evaluation on OASIS-3, a dataset not used during the training of DIST-CLIP whereas the baselines were trained on it, showed that DIST-CLIP produces anatomically plausible outputs with highly competitive SSIM and PSNR (Supplementary Fig. S7). DIST-CLIP/I significantly outperformed HACA3 on both T1w→T2w and FLAIR→T2w tasks ($p < 0.001$), and achieved performance statistically indistinguishable from the TUMSyn baseline on FLAIR translations ($p > 0.92$) whereas TUMSyn slightly overperformed on other settings ($p < 0.01$). Combined with its strong segmentation agreement (Fig. 3c), these findings support robust generalisation beyond the training distribution.

a

Model	Slice → Metadata			3D Scan → Metadata			Metadata → Slice/(*3D Scan*)			Linear Acc.
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	Acc.
BiomedCLIP	1.4	5.0	8.4	2.5	9.8	15.0	3.6	9.5	13.1	39.0
BiomedCLIP (fine-tuned)	50.0	78.5	82.6	67.4	89.1	92.1	38.5	65.8	71.8	75.5
ViT-B/16 (InfoNCE)	65.6	**85.2**	**90.4**	68.8	92.2	94.4	49.3	69.3	76.6	71.3
ViT-S/16	46.7	79.1	84.4	69.0	92.2	95.2	64.6	77.8	80.9	73.6
2D MR-CLIP (ViT-B/16)	**66.0**	77.3	78.3	**78.7**	**94.2**	**95.3**	**90.9**	**93.6**	**94.4**	82.6
3D MR-CLIP	—	—	—	60.2	79.0	82.0	*(79.3)*	*(91.6)*	*(94.0)*	**86.9**

Values in %. Bold: highest score per column. —: not applicable. Parentheses (italics) indicate Metadata → 3D Scan retrieval.

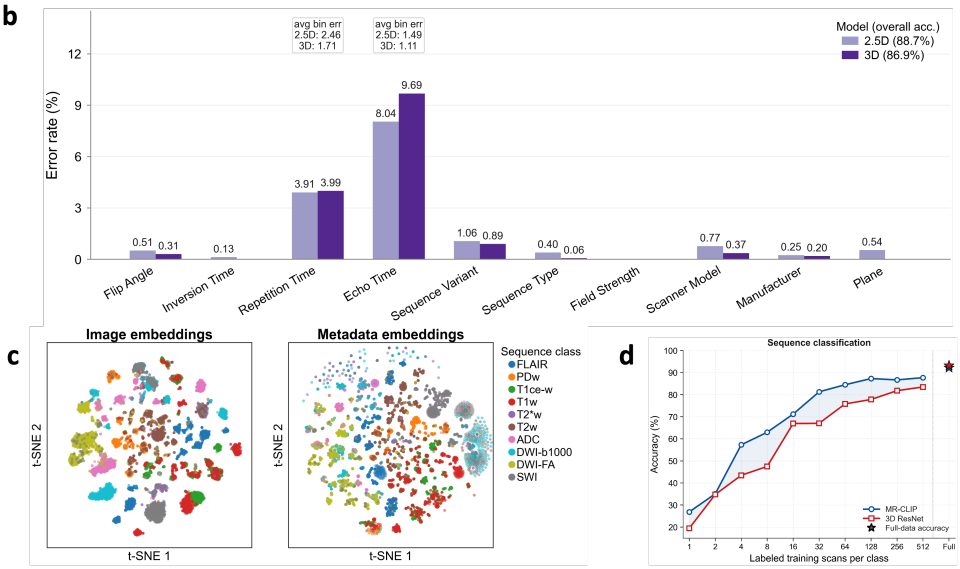


Fig. 5 Metadata-guided representations organise acquisition-dependent latent structure a, Cross-modal retrieval (recall at ranks 1, 5 and 10) for slice-to-metadata, 3D scan-to-metadata and metadata-to-slice matching, with linear-probe accuracy for metadata classification. Highest value per column is shown in bold. **b,** Per-field linear-probe error rates (%) for 2.5D and 3D MR-CLIP across ten DICOM tags; inset values report mean bin-level error for discretised echo time (TE) and repetition time (TR). **c,** t-SNE of image embeddings (left) and metadata embeddings (right) coloured by sequence class. **d,** Few-shot sequence classification accuracy versus labelled training scans per class, comparing a linear classifier on frozen MR-CLIP image embeddings with a supervised 3D ResNet; star denotes full-data accuracy (MR-CLIP significantly outperforms ResNet across few-shot levels, $p < 0.0001$).

2.3 Metadata-guided representations organise acquisition-dependent latent structure

MR-CLIP is designed to understand and encode image contrast by aligning MRI scans with acquisition metadata derived from DICOM tags. The motivation is that MRI appearance is fundamentally determined by acquisition physics (e.g., scanner type, sequence parameters, field strength), yet these factors are often treated as noisy labels or ignored in representation learning. We address this by discretising numerical metadata and grouping acquisitions into contrast definitions as described in Section 4.3. Using these discretised acquisition metadata as supervision signals, MR-CLIP learns

embeddings in which image similarity is structured by imaging protocol rather than manually defined broad sequence labels.

We trained MR-CLIP in three configurations: a two-dimensional model, a slice-aggregated 2.5D model, and a three-dimensional volumetric model, using 169,634 brain volumes from 40,005 subjects spanning 1,415 contrast groups from clinically acquired data with diverse scanner and acquisition settings (Supplementary Fig. S1). The 2D model operated on individual slices, whereas the 2.5D variant aggregated predictions across slices at inference time (e.g., by maximum-vote aggregation for classification tasks) to obtain volume-level predictions. We compared these models against BiomedCLIP [33] and a ViT-B/16 baseline trained with conventional InfoNCE supervision [34, 35].

Cross-modal retrieval on the held-out test set (Fig. 5a) quantifies whether embeddings organise acquisitions by protocol. Two-dimensional MR-CLIP reached 66.0% recall@1 for slice-to-metadata matching and 90.9% for metadata-to-slice matching. At volume level, scan-to-metadata recall@1 reached 78.7% when slice embeddings were aggregated and 60.2% with the native three-dimensional encoder. MR-CLIP consistently outperformed frozen and fine-tuned BiomedCLIP, and in most configurations achieved higher performance than the InfoNCE ViT-B/16 baseline. Metadata-to-slice retrieval exceeded slice-to-metadata retrieval, as expected since many slices share one acquisition description. Linear probing on frozen three-dimensional embeddings gave 86.9% accuracy for grouped metadata classification. Main results used skull-stripped volumes; retaining the skull slightly increased retrieval scores (Supplementary Table S2).

We then trained linear probes to predict the full set of metadata attributes from frozen embeddings and report performance for each individual field in Fig. 5b. Categorical attributes such as imaging plane, field strength, sequence type, and manufacturer were recovered with near-zero error. For 3D MR-CLIP, the largest prediction errors were observed for echo time (TE) and repetition time (TR), where mean offsets reached 1.11 and 1.71 bins, corresponding to approximately 11.1 ms and 855 ms, respectively, which likely reflects discretisation ambiguity rather than missing contrast information. Overall, 2.5D slice aggregation slightly outperformed its 3D counterpart thanks to its strategy of max-vote aggregation across slices within a volume.

t-SNE projections of image and metadata embeddings formed well-aligned clusters that emerged naturally according to major MRI sequence families (e.g., T1w, T2w, FLAIR, and diffusion-weighted imaging; Fig. 5c). In few-shot sequence classification, a linear classifier on frozen MR-CLIP embeddings significantly outperformed an end-to-end 3D ResNet when labels were scarce ($p < 0.0001$ at all evaluated few-shot levels except 2 shots) and matched it with abundant labels (Fig. 5d). Metadata-supervised pretraining therefore enables robust sequence classification from frozen embeddings, including rare or inconsistently annotated protocols, without requiring large manually labelled datasets.

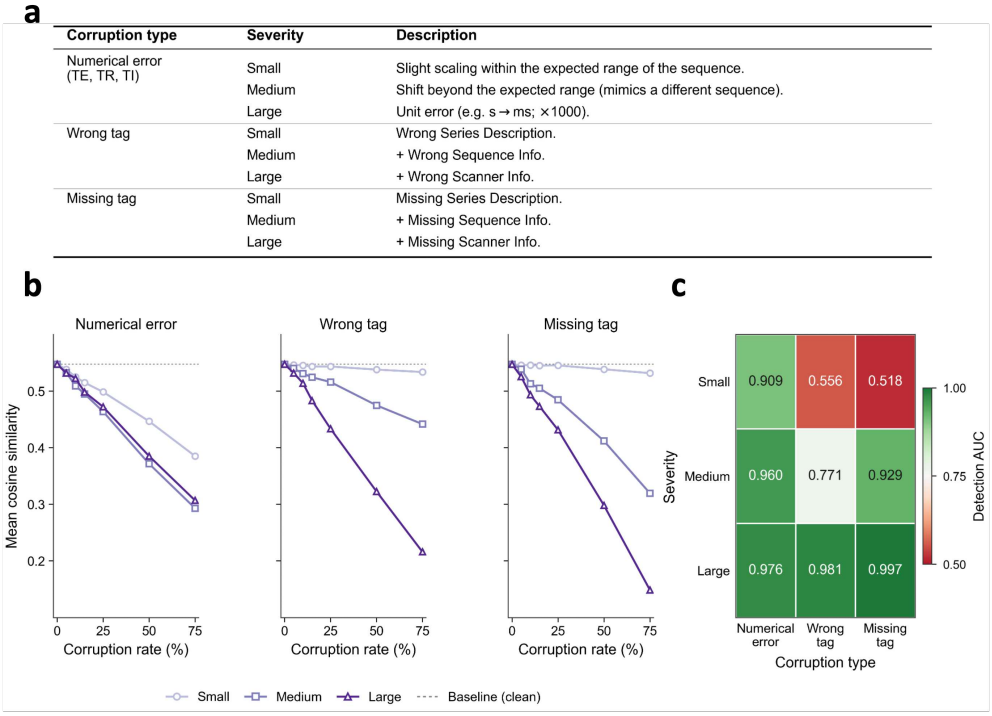


Fig. 6 Unsupervised detection of DICOM metadata corruption via image–metadata embedding consistency. a, Synthetic corruption protocol applied to held-out test metadata, comprising numerical errors; incorrect categorical tags; and missing fields, each at small, medium and large severity. **b,** Mean cosine similarity between MR-CLIP image and metadata embeddings as a function of corruption rate (0–75%) for each corruption type and severity; dotted line denotes clean baseline. **c,** Detection AUC at 50% corruption rate using embedding distance as an anomaly score, evaluated separately for each corruption type and severity (colour scale: 0.5–1.0).

2.4 Image–metadata consistency enables automated quality control

Clinical imaging archives rely on accurate DICOM metadata and consistent image quality for organisation, retrieval, and quality control [36–38]. However, acquisition fields such as timing parameters, sequence labels, and vendor information are frequently missing, inconsistent, or incorrectly recorded. Because MR-CLIP aligns each volume with a metadata prompt, inconsistent tags should reduce image–metadata similarity even when the image appears normal. On held-out test volumes we corrupted metadata at rates from 0% to 75%, perturbing echo time, repetition time and inversion time, swapping categorical tags, or deleting fields at graded severity (Fig. 6a).

Mean cosine similarity between image and metadata embeddings fell monotonically with corruption rate for all classes (Fig. 6b). Numerical timing errors produced the steepest decline, consistent with TE, TR and TI governing contrast. Wrong categorical tags separated clearly by severity; missing sequence or scanner identifiers were most

damaging at high severity. Edits to series description alone had little effect because that field is inherently noisy and excluded from contrast groups during training.

At 50% corrupted volumes, embedding distance classified clean versus corrupted metadata with high area under the receiver operating characteristic curve (AUC) for severe edits (Fig. 6c): 0.997 (large missing tags), 0.981 (large wrong tags), 0.976 (large numerical unit errors). Subtle wrong-tag or missing-tag edits were harder (AUC 0.556 and 0.518). Small numerical shifts were still detected reliably (AUC 0.909–0.976). Across corruption rates from 5 to 75%, detection AUC rose with both error rate and severity for all three corruption classes; large edits remained near-perfect even at low prevalence, whereas small wrong-tag and missing-tag corruptions were weakest at low rates (Supplementary Fig. S6). The same MR-CLIP encoder used for retrieval and harmonisation can therefore flag metadata–image inconsistency without hand-written rules, supporting human review of large archives.

3 Discussion

Multi-centre MRI will remain heterogeneous as scanners are replaced, sequences are shortened for throughput and new acquisition protocols are introduced. MaRaI rests on a simple premise: much of that heterogeneity is not opaque, it is recorded in DICOM tags and coupled to acquisition physics. Learning from metadata at hospital scale, then factorising anatomy from acquisition-dependent appearance offers a practical route to reduce, control and audit acquisition shift (Fig. 1).

Anatomical maps stabilize morphometry across protocols and improve cross-vendor classification, indicating that a shared structural representation can be extracted before comparing pathology across sites. Harmonisation extends the same factorisation to synthesis: fixing anatomical maps (β) while changing contrast or scanner signature is a constructive proof that tissue boundaries are preserved, and ON-Harmony downstream evaluations confirm that harmonised outputs lose scanner identifiability while retaining brain-age signal. The combination of cross-contrast translation, cross-scanner generalisation, and metadata-only target specification with performance close to image-guided synthesis were not previously combined in one harmonisation framework tied to archive-scale contrast learning. Earlier methods typically specialise along a single axis, paired-scanner harmonisation, one contrast direction, or target-image conditioning at inference [10–12, 20, 39, 40]. DIST-CLIP avoids these restrictions by reusing MR-CLIP contrast embeddings with explicit anatomy–style separation, which allows DIST-CLIP/I and DIST-CLIP/T to share a single set of trained weights.

Beyond harmonisation, the same contrast embeddings serve as a scalable tool for organising and auditing large clinical archives. Retrieval by acquisition similarity enables cohort stratification without manual sequence labelling; few-shot classification supports sequence discovery in datasets where series descriptions are inconsistent or absent; and embedding-based flagging identifies volumes where image appearance and DICOM tags disagree. Each of these failure modes has concrete consequences: mislabelled sequences corrupt cohort definitions, undetected drift in echo time after a protocol edit silently shifts feature distributions, and missing vendor fields obstruct federated queries. Embedding-based quality control complements existing rule-based

auditing tools [36–38] by ranking suspicious volumes for human review rather than requiring exhaustive manual inspection. For multi-site trials, this closes an important loop: metadata-conditioned harmonisation can normalise appearance to a protocol specification in the trial charter even when individual sites lack an on-scanner reference scan for every contrast, provided that acquisition tags are trustworthy, which is precisely what consistency checking enforces.

Clinical images contain more than anatomy and acquisition protocol. Pathology, motion and acquisition artefacts also shape image appearance. Although MaRaI explicitly separates anatomy from acquisition-dependent contrast, these additional factors remain only partially disentangled. As a result, harmonisation could inadvertently transfer disease-related appearance as if it were acquisition style, and downstream models may still confound pathology with residual acquisition effects. Extending the framework to explicitly represent pathology and artefacts as separate factors is therefore a natural next step. Such a tri-factor decomposition—anatomy, acquisition and pathology—more closely reflects how radiologists interpret images and may be necessary for disease-centric applications.

Our evidence is strongest for adult brain MRI acquired in routine clinical practice. Whether the same principles extend to paediatric imaging, other organs, or quantitative MRI remains to be explored. Although we validated the approach across multiple datasets and scanner vendors, further evaluation on independent clinical data from a wider range of institutions, patient demographics, acquisition settings, and manufacturers is needed to establish robustness and generalisability. The framework also relies on informative DICOM metadata. While major inconsistencies can be detected through image–metadata agreement, subtle metadata corruption and incomplete protocol documentation remain challenging. Future work should therefore include prospective deployment studies, expert-led quality-control assessments, and validation across broader clinical environments.

Despite these limitations, the central finding remains: acquisition variability need not be treated solely as an unavoidable source of domain shift. By learning from metadata at hospital scale and separating anatomy from acquisition-dependent appearance, variability can be transformed into a structured component of the imaging process that can be modelled, audited and controlled. This perspective opens the door to acquisition-aware foundation models that can explicitly account for and control acquisition variability, an important capability for deployment in heterogeneous real-world imaging systems, and thereby contribute to more robust, equitable, and auditable medical imaging AI.

4 Methods

MaRaI was trained and evaluated on de-identified clinical brain MRI data, with external benchmarks on ON-Harmony, ADNI and OASIS-3. The pipeline comprises shared preprocessing, metadata grouping and prompt construction, MR-CLIP contrastive pretraining, DIST-CLIP harmonisation with frozen MR-CLIP encoders, and evaluation protocols for retrieval, harmonisation metrics, downstream readouts and metadata quality control.

4.1 Datasets and data splits

CLIP training cohort (MR-CLIP). Brain MRI volumes were collected from King's College Hospital (KCH) and Guy's and St Thomas' NHS Foundation Trust (GSTT), comprising 40,005 subjects and 169,634 three-dimensional volumes. DICOM metadata yielded 21,660 unique acquisition configurations, which were consolidated into 1,415 contrast-aware groups after metadata discretisation (see Section 4.3). Data were split at the subject level into training (60%), validation (10%), and test (30%) sets, with an enlarged test set to facilitate a more robust evaluation of harmonization on unseen subjects.

Harmonisation cohort (DIST-CLIP). For cross-contrast synthesis, we assembled paired brain MRI data by selecting imaging sessions in which a subject underwent at least two distinct MRI contrast acquisitions. The resulting dataset comprised 21,115 volumes from 8,466 subjects across 9,280 imaging sessions. Of these sessions, 91.0% contained two acquisitions, 7.8% contained three, and 1.2% contained four. Models were trained on pairs from the MR-CLIP training partition; quantitative harmonisation results in the main figures and Supplementary Fig. S2 are reported on a fixed held-out test set of source–target pairs from MR-CLIP test set partition (counts in Supplementary Table S1).

External evaluation cohorts. Cross-scanner harmonisation was evaluated on the phase A of ON-Harmony travelling-heads dataset [32]: FLAIR volumes from ten healthy participants, each scanned at multiple 3 T sites with systems from GE, Philips and Siemens were used for harmonisation and downstream brain-age and site-classification readouts. Anatomical representation analyses and Alzheimer's disease classification used the T1w scans from Alzheimer's Disease Neuroimaging Initiative (ADNI) [29] for cross-vendor generalisation: classifiers were trained on Siemens acquisitions (n=1042 subjects) and evaluated on GE acquisitions (n=260 subjects). Out-of-distribution harmonisation used OASIS-3 [41] (Supplementary Fig. S7). All data were preprocessed with the same registration, skull-stripping and normalization pipeline as the clinical cohort.

4.2 Image preprocessing

All three-dimensional volumes were rigidly registered to MNI152 space and skull-stripped using SynthStrip [42] unless noted otherwise in ablation experiments (Supplementary Tables S3 and S2). For 2D MR-CLIP training, every second slice from the central 100 slices of each volume was retained. Acquisition plane (axial, coronal or sagittal) was inferred from voxel spacing by selecting the axis with highest in-plane resolution; isotropic volumes defaulted to axial sampling. Slices were resized to 224×224 pixels and intensity-normalized to $[0, 1]$. For harmonisation training, the same subsampling strategy was used whereas all slices were used at test time.

4.3 DICOM metadata processing and contrast grouping

For each acquisition, we extracted DICOM tags that jointly determine image contrast: manufacturer, scanner model, imaging plane, magnetic field strength, sequence type, sequence variant, flip angle, echo time (TE), repetition time (TR) and inversion

time (TI). Series description was included in metadata prompts as auxiliary text but excluded from contrast-group construction owing to site-dependent variability.

Categorical tags were grouped by exact value. Numerical timing parameters were discretised to reduce sensitivity to minor protocol variation: TE and TR were jointly binned on a 20×20 grid spanning the TE–TR ranges observed in the training cohort, and TI was binned separately (most acquisitions lack an inversion pulse), following the grouping strategy introduced in MR-CLIP [23]. Each unique combination of grouped categorical values and discretised timing bins defined one contrast-aware supervisory label used for supervised contrastive learning. Alternative discretisations (for example, 5×5 or 10×10 TE–TR grids) were explored in ablations and a full comparison across grid granularities and transfer settings (including k-means discretisations) is provided in Supplementary Table S4.

Metadata were formatted as natural-language prompts using a fixed sentence template ("A brain MRI, plane ..., Scanner (Manufacturer, Model, Field Strength): (...), Acquisition (Description, Sequence, Variant): (...), Imaging Parameters (Echo Time, Repetition Time, Inversion Time, Flip Angle): (...)"). Exemplar common metadata descriptions are provided in Supplementary Table S5.

4.4 MR-CLIP: metadata-guided contrastive learning

MR-CLIP aligns MRI images and metadata prompts in a shared embedding space using bidirectional supervised contrastive (SupCon) loss [25]. Let z_i denote the Ł_2-normalized embedding of anchor i (image or text). The set of positives $P(i)$ contains all batch embeddings sharing the same contrast-aware label as i, excluding i itself; negatives comprise all remaining embeddings in the batch. For temperature τ,

$$\mathcal{L}_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i^\top z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i^\top z_a/\tau)}, \tag{1}$$

where $A(i)$ is the set of all embeddings in the batch except i. The training objective averages image-to-text and text-to-image SupCon terms:

$$\mathcal{L}_{\text{MR-CLIP}} = \tfrac{1}{2}\left(\mathcal{L}^{\text{img}\to\text{text}} + \mathcal{L}^{\text{text}\to\text{img}}\right). \tag{2}$$

Architectures. The default image encoder is a Vision Transformer (ViT-B/16); metadata were encoded with the paired text transformer. Image and metadata embeddings were Ł_2-normalized to unit length before contrastive learning. We additionally trained a three-dimensional variant with a volumetric ViT image encoder operating on full scans, and a 2.5D variant that aggregated slice-level embeddings for volume-level retrieval and classification.

Training. Models were implemented in PyTorch. Two-dimensional MR-CLIP was trained on three NVIDIA A100 GPUs (40 GB) with per-GPU batch size 3,000 using gradient accumulation and sharded contrastive loss following OpenCLIP [43]. Three-dimensional MR-CLIP used per-GPU batch size 150 with the same sharded-loss strategy. Optimisation used Adam [44] (lr $= 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay 0.2) for 100 epochs with 2,000 warm-up steps. Gradient checkpointing was

enabled. Training augmentations included random affine transforms, random resized crops, Gaussian blur and horizontal flips; patch dropout (0.5) and text dropout (0.2) were applied to the image and text towers, respectively. The SupCon temperature τ followed the OpenCLIP ViT-B/16 configuration used for initialization.

4.5 DIST-CLIP: anatomy–contrast factorisation and harmonisation

DIST-CLIP disentangles each slice into an anatomical map β and a contrast embedding θ, then synthesises a harmonised image whose appearance matches a target contrast while preserving anatomy, building on disentangled MRI synthesis [10, 11, 27] and CLIP-guided style transfer [45, 46].

Anatomy mapper. The source image I_{src} is encoded by a U-Net anatomy mapper with instance normalization in the first layer to produce β_{src}, intended to capture tissue boundaries and global morphology while suppressing contrast-specific intensity patterns.

Contrast encoding. Pre-trained and frozen MR-CLIP encoders [23, 24] supply contrast guidance. The image encoder E_I maps a target image to $\theta_i = E_I(I_{\text{tgt}})$; the metadata encoder E_M maps the target prompt to $\theta_m = E_M(M_{\text{tgt}})$. During training, the style fusion decoder was conditioned on θ_i or θ_m with probability 0.5 per iteration so that either modality can drive inference alone. To further promote disentanglement and flexibility, the target contrast is derived from a randomly sampled brain slice, ensuring that the conditioning signal is not tied to a fixed spatial location and encouraging invariance to anatomical position.

Style fusion decoder and adaptive style transfer. The style fusion decoder (Supplementary Fig. S5) maps the source anatomical representation β_{src} to a harmonised slice while rewriting acquisition-dependent appearance. Architecturally, it is a U-Net decoder with twice the channel width of the anatomy mapper: skip connections from the encoder are merged at each upsampling stage to preserve spatial anatomy, and an adaptive style transfer (AST) block is inserted at every resolution level (bottleneck and each decoder stage).

Each AST block conditions decoder features on the target contrast embedding θ (θ_i from the MR-CLIP image encoder or θ_m from the metadata encoder; Supplementary Fig. S5, right). First, an MLP maps θ to a style query. Multi-head attention then compares this query with feature maps from the previous decoder layer, yielding spatially selective, style-conditioned activations. In parallel, a second MLP predicts channel-wise mean and standard deviation parameters. These parameters drive spatial adaptive instance normalization (AdaIN) [28], which rescales the attended features so local intensity statistics match the target contrast while anatomical layout remains anchored to β_{src}. The bottleneck AST module uses eight-head attention; upsampling AST modules use lighter linear-attention variants for efficiency. Injecting style at multiple scales allows both global contrast and fine texture to be controlled jointly, rather than only at the deepest layer.

Training objectives. Training minimized

$$\mathcal{L}_{\text{total}} = \lambda_\beta \mathcal{L}_\beta + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{global}} \mathcal{L}_{\text{global}} + \lambda_{\text{dir}} \mathcal{L}_{\text{dir}} \quad (3)$$

where $\mathcal{L}_{\text{rec}}$ is an ℓ_1 reconstruction term between the synthesised image I_{rec} and target I_{tgt}. $\mathcal{L}_{\text{perc}}$ is a multi-layer VGG-19 perceptual loss. $\mathcal{L}_{\text{adv}}$ is a patch-based adversarial term that encourages synthesised slices to be locally indistinguishable from real target contrast images. We use a 2D PatchGAN discriminator implemented in MONAI [47], following the Pix2PixHD architecture [48]. To preserve fine-grained anatomical structure while promoting contrast invariance, we adopt the patch-level contrastive alignment loss ($\mathcal{L}_\beta$) from HACA3 [11]. Rather than enforcing identical features across contrasts, which can be overly restrictive, this loss aligns corresponding patches in the source (β_{src}) and target (β_{tgt}) anatomy maps while separating non-corresponding patches. Implemented as a patch-wise InfoNCE objective [34], it encourages structural consistency across contrasts while allowing necessary feature variations. Formally, the loss is defined as:

$$\mathcal{L}_\beta = \mathbb{E}_{\beta_{src},\beta_{tgt}} \sum_{i \in P} \left[-\log \frac{e^{\text{sim}(z_i, z_i')/\tau}}{\sum_{k \in P} e^{\text{sim}(z_i, z_k')/\tau}} \right] \tag{4}$$

where P is the set of all patch locations. For each source patch z_i in β_{src}, z_i' is the corresponding patch in β_{tgt}, while z_k' (for all $k \in P$) represents *all* patches in the target map, $\text{sim}(\cdot,\cdot)$ is the cosine similarity, and τ is the temperature parameter. Global style consistency was enforced by

$$\mathcal{L}_{\text{global}} = \tfrac{1}{2}(1 - \cos(E_I(I_{\text{rec}}), \theta_i)) + \tfrac{1}{2}(1 - \cos(E_I(I_{\text{rec}}), \theta_m)) \tag{5}$$

Directional consistency followed StyleGAN-NADA [45]:

$$\mathcal{L}_{\text{dir}} = 1 - (\Delta I \cdot \Delta M)/(\|\Delta I\| \|\Delta M\|) \tag{6}$$

with $\Delta I = E_I(I_{\text{rec}}) - E_I(I_{\text{src}})$ and $\Delta M = E_M(M_{\text{tgt}}) - E_M(M_{\text{src}})$. Loss weights were selected to keep all objectives on a comparable scale, ensuring balanced optimization and preventing any single term from dominating training: $\lambda_\beta = 0.1$, $\lambda_{\text{rec}} = 10$, $\lambda_{\text{perc}} = \lambda_{\text{adv}} = 1$, and $\lambda_{\text{global}} = \lambda_{\text{dir}} = 1$.

Training. For each training iteration, we sampled a subject with at least two available contrasts and constructed an ordered source–target contrast pair, yielding the corresponding paired volumes. DIST-CLIP was trained for 25 epochs on NVIDIA A100 GPUs using the Adam optimizer (lr $= 10^{-4}$, batch size 25). The MR-CLIP encoders were kept frozen throughout harmonisation training. The anatomy mapper, style fusion decoder, and discriminator were optimized jointly in an end-to-end manner.

DIST-CLIP-2.5D. For volumetric harmonisation we trained a separate DIST-CLIP-2.5D model without the Anatomy mapper due to memory constraints. Contrast guidance used the frozen three-dimensional MR-CLIP image encoder: for each volume, θ was computed once from the target scan and broadcast to all slices before decoding, enforcing slice-consistent contrast within the volume. Training objectives and data partitions matched slice-wise DIST-CLIP/I; only the contrast encoder (3D versus 2D MR-CLIP) and harmonisation network weights differed.

4.6 Evaluation protocols

Cross-modal retrieval and linear probing. We evaluated recall at ranks 1, 5 and 10 (R@1/5/10) for slice-to-metadata, metadata-to-slice (volume), and volume-to-metadata retrieval on the held-out test set. Volume-level image embeddings were obtained by majority vote across slice embeddings (2D/2.5D) or direct volumetric encoding (3D). Linear probes were trained on frozen image embeddings to predict grouped metadata labels and sequence class.

Few-shot sequence classification. A linear classifier was trained on frozen MR-CLIP image embeddings with $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ labelled volumes per sequence class and compared against a supervised 3D ResNet trained end-to-end on the same labelled subsets.

Metadata quality control. On held-out test scans, we introduced synthetic metadata corruptions [24] at rates of 0, 25, 50 and 75% of the test set. Corruptions comprised: (i) *Numerical error*: small scaling of TE/TR/TI within range, medium shifts outside expected sequence ranges, or large unit errors (for example, seconds versus milliseconds); (ii) *Wrong tag*: incorrect series description, sequence fields, and scanner identifiers; and (iii) *Missing tag*: progressively removing series description, sequence information, and scanner fields. Mean cosine similarity between $E_I(I)$ and $E_M(M)$ was computed as a function of corruption rate. Detection performance was summarised by the area under the receiver operating characteristic curve (AUC) at 50% corruption rate, using one minus cosine similarity as an anomaly score.

Harmonisation metrics. Image fidelity was quantified using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) against target slices on the held-out test pairs, reported for every available source–target contrast direction (Supplementary Fig. S2). PSNR and SSIM were computed in intensity space on 224×224 slices after aligning synthesised outputs to target contrast. Volumetric consistency was assessed qualitatively by inspecting axial, sagittal and coronal reformations (Supplementary Fig. S4). For T1w→T2w harmonisation, FreeSurfer SynthSeg [31] was applied to each method's harmonised NIfTI and to the reference T2w scan; per case we computed mean Dice across labelled structures between reference and harmonised parcellations and averaged across cases (held-out clinical set, $n = 20$; OASIS-3 test set, $n = 14$; evaluation scripts in the MaRaI repository). On ON-Harmony [32] ($n = 10$ participants), we additionally evaluated whether harmonisation suppresses scanner-specific signatures while preserving biological signal using brain-age regression and scanner-vendor classification on harmonised versus native slices.

Anatomical representation analysis. Anatomical maps β were extracted with the trained DIST-CLIP anatomy mapper. Segmentation consistency was evaluated on paired cross-contrast subjects from the harmonisation cohort ($n = 53$ subjects with at least three registered acquisitions of distinct contrasts, including T1w, T2w, FLAIR, PD and T2*). For each acquisition we segmented both the native scans used in training and the corresponding anatomical β-maps.

Tissue maps were generated independently for every image with SPM12 unified segmentation [30], yielding discrete labels for grey matter (label 1), white matter (label 2) and cerebrospinal fluid (label 3). Tissue volume (ml) was computed as the labelled voxel count multiplied by voxel volume. For each subject and tissue class, we computed

the coefficient of variation (CV= σ/μ) of volumes across all available acquisitions of that subject, separately for raw slices and β-maps. Lower CV indicates more stable tissue quantification across contrasts.

Cross-scanner Alzheimer's disease classification on ADNI [29] compared models trained on Siemens acquisitions and tested on GE acquisitions using raw intensities versus β-maps as input. Results report Balanced Accuracy on 5-fold Cross-validation.

Baselines. MR-CLIP was compared against frozen and fine-tuned BiomedCLIP [33], ViT-S/16 and ViT-B/16 models trained with standard InfoNCE loss [34], and a supervised 3D ResNet for few-shot sequence classification. Harmonisation was compared against HACA3 [11] and TUMSyn [20], including slice-wise DIST-CLIP/I, DIST-CLIP/T and volumetric DIST-CLIP-2.5D, under the same preprocessing and held-out test partitions. HACA3 and TUMSyn were run using their public implementations with recommended settings. HACA3 is runned with skull and skull stripping is applied during evaluation. All methods received identical registration, skull-stripping, resampling before metric computation.

4.7 Software and reproducibility

Code and pre-trained model weights for MR-CLIP, DIST-CLIP and evaluation scripts are available in the MaRaI repository (https://github.com/myigitavci/MaRaI). Experiments were implemented in PyTorch; software versions, random seeds, full hyperparameter tables, and downstream model architectures are documented in the repository configuration files.

Declarations

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Data-Driven Health (DRIVE-Health; grant EP/Y035216/1) at King's College London, with additional support from deepc GmbH and the Scientific and Technological Research Council of Türkiye (TUBITAK) 2213-A Overseas Graduate Scholarship.

Competing interests

M.Y. Avci's PhD is partially funded by deepc GmbH. The other authors declare no competing interests.

Ethics approval and consent to participate

Approved under UK Health Research Authority generic approval (IRAS ID 349531; REC reference 24/ES/0099).

Data availability

Clinical MRI data from King's College Hospital and Guy's and St Thomas' NHS Foundation Trust are not publicly available owing to governance restrictions. ON-Harmony is available on OpenNeuro (`ds004712`). ADNI and OASIS-3 are available from their respective repositories subject to data-use agreements.

Code availability

Code and pre-trained weights are available at https://github.com/myigitavci/MaRaI.

Author contributions

M.Y.A., P.B., V.F., N.G., P.W., M.Y., S.O. and J.C. designed the study. M.Y.A., P.B., V.F., and J.C. developed methodology and software. M.Y.A., P.B., and N.G. performed experiments and analysis. M.Y.A. and J.C. wrote the manuscript with input from all authors. S.O. and J.C. supervised the project.

Appendix A Supplementary information

A.1 Metadata distribution

Supplementary Fig. S1 summarises DICOM metadata distribution in the dataset.

A.2 Complete cross-contrast harmonisation metrics

Supplementary Fig. S2 reports mean SSIM and PSNR ($\pm$ s.d.) for every available source–target contrast pair in the harmonisation test set, comparing HACA3, TUMSyn, DIST-CLIP/T, DIST-CLIP/I and DIST-CLIP-2.5D. Panel titles indicate the translation direction and the number of test pairs (n). Following common practice in the image synthesis literature, these metrics are computed over the entire image volume, including background regions, to facilitate direct comparison with previously published methods. Hatched bars indicate results that are statistically significantly better than both baseline methods (HACA3 and TUMSyn) ($p < 0.05$).

While most prior studies report SSIM and PSNR over the full image domain, including background voxels, these metrics can be influenced by large homogeneous regions outside the brain. To assess synthesis quality within anatomically relevant structures, we additionally computed SSIM and PSNR exclusively within the brain foreground mask. Supplementary Fig. S3 presents the corresponding foreground-only evaluation for all source–target contrast pairs.

A.3 Volumetric consistency of harmonised outputs

Supplementary Fig. S4 compares axial, sagittal and coronal views for three target contrasts (T2w, FLAIR and T1w) synthesised from a common source. Slice-independent two-dimensional harmonisation shows through-plane discontinuities in orthogonal views, whereas DIST-CLIP-2.5D yields smooth volumetric appearance at the cost of reduced high-frequency texture relative to ground truth.

A.4 Style fusion decoder architecture

Supplementary Fig. S5 summarises the multi-scale style fusion decoder and the internal AST module used for contrast injection.

A.5 Held-out harmonisation test pairs.

Supplementary Table S1 lists the number of source–target contrast pairs in the held-out harmonisation test set for each translation direction (from the DIST-CLIP evaluation protocol).

Table S1 Held-out test source–target pair counts for cross-contrast harmonisation.

Source → target	T1w	T2w	PDw	FLAIR
T1w	–	50	1	26
T2w	50	–	274	366
PDw	1	274	–	41
FLAIR	26	366	41	–

A.6 Skull-stripping ablations.

Main experiments used SynthStrip skull-stripped brain volumes. To quantify sensitivity to this preprocessing choice, we retrained and re-evaluated models on matched data with the skull retained, keeping registration, resampling, train/test splits and evaluation protocols otherwise identical.

We compared three-dimensional MR-CLIP on the held-out test set under skull-stripped versus with-skull (Supplementary Table S2). Retaining the skull slightly improved retrieval scores: volume-to-metadata R@1 increased slightly (62.2% versus 60.2%), along with metadata-to-volume matching (R@1 87.3% versus 79.3%; R@5 96.7% versus 91.6%).

Table S2 Skull-stripping ablation for three-dimensional MR-CLIP cross-modal retrieval on the KCH held-out test set. Recall at ranks 1, 5 and 10 (R@1/5/10, %) for volume-to-metadata (V→M) and metadata-to-volume (M→V) matching. Models were trained and evaluated under the same preprocessing setting in each row.

	V→M			M→V		
Setting	R@1	R@5	R@10	R@1	R@5	R@10
Skull-stripped	60.2	79.0	82.0	79.3	91.6	94.0
With skull	62.2	78.8	82.7	87.3	96.7	97.7

We repeated held-out cross-contrast harmonisation with DIST-CLIP/I under the same pair set (Supplementary Table S3). Skull-stripping markedly improved structural agreement (mean SSIM 0.952 ± 0.020 versus 0.828 ± 0.057 with skull retained), whereas mean PSNR was similar (27.91 ± 3.96 dB versus 27.64 ± 2.55 dB). The larger SSIM gain indicates that non-brain intensity structure in with-skull inputs mainly harms spatial correspondence to the target contrast, not global intensity scale.

Table S3 Skull-stripping ablation on held-out harmonisation (DIST-CLIP/I). Mean ± s.d. over harmonisation pairs present in both the skull-stripped and with-skull.

Setting	SSIM ↑	PSNR (dB) ↑
Skull-stripped	0.952 ± 0.020	27.91 ± 3.96
With skull	0.828 ± 0.057	27.64 ± 2.55

A.7 Grouping strategy of numerical tags.

We investigate how discretising the TE–TR–TI space affects retrieval performance (Supplementary Table S4). We evaluate both in-domain settings (same grid for training and testing) and cross-discretisation generalisation. Overall, coarser groupings consistently improve performance. In particular, a 5×5 TE–TR grid achieves the best results across most metrics, which indicates that reduced label granularity simplifies the alignment task, likely by lowering label noise and increasing intra-class consistency. However, lowering grid size results in loss of fine-grained understanding of the parameter space; therefore 20x20 was chosen in rest of the experiments. We further compare grid-based discretisation with k-means clustering. Although k-means with 20 clusters outperforms the closest grid setting (5×5) for image-to-text and 3D scan-to-text retrieval, it performs worse on text-to-image retrieval. Increasing the number of clusters also leads to slightly degraded performance relative to comparable grid resolutions. These results suggest that k-means is sensitive to the choice of cluster number and tightly coupled to the training distribution, limiting its robustness across acquisition protocols. In contrast, grid-based grouping, while not always optimal, is more stable and likely to generalise better to diverse clinical settings.

Finally, we assess transferability by training MR-CLIP on a fine 20×20 grid and evaluating it on coarser discretisations without retraining. Despite being trained on fine-grained labels, the model maintains strong performance on coarser groupings, demonstrating robustness and flexibility across different settings.

A.8 Metadata quality-control detection performance

Supplementary Fig. S6 reports detection AUC (embedding distance as anomaly score) for numerical timing errors (**a**), incorrect categorical tags (**b**) and missing fields (**c**), evaluated at corruption rates of 5, 10, 15, 25, 50 and 75% and at small, medium and

Table S4 Retrieval performance (%) across different discretisation granularities of TE and TR (TExTR). Top: results when training and evaluating on the same set. Bottom: generalisation performance when transferring from a 20×20 training grid to other discretisations. We report R@1/5/10 for slice-to-metadata, 3D scan-to-metadata, and metadata-to-slice retrieval. Highest values in each column are bolded.

Set (# total classes)	**Slice→Metadata**			**3D Scan→Metadata**			**Metadata→Slice**		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Trained and evaluated on same sets									
40x20 (1770)	62.5	72.1	73.2	73.9	92.8	94.1	88.6	92.4	93.5
20x20 (1415)	66.0	77.3	78.3	78.7	94.2	95.2	90.9	93.6	94.4
20x10 (1017)	69.5	80.4	81.1	82.0	95.4	96.0	94.1	96.1	96.8
10x10 (792)	77.2	84.3	85.2	86.4	96.1	96.6	92.0	96.5	97.6
10x5 (599)	77.3	86.2	86.7	84.3	96.7	97.1	86.6	97.7	98.0
5x5 (488)	69.7	86.9	88.0	89.1	97.0	97.4	**93.7**	**97.8**	**98.1**
k-means(nc=100) (1220)	60.1	70.5	71.6	72.6	93.6	95.2	91.2	93.5	94.5
k-means(nc=50) (826)	71.8	83.0	83.7	83.9	96.1	96.9	93.6	95.6	96.0
k-means(nc=20) (522)	**81.4**	**90.1**	**90.4**	**91.3**	**97.2**	**97.6**	88.7	97.7	97.8
Trained on 20x20 set, evaluated on other sets									
40x20 (1770)	54.1	74.1	76.7	68.0	93.2	94.7	74.1	84.4	88.4
20x20 (1415)	66.0	77.3	78.3	78.7	94.2	95.2	90.9	93.6	94.4
20x10 (1017)	71.1	80.1	80.8	81.7	95.4	96.1	93.2	95.2	95.7
10x10 (792)	76.7	84.7	85.3	85.9	96.1	96.6	93.8	95.7	96.3
10x5 (599)	78.9	86.2	86.6	87.3	96.6	97.1	94.7	96.5	97.0
5x5 (488)	**83.6**	**89.9**	**90.3**	**90.5**	**97.2**	**97.5**	**95.6**	**96.8**	**97.3**

large severity. Performance increased with both corruption rate and severity for all classes; large numerical and missing-tag edits remained highly detectable even at low prevalence, whereas subtle wrong-tag and missing-tag corruptions were hardest at low rates, consistent with the 50% summary in Fig. 6c.

A.9 Out-of-distribution harmonisation on OASIS-3

Supplementary Fig. S7 evaluates zero-shot harmonisation on OASIS-3, which was excluded from DIST-CLIP training; conversely, HACA3 and TUMSyn were trained directly on this cohort. Despite this distribution shift, DIST-CLIP/T and DIST-CLIP/I achieve SSIM and PSNR comparable to, and in some aspects exceeding, the in-distribution baselines across the shown translation directions (FLAIR↔T2w and T1w→T2w). DIST-CLIP/I significantly outperforms HACA3 on both T1w→T2w and FLAIR→T2w tasks ($p < 0.001$ for SSIM and PSNR). On the FLAIR→T2w translation, the zero-shot DIST-CLIP/I achieves performance that is statistically indistinguishable from the in-domain TUMSyn baseline (PSNR $p = 0.99$, SSIM $p = 0.92$).

A.10 Common exemplar metadata descriptions.

Supplementary Table S5 presents the five most frequently occurring metadata descriptions in the clinical dataset. Each entry combines structured DICOM-derived

information, including imaging plane, scanner configuration (manufacturer, model, and field strength), acquisition protocol details (description, sequence, and variant) and key imaging parameters such as echo time (TE), repetition time (TR), inversion time (TI), and flip angle.

Table S5 Top 5 most common metadata descriptions in the clinical dataset.

A brain MRI, ***plane*** axial, ***Scanner*** (***Manufacturer***, ***Model***, ***Field Strength***): (GE, Signa_HDxt, 1.5), ***Acquisition*** (***Description***, ***Sequence***, ***Variant***): (Ax T2 FLAIR, SE_IR, SK), ***Imaging Parameters*** (***Echo Time***, ***Repetition Time***, ***Inversion Time***, ***Flip Angle***): (0.129, 8.002, 2, 90)
A brain MRI, ***plane*** axial, ***Scanner*** (***Manufacturer***, ***Model***, ***Field Strength***): (GE, Signa_HDxt, 1.5), ***Acquisition*** (***Description***, ***Sequence***, ***Variant***): (Ax PD/T2, SE, SK), ***Imaging Parameters*** (***Echo Time***, ***Repetition Time***, ***Inversion Time***, ***Flip Angle***): (0.0203, 5.9, NONE, 90)
A brain MRI, ***plane*** axial, ***Scanner*** (***Manufacturer***, ***Model***, ***Field Strength***): (Siemens, Aera, 1.5), ***Acquisition*** (***Description***, ***Sequence***, ***Variant***): (t2_tse_tra_384_p2, SE, SK_SP_OSP), ***Imaging Parameters*** (***Echo Time***, ***Repetition Time***, ***Inversion Time***, ***Flip Angle***): (0.087, 3.96, NONE, 150)
A brain MRI, ***plane*** axial, ***Scanner*** (***Manufacturer***, ***Model***, ***Field Strength***): (Siemens, Avanto, 1.5), ***Acquisition*** (***Description***, ***Sequence***, ***Variant***): (pd+t2_tse_tra, SE, SK_SP_OSP), ***Imaging Parameters*** (***Echo Time***, ***Repetition Time***, ***Inversion Time***, ***Flip Angle***): (0.014, 3.68, NONE, 150)
A brain MRI, ***plane*** coronal, ***Scanner*** (***Manufacturer***, ***Model***, ***Field Strength***): (GE, SIGNA_HDx, 1.5), ***Acquisition*** (***Description***, ***Sequence***, ***Variant***): (Cor T1, SE, NONE), ***Imaging Parameters*** (***Echo Time***, ***Repetition Time***, ***Inversion Time***, ***Flip Angle***): (0.02, 0.4, NONE, 90)

References

[1] Kushol, R., Parnianpour, P., Wilman, A. H., Kalra, S. & Yang, Y.-H. Effects of MRI scanner manufacturers in classification tasks with deep learning models. *Scientific Reports* **13**, 16791 (2023).

[2] Yan, W. *et al.* MRI manufacturer shift and adaptation: Increasing the generalizability of deep learning segmentation for MR images acquired with different scanners. *Radiology: Artificial Intelligence* **2**, e190195 (2020). URL https://doi.org/10.1148/ryai.2020190195.

[3] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine* **17**, 195 (2019).

[4] Sendra-Balcells, C. *et al.* Domain generalization in deep learning for contrast-enhanced imaging. *Computers in Biology and Medicine* **149**, 106052 (2022).

[5] Posselt, C. *et al.* Simulation of acquisition shifts in T2 weighted fluid-attenuated inversion recovery magnetic resonance images to stress test artificial intelligence segmentation networks. *Journal of Medical Imaging* **11**, 024013 (2024). URL https://doi.org/10.1117/1.JMI.11.2.024013.

[6] Glocker, B., Robinson, R., Castro, D. C., Dou, Q. & Konukoglu, E. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. Eprint arXiv:1910.04597 (2019). arXiv:1910.04597.

[7] Dinsdale, N. K., Jenkinson, M. & Namburete, A. I. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage* **228**, 117689 (2021).

[8] Kamnitsas, K. *et al.* Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *Information Processing in Medical Imaging* 597–609 (2017).

[9] Fortin, J.-P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161**, 149–170 (2017).

[10] Dewey, B. E. *et al.* A disentangled latent space for cross-site MRI harmonization. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 720–729 (2020).

[11] Zuo, L. *et al.* HACA3: A unified approach for multi-site MR image harmonization. *Computerized Medical Imaging and Graphics* **109**, 102285 (2023). URL https://www.sciencedirect.com/science/article/pii/S0895611123001039.

[12] Modanwal, G., Vellal, A., Buda, M. & Mazurowski, M. A. MRI image harmonization using cycle-consistent generative adversarial network. *Medical Imaging*

2020: Computer-Aided Diagnosis **11314**, 1131413 (2020).

[13] Yang, Q., Shomal-Zadeh, F. & Gholipour, A. Harmonization in magnetic resonance imaging: A survey of acquisition, image-level, and feature-level methods. *Medical Image Analysis* **111**, 104066 (2026). URL https://www.sciencedirect.com/science/article/pii/S1361841526001349.

[14] National Electrical Manufacturers Association. Digital imaging and communications in medicine (DICOM) standard. https://www.dicomstandard.org/ (2025). NEMA PS3 / ISO 12052.

[15] Gauriau, R. *et al.* Using DICOM metadata for radiological image series categorization: A feasibility study on large clinical brain MRI datasets. *Journal of Digital Imaging* **33**, 747–762 (2020).

[16] Liang, S. *et al.* Magnetic Resonance Imaging sequence identification using a metadata learning approach. *Frontiers in Neuroinformatics* **15**, 622951 (2021).

[17] Du, R. & Vardhanabhuti, V. 3D-RADNet: Extracting labels from DICOM metadata for training general medical domain deep 3D convolution neural networks. *Proceedings of the Third Conference on Medical Imaging with Deep Learning* **121**, 174–192 (2020).

[18] Cluceru, J., Lupo, J. M., Interian, Y., Bove, R. & Crane, J. C. Improving the automatic classification of brain MRI acquisition contrast with machine learning. *Journal of Digital Imaging* **36**, 289–305 (2023).

[19] Chung, H. *et al.* ContextMRI: Enhancing compressed sensing MRI through metadata conditioning. Eprint arXiv:2501.04284 (2025). arXiv:2501.04284.

[20] Wang, Y. *et al.* Toward general text-guided multimodal brain MRI synthesis for diagnosis and medical image analysis. *Cell Reports Medicine* **6**, 102182 (2025).

[21] Holland, R. *et al.* Metadata-enhanced contrastive learning from retinal optical coherence tomography images. *Medical Image Analysis* **97**, 103296 (2024).

[22] Gueld, M. O. *et al.* Quality of DICOM header information for image categorization. *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation* **4685**, 280–287 (2002).

[23] Avci, M. Y. *et al.* MR-CLIP: Efficient metadata-guided learning of MRI contrast representations. *Machine Learning in Medical Imaging* 85–94 (2026).

[24] Avci, M. Y. *et al.* Metadata-aligned 3D MRI representations for contrast understanding and quality control. *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)* 1–4 (2026).

[25] Khosla, P. *et al.* Supervised contrastive learning. *Advances in neural information processing systems* **33**, 18661–18673 (2020).

[26] Avci, M. Y. *et al.* DIST-CLIP: Arbitrary metadata and image guided MRI harmonization via disentangled anatomy-contrast representations. Eprint arXiv:2501.04284 (2025). arXiv:2512.07674.

[27] Ouyang, J., Adeli, E., Pohl, K. M., Zhao, Q. & Zaharchuk, G. Representation disentanglement for multi-modal brain MRI analysis. *International conference on information processing in medical imaging* 321–333 (2021).

[28] Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE international conference on computer vision* 1501–1510 (2017).

[29] Jack Jr., C. R. *et al.* The alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* **27**, 685–691 (2008).

[30] Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005). URL https://www.sciencedirect.com/science/article/pii/S1053811905001102.

[31] Billot, B. *et al.* SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis* **86**, 102789 (2023). URL https://www.sciencedirect.com/science/article/pii/S1361841523000506.

[32] Warrington, S. *et al.* A multi-site, multi-modal travelling-heads resource for brain MRI harmonisation. *Scientific Data* **12**, 609 (2025).

[33] Zhang, S. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025).

[34] van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Eprint arXiv:1807.03748 (2019). arXiv:1807.03748.

[35] Radford, A. *et al.* Learning transferable visual models from natural language supervision. *International conference on machine learning* 8748–8763 (2021).

[36] Sinha, H. & Raamana, P. R. Solving the pervasive problem of protocol non-compliance in MRI using an open-source tool mrQA. *Neuroinformatics* **22**, 297–315 (2024).

[37] Sadri, A. R. *et al.* MRQy—an open-source tool for quality control of MR imaging data. *Medical physics* **47**, 6029–6038 (2020).

[38] Keaveney, S. *et al.* An open-source repository-based tool for quality control of imaging protocol compliance: demonstration in a multicentre MRI study. *British Journal of Radiology* **98**, 1236–1244 (2025).

[39] Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision* 2223–2232 (2017).

[40] Roca, V., Kuchcinski, G., Pruvo, J.-P., Manouvriez, D. & Lopes, R. IGUANe: A 3D generalizable CycleGAN for multicenter harmonization of brain MR images. *Medical Image Analysis* **99**, 103388 (2025). URL https://www.sciencedirect.com/science/article/pii/S136184152400313X.

[41] LaMontagne, P. J. *et al.* OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer's Disease. *Alzheimer's & Dementia* **14**, P1097–P1097 (2018).

[42] Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. SynthStrip: skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022). URL https://www.sciencedirect.com/science/article/pii/S1053811922005900.

[43] Ilharco, G. *et al.* OpenCLIP. Zenodo software, version 0.1 (2021).

[44] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[45] Gal, R. *et al.* StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* **41**, 1–13 (2022).

[46] Kwon, G. & Ye, J. C. CLIPStyler: Image style transfer with a single text condition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 18062–18071 (2022).

[47] Cardoso, M. J. *et al.* MONAI: An open-source framework for deep learning in healthcare (2022). URL https://arxiv.org/abs/2211.02701. arXiv:2211.02701.

[48] Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* 1125–1134 (2017).

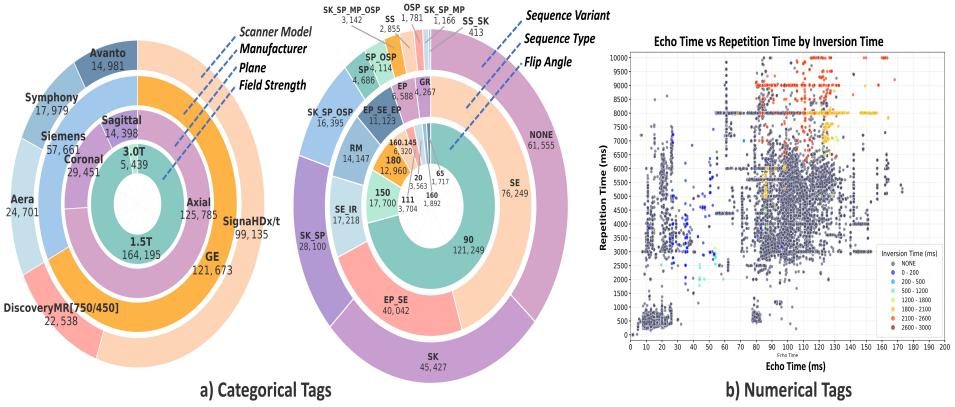


Fig. S1 Metadata distribution in the clinical cohort. a, Categorical tags: nested distributions of imaging plane, field strength, manufacturer, scanner model, flip angle, sequence type and sequence variant. **b,** Numerical tags: echo time (TE) versus repetition time (TR), with inversion time (TI) indicated by colour.

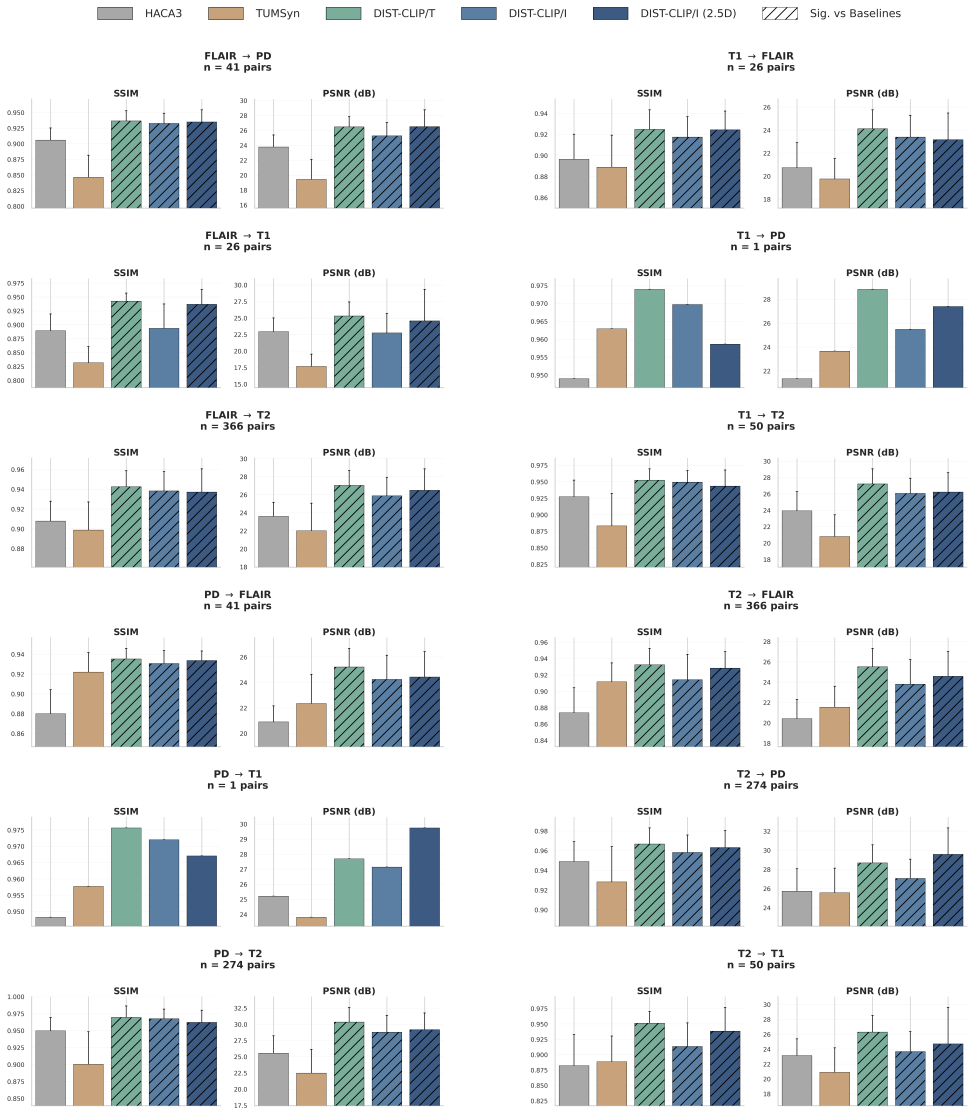


Fig. S2 Complete cross-contrast harmonisation benchmarks. SSIM and PSNR (dB) for all bidirectional translation tasks among T1w, T2w, PDw and FLAIR sequences. Hatched bars denote methods that achieve statistically significant improvements over both HACA3 and TUMSyn ($p < 0.05$).

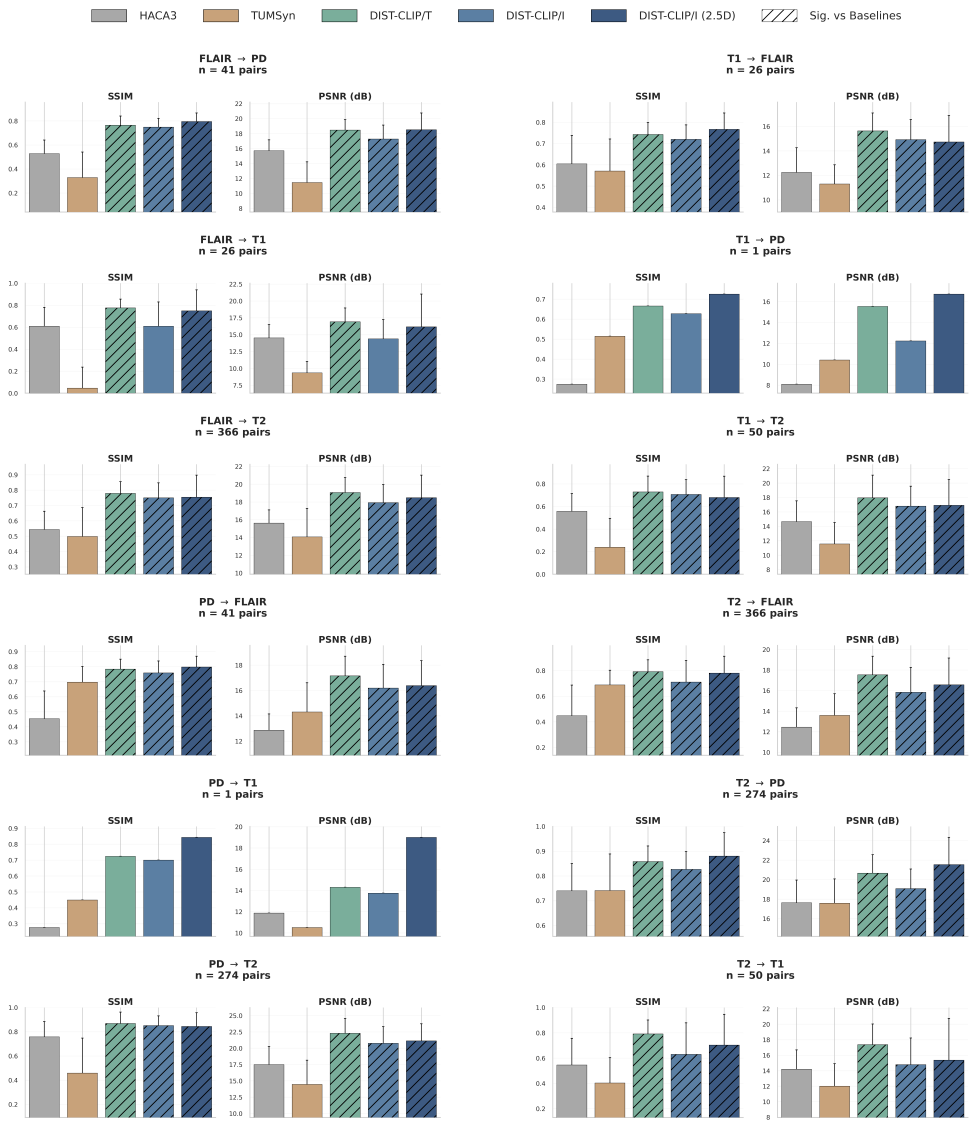


Fig. S3 Foreground-only cross-contrast harmonisation benchmarks. SSIM and PSNR (dB) computed exclusively within the brain foreground mask for all bidirectional translation tasks among T1w, T2w, PDw and FLAIR sequences. Hatched bars denote methods that achieve statistically significant improvements over both HACA3 and TUMSyn ($p < 0.05$).

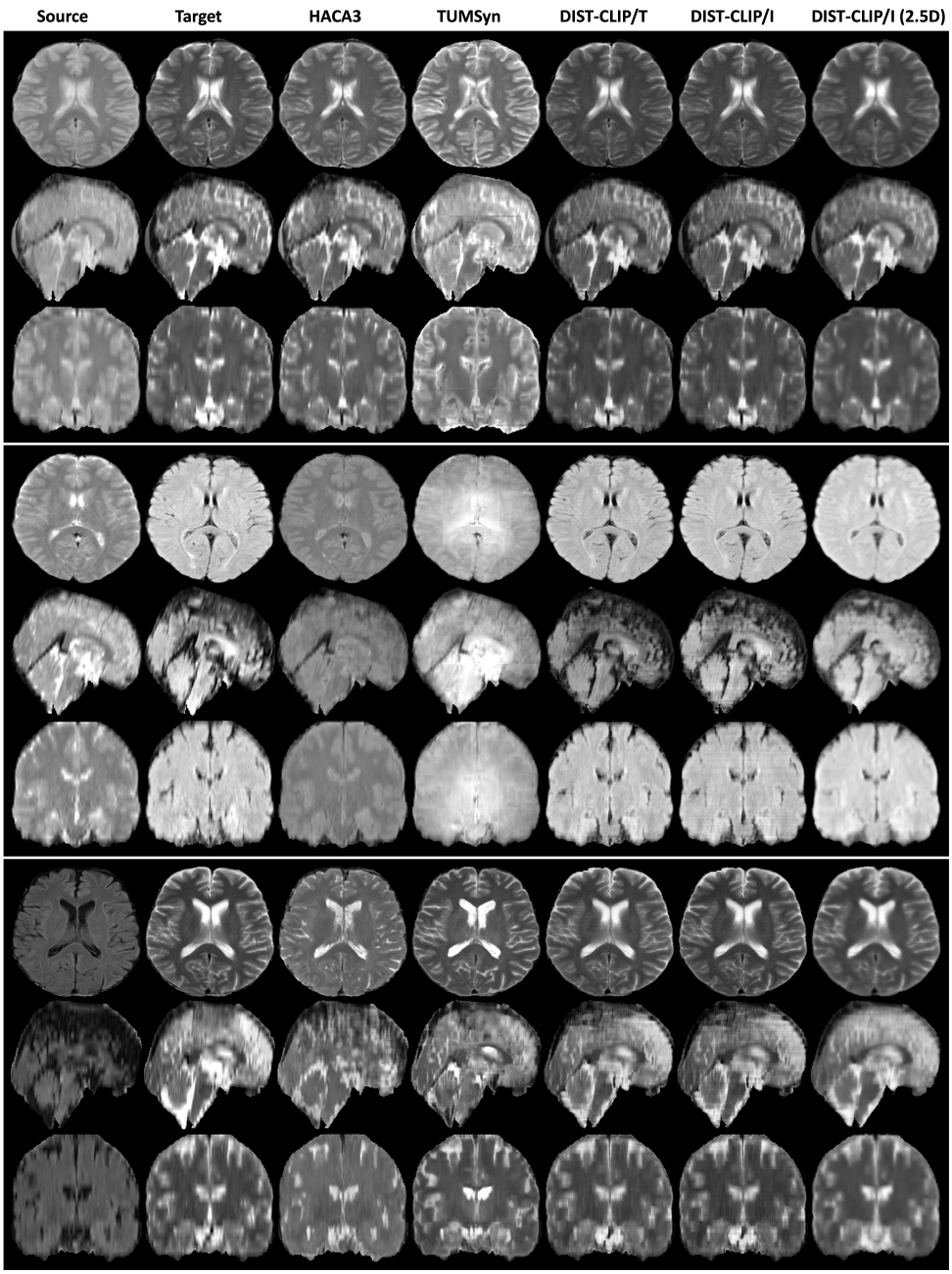


Fig. S4 Volumetric consistency of cross-contrast harmonisation. For each target contrast block, columns show ground truth and harmonised volumes in axial, sagittal and coronal planes for representative baselines and DIST-CLIP-2.5D variant. Stair-step artefacts in orthogonal views indicate slice-wise inconsistency of purely two-dimensional synthesis; DIST-CLIP-2.5D reduces these artefacts while producing smoother intensity profiles.

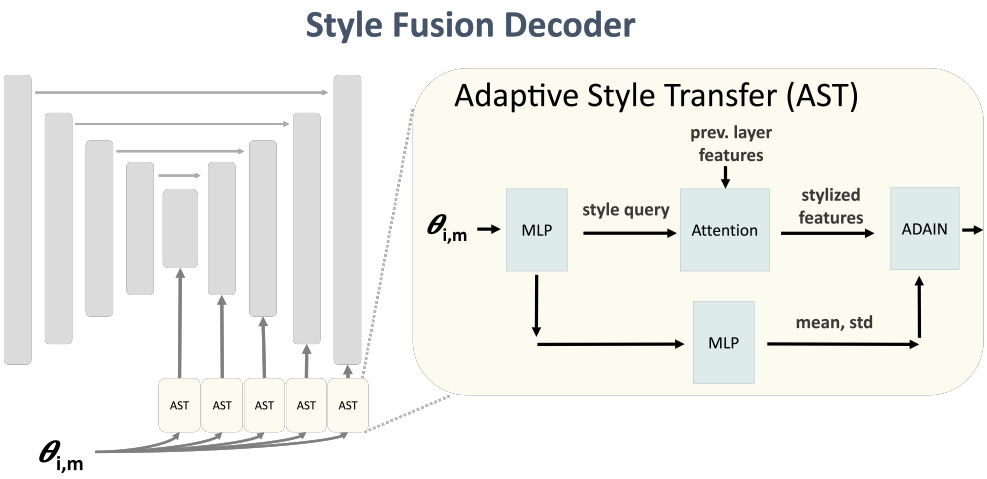


Fig. S5 Style fusion decoder and adaptive style transfer (AST). Left, Style Fusion Decoder pathway: anatomical features from β_{src} are manipulated through successive decoder blocks with encoder skip connections; an AST block at each resolution injects the target contrast embedding θ. **Right,** AST detail: θ is mapped by an MLP to a style query for attention over previous-layer features; a parallel MLP predicts AdaIN mean and standard deviation parameters that modulate the attended features to match target contrast appearance.

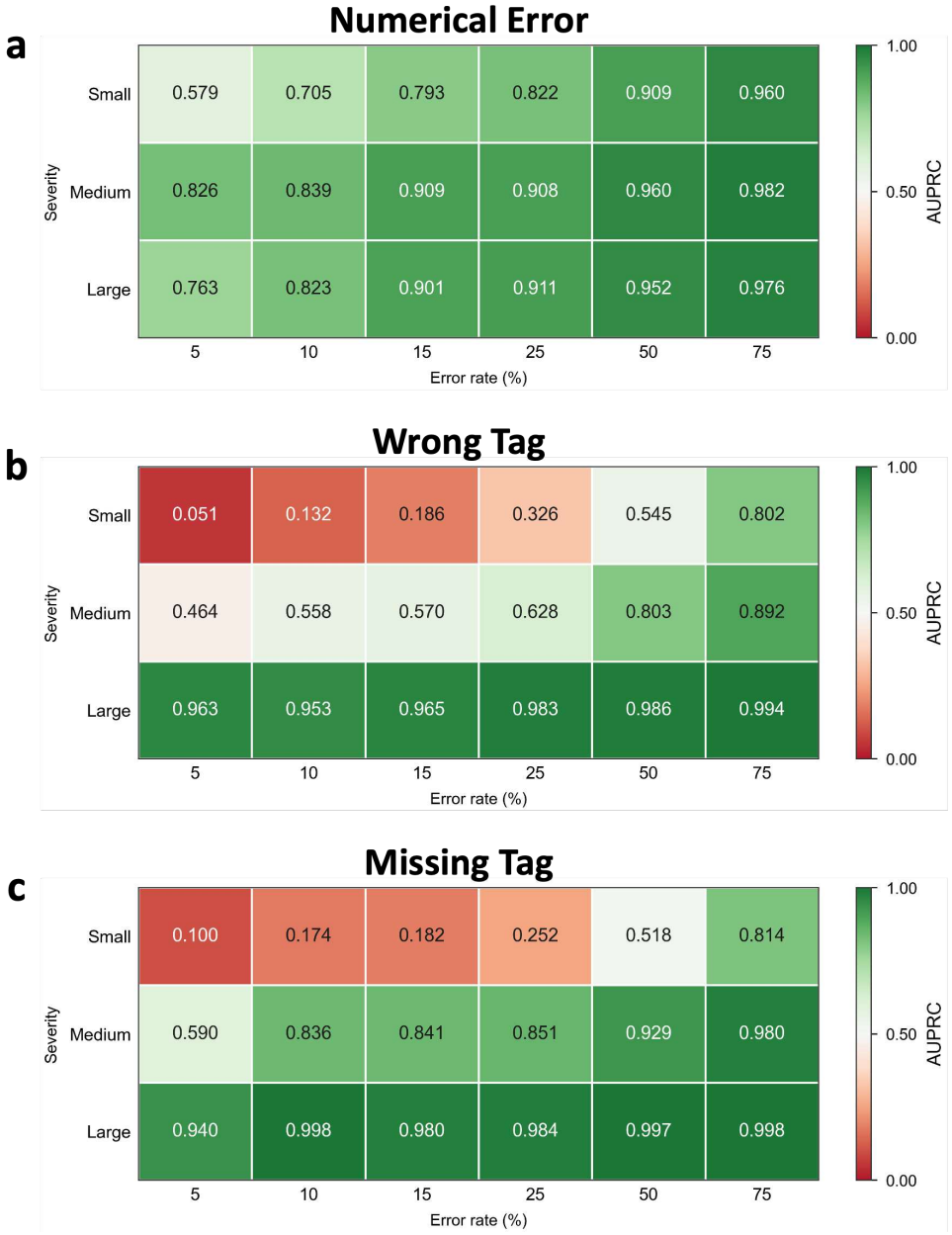


Fig. S6 Metadata corruption detection across corruption rates. Detection AUC using one minus cosine similarity between MR-CLIP image and metadata embeddings as an anomaly score, evaluated on held-out test volumes with synthetic metadata corruptions at rates of 5–75% (columns) and small, medium or large severity (rows). **a,** Numerical errors in echo time, repetition time and inversion time. **b,** Incorrect categorical tags (series description, sequence fields, scanner identifiers). **c,** Missing tags (progressive removal of the same fields).

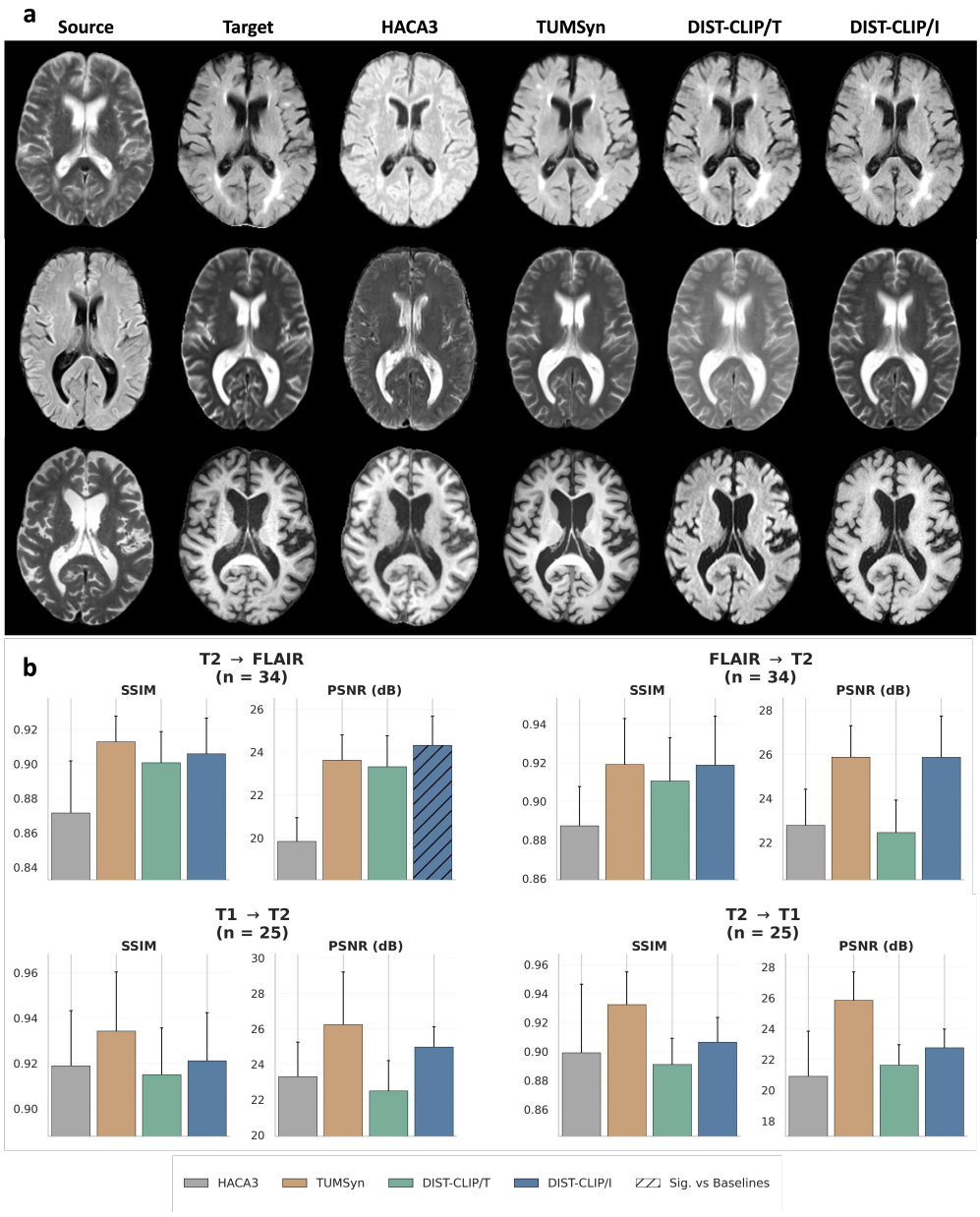


Fig. S7 Zero-shot harmonisation on OASIS-3. a, Representative FLAIR→T2w, T2w→FLAIR and T1w→T2w examples (source, target, HACA3, TUMSyn, DIST-CLIP/T and DIST-CLIP/I). **b,** Mean SSIM and PSNR (±s.d.; n per panel) for the same translation directions. Hatched results indicate statistically significant improvements over the baselines ($p < 0.05$).