Deep Learning Models for Automated Quality Assessment and Artifact Reduction in Ultra-Low-Field Pediatric Brain MRI

Rameez Ur Rehman Siddiqui[1,2⋆] and Samiya Jabbar[3]

[1] Center of Excellence for Trauma and Emergencies, The Aga Khan University, Karachi, Sindh, Pakistan

[2] UHasselt, Transportation Research Institute (IMOB),Hasselt, Belgium.

[3] Department of Computer Science & Information Technology, NED University of Engineering and Technology, Pakistan

rameez.siddiqui2@gmail.com

Abstract. Quality control for magnetic resonance imaging (MRI) usually focuses on one task: catching bad scans. In low-resource pediatric settings, however, a flagged scan often cannot simply be re-acquired, so being able to *improve* a marginal scan matters as much as being able to *detect* what is wrong with it. We address both halves of this problem with two independently developed models, sharing only a common cohort of ultra-low-field (0.064 T) Hyperfine SWOOP pediatric brain MRI, a portable, sedation-free scanner increasingly used where conventional high-field MRI is unavailable. Our first model screens each scan for seven common artifact types (Noise, Zipper, Positioning, Banding, Motion, Contrast, Distortion) on a three-level severity scale (none, mild, severe). Because severity is ordered rather than merely categorical, we use a CORN (Conditional Ordinal Regression for Neural networks) ordinal loss applied to a multi-head 2.5D ResNet18 classifier, trained across five patient-level splits with a post-hoc calibration step that corrects the decision thresholds. Our second model attempts to *fix*, rather than only flag, artifacts – noise, motion blur, zipper-like ghosting, banding, and contrast loss – using a residual U-Net trained to restore artificially degraded versions of real scans, since no genuinely paired clean/degraded acquisitions exist for the same patient. On the official challenge test set the detection model reaches an accuracy of 0.837 and a weighted F1 of 0.808 (micro F1 0.837, macro F1 0.576; patient-level cross-validation gave an accuracy of 0.833 and a weighted F1 of 0.832, with a composite score of 0.836 after calibration. Weighted metrics transfer closely between the two, while macro metrics do not, showing where the ceiling on rare, clinically important cases currently sits. The enhancement model reaches PSNR 33.9 dB, SSIM 0.895 and LPIPS 0.046 on an internal validation cohort under fixed severity-1 noise and motion degradation, and a Fréchet Inception Distance of 124.4 and Fréchet Radiomics Distance of 7.5 under official evaluation on a separate external cohort of real scans. The two models are trained and evaluated separately; we outline, but do not evaluate, how they could be composed into an acquire–assess–enhance–reassess quality-control loop for point-of-care low-field pediatric imaging.

⋆ Corresponding author

Keywords: Low-field MRI · Pediatric neuroimaging · Deep learning · Computer vision · Image quality assessment · Artifact reduction · Hyperfine

1 Introduction

Accurately characterizing structural changes in the developing brain is critical for understanding healthy neurodevelopment and for early identification of neurodevelopmental disorders. This makes reliable pediatric MRI quality control especially important, yet deep-learning quality-control tools built for adult, high-field MRI generally do not transfer well to pediatric scans: children's brains have poorer gray/white matter contrast and change anatomically much faster than adult brains [1]. Portable ultra-low-field (ULF, 0.064 T) scanners such as the Hyperfine SWOOP were designed to close this gap: despite lower image quality than 1.5 T/3 T systems, they are inexpensive, portable, and critically for children do not require sedation.

Quality control for such scans naturally splits into two complementary tasks: *detecting* artifacts, so a scan can be flagged, re-acquired, or routed to a specialist; and *reducing* artifacts where possible, so a marginal scan does not have to be discarded. Most prior work addresses only one of these tasks in isolation. The LISA 2026 Challenge poses them as two separate tasks 1a (artifact detection) and 1b (artifact reduction) over a shared, recently released dataset of ULF pediatric brain MRI collected across multiple low-resource clinical sites [14]. We entered both, developing an independent model for each under its own task protocol; the two share only the imaging cohort described in Sect. 2.1. The two tasks are presented together because they address complementary halves of one quality-control problem on a shared cohort, not because the models are jointly trained or evaluated end-to-end.

Prior entries to this challenge have approached quality assessment in several ways: auxiliary prediction of the acquisition plane alongside severity scoring [16], synthetic artifact generation to enlarge under-represented severity classes [17], multi-label architectures scoring all categories jointly [18], gradient-based loss reweighting with class-balanced batch construction [19], and view-conditional slice-level models with per-slice probability aggregation [20]. Two features distinguish our approach. First, severity is modelled as an ordinal quantity throughout: prior entries have treated the three levels as nominal classes, and where ordinal losses have been examined they were reported as inconclusive [19] or identified as future work [20]. Second, we address artifact reduction as well as detection, and reuse the synthetic-degradation idea [17] for restoration supervision rather than for augmentation. We make no claim of architectural novelty in either model; the contribution is the adaptation of established components to the constraints of this cohort small sample, severe class imbalance, ordered labels, no paired restoration targets and the reporting of what that adaptation does and does not achieve.

This paper contributes two models built on that shared imaging cohort: a multi-head 2.5D ResNet18 classifier using a CORN ordinal loss [15] for detection (Sects. 2.2 – 2.7), and a residual U-Net trained with a self-supervised degradation-restoration strategy for reduction (Sects. 2.8 – 2.11), since no paired clean/degraded scans of the same patient exist. We evaluate detection by patient-level cross-validation and enhancement on an internal validation split plus an external test cohort, and close by discussing how the

two models could in principle be composed into an acquire–assess–enhance–reassess quality-control loop a design proposal arising from having addressed both tasks, not an implemented or evaluated system.

2 Methods

We first describe the shared imaging cohort (Sect. 2.1) that both models draw on, then present the two independently developed models in turn: artifact detection (Task 1a, Sects. 2.2–2.7), followed by artifact reduction (Task 1b, Sects. 2.8–2.11). The two were trained separately, with no shared parameters, no joint objective, and no data passed between them. An overview of both task pipelines is shown in Fig. 1.

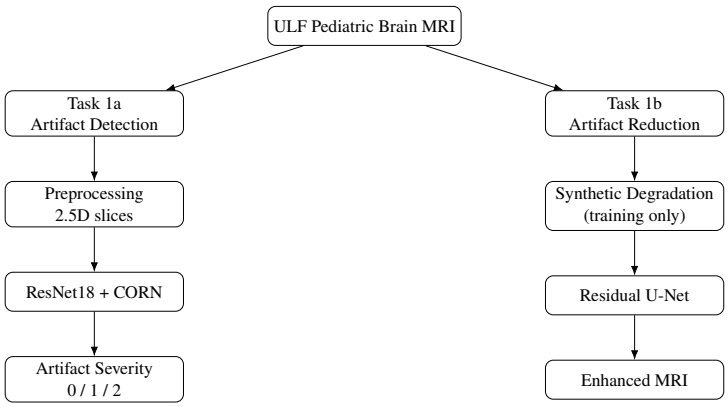

Fig. 1. Overview of the two independently developed pipelines. Task 1a performs artifact detection and severity assessment, while Task 1b performs artifact reduction.

2.1 Dataset

The dataset, provided by the challenge organizers, comprises 531 labeled 0.064 T T2-weighted Hyperfine SWOOP images (`.nii.gz`) from 243 unique patients, orthogonally acquired along up to three views per patient (176 axial, 175 sagittal, 180 coronal). Not every patient has all three views: about 41% have only one, with the rest having two or three. Acquisition used a spin-echo sequence (TR 1.5 s, TE 5 ms, TI 400 ms) at three clinical sites: Kawempe National Referral Hospital, Makerere University (Kampala, Uganda); CUBIC, University of Cape Town (South Africa); and the Advanced Baby Imaging Lab, Rhode Island Hospital / Warren Alpert Medical School, Brown University (Providence, RI, USA). Each image is labeled across the seven artifact categories on an ordinal 0–2 severity scale (0: no visible artifact; 1: mild impact on tissue distinguishability; 2: severe impact) by an expert medical image evaluator. Table 1 shows

the per-artifact class distribution, which is heavily skewed toward severity 0 in every category, most severely for Banding (only 21/531 images at severity 1 or 2).

For the enhancement task, the same images are partitioned by their original Noise/Motion severity labels into four groups – Clean (301), Noise-only (62), Motion-only (137), and Noise+Motion (31) – which drive the synthetic-degradation strategy of Sect. 2.8. An additional 114-image external test cohort (unlabeled) is shared across both tasks and held out separately from the training images described above.

Table 1. Per-artifact severity class distribution (531 labeled images, 243 patients).

Artifact	Severity 0	Severity 1	Severity 2
Noise	438	42	51
Zipper	353	142	36
Positioning	462	45	24
Banding	510	9	12
Motion	363	93	75
Contrast	335	165	31
Distortion	388	95	48

2.2 Quality-Assessment Preprocessing

Rather than processing a full 3D volume, we represent each scan with a lightweight 2.5D input, reusing an ImageNet-pretrained 2D backbone (Sect. 2.3) despite the modest dataset size: nine evenly-spaced slices along the acquisition-view axis (axial/coronal/sagittal), each repeated to three channels and normalized with standard ImageNet statistics. Volumes are intensity-normalized by clipping to the 1st/99th percentile of non-zero voxels and min–max scaling, then resampled to $112 \times 112 \times 40$ before slice extraction. During training only, we apply view-aware, label-safe augmentation: mild gamma jitter on the raw volume, plus a horizontal flip applied only to axial/coronal volumes (approximately bilaterally symmetric, with magnitude- rather than direction-based severity labels) and never to sagittal volumes. Rotation, color-jitter, and perspective transforms are excluded because they can fabricate or erase the exact Positioning/Contrast/Distortion signal being classified.

2.3 Quality-Assessment Model Architecture

Our classifier uses a ResNet18 backbone pretrained on ImageNet, shared across all nine slices of a volume: each slice is encoded independently to a 512-dimensional feature vector, and the nine per-slice features are mean-pooled into a single volume-level representation – hence "2.5D", approximating a view of the whole volume without true 3D convolutions. The pooled feature is concatenated with a 3-dimensional acquisition-view one-hot indicator and passed through a shared fully-connected layer ($512+3 \rightarrow 256$) with ReLU and dropout (0.3).

Each of the seven artifact heads is built to respect the fact that severity is ordered, not just categorical: mistaking a severity-2 scan for severity-0 is a worse error than mistaking it for severity-1. We use a CORN (Conditional Ordinal Regression for Neural networks) parameterization [15]: each head predicts *two* threshold probabilities, $P(\text{severity} > 0)$ and $P(\text{severity} > 1 \mid \text{severity} \geq 1)$, instead of three independent class logits. Decoding takes the cumulative product of the two sigmoid outputs and counts how many exceed 0.5, which is rank-consistent by construction – unlike a nominal 3-way softmax, which has no mechanism preventing invalid orderings (e.g. "severity > 1 but not severity > 0").

2.4 Quality-Assessment Training Procedure

To get a reliable estimate of performance from this small a cohort, we use patient-level 5-fold cross-validation, stratified with a multilabel-stratified fold split [5] on per-artifact presence. Splitting by patient, not by file, matters: a file-level split can let two views of the same patient land in different folds, leaking information and inflating the apparent score (Sect. 4). Each fold is trained with a differential-learning-rate Adam optimizer (backbone 1×10^{-4}, head 1×10^{-3}, weight decay 1×10^{-4}) under cosine-annealing decay [3] ($T_{\max} = 40$ epochs), for up to 50 epochs with early stopping (patience 15) on the validation composite score (Sect. 2.7). The loss is the CORN ordinal loss of Sect. 2.3: each threshold task is a binary cross-entropy with severity-capped inverse-frequency weights (caps 1.5/1.5/3.0 across the three severity levels), computed per task rather than per nominal class, with the second (severity>1) task trained only on samples with severity ≥ 1. To further counter how scarce severe cases are (Table 1), training also oversamples severity-2 images via a weighted random sampler, capped at 5×. Gradients are clipped to a maximum norm of 5.0; batch size is 8.

2.5 Quality-Assessment Decision Calibration

Even with CORN's ordinal parameterization and severity-capped task weighting, the trained model still tends to under-predict severity 1/2, simply because severity 0 remains the majority class in every artifact. We correct for this after training rather than during it. After the 5-fold training, we pool the out-of-fold threshold probabilities – each patient scored only by the fold that held them out – and fit a per-artifact, per-threshold additive offset (applied in logit space before the sigmoid/cumulative-product decode) via coordinate-ascent search over a coarse grid. Offsets are chosen to directly maximize the *composite evaluation score* (Sect. 2.7) on the pooled out-of-fold predictions: in earlier development, calibrating toward macro-F1 instead measurably traded away the majority-class weighted accuracy the composite rewards. Because the official evaluation gives macro metrics substantially more weight than our composite does, this was a trade-off rather than a clear improvement, and Sect. 4 revisits it against the official results. To avoid overfitting the offsets to one pooled sample, each category's offset is fit by leave-one-fold-out cross-validation and the final offset is the per-dimension median across the five fits. Note that the offsets are nonetheless fit on the same pooled out-of-fold predictions used to report cross-validation performance; leave-one-fold-out fitting reduces but does not eliminate this. Sect. 3.1 reports the official evaluation as an independent check. This

calibration step improves the pooled out-of-fold composite score (numbers in Sect. 3) at essentially no cost to the metrics that dominate it.

2.6 Quality-Assessment Inference and Ensembling

For a new scan we generate predictions by ensembling all five fold models. Test-time augmentation mirrors the view-aware training augmentation of Sect. 2.2: axial/coronal volumes are evaluated both unflipped and horizontally flipped (predictions un-flipped back before averaging), while sagittal volumes are evaluated only unflipped, since flipping them was never part of their training distribution. Threshold probabilities are averaged across all (model, augmentation) combinations in logit space; the per-artifact, per-threshold decision-calibration offsets (Sect. 2.5) are added before the cumulative-product decode, and the resulting ordinal class is taken as the final severity prediction per artifact, per patient.

2.7 Quality-Assessment Evaluation Protocol

We report five metrics – accuracy, F1, F2, precision, and recall – on pooled out-of-fold predictions across all seven categories, so every prediction is scored only by the fold that held its patient out. We report micro-, macro-, and weighted-averaged variants of F1/F2/precision/recall: since severity 0 is the majority class everywhere (Table 1), weighted/micro metrics summarize overall accuracy, while macro metrics weight all three severity levels equally, better reflecting our ability to catch the rare but clinically important severity-1/2 cases. We summarize these into a single composite score, defined as the mean of weighted accuracy, F1, F2, precision, and recall; we additionally report mean severe-case (severity-2) recall per artifact as a clinical-safety diagnostic this composite score does not itself capture. Note that for single-label classification, micro-averaged precision, recall, F1, and F2 equal accuracy. The composite score is our own internal summary, not the challenge's ranking procedure; the official evaluation reports fourteen metrics – micro-, macro-, and weighted-averaged F1, F2, precision, and recall, plus accuracy – and therefore weights macro performance considerably more heavily. We report the composite because it is what our calibration step targets (Sect. 2.5), and report the full official metric set alongside it in Sect. 3.1.

2.8 Enhancement: Self-Supervised Training Strategy

Training a model to remove artifacts normally calls for paired clean/degraded scans of the same patient, which this dataset lacks. We work around this with a self-supervised degradation-restoration strategy: every training image is synthetically degraded, regardless of its real quality label, and the network is trained to reconstruct the pre-degradation image under the combined loss of Sect. 2.9. Classical denoising methods such as non-local means [4] operate without training data but assume a noise model and do not address motion, banding, or contrast degradation; we instead learn a single restoration function covering five artifact categories. We synthesize five of the seven QA artifact categories (Sect. 2.1): Rician noise [12]; directional motion blur; a zipper-like ghosting

artifact via attenuation of randomly selected k-space lines before the inverse transform; a banding artifact via multiplicative low-frequency sinusoidal modulation along one spatial axis; and a contrast artifact via gamma compression. Distortion and Positioning are not synthesized: geometric dewarping is ill-posed to learn from synthetic pairs alone, and Positioning reflects anatomical framing rather than a pixel-level degradation. For each synthesized category, the severity range applied is conditioned on the image's own *real* label for that category: severity-0 images draw synthetic severity uniformly from $\{0, 1, 2\}$, while severity-1/2 images draw from $\{1, 2\}$, reinforcing the degradation type already present rather than suppressing it. This teaches the network a general artifact-removal function that is then applied, without further synthetic degradation, to real scans at inference – a methodological approximation rather than direct supervision against true clean references (Sect. 4).

2.9 Enhancement Model Architecture

We use a residual U-Net (ResUNet), based on the U-Net encoder-decoder architecture [6]: four encoder stages (64/128/256/512 channels), each a residual block of two 3×3 convolutions with instance normalization and GELU activation, max-pooling between stages, a symmetric decoder with transposed-convolution upsampling and skip connections, and a residual output head, $\sigma(x + f(x))$ – the network predicts an additive correction rather than the image outright, an easier target to learn than full reconstruction. The model operates on single-channel 2D slices and has 32.4M trainable parameters. During training it sees one mid-slice per volume (the pipeline's earlier, independently-validated best-performing configuration; Sect. 4); at inference it is applied to every slice of a volume, as described in Sect. 2.10. Because the metrics we ultimately care about (Sect. 2.11) are mostly perceptual rather than purely pixel-wise, the training loss reflects that mix: $L = 0.20\, L_{\mathrm{MSE}} + 0.25\, L_{\mathrm{SSIM}} + 0.40\, L_{\mathrm{LPIPS}} + 0.15\, L_{\mathrm{FFT}}$, where $L_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}$ [9] uses an 11×11 pooling window, L_{LPIPS} is the AlexNet-backbone LPIPS perceptual distance [7], and L_{FFT} is an L_1 loss on the 2D Fourier magnitude that discourages the network from over-smoothing away real high-frequency structure.

2.10 Enhancement Training Procedure

We use a patient-level 85%/15% train/validation split (206/37 patients, 448/83 images) and train with AdamW [2] (learning rate 1×10^{-4}, weight decay 5×10^{-5}) under cosine-annealing decay [3] ($T_{\max} = 80$ epochs, $\eta_{\min} = 10^{-6}$), for up to 80 epochs with early stopping (patience 12) on validation combined-loss. Gradients are clipped to a maximum norm of 1.0; batch size is 6; input images are resized to 256×256. We fix `cudnn.deterministic=True` and a seeded DataLoader throughout, after finding that without them, identical runs converged to different checkpoints with materially different metrics (Sect. 4). At inference the model is applied to a full volume slice-by-slice. Each slice is intensity normalized independently by min–max scaling, resized to 256×256, and passed through the network three times – unaugmented, horizontally flipped, and vertically flipped, each un-flipped before averaging – since flip-based TTA carries no label-direction ambiguity for a pixel-wise restoration target, unlike detection (Sect. 2.2). The averaged output is resized back to the slice's native in-plane dimensions and rescaled

to that slice's original intensity range. Enhanced slices are reassembled into a volume, the header's scale/intercept is inverted, and the result is cast to the input data type and written out with the original affine and header, so the enhanced scan is geometrically identical to the input. Slices are taken along a single fixed volume axis, and no constraint couples adjacent slices: each is normalized and restored independently, so through-plane consistency is not enforced by the procedure (Sect. 4).

2.11 Enhancement Evaluation Protocol

We report two complementary evaluations. First, an internal validation benchmark on 83 previously-unseen images: each is degraded with a fixed, moderate synthetic severity (noise severity 1, motion severity 1) for comparability across checkpoints, then scored against its own pre-degradation version with PSNR, SSIM, LPIPS, and the no-reference metrics BRISQUE [10] and CLIP-IQA [11], with images intensity-normalized to $[0, 1]$. These 83 images form the validation split of Sect. 2.10 and were used for early stopping and checkpoint selection, so this benchmark is a validation estimate rather than an untouched test result. Second, evaluation on the external test cohort of real scans (no synthetic degradation), scored by the challenge organizers' official evaluation pipeline rather than computed locally. This reports Fréchet Inception Distance (FID) [8] and Fréchet Radiomics Distance (FRD) [13] at the distribution level; PSNR (native intensity scale, not comparable to the $[0, 1]$-normalized internal-benchmark PSNR) and LPIPS against a shared artifact-free reference bank, which is unpaired and therefore not a per-image fidelity measurement; and the no-reference metrics BRISQUE and CLIP-IQA, reported both for the enhanced output and as a per-case difference relative to the corresponding unenhanced input. FID and FRD are population-level metrics requiring hundreds of images to estimate reliably, so we report them only from the external-cohort evaluation.

3 Results

3.1 Quality-Assessment Results

Unless stated otherwise, numbers in this subsection are computed on pooled out-of-fold predictions from the patient-level 5-fold cross-validation described in Sect. 2.7, so no patient is ever scored by a model that saw their images during training. Table 2 summarizes performance per fold before decision calibration, and Table 3 breaks the results down per artifact, pooled across folds. Calibration improves the composite score from 0.8322 to 0.8359 at essentially no cost to the weighted/micro metrics that dominate it (Sect. 2.5). Table 4 compares these cross-validation estimates against the official challenge evaluation on the held-out test set. Weighted and micro metrics transfer closely — accuracy and recall are marginally higher officially (0.837 vs. 0.8332), weighted F1 slightly lower (0.808 vs. 0.8315) — indicating that patient-level cross-validation gave a well-calibrated estimate of held-out performance on these metrics. Macro metrics do not transfer as well: macro F1 falls from 0.6953 to 0.576 and macro recall from 0.6829 to 0.613. Recomputing our composite on the official metrics gives 0.827, against 0.8359 out-of-fold. Mean severe-case (severity-2) recall across the seven artifacts is 0.571, with per-artifact values in Table 3; see Sect. 4 for interpretation.

Table 2. Per-fold cross-validation summary

Fold	Val. images	Composite score
1	106	0.8349
2	108	0.8280
3	104	0.8258
4	112	0.8401
5	101	0.8321
Mean		**0.8322**

Table 3. Per-artifact detection performance on pooled out-of-fold predictions. Weighted recall equals accuracy and is omitted; n(Sev-2) is the number of severity-2 cases. The final row is pooled across artifacts

Artifact	Accuracy	$\text{Precision}_{\text{weighted}}$	$\text{F1}_{\text{weighted}}$	n(Sev-2)	Sev-2 Recall
Noise	0.9171	0.9093	0.9127	51	0.9020
Zipper	0.8267	0.8236	0.8247	36	0.6111
Positioning	0.8663	0.8625	0.8638	24	0.3333
Banding	0.9661	0.9640	0.9647	12	0.6667
Motion	0.7514	0.7467	0.7489	75	0.7867
Contrast	0.7646	0.7585	0.7609	31	0.3871
Distortion	0.7401	0.7478	0.7418	48	0.3125
Pooled	**0.8332**	**0.8304**	**0.8315**	**277**	**0.5713**

Table 4. Cross-validation estimate vs. official challenge evaluation. Official values are from the calibrated model.

Metric	Micro		Macro		Weighted	
	OOF	Official	OOF	Official	OOF	Official
Accuracy	0.8332	0.837	0.6829	0.613	0.8332	0.837
F1	0.8332	0.837	0.6953	0.576	0.8315	0.808
F2	0.8332	0.837	0.6877	0.588	0.8325	0.822
Precision	0.8332	0.837	0.7096	0.635	0.8304	0.833
Recall	0.8332	0.837	0.6829	0.613	0.8332	0.837

3.2 Enhancement Results

As with detection, we report enhancement quality at two levels: a controlled internal benchmark where the degradation is known, and a harder external cohort of real scans with no ground truth. Table 5 reports the internal validation benchmark (83 validation images, fixed moderate synthetic degradation, never seen in training) at the selected checkpoint (validation-loss minimum, epoch 66 of 78; training early-stopped at 78 of 80 epochs). Table 6 reports the official challenge evaluation on the external test cohort for our enhancement model (same architecture and training recipe as Sect. 2.9–2.10). The official pipeline additionally reports per-case differences relative to the unenhanced inputs for the two no-reference metrics. Since BRISQUE and CLIP-IQA are oppositely signed by construction lower is better for the former, higher for the latter – the two differences carry opposite implications for enhancement quality, and each spans both positive and negative values across the cohort. As noted in Sect. 2.11, PSNR in the two tables is computed on different intensity scales and should not be compared directly across them.Representative enhancement examples are shown in Fig. 2.

Table 5. Internal validation benchmark

Metric	Noise + Motion, sev 1			All five, sev 1
	Mean	Min	Max	Mean
PSNR (dB)	33.888	29.147	36.376	33.575
SSIM	0.895	0.755	0.939	0.891
LPIPS	0.046	0.022	0.120	0.048
BRISQUE	54.65	38.89	81.76	55.37
CLIP-IQA	0.482	0.225	0.719	0.478

Table 6. External test-cohort metrics, official challenge evaluation. Δ rows are per-case differences vs. the unenhanced input; the sign convention is undocumented.

Metric	Mean	Min	Max
FID	124.429	95.253	231.224
FRD		7.493	
PSNR (dB)	12.072	11.226	12.908
LPIPS	0.464	0.425	0.561
BRISQUE (enhanced)	65.870	47.318	80.521
BRISQUE (Δ, per case)	5.868	−12.684	20.519
CLIP-IQA (enhanced)	0.446	0.256	0.602
CLIP-IQA (Δ, per case)	0.068	−0.122	0.225

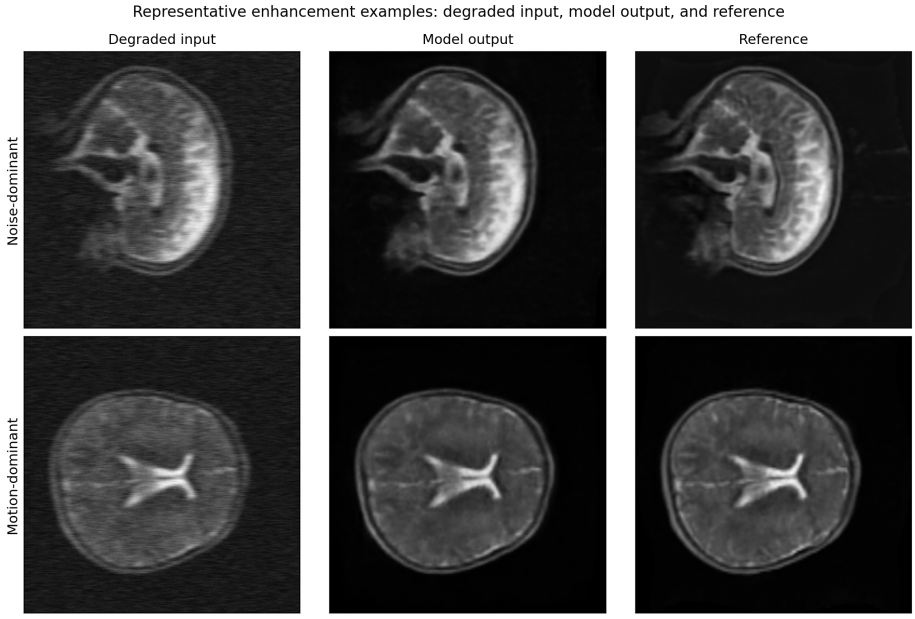

Fig. 2. Representative enhancement examples: input, output, and reference.

4 Discussion

Quality assessment. Our detection ensemble reaches an accuracy of 0.833 and a weighted F1 of 0.832, with a composite score of 0.836 after calibration, from only 531 images across 243 patients, showing that a pretrained 2.5D backbone with an ordinal-aware loss can partly offset a small, imbalanced dataset. The more telling number is the gap between weighted metrics (∼0.83) and macro metrics (∼0.69, Table 4): the model is still far better at recognizing "no artifact" than at telling mild from severe cases, exactly where QA decisions matter clinically, and severe-case recall (0.571 mean, as low as 0.313 for Distortion) makes the same point directly. This gap is narrowest for Banding, but with only 12 severity-2 cases that estimate is too imprecise to interpret; the gap is widest, and best supported, for Motion, Contrast, and Distortion (Table 3). Decision calibration (Sect. 2.5) closes part of the remaining gap (0.8322→0.8359) by correcting a decision-boundary bias rather than a representation problem. It was retargeted at the composite score after an earlier version calibrating toward macro-F1 traded away majority-class accuracy but the official evaluation puts that choice in a different light. Optimizing a weighted-only summary therefore improved the number we were tracking while leaving the half of the official evaluation where our model is weakest unaddressed. The residual gap points to class imbalance and the limits of a 2.5D, slice-pooled representation, rather than the ordinal loss itself, as the more promising targets for further gains.

Enhancement. Two practical lessons stood out. First, metrics can disagree: a multi-slice variant achieved a better FID but a consistently worse CLIP-IQA (0.37–0.38 vs. 0.426) — closer to the reference distribution is not the same as better-looking — so we now check all five metrics together before trusting a new configuration. The official per-case differences make the same point on real scans. Because BRISQUE and CLIP-IQA are oppositely signed — lower is better for the former, higher for the latter — a common-signed pair of differences (5.868 and 0.068) carries opposite implications whichever way the pipeline subtracts; and both span positive and negative values across the cohort (BRISQUE −12.684 to 20.519, CLIP-IQA −0.122 to 0.225), so neither describes a uniform effect (Table 6) and the means conceal that.

Second, reproducibility has to be engineered in: without fixed cuDNN and DataLoader seeds, two runs of identical code converged to different checkpoints with meaningfully different scores (FID 133 vs. 126), so determinism is now fixed in every reported run (Sect. 2.10). Widening the synthetic-degradation repertoire from two artifact categories to five (Sect. 2.8), together with flip-based TTA, improved most external-cohort metrics (FID 133.4→124.4, CLIP-IQA 0.426→0.446, FRD 8.4→7.5) at the cost of BRISQUE and a flat PSNR, consistent with the same lesson that no single metric tells the whole story. Restoration quality is also comparable under simultaneous corruption by all five synthesized categories (Table 5), indicating the training objective generalizes beyond the two categories Task 1b targets.

Several limitations bound what the Task 1b evaluation can support. The model's training target is a pre-degradation proxy rather than a verified clean reference, since no paired clean/degraded scans exist. The internal benchmark uses synthetic degradation on a split that also determined checkpoint selection, so transfer to real artifacts is not established by it, and at 83 images it is too small to estimate population-level metrics (FID, FRD) reliably on its own. The official evaluation reports per-case differences relative to the unenhanced inputs, but does not document the sign convention, so we report their magnitude and spread without inferring a direction of change; establishing whether enhancement improves real scans on these metrics requires an input baseline computed under a known convention. And because restoration is applied slice-by-slice with per-slice intensity normalization, volumetric consistency along the through-plane direction is neither enforced nor measured; for anisotropic ULF acquisitions this is a plausible source of inter-slice variation that a 3D or 2.5D formulation would avoid.

Composing the two models: The two models were developed, trained, and evaluated entirely independently, one per task; no result reported above reflects the two operating in sequence. Having built both, however, a composition is apparent: detection (Sects. 2.2–2.7) flags which artifacts and at what severity are present, while enhancement (Sects. 2.8–2.11) reduces five of those seven categories. Closing the loop — acquire, score, enhance the flagged scans, then re-score the enhanced output with the same detection model — would give an automatic, artifact-specific acceptance check. We report no end-to-end evaluation of this loop, and whether enhancement measurably improves the detection model's predicted severity is untested; establishing that is the immediate next step.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ravi, D., Alzheimer's Disease Neuroimaging Initiative, Barkhof, F., Alexander, D.C., Puglisi, L., Parker, G.J.M., Eshaghi, A.: An efficient semi-supervised quality control system trained using physics-based MRI-artifact generators and adversarial training. Med. Image Anal. **91**, 103033 (2024)
2. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. ICLR (2019)
3. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: Proc. ICLR (2017)
4. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: Proc. CVPR, vol. 2, pp. 60–65 (2005)
5. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multi-label data. In: Proc. ECML PKDD. LNCS, vol. 6913, pp. 145–158. Springer (2011)
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Proc. MICCAI. LNCS, vol. 9351, pp. 234–241. Springer (2015)
7. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. CVPR, pp. 586–595 (2018)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc. NeurIPS, pp. 6626–6637 (2017)
9. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
10. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012)
11. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proc. AAAI, pp. 2555–2563 (2023)
12. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magn. Reson. Med. **34**(6), 910–914 (1995)
13. Konz, N., Osuala, R., Verma, P., et al.: Fréchet radiomic distance (FRD): a versatile metric for comparing medical imaging datasets. Med. Image Anal. **110**, 103943 (2026)
14. LISA 2026 Challenge: Low-field pediatric brain MRI segmentation and quality assurance. Synapse, syn72118611 (2026). https://www.synapse.org/Synapse:syn72118611
15. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Anal. Appl. **26**, 941–955 (2023)
16. Kim, H., Seo, J., Ryu, S., Park, J.H., On, S., Choi, J.: Axis-guided quality assessment and multi-label hippocampal and ventricular segmentation in low-resolution pediatric brain MRI. In: LISA 2024. LNCS, vol. 15515, pp. 53–62. Springer (2025)
17. Sundaresan, V., Dinsdale, N.K.: Automated quality assessment using appearance-based simulations and hippocampus segmentation on low-field paediatric brain MR images. In: LISA 2024, pp. 41–52. Springer (2024)
18. Zhu, Y., Jiang, H., Cai, R., Chen, G.: Multi-label MambaOut for quality assessment of low-field pediatric brain MR images. In: LISA 2024, pp. 3–11. Springer (2024)
19. Almsouti, A., Khamitova, A., Taratynova, D., Yaqub, M.: BRIQA: balanced reweighting in image quality assessment of pediatric brain MRI. In: LISA 2025. LNCS, vol. 16411, pp. 3–14. Springer (2026)
20. Lazo-Quispe, C., Espinoza-Chamorro, R.: Robust multi-label classification of MRI artifacts in low-field neonatal brain imaging via view-conditional dual-task learning. In: LISA 2025. LNCS, vol. 16411, pp. 15–25. Springer (2026)