A 2.5D Ordinal Quality Assessment and Metadata-Preserving Residual Restoration Framework for Ultra-Low-Field Pediatric Brain MRI

Jongjun Won and Changseon Park

Research Institute, NEUROPHET Inc., Seoul 06234, Republic of Korea
{wjj910,changseon}@neurophet.com

Abstract. Ultra-low-field (ULF) MRI enables portable, low-cost, point-of-care neuroimaging, but its low signal-to-noise ratio, reduced tissue contrast, anisotropic resolution, and acquisition artifacts hinder automated analysis of pediatric brain scans. We present a unified framework for Task 1 of the LISA 2026 Challenge, covering both automated artifact assessment (Task 1a) and image quality improvement (Task 1b). For Task 1a we formulate multi-artifact grading as an ordinal problem: a 2.5D ConvNeXt–GRU model predicts three-level severity for seven artifacts through a cumulative-link head, with a subject-wise 5-fold ensemble. For Task 1b, where paired clean targets are unavailable, we train a metadata-preserving residual 3D U-Net on online-synthesized degraded–clean pairs; it predicts a bounded, identity-initialized correction regularized by the frozen Task 1a model as an in-domain critic restricted to intensity-domain artifacts, and restored volumes are written back into the original NIfTI grid unchanged. On held-out proxy validation the restorer reduces FID from 106.8 to 41.6, pooled LPIPS from 0.484 to 0.290, and a Fréchet Radiomics Distance proxy from 330 to ≈ 1, while an independent in-domain grader indicates reduced artifact severity across all seven classes; all 114 challenge validation outputs pass an exact metadata-preservation check. Code is available at https://github.com/Krying/LISA_2026-task1-a.

Keywords: Ultra-low-field MRI · Pediatric brain MRI · Image quality assessment · Ordinal classification · Image restoration · Metadata preservation.

1 Introduction

Magnetic resonance imaging (MRI) is central to studying pediatric brain development, but conventional high-field systems are expensive, immobile, and infrastructure-intensive, and pediatric imaging is further complicated by scan tolerability, acoustic burden, and motion. Ultra-low-field (ULF) MRI offers a portable, low-cost, point-of-care alternative [3, 4]. Its reduced field strength,

however, lowers signal-to-noise ratio and tissue contrast, yields anisotropic resolution, and introduces acquisition artifacts—noise, motion, zipper, banding, contrast degradation, distortion, and positioning errors—amplified by pediatric motion and heterogeneous acquisition quality, so reliable quality assessment and quality improvement are both essential for practical use.

The LISA 2026 Challenge addresses these needs through two complementary tasks [1, 2]. Task 1a is automated artifact assessment—predicting ordinal severity for multiple artifact categories; unlike standard classification, severity is inherently ordered and severe grades are strongly underrepresented. Task 1b is image quality improvement: reduce artifacts while preserving clinically relevant anatomy and the spatial metadata of the original NIfTI image. Restoration is hard because paired high-quality targets are unavailable and outputs must remain faithful to the measured ULF acquisition rather than hallucinate clean-looking anatomy. Prior work uses no-reference quality metrics and deep artifact classification [9, 10], with view-aware, class-imbalance-aware designs for low-field pediatric MRI [5]; for restoration, generative enhancement can improve perceptual quality but may alter anatomy when optimized for distributional similarity alone [19], so conservative enhancement is preferable in this clinical setting.

We present a unified framework for both tasks. For Task 1a we use a 2.5D volume-level ordinal classifier: each volume becomes a sequence of adjacent-slice stacks, an ImageNet-pretrained ConvNeXt encoder [17] extracts slice features, a bidirectional recurrent module aggregates them, and a cumulative-link head predicts seven ordinal severities, with a subject-wise 5-fold ensemble. For Task 1b we train a metadata-preserving residual 3D U-Net [15, 16] on online-synthesized degraded–clean pairs from a strict-clean subset; it predicts a bounded, identity-initialized correction, so passthrough is the default and anatomical alteration is structurally limited, and restored voxels are written back into the original NIfTI grid unchanged in geometry and data type. The two tasks are coupled: the frozen Task 1a model is reused as an in-domain artifact critic that penalizes only intensity-domain artifacts during Task 1b training, excluding geometry labels that could incentivize anatomical warping.

2 Methods

2.1 Dataset

We used the official LISA 2026 Task 1 quality dataset, which consists of ultra-low-field (0.064 T) pediatric brain MRI volumes with image-level artifact annotations. The metadata contained 531 NIfTI volumes from 243 subjects. Each subject was scanned in up to three orthogonal orientations, resulting in 176 axial, 180 coronal, and 175 sagittal volumes. Each volume was labeled for seven artifact categories: Noise, Zipper, Positioning, Banding, Motion, Contrast, and Distortion. For each artifact, the label was provided on a three-level ordinal severity scale, where 0 indicates absence of the artifact, 1 indicates moderate severity, and 2 indicates severe severity.

Table 1. Dataset summary used by the final restoration run. A volume is *strict-clean* when all seven artifact labels are zero.

Split	Volumes	Subjects	Axial	Coronal	Sagittal
Strict-clean training pool	126	80	36	62	28
Train split	98	64	28	50	20
Held-out clean validation	28	16	8	12	8
Challenge validation inputs	114	–	36	43	35

For Task 1a, all 531 labeled volumes were used for multi-artifact severity assessment. Because multiple acquisitions from the same subject may be present in different orientations, we split the data at the subject level to avoid leakage between training and validation folds. Because the labeled set is small and strongly imbalanced, we compared 4- and 5-fold subject-wise cross-validation; the 5-fold ensemble gave the best validation performance and was used for the final Task 1a system.

For Task 1b, quality improvement is required without paired degraded–clean images, so we used the artifact labels to define a clean target domain from the same dataset. A volume was considered *strict-clean* only when all seven artifact labels were zero. This criterion yielded 126 strict-clean volumes from 80 subjects. These volumes were used as clean restoration targets, while degraded–clean training pairs were generated online through synthetic artifact simulation. The strict-clean subset was split at the subject level into 98 training volumes from 64 subjects and 28 held-out validation volumes from 16 subjects (Table 1). The official challenge validation set for Task 1b contained 114 unlabeled input volumes (36 axial, 43 coronal, 35 sagittal), used only for generating the final restoration submission.

2.2 Preprocessing

Preprocessing is task-specific: Task 1a needs volume-level severity prediction, Task 1b voxel-level restoration with exact metadata preservation.

For Task 1a, each volume is loaded as a 3D float image and robustly normalized by clipping non-background intensities to robust percentile limits. The through-plane axis is the one with the largest voxel spacing, so each slice keeps high-resolution in-plane anatomy; we sample 32 locations uniformly along it and stack three adjacent slices as channels, resize to 224×224, and normalize with ImageNet statistics. Training uses light augmentation (slice-location jitter, random flips); inference is deterministic.

For Task 1b, each strict-clean volume is loaded in physical units and normalized to $[0, 1]$ using foreground percentiles (the 0.5th/99.5th of voxels above the volume minimum). The largest-spacing axis is moved first, giving a coarse-axis-first (S, H, W) layout with the two high-resolution axes last; training crops or zero-pads to $48 \times 160 \times 160$ (centered at validation). Outputs are mapped back to physical intensity using the input's percentile statistics.

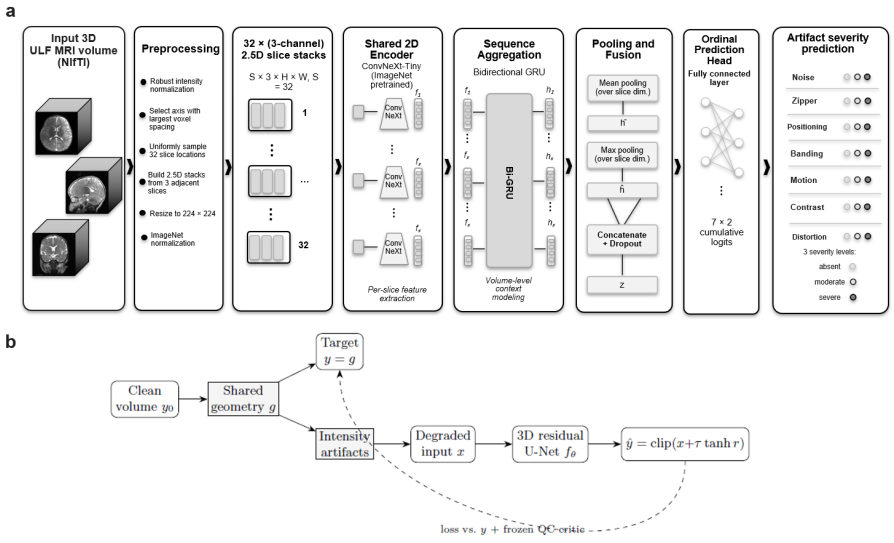

Fig. 1. Overview of the proposed framework. **(a)** Task 1a: each 3D ULF MRI volume is normalized and converted into a sequence of 32 three-channel 2.5D slice stacks; a shared ImageNet-pretrained ConvNeXt-Tiny encoder extracts per-slice features that a bidirectional GRU aggregates into volume-level context, and an ordinal head predicts 7×2 cumulative logits, i.e. three severity levels (absent/moderate/severe) for each of the seven artifacts. **(b)** Task 1b training: a shared spatial transform g defines the geometry of both the clean target $y = g(y_0)$ and the input, intensity/acquisition artifacts are applied to the input only, and a bounded, identity-initialized 3D residual U-Net predicts $\hat{y} = \mathrm{clip}(x + \tau \tanh r)$; the frozen Task 1a model acts as an in-domain quality-control critic in the loss.

2.3 Model Architecture

Our framework consists of two task-specific networks: a 2.5D ordinal classifier for Task 1a and a metadata-preserving residual 3D restoration network for Task 1b, connected through a quality-assessment-to-restoration design in which the Task 1a classifier is reused as a frozen in-domain artifact critic for Task 1b. An overview of both pipelines is shown in Fig. 1.

Task 1a: 2.5D Ordinal Artifact Classifier. The input to the network was a fixed-length sequence of 2.5D slice stacks. Given an input volume represented as $x \in \mathbb{R}^{S\times3\times H\times W}$ with $S = 32$, each slice stack was treated as a three-channel image and processed independently by a shared 2D convolutional encoder. The encoder was an ImageNet-pretrained ConvNeXt-Tiny model [17] with its classification layer removed, followed by global average pooling to obtain a per-stack feature vector. The resulting sequence of slice-level embeddings was passed to a bidirectional gated recurrent unit (GRU) to aggregate contextual information across the volume. The output sequence was summarized by both mean pooling

(capturing the global artifact tendency) and max pooling (capturing strong localized artifact responses); the two pooled vectors were concatenated and passed through dropout and a fully connected head.

For the main ordinal model, the head produced two logits per artifact, the cumulative probabilities $P(y > 0)$ and $P(y > 1)$, decoded into three ordered classes as $p_0 = 1 - P(y > 0)$, $p_1 = P(y > 0) - P(y > 1)$, $p_2 = P(y > 1)$. The final output was a 7×3 probability tensor (one distribution per artifact), averaged across the 5-fold ensemble at inference.

Task 1b: Metadata-Preserving Residual 3D Restoration Network. For Task 1b we used a full-volume 3D residual U-Net [15, 16] with replication-padded 3D convolutions, GroupNorm [18], and PReLU activations, a bottleneck bridge, and transposed-convolution decoder blocks with skip connections. The final model used four downsampling stages, base width 32, depth 8, and approximately 17.9M trainable parameters. Instead of predicting a restored image directly, the network predicted a bounded residual correction. Given a degraded input volume x, the enhanced output was

$$\hat{y} = \mathrm{clip}\big(x + \tau \tanh(h_\theta(x)),\, 0,\, 1\big), \qquad \tau = 0.5, \tag{1}$$

where $h_\theta(x)$ is the U-Net head. The final $1 \times 1 \times 1$ convolution was initialized with zero weights and bias, so at initialization the model performed an exact identity mapping. This made passthrough the default behavior and constrained the model to bounded corrections rather than synthesizing a new image, reducing the risk of excessive anatomical alteration.

Task Coupling Through an Artifact Critic. The Task 1a classifier was reused as a frozen in-domain artifact critic during Task 1b training. Given the restored volume, the critic estimated expected ordinal severity scores for intensity-domain artifacts only—Noise, Motion, Contrast, Zipper, and Banding. Positioning and Distortion were excluded because penalizing geometry-related labels could encourage the restoration model to alter anatomy or spatial structure. This coupling provided a task-specific quality signal while maintaining conservative, metadata-preserving restoration behavior.

2.4 Loss Function

Task 1a: Ordinal Artifact Classification Loss. Because the three severity levels are ordered, we used an ordinal binary cross-entropy loss rather than treating the classes as nominal. For each artifact k, the model predicts two cumulative logits for the thresholds $y_k > 0$ and $y_k > 1$. Given the ground-truth label $y_k \in \{0, 1, 2\}$, we form two binary ordinal targets $t_{k,1} = \mathbb{1}(y_k > 0)$ and $t_{k,2} = \mathbb{1}(y_k > 1)$, so that class 0 is encoded as $(0, 0)$, class 1 as $(1, 0)$, and class 2 as $(1, 1)$. Let $z_{k,m}$ be the predicted logit for artifact k and threshold m. The

ordinal loss is a weighted binary cross-entropy over all artifacts and thresholds,

$$\mathcal{L}_{\text{ord}} = \frac{1}{7 \times 2} \sum_{k=1}^{7} \sum_{m=1}^{2} \text{BCEWithLogits}\big(z_{k,m},\, t_{k,m};\, w_{k,m}\big), \tag{2}$$

where $w_{k,m}$ is a positive-class weight computed from the training fold to compensate for the strong imbalance between absent, moderate, and severe labels (the cumulative probabilities are decoded into three classes as above).

Task 1b: Restoration Loss. The restoration network was trained on synthetically generated degraded–clean pairs. The primary objective combined a foreground-weighted Charbonnier loss, a volumetric structural similarity loss, and a 3D gradient consistency term,

$$\mathcal{L}_{\text{rest}} = \mathcal{L}^{\text{fg}}_{\text{charb}} + \lambda_s \, \mathcal{L}_{\text{SSIM3D}} + \lambda_g \, \mathcal{L}_{\nabla}. \tag{3}$$

The foreground-weighted Charbonnier loss is

$$\mathcal{L}^{\text{fg}}_{\text{charb}} = \frac{\sum_i w_i \sqrt{(\hat{y}_i - y_i)^2 + \epsilon^2}}{\sum_i w_i}, \qquad w_i = 1 + 4 \cdot \mathbb{1}(y_i > 0.05), \tag{4}$$

where y is the clean target, $\hat{y}$ the restored output, and foreground voxels are upweighted so the model does not spend capacity denoising background air at the expense of brain tissue. The structural term is a volumetric SSIM loss $\mathcal{L}_{\text{SSIM3D}} = 1 - \text{SSIM3D}(\hat{y}, y)$ [14], and the gradient term $\mathcal{L}_{\nabla} = \frac{1}{3} \sum_{d \in \{x,y,z\}} \|\nabla_d \hat{y} - \nabla_d y\|_1$ matches finite-difference gradients along the three axes to preserve edges and discourage over-smoothing; we set $\lambda_s = 1.0$ and $\lambda_g = 0.5$. After a warm-up we add an auxiliary term, $\mathcal{L} = \mathcal{L}_{\text{rest}} + \lambda_a \mathcal{L}_{\text{artifact}}$ with $\lambda_a = 0.05$, where $\mathcal{L}_{\text{artifact}}$ is the mean expected ordinal severity over the intensity-domain artifacts (Noise, Motion, Contrast, Zipper, Banding) from the frozen Task 1a critic; Positioning and Distortion are excluded so the critic never incentivizes anatomical warping.

2.5 Synthetic Degradation for Task 1b

Because paired clean–degraded examples are unavailable, we train on unlimited synthetic pairs from strict-clean images. Shared spatial augmentation—flips, small in-plane affine transforms, and mild elastic deformation—is applied identically to input and target, adding robustness without asking the model to change geometry. Input-only intensity/acquisition artifacts then corrupt the input only: Rician noise [13], k-space line-motion, smooth bias fields, zipper lines, k-space spike banding, and low-frequency clouding. For the final run we disabled input-only contrast and excluded geometric distortion, since learning to undo geometry conflicts with metadata preservation; severities are sampled from ranges motivated by the challenge labels.

2.6 Experimental Details

Task 1a training. For each of the subject-wise 5 folds, the model (ConvNeXt-Tiny encoder, bidirectional GRU, ordinal head) was trained on four folds and validated on the fifth, using sequences of 32 2.5D slice stacks at 224×224. We used AdamW [20] (learning rate 2×10^{-4}, weight decay 1×10^{-4}), batch size 40, cosine annealing, mixed precision, and gradient clipping (max norm 1.0); the best checkpoint per fold was selected by the validation mean of accuracy, precision, recall, F1, and F2, and the 5-fold probabilities were averaged.

Task 1b training. The restorer was trained on normalized $48 \times 160 \times 160$ crops for 500 epochs using AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-4}), batch size 16, mixed precision, and cosine annealing to zero. The frozen artifact critic was activated only after a warm-up, so structure-preserving restoration settled before the artifact-aware signal.

Task 1b model selection and inference. Because Task 1b does not provide paired clean targets for the challenge validation images, model selection was performed on a subject-disjoint held-out clean split with deterministic synthetic degradation. Proxy metrics—paired PSNR, pooled unpaired PSNR, LPIPS [8], BRISQUE [9], CLIP-IQA [10], FID [7], and a Fréchet Radiomics Distance (FRD) proxy [6, 11, 12]—were computed periodically, and the final checkpoint was chosen by the pooled LPIPS proxy. At inference each volume was processed whole with the same percentile normalization, the internal axis reordering and padding were reversed, and the denormalized image was written into the original NIfTI grid—preserving shape, spacing, orientation, affine, and qform/sform metadata, with integer inputs rounded and clipped to the stored dtype range.

3 Results

3.1 Task 1a

We evaluated Task 1a with subject-wise cross-validation on the 531 labeled volumes, scoring the out-of-fold predictions with the challenge's micro-, macro-, and weighted-averaged metric definitions. The seven per-artifact predictions are pooled into a single three-class ($\{0, 1, 2\}$) problem; consequently the micro-averaged F1, F2, precision, recall, and accuracy coincide, and we report this shared value as Accuracy. Table 2 compares the encoder and fold configurations we explored (metrics averaged over folds).

All configurations reach comparable pooled accuracy (0.827–0.835) but much lower macro-F1 (≈ 0.68), reflecting the class imbalance: the absent class dominates, so moderate and severe grades are harder to recall. The ConvNeXt-Tiny ordinal model with a subject-wise 5-fold ensemble gives the best accuracy (0.835) and macro-F1 (0.690), outperforming both the 4-fold ConvNeXt-Tiny runs and the larger ConvNeXt-Base encoder; since ensembling is generally expected to

Table 2. Task 1a subject-wise cross-validation on the labeled set, using the challenge's pooled three-class metric definitions (Accuracy equals the micro-averaged F1/F2/precision/recall). All encoders are ImageNet-pretrained ConvNeXt and use the ordinal cumulative-link loss; ConvNeXt-Tiny with a 5-fold ensemble is the final submission. Best per column in bold.

Encoder	Folds	Accuracy	$F1_{macro}$	P_{macro}	R_{macro}	$F1_{wt}$
ConvNeXt-Tiny (30 ep)	4	0.827	0.678	0.668	**0.700**	0.825
ConvNeXt-Tiny (300 ep)	4	0.832	0.683	0.714	0.663	0.830
ConvNeXt-Tiny	5	**0.835**	**0.690**	**0.714**	0.675	**0.833**
ConvNeXt-Base	5	0.831	0.685	0.713	0.666	0.830

Table 3. Per-artifact out-of-fold results for the final ConvNeXt-Tiny ordinal 5-fold model (per-fold mean). $n_{\geq 1}/n_2$: moderate-or-severe / severe counts among the 531 volumes.

Artifact	$n_{\geq 1}$	n_2	Accuracy	$F1_{macro}$
Noise	93	51	0.927	**0.772**
Zipper	178	36	0.840	0.732
Positioning	69	24	0.881	0.529
Banding	21	12	**0.959**	0.546
Motion	168	75	0.751	0.658
Contrast	196	31	0.770	0.654
Distortion	143	48	0.717	0.512

help in the low-data regime, this 5-fold ConvNeXt-Tiny ordinal ensemble is our final submission. These in-distribution values are optimistic relative to the official leaderboard, where our best submission reached accuracy 0.843 and macro-F1 0.639 (weighted-F1 0.826): the pooled accuracy transfers closely while the lower macro-F1 confirms the in-distribution macro scores were mildly optimistic. Fig. 2 shows predictions across the three orientations.

Table 3 breaks the final model down by artifact. High accuracy on rare, well-separated artifacts (Banding, Noise) masks a low macro-F1 on the hardest classes—Positioning, Banding, and especially Distortion ($F1_{macro}$ 0.51), whose severe grades are scarce or ambiguous—confirming that pooled accuracy is dominated by easy negatives while minority severe grades drive the residual error.

3.2 Task 1b

Table 4 reports the proxy suite on the subject-disjoint held-out split (deterministic degradation) for the metadata-faithful passthrough and the selected and final checkpoints. The model substantially improved the challenge's emphasized metrics: FID $106.8 \rightarrow 41.6$, pooled LPIPS $0.484 \rightarrow 0.290$, and FRD proxy $330.0 \rightarrow \approx 1$, while BRISQUE and CLIP-IQA stayed flat-to-slightly-improved and the unpaired pooled PSNR edged up ($16.03 \rightarrow 16.38$). The FRD is a *proxy*: radiomics

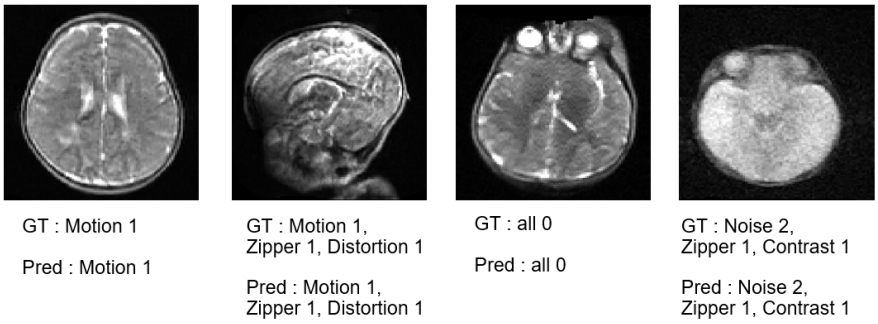

Fig. 2. Representative Task 1a predictions on held-out validation cases across the three acquisition orientations. Ground-truth (GT) and predicted (Pred) artifact severities agree in each example, including the artifact-free (all-zero) case and multi-artifact cases.

Table 4. Internal held-out proxy metrics for Task 1b on the subject-disjoint clean validation split with deterministic synthetic degradation. Arrows indicate the better direction; the passthrough column corresponds to the metadata-faithful input.

Metric	Passthrough	Selected	Final
FID (↓)	106.8	42.8	**41.6**
LPIPS, pooled (↓)	0.484	**0.288**	0.290
FRD proxy (↓)	330.0	**0.88**	1.05
BRISQUE (↓)	51.7	51.2	**50.2**
CLIP-IQA (↑)	0.352	0.371	**0.372**
PSNR, pooled unpaired (↑)	16.03	16.37	**16.38**
PSNR, paired masked (↑)	**31.75**	28.70	29.15

features are standardized per feature on a small reference bank (16 held-out subjects), so its absolute scale is sensitive to feature normalization and sample size, and the near-1 value should be read as a large relative improvement rather than an absolute distance.

The only metric where the model did not exceed passthrough was the paired masked PSNR (31.75 vs. 29.15 dB), reflecting the perception–distortion tradeoff [19]: the challenge scores an unpaired PSNR against a bank, and the bounded residual keeps every output within a few decibels of its input, so this is expected under our conservative design. To check that the gains reflect genuine artifact removal rather than smoothing, we scored the restored volumes with the in-domain grader (the Task 1a model); to keep this independent of the training objective, the restorer here is an earlier checkpoint (base width 24) trained without the critic, so the grader is not scoring a network optimized against it (Table 5).

The grader's predicted severity decreased for all seven artifacts, including the two stated Task 1b targets (Noise $0.196 \to 0.057$, Motion $0.282 \to 0.201$), a

Table 5. Independent in-domain grader validation on the 114 real challenge validation volumes: mean predicted artifact severity (lower is cleaner). To avoid circularity the restorer here is an earlier checkpoint trained without the artifact critic. The restoration model lowers all seven severities, whereas a naive Gaussian blur raises the total and the intensity targets.

Artifact	Input	Restored	Gaussian blur ($\sigma = 1$)
Noise	0.196	**0.057**	0.203
Motion	0.282	**0.201**	0.566
Contrast	0.265	**0.205**	0.497
Total (7 artifacts)	1.581	**1.017**	2.220

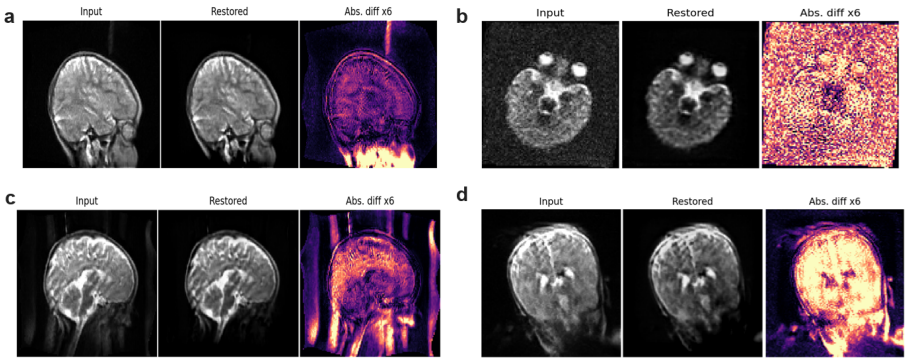

Fig. 3. Qualitative Task 1b restoration on challenge validation volumes **(a–d)**. For each case: the input (left), the restored output (middle), and the absolute difference magnified 6× (right). The bounded residual concentrates its edits on plausible intensity artifacts while leaving anatomy and image geometry intact; the difference maps are dominated by background and skull regions rather than by brain structure.

∼36% reduction in total; a naive Gaussian blur instead raised the total to 2.22 and increased five of seven severities (all three intensity targets), consistent with structural improvement rather than smoothing. Finally, all 114 official validation volumes passed an exact metadata check (114/114: shape, spacing, orientation, affine, qform/sform, dtype). Fig. 3 shows qualitative restorations with magnified difference maps.

4 Discussion

The two tasks share one design principle: predictions and edits should respect the ordered, imbalanced structure of the artifact labels and the measured ULF acquisition. For Task 1a, the cumulative-link ordinal head and 2.5D aggregation encode severity order directly, better matching the evaluation than nominal classification. For Task 1b, the unpaired reference-bank evaluation rewards clean-looking images, but a model can raise distributional or no-reference scores

by adding structure unsupported by the ULF measurement; the residual cap, identity initialization, foreground-weighted fidelity, and exclusion of geometry artifacts from the critic all curb this risk. We also treat metadata preservation as a first-class objective, writing into the input grid rather than resampling, since reorientation, isotropic resampling, or a fresh NIfTI header can silently alter spatial metadata.

Several limitations remain. For Task 1a we report cross-validation trends without fold-wise confidence intervals or significance tests, relying instead on the independent held-out leaderboard as the arbiter of generalization. For Task 1b, our proxy harness aids model selection but is not the official server, and the main gap is sim-to-real: training uses only 126 strict-clean volumes with synthetic corruption evaluated against synthetic targets, so whether the improvements transfer to real ULF artifacts remains the key open question—although the in-domain grader (Table 5), applied to the 114 real validation volumes, already shows severity reductions on real data. Metadata preservation guarantees geometric fidelity but not anatomical fidelity per se; conservatism is instead enforced by the bounded residual, identity initialization, and geometry-excluded critic, with difference maps concentrated outside central brain tissue (Fig. 3) and rigorous anatomical validation left to future work, alongside better sim-to-real calibration and larger reference banks.

5 Conclusion

We presented a unified, metadata-preserving framework for LISA 2026 Task 1: a 2.5D ordinal artifact classifier (Task 1a) and an input-anchored bounded residual 3D U-Net regularized by the frozen Task 1a critic (Task 1b). It improves the perceptual and distributional proxy metrics over passthrough while structurally discouraging hallucination; an independent grader suggests these gains are largely structural, and all outputs preserve the required spatial metadata. Because Task 1b is evaluated only on proxy metrics and synthetically degraded scans, restoration performance on genuinely corrupted ULF acquisitions remains to be validated.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Lepore, N., Linguraru, M.G. (eds.): Low Field Pediatric Brain MRI Segmentation and Quality Assurance: Second MICCAI Challenge (LISA 2025). Lecture Notes in Computer Science, vol. 16411. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-14417-1
2. Low-field Pediatric Brain MRI Segmentation and Quality Assurance (LISA) Challenge, MICCAI 2026. https://openreview.net/group?id=MICCAI.org/2026/Workshop/LISA (2026)
3. Arnold, T.C., Freeman, C.W., Litt, B., Stein, J.M.: Low-field MRI: clinical promise and challenges. Journal of Magnetic Resonance Imaging **57**(1), 25–44 (2023)

4. Iglesias, J.E., et al.: Quantitative brain morphometry of portable low-field-strength MRI using super-resolution machine learning. Radiology **306**(3), e220522 (2023)
5. 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. Lecture Notes in Computer Science, vol. 15515, pp. 53–62. Springer, Cham (2025)
6. Thummerer, A., et al.: SynthRAD2023 grand challenge dataset: generating synthetic CT for radiotherapy. Medical Physics **50**(7), 4664–4674 (2023)
7. 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: Advances in Neural Information Processing Systems (NeurIPS), pp. 6626–6637 (2017)
8. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018)
9. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
10. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: AAAI Conference on Artificial Intelligence, pp. 2555–2563 (2023)
11. van Griethuysen, J.J.M., et al.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), e104–e107 (2017)
12. Konz, N., Osuala, R., Verma, P., et al.: Fréchet Radiomic Distance (FRD): a versatile metric for comparing medical imaging datasets. arXiv:2412.01496 (2025)
13. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magnetic Resonance in Medicine **34**(6), 910–914 (1995)
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)
16. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. Lecture Notes in Computer Science, vol. 9901, pp. 424–432. Springer (2016)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986 (2022)
18. Wu, Y., He, K.: Group normalization. In: European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
19. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6228–6237 (2018)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)