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.

Anatomy-preserving enhancement of fetal brain MRI without clean references

Qiyuan Tian

qiyuantian@tsinghua.edu.cn

Tsinghua University https://orcid.org/0000-0002-8350-5295

Yingqi Hao

Tsinghua University

Mingxuan Liu

Tsinghua University

Nona Law

Yi Liao

West China Second University Hospital https://orcid.org/0000-0001-8926-8692

Hongjia Yang

Tsinghua University

Xiaotian Hu

Tsinghua University

Juncheng Zhu

West China Second University Hospital

Kasidit Anmahapong

Tsinghua University

Yijin Li

Tsinghua University

Zihan Li

Tsinghua University

Yifei Chen

Tsinghua University

Nan Sun

The Third Affiliated Hospital of Zhengzhou University

Rong Hu

Qujing Maternal and Child Health Hospital

Min Kang

Sichuan Provincial Woman's and Children's Hospital, The Affiliated Women's and Children's Hospital of Chengdu Medical College

Yan Song

Sichuan Provincial Woman’s and Children’s Hospital, The Affiliated Women’s and Children’s Hospital of Chengdu Medical College

Hua Lai

UESTC Chengdu Women's and Children's Central Hospital

Xiaoling Zhou

University of Electronic Science and Technology of China

Gang Ning

West China Second University Hospital, Sichuan University

Haibo Qu

West China Second University Hospital, Sichuan University

Article

Keywords:

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10372146/v1

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

Additional Declarations: There is **NO** Competing Interest.

Anatomy-preserving enhancement of fetal brain MRI without clean references

Yingqi Hao[1, 2#], Mingxuan Liu[2#], Yitong Luo[2#], Yi Liao[1, 3#], Hongjia Yang[2], Xiaotian Hu[2], Juncheng Zhu[1, 3], Kasidit Anmahapong[2], Yijin Li[2], Zihan Li[2], Yifei Chen[2], Nan Sun[4], Rong Hu[1,3], Min Kang[5], Yan Song[5], Hua Lai[6], Xiaoling Zhou[6], Gang Ning[1,3], Haibo Qu[1, 3*], Qiyuan Tian[2*]

[1] Department of Radiology, West China Second University Hospital, Sichuan University, Chengdu, Sichuan Province, China;

[2] School of Biomedical Engineering, Tsinghua University, Beijing, China;

[3] Key Laboratory of Birth Defects and Related Diseases of Women and Children, Ministry of Education, Chengdu, Sichuan Province, China;

[4] Department of Radiology, The Third Affiliated Hospital of Zhengzhou University;

[5] Department of Radiology, Sichuan Provincial Woman's and Children's Hospital, The Affiliated Women's and Children's Hospital of Chengdu Medical College, Chengdu, China;

[6] Chengdu Women's and Children's Central Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China.

[#] These authors contributed equally.

[*] Corresponding author: Qiyuan Tian, PhD, Center for Biomedical Imaging Research, Tsinghua University 30 Shuangqing Road, Haidian District, Beijing, China, PA100084 (qiyuantian@tsinghua.edu.cn). Haibo Qu, MD, No.20, Section 3, Renmin South Road, Chengdu, Sichuan Province, China, PA610041(windowsqhb@126.com).

Abstract

Fetal brain MRI plays a crucial role in prenatal diagnosis, yet inevitable fetal motion and slice-to-volume reconstruction (SVR) errors introduce corruptions including noise and artifacts that degrade diagnostic reliability. Existing image enhancement methods either require noise-free reference data unavailable in fetal MRI or assume pixel-level independent noise that fundamentally mismatches the spatially coarse noise in this modality. To address these limitations, this study introduces FetMRIE, the first unsupervised image enhancement framework for fetal brain MRI via adaptive state-matching denoising diffusion. The proposed two-stage architecture combines variational autoencoder-based initial denoising with diffusion-based structural restoration, effectively balancing noise suppression and anatomical fidelity. Experiments on 2,481 fetuses across eight multi-center datasets, acquired with diverse T_2-weighted sequences (i.e., ss-FSE, TSE-SSH, HASTE, and B-TFE) on 1.5-T and 3.0-T scanners and comprising both normative and pathological cases (i.e., ventriculomegaly and germinal matrix-intraventricular hemorrhage), demonstrated that FetMRIE achieves state-of-the-art performance on unsupervised PSNR (u-PSNR), unsupervised MSE (u-MSE), tissue contrast t-score (TCT), and cross-tissue noise consistency (CNC) with robust cross-domain generalizability. Four radiologists with strong inter-rater agreement (correlation coefficients: 0.720–0.871) consistently rated FetMRIE highest in noise suppression, anatomical fidelity, and overall image quality. Three downstream tasks (i.e., gestational age prediction, anomaly detection, and biometry measurement) further demonstrated that FetMRIE improves the accuracy and reliability of automated clinical analyses by enhancing normative anatomical clarity while faithfully preserving pathological signatures.

1 Introduction

Fetal brain magnetic resonance imaging (MRI) plays a crucial role in prenatal diagnosis, due to its non-invasive safety and superior tissue contrast compared to Ultrasound (US)[1–4]. However, due to the inevitable fetal motion, 3D MR sequences are not applicable on fetal MRI[5,6]. To address this problem, clinical fetal MRI typically employs ultrafast single-shot sequences, yielding multiple stacks of 2D T_2-weighted thick-slice images from three orthogonal anatomical orientations. To obtain super-resolution 3D brain volumes, slice-to-volume reconstruction (SVR) technologies must be employed to integrate these orthogonal 2D acquisitions. Despite optimization-based (e.g., NiftyMIC[7]) and deep learning-based (e.g., NesVoR[8]) SVR approaches successfully reconstruct super-resolution 3D brain volumes, the fetal movements and maternal respirations inevitably introduce various potential corruptions, such as topological inconsistencies, geometric distortions and high noise levels[9–11]. These noises and artifacts severely degrade diagnostic reliability and have a negative impact on diverse downstream analyses[12,13]. Consequently, anatomically faithful post-processing image enhancement for fetal brain MRI is of great importance.

However, developing effective deep learning-based enhancement algorithms for fetal brain MRI remains substantially more challenging than traditional MRI denoising tasks. Specifically, supervised denoising methods are trained on paired noisy and noise-free images[14–16], but the latter are absent for fetal MRI. Self-supervised approaches such as Self2Self[17], Zero-shot Noise2Noise (ZS-N2N)[18] and Pixel2Pixel[19] eliminate the need for noise-free references. However, these methods are generally developed under the assumption that noise occurs at the pixel-level (e.g. Gaussian and Rician noise)[20], whereas fetal MRI is typically affected by diverse and non-traditional forms of degradation[9]. Therefore, neither type of method is directly applicable to fetal

MRI enhancement. Additionally, due to the absence of ground truth (GT), common metrics such as PSNR and SSIM cannot be computed, which makes quantitative evaluation of enhancement performance challenging[21].

To the best of our knowledge, the recently proposed BME-X[15] is the only existing method specifically designed for MRI enhancement across the lifespan and applicable to fetal images. It consists of a tissue classification network that predicts tissue types and a tissue aware enhancement network that generates high quality images. However, its evaluation on fetal brain MRI remains insufficient. Specifically, (1) the original study for evaluating fetal MRI enhancement relied solely on the tissue contrast t-score (TCT)[22] between white matter (WM) and gray matter (GM), without reader assessment by experienced radiologists; (2) it lacked evaluation on pathological cases; and (3) it was validated only on an in-distribution (ID) dataset. More critically, because BME-X employs a three-category tissue classification network for WM, GM, and cerebrospinal fluid (CSF), together with tissue-aware enhancement that drives each tissue class toward uniform intensity, it often introduces severe over smoothing, resulting in the loss of critical anatomical features such as cortical folding patterns and deep gray matter (dGM) boundaries. Thus, there remains an urgent need for a fully unsupervised framework that can balance noise suppression with faithful preservation of anatomical structures.

To address these limitations, we propose a fully unsupervised fetal brain MR image enhancement (FetMRIE) framework without clean reference, based on adaptive state-matching denoising diffusion. Our two-stage framework first suppresses dominant noise through latent space

regularization, and then restores blurred anatomical details via adaptive structure-preserving reconstruction. The main contributions are summarized as follows:

1) FetMRIE is the first unsupervised image enhancement framework for fetal brain MRI, which combines VAE-based initial denoising with adaptive state-matching diffusion-based restoration to break the trade-off between noise removal and structural fidelity.
2) FetMRIE achieves state-of-the-art (SOTA) performance on four reference-free quantitative metrics (unsupervised PSNR [u-PSNR], unsupervised MSE [u-MSE], tissue contrast t-score [TCT], and cross-tissue noise consistency [CNC]), evaluated on an unprecedented cohort of 2,481 fetuses across eight datasets encompassing different centers, sequences, and both normative and pathological fetal brains.
3) FetMRIE was consistently rated highest in noise suppression, anatomical fidelity, and overall image quality by four radiologists with strong inter-rater agreement (correlation coefficients: 0.720–0.871) in a reader study.
4) FetMRIE further improves the accuracy and reliability of automated clinical analyses across three clinically critical downstream tasks, including gestational age prediction, anomaly detection, and biometry measurement.

2. Results

2.1 Quantitative Evaluation

2.1.1 uMSE and uPSNR

FetMRIE achieved the lowest uMSE on six out of eight datasets and the highest uPSNR on seven out of eight datasets (Table 2), demonstrating SOTA reference-free image fidelity with robust cross-domain generalizability (Figure 2, Figure 3). For internal validation, FetMRIE achieved the best uMSE ($\times 10^{-3}$) and uPSNR (dB) on WCT-Normal-Test (0.587 and 33.02), WCT-VM (0.435 and 34.06), and WCB (1.542 and 33.01), with the sole exception of WCT-GMH where BM4D yielded a lower uMSE (0.768 vs. 0.950) but a lower uPSNR (31.63 vs. 33.05). For external validation, FetMRIE consistently ranked first on LFC (0.813 and 31.49), CSH-Normal (0.489 and 33.68), and CSH-GMH (0.515 and 33.12), while on FeTA, Pixel2Pixel achieved a marginally lower uMSE (0.381 vs. 0.512) and higher uPSNR (34.94 vs. 33.37). Notably, FetMRIE maintained low variance across all datasets (e.g., uPSNR standard deviations ranging from 1.46 to 4.48 dB), indicating stable enhancement regardless of center, sequence, or pathology.

In contrast, competing methods exhibited unstable performance across datasets (Table 2). Specifically, while BM4D performed competitively on some datasets such as WCT-GMH and FeTA, it showed notable performance degradation on the remaining datasets. BME-X exhibited severe degradation on the WCB dataset, yielding the worst scores (uMSE [$\times 10^{-3}$]: 16.06×10^{-3}, uPSNR [dB]: 18.35). Pixel2Pixel achieved competitive results on FeTA but fell substantially behind FetMRIE on other datasets (e.g., uMSE [$\times 10^{-3}$]: 6.136 vs. 0.813 on LFC; uPSNR [dB]: 24.31 vs. 34.06 on WCT-VM).

2.1.2 Tissue Contrast T-score

FetMRIE achieved the highest average TCT on all eight datasets (Table 3), demonstrating SOTA tissue contrast enhancement with robust cross-domain generalizability (Figures 2 and 3). For internal validation, FetMRIE yielded the best average TCT on WCT-Normal-Test (1.134, +3.9% vs. original), WCT-VM (1.255, +1.9%), WCT-GMH (1.131, +3.7%), and WCB (1.115, +5.3%). For external validation, FetMRIE consistently ranked first on FeTA (1.251, +2.2%), LFC (1.165, +9.3%), CSH-Normal (1.250, +3.1%), and CSH-GMH (1.161, +2.0%).

In contrast, competing methods exhibited heterogeneous performance across tissue types that prevented them from achieving overall contrast improvement (Table 3). Specifically, BME-X generally achieved the highest TCT (WM vs. cGM) across most datasets (e.g., 1.558 on WCT-Normal-Test and 1.600 on WCT-VM) but consistently yielded the lowest TCT (WM vs. dGM) (e.g., 0.559 and 0.618, respectively), potentially because its three-category tissue classification conflated dGM with WM, driving both regions toward uniform intensity and thereby collapsing their contrast (Figure 2 and Figure 3). Conversely, BM4D tended to achieve the highest TCT (WM vs. dGM) on several datasets (e.g., 1.240 on WCB and 1.368 on LFC) but performed poorly on TCT (WM vs. cGM) (e.g., 0.883 and 0.885, respectively), as its collaborative filtering may introduce intensity distortions in cortical regions (Figure 3e, 3g).

2.1.3 Cross-Tissue Noise Consistency

FetMRIE achieved the lowest CNC on all eight datasets (Table 3), demonstrating SOTA performance in spatially uniform noise suppression with robust cross-domain generalizability (Figure 2 and Figure 3). For internal validation, FetMRIE yielded the best CNC on WCT-Normal-Test (0.075), WCT-VM (0.059), WCT-GMH (0.036), and WCB (0.019), representing reductions

of 81.8%, 79.8%, 83.6%, and 93.0% relative to the next-best competing method (BM4D), respectively. For external validation, FetMRIE consistently ranked first on FeTA (0.366), LFC (0.137), CSH-Normal (0.163), and CSH-GMH (0.064).

2.2 Reader Assessment

FetMRIE achieved the highest mean scores from radiologists across all three evaluated datasets (Table 4). Specifically, FetMRIE obtained mean scores of 3.72 on WCT-Normal-Test, 3.71 on WCB, and 3.81 on the external LFC dataset, with consistently high ratings from both junior and senior radiologists in noise suppression, anatomical fidelity, and overall image quality. Notably, BME-X consistently scored near 1.0 (the minimum score) across all datasets and evaluators, confirming the severe over-smoothing and structural alteration observed in the quantitative evaluations (Table 2, Table 3, Figure 2, Figure 3).

To validate the reliability of these subjective evaluations, pairwise Pearson's correlation coefficients were computed among the four radiologists (Figure 4). The aggregate correlation matrix showed pairwise Pearson's correlation coefficients ranging from 0.720 to 0.871 among the four radiologists, with consistently high agreement maintained across different experience levels (junior vs. senior) and all three evaluated datasets.

2.3 Downstream Tasks

2.3.1 Gestational Age Prediction

FetMRIE achieved the lowest AAD of 0.720±0.734 weeks on the WCT-Normal-Test dataset (Table 5), outperforming the original noisy images (0.773) and all competing methods, including BME-X (0.723 weeks), ZS-N2N (0.728 weeks), BM4D (0.730 weeks), and Pixel2Pixel (0.748

weeks). The lower prediction error reflects that FetMRIE faithfully preserves normative anatomical features crucial for assessing fetal neurodevelopment.

2.3.2 Anomaly Detection

FetMRIE achieved the highest AUROC and AUPR for both VM and GMH-IVH anomaly detection tasks (Table 5, Figure 5), demonstrating SOTA performance in pathological discrimination. For GMH-IVH detection, FetMRIE achieved an AUROC of 0.843, outperforming the baseline using the original images (AUROC: 0.828) and all competing methods, including Pixel2Pixel (AUROC: 0.774), ZS-N2N (AUROC: 0.758), BM4D (AUROC: 0.720), and BME-X (AUROC: 0.677). For VM detection, FetMRIE similarly yielded the highest AUROC of 0.757 and AUPR of 0.821, surpassing all competing methods and the baseline using the original images (AUROC: 0.716, AUPR: 0.788). In contrast, methods prone to over-smoothing exhibited markedly degraded anomaly detection performance (Table 5). Specifically, BME-X and BM4D yielded lower AADs on pathological cases compared to using the original images (e.g., BME-X: 1.355 vs. 2.077 on WCT-GMH; BM4D: 1.267 vs. 1.397 on WCT-VM), indicating that clinically meaningful pathological signatures were blurred or removed during enhancement, which directly translated into reduced AUROC values (e.g., BME-X: 0.677 and 0.651 for GMH and VM, respectively).

2.3.3 Biometry Measurement

FetMRIE achieved the lowest average MAE of 0.982 mm across the three measurements for TCD, VAD, and VH on the LFC dataset (Figure 6, Table 6), outperforming the baseline using the original noisy images (1.187 mm) and all competing methods, including BM4D (1.150 mm), ZS-N2N (1.178 mm), Pixel2Pixel (1.206 mm), and BME-X (1.261 mm). For each measurement, FetMRIE

yielded the best performance for TCD (1.295 mm) and VH (0.932 mm), demonstrating faithful preservation of structural boundaries critical for cerebellar and vermian morphometry.

In contrast, several competing enhancement methods yielded higher measurement errors than the baseline using the original images (Table 6). Specifically, BME-X exhibited the worst average MAE (1.261 mm), with degraded performance across all three measurements (e.g., TCD: 2.015 vs. 1.733 mm for the baseline), likely due to its over-smoothing, which blurs essential anatomical boundaries required for accurate landmark localization (Figure 6). Pixel2Pixel similarly produced elevated errors (1.206 mm), as its pixel-level denoising assumption may produce inaccurate edges that mislead the automated measurement model.

2.4 Ablation Study

The full FetMRIE model consistently achieved the best enhancement performance (uMSE and uPSNR) across all datasets (Table 7, Figure 7), demonstrating that both the diffusion-based structural restoration (Stage II) and the adaptive state-matching mechanism are essential for the final performance.

Removing the diffusion-based structural restoration (w/o Stage II) severely degraded enhancement performance, particularly on pathological and external datasets. Specifically, on the WCT-GMH dataset, ablating Stage II caused the uMSE ($\times 10^{-3}$) to increase from 0.950 to 19.68 and the uPSNR to drop from 33.05 dB to 22.50 dB sharply. A similar performance drop was observed on the external FeTA dataset (uMSE [$\times 10^{-3}$]: 0.512 vs. 3.503; uPSNR [dB]: 33.37 vs. 28.27).

Qualitatively, retaining only the VAE for denoising resulted in noticeable over-smoothing, which severely blurred the boundaries of clinically critical pathological lesions (Figure 7c).

Similarly, ablating the adaptive state-matching mechanism (w/o state matching) led to inferior quantitative metrics and a noticeable loss of fine anatomical details. For example, replacing the adaptive timestep estimation with a fixed denoising schedule on the WCB dataset increased the uMSE ($\times 10^{-3}$) from 1.542 to 3.368 and decreased the uPSNR from 33.01 dB to 32.18 dB. Qualitatively, omitting the state matching module compromised delicate structural features, such as fine cortical folds and cerebellar boundaries (Figure 7a, 7b).

3 Discussion

Our proposed FetMRIE enhancement framework does not require clean reference data, yet achieves uncompromised enhancement performance, effectively addressing the inherent scarcity of noise-free fetal brain MRI data and thereby substantially improving its clinical feasibility and accessibility. FetMRIE achieved SOTA results against four diverse competing methods spanning traditional filtering (BM4D), foundation-model-based enhancement (BME-X), and zero-shot denoising (ZS-N2N, Pixel2Pixel), on all four reference-free quantitative metrics (u-PSNR, u-MSE, TCT, and CNC) across eight multi-center datasets encompassing diverse scanners, sequences, and both normative and pathological conditions, and was consistently rated highest by four radiologists with strong inter-rater agreement (correlation coefficients: 0.720–0.871). More importantly, when integrated as a preprocessing step, FetMRIE consistently improved three clinically critical downstream tasks, achieving the lowest GA prediction error (AAD: 0.720 week), the highest anomaly detection AUROC (GMH-IVH: 0.843; VM: 0.757), and the lowest biometry MAE (0.982 mm).

The superior performance of FetMRIE stems from its two-stage design that decomposes the enhancement into noise suppression and structural restoration. Specifically, in Stage I, FetMRIE leverages the VAE's latent bottleneck to compress noisy inputs into a KL-regularized anatomical representation, filtering out dominant noise during reconstruction. Then, in Stage II, FetMRIE's diffusion model restores fine-grained structures (e.g., cortical folds and lesion boundaries) that are inevitably lost during Stage I compression, by learning to map the clean but over-smoothed VAE output to the structurally faithful yet noisy original image. The adaptive diffusion state matching mechanism further enables case-specific restoration by dynamically adjusting the reverse timestep

to each sample's noise level, naturally accommodating the heterogeneous degradation across subjects, protocols, and centers. The ablation study further confirmed the necessity of each component directly (Table 7, Figure 7). Removing Stage II reduced u-PSNR by 3.3%–46.9% across the eight datasets, with the most severe degradation on pathological cases and visible blurring of hemorrhagic lesion (Figure 7c). Disabling state matching similarly degraded u-PSNR by 2.0%–23.0%, with compromised cortical folds and cerebellar boundaries (Figure 7a, Figure 7b).

Compared to FetMRIE, competing methods exhibited distinct failure modes rooted in assumptions that are fundamentally violated in fetal brain MRI. Specifically, BM4D requires the grouping of strictly similar 3D patches for collaborative filtering, a prior frequently invalidated by spatially varying artifacts induced by fetal motion. Consequently, BM4D exhibited tissue-biased enhancement with u-PSNR 4%–37% lower than FetMRIE across datasets (Table 2, Table 3). BME-X's three-category tissue classification conflated WM and dGM toward uniform intensity, yielding 33%–49% lower TCT between WM and dGM (Table 3). More critically, BME-X erased pathological signatures such as hemorrhage (Figures 2e, 3c), which led to near-minimum radiologist scores (i.e., 1.0, Table 4). ZS-N2N and Pixel2Pixel both assume pixel-level independence of noise, which fundamentally mismatches the spatially correlated corruptions from fetal motion and SVR. As a result, ZS-N2N failed to effectively suppress noise, yielding substantially higher CNC values than FetMRIE (Table 3), whereas Pixel2Pixel exhibited significant degradation in both u-PSNR (7%–29% decrease) and CNC, accompanied by pronounced tissue-dependent residual artifacts (Table 2, Table 3). In summary, none of the

competing methods could simultaneously ensure robust noise suppression and preserve anatomical fidelity.

As fetal brain MRI becomes increasingly prevalent in prenatal diagnosis and developmental neuroscience research[1,23,24], ensuring high image quality has become clinically essential. Importantly, the clinical value of image enhancement should ultimately be judged not by visual improvement alone, but by its impact on downstream analytical tasks[15,20]. For example, a recent study demonstrated that noise, artifacts, and degraded tissue contrast systematically distort normative centile estimates, with the largest effects observed at clinically relevant outer centiles (1st–10th) used to detect atypical neurodevelopment[13]. This bias cannot be mitigated by increasing sample size, highlighting the importance of effective image enhancement as a preprocessing step. Accordingly, FetMRIE was evaluated across three clinically relevant downstream tasks, achieving consistent improvements that collectively demonstrate its practical clinical utility. Specifically, FetMRIE achieved the lowest GA prediction error, confirming that the visibility of rapidly evolving developmental features such as ventricular contours, parenchymal boundaries, and expanding GM/WM volumes[25] was effectively enhanced. For anomaly detection, FetMRIE attained the highest AUROC for both GMH-IVH and VM, indicating that it preserved pathological signatures that were erased by over-smoothing competitors. For automated biometry, FetMRIE achieved the lowest average MAE because of its superior boundary fidelity for precise landmark localization. For example, accurate TCD measurement relies on clear delineation of cerebellar hemisphere margins from the surrounding CSF, while VAD and VH depend on well-delineated vermis contours[26]. In contrast, over-smoothing methods (e.g., BME-X and Pixel2Pixel) blur these critical boundaries, increasing errors beyond the original baseline.

Beyond these specific improvements, the broader significance of FetMRIE lies in enabling automated, large-scale quantitative fetal brain analysis and prenatal diagnosis. Specifically, while deep learning-based quantitative morphometry and disease diagnosis have become well established in adult and neonatal neuroimaging[27–30], these applications remain underexplored in fetal MRI, largely because of heterogeneous image quality and low tissue contrast, which result in poorly delineated anatomical structures and thereby limit the performance of automated algorithms[3,12,31,32]. By faithfully restoring tissue contrast and boundary sharpness, FetMRIE is expected to benefit not only the biometry measurements validated in this study but also a wider range of boundary-sensitive applications, such as automated tissue segmentation, cortical surface reconstruction, and volumetric parcellation. By preserving pathological signatures, as demonstrated in VM (morphological alterations) and GMH-IVH (regional intensity variations), FetMRIE may be applicable to the diagnosis of other fetal pathologies such as fetal growth restriction (FGR)[33], Dandy-Walker malformation (DWM)[34], and agenesis of the corpus callosum (ACC)[35]. Crucially, the practical barriers and costs to realizing these applications are minimal. As a post-processing step, FetMRIE can be seamlessly integrated into established clinical and research workflows without requiring modification of existing acquisition protocols or SVR pipelines. It may also reduce the need for repeat scans due to suboptimal image quality, thereby shortening scan time, reducing maternal discomfort, and improving scanning throughput.

The deployment flexibility and high scalability of FetMRIE are enabled by its unsupervised learning paradigm. Unlike supervised methods (e.g., DUNCUN[14]) that rely on unattainable noise-free reference data, or zero-shot self-supervised approaches (e.g., ZS-N2N, Pixel2Pixel, and

Self2Self) that require computationally expensive per-image optimization, FetMRIE requires only original noisy images for training to achieve efficient inference across different cohorts without any scanner-specific fine-tuning. This cross-domain robustness indicates that FetMRIE can be retrospectively applied to previously acquired datasets in multi-site repositories (e.g., dHCP[36] and FeTA challenges[3,11]) to standardize image quality without requiring re-scanning or re-acquisition. Moreover, FetMRIE holds significant promise for resource-limited areas where dedicated fetal MRI expertise and optimized acquisition protocols are scarce, thereby expanding access to reliable quantitative fetal brain assessment. Most importantly, the proposed unsupervised framework is not limited to T2-weighted fetal brain MRI. Future extensions to other contrasts (e.g., T1-weighted and diffusion-weighted imaging) and broader demographics (neonatal, pediatric, and adult) would position FetMRIE as a generalizable MRI enhancement foundation model across the lifespan.

There are several limitations of this study. First, although validation data from eight datasets across three clinical centers and four sequence types demonstrate the generalizability of FetMRIE (Table 1), the model was trained exclusively on data from a single institution. Future work will explore multi-center training and domain adaptation strategies to further enhance cross-site robustness and clinical impact. Second, the two-stage architecture involving sequential VAE and diffusion-based processing of overlapping 3D patches introduces considerable computational cost. Future work will explore more advanced and lightweight neural network architectures, as well as optimized inference strategies, to improve computational efficiency for clinical deployment.

In conclusion, we propose FetMRIE and demonstrate its effectiveness for unsupervised fetal brain MRI enhancement without clean reference data. FetMRIE achieves state-of-the-art performance

across comprehensive quantitative metrics (uMSE, uPSNR, TCT, and CNC) and in radiologist evaluations, effectively balancing noise suppression and anatomical fidelity while eliminating the need for noise-free training data. Furthermore, it significantly improves the accuracy and robustness of automated downstream tasks, including gestational age prediction, pathological anomaly detection, and biometry measurement, highlighting its clinical value in both routine practice and large-scale research. Overall, FetMRIE provides a robust framework for automated fetal brain MRI preprocessing that can be readily adopted in clinical and research workflows.

4 Methods

4.1 FetMRIE Pipeline

Given a noisy input 3D fetal brain MRI volume $x \in R^{D\times H\times W}$ corrupted by acquisition and SVR noise ϵ (i.e., $x = s + \epsilon$), where s is the targeted noise-free image, FetMRIE enhances x through a two-stage framework. (1) Stage I: A variational autoencoder (VAE)[37] compresses x into a regularized latent representation, producing an initial denoised estimate s_0[38,39]. (2) Stage II: A denoising diffusion model (DDPM) [30,40] refines s_0 into the final enhanced output $\hat{s}$ via adaptive state-matching conditional generation, restoring anatomical structures while preserving tissue contrast.

4.1.1 Stage I: VAE Initial Denoising

VAEs have been explored for fully unsupervised image denoising, where the KL-regularized latent bottleneck provides a rudimentary ability to separate signal from noise, even without an explicit noise model or clean reference data [39]. Our recent study further showed that, for unsupervised fetal brain MRI enhancement, a 3D convolutional VAE achieves superior performance in tissue contrast preservation and lesion fidelity compared to BME-X[38]. Motivated by these findings, a VAE was adopted in this study as the first stage of FetMRIE (Figure 1a), where latent space regularization compels the encoder to learn a compact manifold of brain anatomy, initially filtering stochastic noise during reconstruction.

Specifically, the VAE consists of an encoder $q_\varphi(z|x)$ that maps the input to a latent distribution $\mathcal{N}(\mu, \sigma^2)$, and a decoder $p_\theta(s_0|z)$ that reconstructs the denoised volume. The latent variable z

is sampled via the reparameterization strategy: $z = \mu + \sigma \cdot \epsilon'$, where $\epsilon' \sim \mathcal{N}(0, I)$. The training objective combines reconstruction fidelity, perceptual consistency, and latent regularization:

$$\mathcal{L}_{stage1} = \|s_0 - x\|_1 + \lambda_p \mathcal{L}_{perceptual}(s_0, x) + \lambda_{KL} KL(\, q_\varphi(z|x) \| \mathcal{N}(0, I)),$$

where $\mathcal{L}_{perceptual}$ is computed via a pretrained 3D SqueezeNet encoder to enforce structural consistency beyond pixel-wise metrics, with $\lambda_p = 10^{-3}$ and $\lambda_{KL} = 1.0$.

4.1.2 Stage II: Diffusion-Based Structural Restoration

Although the VAE effectively reduces noise through latent space regularization, the compression process may lead to over-smoothing of delicate anatomical details, such as cortical folds and ventricular boundaries. Accordingly, a subsequent diffusion-based refinement stage (Figure 1a) is introduced for recovering these fine-grained structures while preserving structural fidelity.

To this end, a denoising diffusion probabilistic model (DDPM) is employed to restore anatomical features that are degraded during the VAE compression process. In contrast to standard DDPMs, which simulate the complete forward trajectory from clean data to Gaussian noise over $T = 1000$ timesteps, the proposed method leverages the prior information obtained in Stage I to operate within a substantially reduced noise regime. Specifically, the diffusion model is trained only on early timesteps corresponding to the residual noise level after VAE denoising, while during inference, an adaptive state-matching mechanism dynamically determines the reverse starting point for each test sample, enabling case-specific restoration.

(1) Unconditional Training on Early Timesteps: To determine the effective noise range for training, the residual noise $\epsilon_{\text{res}} = x - s_0$ is analyzed after obtaining s_0 from the VAE. The standard

deviation $\sigma_{\text{res}} = \text{std}(\epsilon_{\text{res}})$ is then aligned with a corresponding diffusion timestep according to the noise schedule coefficients:

$$t^* = \arg\min_{t\in[1,T]} \left|\sqrt{1-\bar{\alpha}_t} - \sigma_{\text{res}}\right|,$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ denotes the cumulative product of the variance schedule, with $\beta_1 = 10^{-4}$ and $\beta_{1000} = 1.95 \times 10^{-2}$ defined by a scaled linear schedule. By evaluating σ_{res} for all samples in the training dataset and taking the maximum value, σ_{res} is observed to be bounded by approximately 0.40, corresponding to an effective timestep upper bound of $T_{max} \approx 300$. This finding suggests that s_0 already lies in a partially denoised regime, such that the diffusion model only needs to handle moderate noise levels. Accordingly, the UNet ϵ_θ is trained solely on timesteps $t \in [1, T_{max}]$, rather than the complete interval $[1, 1000]$. The training objective follows the standard DDPM formulation:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t\sim\mathcal{U}(1,T_{max}),\epsilon\sim\mathcal{N}(0,I)}[\left\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}s_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) - \epsilon\right\|_2^2],$$

where s_0 is the pseudo ground truth and ϵ is sampled from a standard Gaussian.

(2) Conditional Inference via State Matching. During inference, the diffusion model performs conditional denoising, where the reverse process is adjusted according to the noise level of each individual sample instead of using a fixed denoising schedule. Specifically, for a test input x, the VAE first generates s_0, after which the corresponding timestep t^* is estimated according to the nearest timestep. The reverse diffusion is then executed from t^* down to 0 (rather than starting from T_{max} for all inputs), following the DDPM posterior sampling:

$$s_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(s_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(s_t, t)\right) + \sigma_t w\,,$$

where $w \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, $\sigma_t^2 = \beta_t(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)$ represents the posterior variance, and the initial sample is set to $s_{t^*} = x$. This adaptive strategy enables samples with lower residual noise to undergo fewer diffusion steps, while inputs with higher noise levels require a longer reverse trajectory, thereby achieving case-specific restoration.

4.1.3 3D Volumetric Processing and Aggregation

To process full 3D fetal brain MRI volumes, FetMRIE partitions each input into overlapping patches of 64^3 voxels, which are sequentially processed by the VAE (Stage I) and the adaptive diffusion model (Stage II). Finally, the restored patches are spatially aggregated to reconstruct the complete enhanced volume, where predictions within the overlapping regions are averaged to ensure smooth transitions.

4.2 Experiment Dataset

To comprehensively evaluate FetMRIE, the largest and most diverse fetal brain MRI benchmark to date was retrospectively assembled, comprising six private datasets from three independent clinical centers and two well-established public datasets, totaling 2,481 cases (Table 1, Figure 1b).

4.2.1 Private Datasets

For this retrospective study, the institutional review boards (IRBs) of West China Second University Hospital (WCSUH) of Sichuan University, Chengdu Women's and Children's Central Hospital (CWCCH), and Sichuan Provincial Woman's and Children's Hospital (SPWCH) approved the use and analyses of fetal MRI data, respectively. Because of the retrospective design of the study, the requirement for informed consent was waived.

Data used for model training and internal validation were acquired from pregnant women referred to the Radiology Department of WCSUH for fetal MRI examinations. For each fetus, 2D T_2-weighted stacks were acquired in axial, coronal, and sagittal orientations using single-shot turbo spin-echo (SS-TSE) sequences (i.e., ss-FSE, TSE-SSH, and HASTE) on both 1.5-T and 3.0-T scanners. The dataset, hereafter referred to as WCT, includes 1,316 fetuses with typical development (WCT-Normal), 385 fetuses diagnosed with ventriculomegaly (WCT-VM), and 52 fetuses diagnosed with germinal matrix-intraventricular hemorrhage (WCT-GMH). Additionally, to evaluate the out-of-distribution (OoD) generalization capability of FetMRIE across different acquisition protocols, an additional dataset (hereafter referred to as WCB, 516 cases) was collected from the same institution using balanced turbo field echo (B-TFE) sequences on a 1.5-T scanner.

For external validation, data were acquired at CWCCH and SPWCH on 1.5-T or 3.0-T scanners using TSE-SSH or ss-FSE or HASTE sequences, comprising 53 fetuses with typical development (hereafter referred to as CSH-Normal) and 17 fetuses with GMH-IVH (hereafter referred to as CSH-GMH).

4.2.2 Public Dataset

To further demonstrate the cross-domain generalization of FetMRIE, two publicly accessible fetal brain datasets were used for additional external validation, including the Fetal Tissue Annotation (FeTA) Challenge dataset[11] and the LFC dataset from Shenzhen Baoan Women's and Children's Hospital[26]. Both datasets were acquired using ss-FSE sequences on 1.5-T and/or 3.0-T scanners. An expert radiologist (K.A.) reviewed all candidate cases and excluded those exhibiting extreme, unrecoverable signal degradation or severe artifacts that completely obscured underlying

anatomical structures, following the quality control standard provided by FetMRQC[10]. After this filtering, 48 FeTA cases and 94 LFC cases were retained for evaluation.

4.2.3 Dataset Splitting

The WCT-Normal dataset (n = 1,316) was randomly divided at the subject level into a training set (WCT-Normal-Train, n = 543) for model development and a held-out test dataset (WCT-Normal-Test, n = 773) for internal validation, together with the WCT-VM (n = 385), WCT-GMH (n = 52), and WCB (n = 516) datasets from the same institution. The external datasets, including CSH-Normal (n = 53), CSH-GMH (n = 17), FeTA (n = 48), and LFC (n = 94), were used exclusively for external validation.

4.3 Dataset Preprocessing

For raw 2D orthogonal stacks from the private datasets, SVR was performed using NiftyMIC [7] for the WCT-GMH dataset and NesVoR[8] for other datasets (Table 1), yielding super-resolved volumetric brain volumes with a spatial resolution of 0.8 × 0.8 × 0.8 mm^3. To ensure spatial consistency across all cohorts, the reconstructed volumes from the public FeTA and LFC datasets were subsequently resampled to the same isotropic resolution using trilinear interpolation.

4.4 Comparison Methods.

FetMRIE was compared against four SOTA denoising and enhancement methods, which do not require clean reference data:

(1) BME-X[15]: A foundation model for MR image enhancement and improved downstream tasks such as segmentation. A publicly available Docker container was directly applied for inference.

(2) BM4D[41]: A training-free block-matching collaborative filtering method for 3D volumetric denoising.

(3) ZS-N2N[18]: A zero-shot Noise2Noise approach that performs single-image blind denoising without any training dataset.

(4) Pixel2Pixel[19]: A pixel-level self-supervised denoising method leveraging masked reconstruction for zero-shot image enhancement.

4.5 Downstream Tasks

To further validate the clinical utility of FetMRIE, three downstream tasks, including gestational age (GA) prediction, anomaly detection, and biometry measurement, were established to compare the enhanced results of FetMRIE with those of the competing methods (Figure 1 c-e).

4.5.1 Gestational Age Prediction

GA is a crucial metric for assessing fetal neurodevelopment, and accurate automated GA prediction relies heavily on the high-quality anatomical detail of fetal brain MRI [42]. To evaluate the impact of image enhancement on normative neurodevelopmental assessment, a standard single-slice GA-prediction baseline[25] was trained and evaluated under strictly identical experimental configurations (Section 2.9), where the sole independent variable was the enhancement method applied to the input images. For model training, 70% of the WCT-Normal-Test (n = 541) was used, whereas the remaining 30% WCT-Normal-Test (n = 232) was used for testing. Predictive accuracy was quantified using the Absolute Age Difference (AAD) between the predicted and biological GA (Figure 1c).

4.5.2 Anomaly Detection

GA prediction is widely used for fetal brain anomaly detection[25,43,44], as pathological conditions disrupt typical brain development trajectories, resulting in pronounced deviations in age estimation. To verify whether different enhancement methods can faithfully preserve clinically meaningful pathological signatures, the pre-trained GA prediction model (Section 4.5.1) was used to extract the predicted GA of pathological cases in the WCT-VM and WCT-GMH datasets. Subsequently, AAD was used as the anomaly score to distinguish normal fetuses (WCT-Normal-Test) from pathological ones. The anomaly detection performance was measured by the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR), with 95% confidence intervals (CIs) calculated via bootstrapping with 1000 resamples (Figure 1d).

4.5.3 Biometry Measurement

Accurate biometric measurement is essential for assessing fetal growth and detecting morphological abnormalities, and it heavily depends on the clarity of structural boundaries in MRI[45]. To evaluate whether the enhancement algorithms could improve the reliability of clinical measurements, automated fetal brain biometry was used as a downstream quantitative proxy. Specifically, the public LFC dataset, which provides GT annotations for Transverse Cerebellar Diameter (TCD), Vermis Anteroposterior Diameter (VAD), and Vermis Height (VH), was split at a 7:3 training-testing ratio. For each enhancement method, a separate 3D Spatial Configuration Network (3D SCN) landmark detection model[46,47] was trained and evaluated under otherwise identical configurations (Section 2.9). The performance was quantified by the Mean Absolute Error (MAE) between predicted measurements and GT, and lower MAE values indicate clearer structural edges and superior enhancement fidelity (Figure 1e).

4.6 Ablation Study

To evaluate the specific contribution of each architectural component in FetMRIE, a comprehensive ablation study was conducted. Two ablation configurations were examined:

(1) Stage II (diffusion-based structural restoration) was removed, retaining only the Stage I VAE for initial denoising, to assess the necessity of diffusion-based refinement in recovering anatomical details that may be over-smoothed during VAE compression.

(2) The adaptive state-matching mechanism was disabled during inference, replacing the sample-specific dynamic timestep estimation with a fixed denoising schedule starting from the maximum timestep (T_{max}), to evaluate the effectiveness of the proposed adaptive state-matching strategy in achieving case-specific restoration.

4.7 Evaluation Metrics for Unsupervised Enhancement

Evaluating unsupervised enhancement methods is challenging due to the absence of GT images, which precludes the computation of conventional supervised metrics such as MSE, PSNR, and structural similarity (SSIM). To solve the problem, four complementary reference-free metrics were used, including uMSE and uPSNR as unbiased surrogates for traditional reconstruction fidelity measures[48], TCT for quantifying inter-tissue contrast improvement[15,22], and our proposed cross-tissue noise consistency (CNC) for assessing the spatial uniformity of noise suppression across anatomical regions.

4.7.1 Unsupervised MSE and PSNR

The uMSE is an unbiased estimator of the supervised MSE, calculated from noisy images via spatial subsampling[48]. Given a 2D noisy image $x \in R^{D\times H}$ and its denoised output y, x is first

subsampled into four 2×2 sub-images $a, b, c, d \in R^{D/2 \times H/2}$ following the strategy of Neighbor2Neighbor[49], and y is correspondingly subsampled to yield $\hat{y}$. The uMSE is then computed as:

$$\text{uMSE} = \frac{1}{n}\sum_i \left[(a_i - \hat{y}_i)^2 - \frac{1}{2}(b_i - c_i)^2\right],$$

where $n = \frac{D}{2} \times \frac{H}{2}$ is the number of pixels in each sub-image and the second term serves as a correction that removes the noise-induced bias, rendering uMSE an asymptotically consistent estimator of the supervised MSE. The uPSNR is subsequently derived as:

$$\text{uPSNR} = 10 \lg\left(\frac{M^2}{\text{uMSE}}\right),$$

where M is the maximum possible value of the signal of interest.

Notably, ZS-N2N was excluded from this evaluation, as its outputs are nearly identical to the noisy inputs (Figures 2 and 3), causing the signal and bias-correction terms in uMSE to be nearly equal and resulting in near-zero or negative uMSE values (thereby precluding uPSNR calculation), which reflect minimal denoising rather than superior performance.

4.7.2 Tissue Contrast T-score

The tissue contrast t-score (TCT) [15,22] is widely used for quantifying the separability between distinct tissue classes relative to their intra-tissue intensity variability:

$$\mathrm{TCT} = \frac{|\mu_{GM} - \mu_{WM}|}{\sqrt{\sigma_{GM}^2 + \sigma_{WM}^2}},$$

where μ_{GM}, μ_{WM} and σ_{GM}^2, σ_{WM}^2 are the means and variances of GM, including cortical GM (cGM) and deep GM (dGM), and WM intensities. A higher TCT after enhancement indicates improved inter-tissue contrast and reduced intra-tissue noise. In this work, tissue segmentation masks required for TCT computation were obtained by applying a trustworthy AI model [50] to each fetal brain volume.

4.7.3 Cross-Tissue Noise Consistency

The Cross-Tissue Noise Consistency (CNC) was proposed to assess whether an enhancement model achieves spatially consistent noise suppression across anatomical regions. Since the noise of MRI should be independent of tissue regions [51], an ideal enhancement algorithm should yield statistically uniform residuals (defined as the voxel-wise differences between the input and the enhanced output) across all tissue regions. Specifically, the CNC is defined as the average pairwise Kullback-Leibler (KL) divergence among all tissue compartments:

$$\mathrm{CNC} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i}^{M} \mathrm{KL}\left(\mathrm{Tissue}_i, \mathrm{Tissue}_j\right),$$

where *M* is the number of tissue types. The KL divergence between two tissue types is:

$$\mathrm{KL}\left(\mathrm{Tissue}_i, \mathrm{Tissue}_j\right) = \frac{1}{2}\left[\log \frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2 + (\mu_i - \mu_j)^2}{\sigma_j^2} - 1\right],$$

where μ_i, μ_i and σ_i, σ_j are the means and variances of residual values within tissue type *i* and *j*, respectively. Tissue segmentation masks required for compartment-wise analysis were also

obtained using the trustworthy AI model[50]. A lower CNC value indicates more spatially uniform noise suppression across anatomical structures, reflecting higher enhancement performance.

4.8 Reader Assessment

To further evaluate FetMRIE's clinical value, a comprehensive reader assessment was conducted. Four radiologists (two junior and two senior) independently scored the enhanced image quality from all methods (excluding ZS-N2N due to outputs nearly indistinguishable from the inputs as shown in Figures 2 and 3) on three datasets (WCT-Normal-Test, WCB, and LFC), with 50 cases randomly selected from each dataset. For each case, the original noisy volume was provided alongside the enhanced outputs, enabling the radiologists to assess whether the enhancement introduced subtle or severe alterations to anatomical structures. Following the quality control protocol of $FetMRQC_{SR}$[9], each radiologist assigned a score ranging from 1 (poorest) to 4 (best) based on noise suppression, anatomical fidelity, and overall image quality. Additionally, inter-rater agreement was quantified by computing pairwise Pearson's correlation coefficients (CCs) among the four radiologists.

4.9 Model Training Details

The proposed FetMRIE model was trained on four NVIDIA A800 GPUs. Each brain volume was divided into 8-12 overlapping 64×64×64 patches, which were assembled after enhancement. The AdamW optimizer used a learning rate of 10^{-4}, betas of 0.9 and 0.999, epsilon set to 10^{-8}, a weight decay of 0.01, and a cosine annealing learning rate scheduler. The network was trained for 200 epochs with a batch size of 4. Additionally, mixed precision training was employed with a gradient scaler to prevent gradient underflow under FP16, improving training speed and reducing memory usage.

The ResNet101 model for downstream GA prediction and anomaly detection tasks was trained on an NVIDIA A800 GPU. All input images were resized to 128×192 pixels. The AdamW optimizer used a learning rate of 10^{-4}, betas of 0.9 and 0.999, epsilon set to 10^{-8}, and a weight decay of 10^{-5}. The networks were trained for 500 epochs with a batch size of 32.

The 3D SCN model for the downstream biometry measurement task was trained on an NVIDIA A800 GPU. The AdamW optimizer used a learning rate of 10^{-4}, betas of 0.9 and 0.999, a weight decay of 0.01, and a cosine annealing learning rate scheduler. The network was trained for 250 epochs with a batch size of 2. MSE loss was employed as the loss function, with target heatmaps generated using a Gaussian filter with a kernel size of 7 and a standard deviation of 3.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 82572198), the Scientific Research Project of Sichuan Medical Association (Grant No. 2024HR130), Science and Technology Department of Sichuan Province (Grant No. 25SYSX0255), Tsinghua University Startup Fund, and Tsinghua University Dushi Program (Grant No. 20241080026, 20251080056).

Data availability

Raw fetal brain MRI data cannot be shared due to the patient privacy constraints. Public fetal brain MRI datasets are available: FeTA Dataset (Only Training sets are available, testing sets are unavailable: (https://fetachallenge.github.io/pages/Data_download); LFC Dataset: (https://drive.google.com/drive/folders/1PVOi5dX4SSjbZ4TqjRPZhfewj7fgFY9p) and (https://github.com/lhaof/LFC).

Code availability

The source codes of FetMRIE are publicly available (https://github.com/yingqihao2022/Fetal_MRI_Image_Enhancement-). The source codes of BME-X are publicly available (https://github.com/DBC-Lab/Brain_MRI_Enhancement). The source codes of ZS-N2N are public available (https://github.com/herryi/ZS-N2N). The source codes of Pixel2Pixel are publicly available (https://github.com/qingma2016/Pixel2Pixel).

Reference

1. Griffiths, P. D. *et al.* Use of MRI in the diagnosis of fetal brain abnormalities in utero (MERIDIAN): a multicentre, prospective cohort study. *The Lancet* **389**, 538–546 (2017).
2. Liu, M. *et al.* Quality-label-free fetal brain MRI quality control based on image orientation recognition uncertainty. *Medical Image Analysis* **110**, 103994 (2026).
3. Payette, K. *et al.* Multi-Center Fetal Brain Tissue Annotation (FeTA) Challenge 2022 Results. *IEEE Trans. Med. Imaging* **44**, 1257–1272 (2025).
4. Siala, S. & Guimaraes, C. V. Fundamentals of fetal brain MRI: indications, technique, and normal anatomy. *Pediatr Radiol* https://doi.org/10.1007/s00247-025-06240-3 (2025) doi:10.1007/s00247-025-06240-3.
5. Gholipour, A., Estroff, J. A., Barnewolt, C. E., Connolly, S. A. & Warfield, S. K. Fetal brain volumetry through MRI volumetric reconstruction and segmentation. *Int J CARS* **6**, 329–339 (2011).
6. Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M. A., Hajnal, J. V. & Schnabel, J. A. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical Image Analysis* **16**, 1550–1564 (2012).
7. Ebner, M. *et al.* An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI. *NeuroImage* **206**, 116324 (2020).
8. Xu, J. *et al.* NeSVoR: Implicit Neural Representation for Slice-to-Volume Reconstruction in MRI. *IEEE Trans. Med. Imaging* **42**, 1707–1719 (2023).
9. Sanchez, T. *et al.* Automatic quality control in multi-centric fetal brain MRI super-resolution reconstruction. Preprint at https://doi.org/10.48550/arXiv.2503.10156 (2025).
10. Sanchez, T. *et al.* FetMRQC: A robust quality control system for multi-centric fetal brain MRI. *Medical Image Analysis* **97**, 103282 (2024).
11. Zalevskyi, V. *et al.* Advances in automated fetal brain MRI segmentation and biometry: Insights from the FeTA 2024 challenge. *Medical Image Analysis* **109**, 103941 (2026).

12. Li, H. *et al.* FetalCSR: Multi-input Attention Fusion Network for Neural ODE-based Fetal Cortical Surface Reconstruction. in *ICLR 2025 Workshop AI4CHL* (Singapore, Singapore, 2025).
13. Sanchez, T. *et al.* Data quality biases normative models derived from fetal brain MRI. *bioRxiv* 2026.01.22.700996 (2026).
14. Liu, S. *et al.* Learning MRI artefact removal with unpaired data. *Nat Mach Intell* **3**, 60–67 (2021).
15. Sun, Y., Wang, L., Gang Li, Lin, W. & Wang, L. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nat. Biomed. Eng* **9**, 521–538 (2024).
16. Zhang, L. *et al.* Iterative learning for joint image denoising and motion artifact correction of 3D brain MRI. *Medical Image Analysis* **108**, 103872 (2026).
17. Quan, Y., Chen, M., Pang, T. & Ji, H. Self2self with dropout: Learning self-supervised denoising from single image. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 1890–1898 (2020).
18. Mansour, Y. & Heckel, R. Zero-Shot Noise2Noise: Efficient Image Denoising without any Data. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14018–14027 (IEEE, Vancouver, BC, Canada, 2023). doi:10.1109/CVPR52729.2023.01347.
19. Ma, Q. *et al.* Pixel2Pixel: A Pixelwise Approach for Zero-Shot Single Image Denoising. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**, 4614–4629 (2025).
20. Li, Z. *et al.* Noise2Average: an iterative residual learning strategy for image denoising without clean data. *Imaging Neuroscience* https://doi.org/10.1162/IMAG.a.1163 (2026) doi:10.1162/IMAG.a.1163.
21. Zhu, P., Liu, S., Liu, Y. & Yap, P.-T. METER: Multi-task efficient transformer for no-reference image quality assessment. *Appl Intell* **53**, 29974–29990 (2023).
22. Duffy, B. A., Zhang, W., Tang, H. & Zhao, L. Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion. *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)* https:// openreview.net/forum?id=H1hWfZnjM (2018).

23. Ljubic, A. *et al.* Ultrasound vs MRI in Diagnosis of Fetal and Maternal Complications. *Donald School Journal of Ultrasound in Obstetrics and Gynecology* **5**, 231–242 (2011).
24. Van Ooijen, I. M. *et al.* Head to Head: Comparing Intracranial Volumes using 3D Ultrasound and Fetal MRI. *NeuroImage* 121633 (2025) doi:10.1016/j.neuroimage.2025.121633.
25. Shi, W. *et al.* Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. *NeuroImage* **223**, 117316 (2020).
26. Gong, H. *et al.* Fetal Cerebellum Landmark Detection Based on 3D MRI: Method and Benchmark. *IEEE J. Biomed. Health Inform.* **29**, 5712–5721 (2025).
27. Li, Z. *et al.* Diffusion MRI data analysis assisted by deep learning synthesized anatomical images (DeepAnat). *Medical Image Analysis* **86**, 102744 (2023).
28. Tian, Q. *et al.* DeepDTI: High-fidelity six-direction diffusion tensor imaging using deep learning. *NeuroImage* **219**, 117017 (2020).
29. Wu, C. *et al.* Towards practical Alzheimer's Disease diagnosis: A lightweight and interpretable spiking neural model. *Biomedical Signal Processing and Control* **117**, 109595 (2026).
30. Zhu, P., Liu, C., Fu, Y., Chen, N. & Qiu, A. Cycle-conditional diffusion model for noise correction of diffusion-weighted images using unpaired data. *Medical Image Analysis* **103**, 103579 (2025).
31. Dong, X. *et al.* Cycle-Consistent Learning for Fetal Cortical Surface Reconstruction. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds Linguraru, M. G. et al.) 212–222 (Springer Nature Switzerland, Cham, 2024). doi:10.1007/978-3-031-72104-5_21.
32. Lee, J., Bae, W., Ji, U.-G., Yang, H. & Lee, J.-M. Fetal and Neonatal Cortical Surface Reconstruction with Anatomical Normal-guidance and Perceptual Enhancements. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 7759–7768 (2026).
33. Nardozza, L. M. M. *et al.* Fetal growth restriction: current knowledge. *Archives of Gynecology and Obstetrics* **295**, 1061–1077 (2017).
34. Akiyama, S. *et al.* Regional brain development in fetuses with Dandy-Walker malformation: A volumetric fetal brain magnetic resonance imaging study. *PLOS ONE* **17**, 1–19 (2022).

35. Santo, S. *et al.* Counseling in fetal medicine: agenesis of the corpus callosum. *Ultrasound in Obstetrics & Gynecology* **40**, 513–521 (2012).
36. Edwards, A. D. *et al.* The Developing Human Connectome Project Neonatal Data Release. *Front. Neurosci.* **16**, 886772 (2022).
37. Kingma, D. P. & Welling, M. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning* **12**, 307–392 (2019).
38. Hao, Y. *et al.* Comprehensive Evaluation of Unsupervised Image Enhancement for Volumetric Fetal Brain MRI. *Medical Imaging with Deep Learning (2025)* https://openreview.net/pdf/ce0e28fc1763d4bce5dd1bffb47fdd186307befc.pdf (2025).
39. Prakash, M., Krull, A. & Jug, F. Fully Unsupervised Diversity Denoising with Convolutional Variational Autoencoders. in *International Conference on Learning Representations* (2021).
40. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at https://doi.org/10.48550/arXiv.2006.11239 (2020).
41. Maggioni, M., Katkovnik, V., Egiazarian, K. & Foi, A. Nonlocal Transform-Domain Filter for Volumetric Data Denoising and Reconstruction. *IEEE Trans. on Image Process.* **22**, 119–133 (2013).
42. Shen, L. *et al.* Attention-guided deep learning for gestational age prediction using fetal brain MRI. *Scientific Reports* **12**, 1408 (2022).
43. You, S. *et al.* Conditional deep generative normative modeling for structural and developmental anomaly detection in the fetal brain. *Neuroimage* **319**, 121442 (2025).
44. Yun, H. J. *et al.* Deep Learning–based Brain Age Prediction Using MRI to Identify Fetuses with Cerebral Ventriculomegaly. *Radiology: Artificial Intelligence* https://doi.org/10.1148/ryai.240115 (2025) doi:10.1148/ryai.240115.
45. Gafner, M. *et al.* Fetal Brain Biometry: Is there an Agreement among Ultrasound, MRI and the Measurements at Birth? *European Journal of Radiology* **133**, 109369 (2020).
46. Li, Y. *et al.* Anatomy-guided Test-Time Adaptation for Automated Fetal Brain MRI Morphometry. in *Medical Imaging with Deep Learning - Short Papers* (2025).

47. Payer, C., Štern, D., Bischof, H. & Urschler, M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical Image Analysis* **54**, 207–219 (2019).
48. Morales, A. M. *et al.* Evaluating unsupervised denoising requires unsupervised metrics. in *Proceedings of the 40th International Conference on Machine Learning* (JMLR.org, 2023).
49. Huang, T., Li, S., Jia, X., Lu, H. & Liu, J. Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14776–14785 (IEEE, Nashville, TN, USA, 2021). doi:10.1109/CVPR46437.2021.01454.
50. Fidon, L. *et al.* A Dempster-Shafer Approach to Trustworthy AI With Application to Fetal Brain MRI Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 3784–3795 (2024).
51. Pfaff, L. *et al.* Self-supervised MRI denoising: leveraging Stein's unbiased risk estimator and spatially resolved noise maps. *Sci Rep* **13**, 22629 (2023).

Table 1. Dataset demographics. Eight fetal brain MRI datasets are used in this study, totaling 2,481 cases across six private datasets from three clinical centers and two public datasets. Gestational age is reported as mean ± standard deviation in weeks. NA: not available.

Dataset		Number	Reconstruction Method	Gestation Age	Sequence Type	Scanner Type
WCT	**Normal**	1316	NesVoR	31.09±3.64	ss-FSE and TSE-SSH and HASTE	1.5-T and 3.0-T
	GMH-IVH	52	NiftyMIC	28.39±3.34		
	VM	385	NesVoR	30.27±4.45		
WCB		516	NesVoR	30.92±3.60	B-TFE	1.5-T
LFC		94	NesVoR	31.24±3.45	ss-FSE	1.5-T
FeTA		48	MIALSRTK/IRTK-Simple	NA	ss-FSE	1.5-T and 3.0-T
CSH	**Normal**	53	NesVoR	30.12±3.40	TSE-SSH and ss-FSE and HASTE	1.5-T and 3.0-T
	GMH-IVH	17	NesVoR	28.61±3.19		

Table 2. Unsupervised image fidelity metrics. uMSE ($\times 10^{-3}$; lower is better) and uPSNR (in dB; higher is better) are reported as mean ± standard deviation. The best result for each metric is highlighted in red. ZS-N2N is excluded because its outputs are nearly identical to the noisy inputs, resulting in degenerate uMSE values that preclude meaningful comparison (Section 2.7.1).

Methods	WCT-Normal-Test	WCT-VM	WCT-GMH	WCB	FeTA	LFC	CSH-Normal	CSH-GMH
	Internal Validation				External Validation			
uMSE ($\times 10^{-3}$)								
BM4D	1.736±1.136	6.234±3.151	0.768±0.353	1.638±0.958	0.513±0.252	11.02±3.983	2.597±1.407	3.449±1.964
BME-X	2.418±1.589	3.216±1.909	2.976±1.973	16.06±7.067	1.773±1.078	2.495±1.455	2.616±1.553	3.101±1.454
Pixel2Pixel	1.110±0.829	4.220±2.096	2.074±2.649	4.037±3.029	0.381±0.251	6.136±2.572	2.155±1.099	2.151±0.903
FetMRIE	0.587±0.788	0.435±0.199	0.950±1.320	1.542±0.633	0.512±0.235	0.813±0.434	0.489±0.258	0.515±0.163
uPSNR (dB)								
BM4D	28.42±2.68	22.68±2.50	31.63±2.10	28.55±2.50	33.43±2.16	19.90±1.75	26.56±2.62	25.23±2.23
BME-X	26.89±2.40	25.59±2.38	26.09±2.64	18.35±1.90	28.18±2.31	26.64±2.27	26.52±2.45	25.58±2.12
Pixel2Pixel	30.73±2.61	24.31±2.33	29.31±4.33	25.21±3.57	34.94±2.47	22.56±2.09	27.21±2.19	27.00±1.62
FetMRIE	33.02±3.46	34.06±2.00	33.05±4.48	33.01±3.95	33.37±2.09	31.49±2.30	33.68±2.29	33.12±1.46

Table 3. Tissue contrast and noise consistency metrics across internal and external validation datasets. TCT (mean ± standard deviation) are reported for WM versus cGM, WM versus dGM, and their average. CNC (mean ± standard deviation) quantifies the spatial uniformity of noise suppression across anatomical regions. Higher TCT and lower CNC indicate superior enhancement performance. The best result for each metric is highlighted in red. NA: not available.

Method	WCT-Normal-Test				WCT-GMH			
	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC
Original	1.394±0.227	0.786±0.154	1.090±0.160	NA	1.254±0.273	0.928±0.168	1.091±0.188	NA
BM4D	1.340±0.244	0.873±0.187	1.107±0.179	0.412±0.139	1.211±0.274	1.027±0.192	1.119±0.200	0.219±0.096
BME-X	1.558±0.237	0.559±0.096	1.058±0.133	0.889±1.393	1.280±0.256	0.573±0.123	0.927±0.145	0.294±0.145
Pixel2Pixel	1.400±0.244	0.824±0.164	1.112±0.171	0.975±0.334	1.259±0.292	0.983±0.180	1.121±0.205	0.333±0.223
ZS-N2N	1.398±0.228	0.789±0.154	1.093±0.160	11.65±11.58	1.258±0.276	0.931±0.168	1.094±0.190	0.726±0.639
FetMRIE	1.435±0.235	0.833±0.163	1.134±0.167	0.075±0.052	1.286±0.282	0.976±0.173	1.131±0.196	0.036±0.040
Method	**WCT-VM**				**WCB**			
	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC
Original	1.448±0.221	1.013±0.167	1.231±0.159	NA	1.045±0.260	1.072±0.194	1.059±0.209	NA
BM4D	1.330±0.231	1.097±0.204	1.214±0.180	0.292±0.095	0.883±0.280	1.240±0.195	1.062±0.216	0.271±0.084
BME-X	1.600±0.205	0.618±0.085	1.109±0.099	0.451±0.218	1.382±0.317	0.601±0.115	0.992±0.184	0.648±0.247
Pixel2Pixel	1.443±0.238	1.051±0.182	1.247±0.174	0.632±0.213	0.953±0.277	1.143±0.192	1.048±0.217	0.399±0.197
ZS-N2N	1.410±0.217	0.992±0.161	1.201±0.155	15.09±8.305	1.046±0.261	1.074±0.193	1.060±0.209	20.93±12.15
FetMRIE	1.461±0.222	1.048±0.176	1.255±0.164	0.059±0.034	1.061±0.264	1.169±0.188	1.115±0.208	0.019±0.015
Method	**FeTA**				**LFC**			
	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC
Original	1.252±0.298	1.196±0.288	1.224±0.243	NA	1.005±0.158	1.127±0.158	1.066±0.141	NA
BM4D	1.084±0.296	1.282±0.325	1.183±0.258	0.463±0.177	0.885±0.186	1.368±0.214	1.126±0.182	0.382±0.095
BME-X	1.658±0.218	0.672±0.179	1.165±0.141	0.772±0.229	1.301±0.207	0.802±0.107	1.051±0.119	0.472±0.201
Pixel2Pixel	1.203±0.316	1.224±0.295	1.213±0.257	0.946±0.391	1.011±0.183	1.238±0.175	1.125±0.158	0.230±0.208
ZS-N2N	1.250±0.299	1.198±0.288	1.224±0.244	7.000±4.471	1.003±0.158	1.130±0.159	1.067±0.141	1.712±1.549
FetMRIE	1.270±0.306	1.233±0.294	1.251±0.250	0.366±0.154	1.065±0.176	1.265±0.181	1.165±0.159	0.137±0.006
Method	**CSH Normal**				**CSH-GMH**			
	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC	TCT WM~cGM	TCT WM~dGM	TCT Average	CNC
Original	1.339±0.238	1.084±0.225	1.212±0.193	NA	1.220±0.248	1.055±0.248	1.138±0.224	NA
BM4D	1.246±0.251	1.183±0.255	1.214±0.207	0.472±0.123	1.131±0.238	1.144±0.282	1.137±0.237	0.247±0.101
BME-X	1.508±0.224	0.661±0.109	1.084±0.117	0.648±0.207	1.299±0.258	0.667±0.138	0.983±0.135	0.359±0.163
Pixel2Pixel	1.316±0.257	1.131±0.229	1.224±0.200	1.486±0.401	1.222±0.257	1.099±0.258	1.160±0.233	0.448±0.196
ZS-N2N	1.338±0.239	1.086±0.225	1.212±0.193	22.63±7.724	1.220±0.248	1.056±0.247	1.138±0.224	2.294±2.163
FetMRIE	1.365±0.244	1.136±0.232	1.250±0.200	0.163±0.051	1.239±0.254	1.084±0.256	1.161±0.231	0.064±0.030

Table 4. Results of the subjective reader assessment. Four radiologists (two junior and two senior) independently scored the enhanced image quality on three datasets (WCT-Normal-Test, WCB, and LFC), with 50 randomly selected cases per dataset. Each radiologist assigned scores ranging from 1 (poorest) to 4 (best) based on noise suppression, anatomical fidelity, and overall image quality. Scores are reported as mean ± standard deviation. ZS-N2N was excluded because its outputs were visually indistinguishable from the original images. The best average result for each dataset is highlighted in red.

WCT-Normal-Test	**Junior Radiologist 1**	**Junior Radiologist 2**	**Senior Radiologist 1**	**Senior Radiologist 2**	**Average**
BM4D	2.04±0.28	2.04±0.28	2.68±0.77	2.32±0.59	2.38±0.23
BME-X	1.00±0.00	1.06±0.42	1.00±0.00	1.06±0.31	1.03±0.03
Pixel2Pixel	3.92±0.27	2.54±0.65	2.72±0.73	3.10±0.79	3.07±0.53
FetMRIE	3.82±0.44	3.90±0.30	3.64±0.56	3.54±0.68	3.72±0.14
WCB	**Junior Radiologist 1**	**Junior Radiologist 2**	**Senior Radiologist 1**	**Senior Radiologist 2**	**Average**
BM4D	2.22±0.45	2.40±0.57	2.14±0.35	2.04±0.45	2.20±0.13
BME-X	1.00±0.00	1.00±0.00	1.00±0.00	1.06±0.31	1.02±0.03
Pixel2Pixel	3.04±0.57	2.44±0.54	3.04±0.57	3.54±0.61	2.95±0.60
FetMRIE	3.70±0.51	3.98±0.14	3.86±0.40	3.32±0.59	3.71±0.25
LFC	**Junior Radiologist 1**	**Junior Radiologist 2**	**Senior Radiologist 1**	**Senior Radiologist 2**	**Average**
BM4D	2.08±0.27	2.04±0.20	2.06±0.31	2.46±0.65	2.16±0.17
BME-X	1.00±0.00	1.00±0.00	1.00±0.00	1.34±0.69	1.08±0.15
Pixel2Pixel	3.94±0.24	2.84±0.37	2.68±0.47	2.34±0.85	2.95±0.60
FetMRIE	3.52±0.65	4.00±0.00	3.86±0.35	3.86±0.45	3.81±0.18

Table 5. Quantitative comparison of gestational age prediction and anomaly detection performance. The gestational age (GA) prediction performance is evaluated using the absolute age difference (AAD) between predicted and biological GA. For normative cases (WCT-Normal-Test), lower AAD indicates more accurate prediction; for pathological cases (WCT-GMH and WCT-VM), higher AAD reflects better preservation of anomaly-associated developmental deviations. Anomaly detection performance (Normal vs. GMH, Normal vs. VM) is quantified by AUROC and AUPR, with 95% confidence intervals (in brackets) calculated via bootstrap with 1000 resamples. The best results for each metric are highlighted in red.

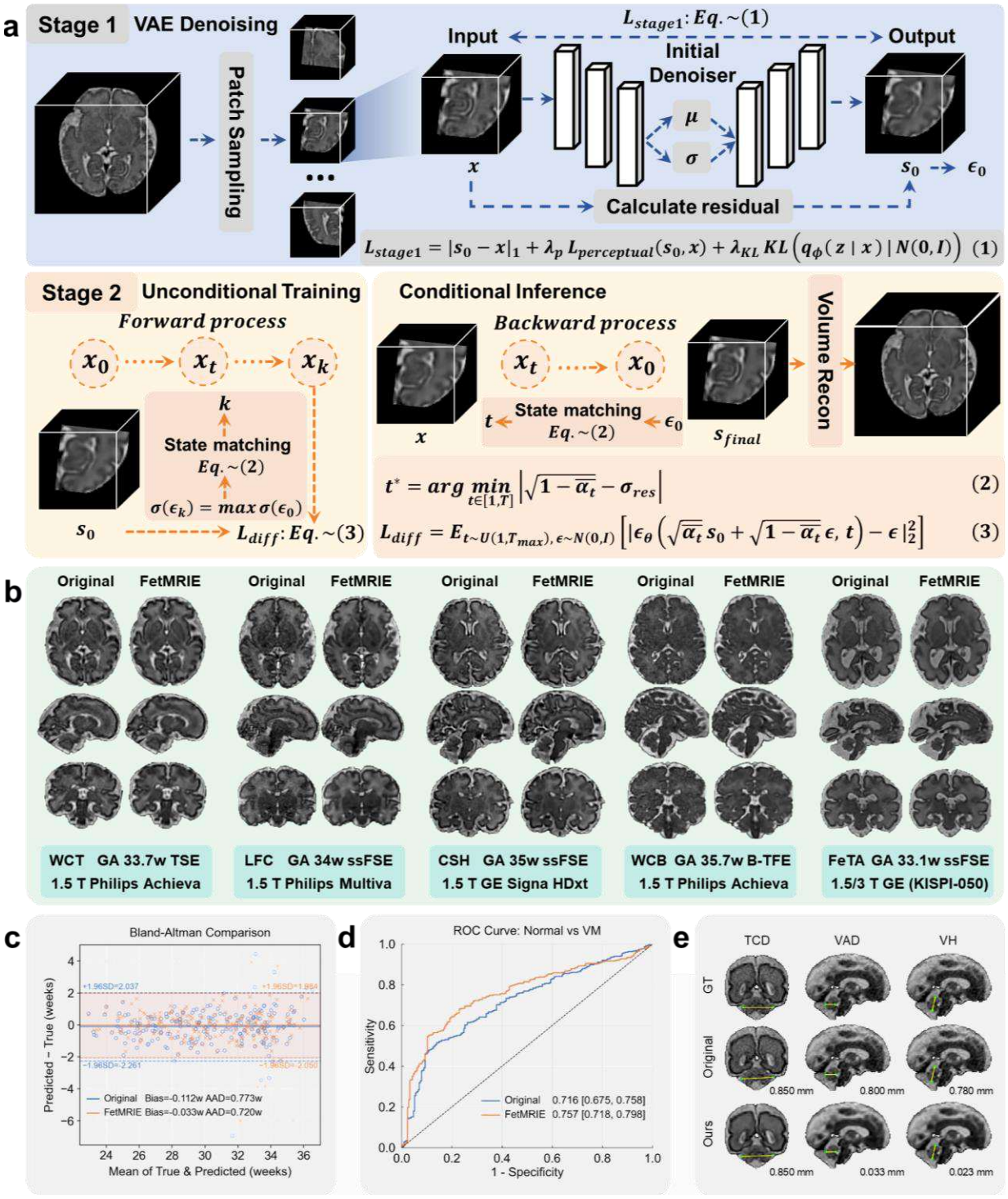
Method	Gestational Age Prediction			Anomaly Detection			
	WCT-Normal-Test	WCT-GMH	WCT-VM	Normal~GMH		Normal~VM	
				AUROC	AUPR	AUROC	AUPR
Original	0.773±0.784	2.077±1.427	1.397±0.944	0.828 [0.756, 0.889]	0.559 [0.415, 0.708]	0.716 [0.675, 0.758]	0.788 [0.740, 0.838]
BM4D	0.730±0.723	1.424±1.214	1.267±0.939	0.720 [0.629, 0.797]	0.405 [0.279, 0.549]	0.684 [0.642, 0.724]	0.772 [0.718, 0.823]
BME-X	0.723±0.738	1.355±1.266	1.083±0.783	0.677 [0.581, 0.764]	0.375 [0.254, 0.511]	0.651 [0.607, 0.694]	0.740 [0.688, 0.793]
ZS-N2N	0.728±0.738	1.647±1.296	1.345±0.909	0.758 [0.669, 0.840]	0.484 [0.343, 0.647]	0.729 [0.679, 0.773]	0.713 [0.649, 0.783]
Pixel2Pixel	0.748±0.720	1.684±1.406	1.337±0.889	0.774 [0.689, 0.845]	0.518 [0.373, 0.657]	0.710 [0.664, 0.753]	0.788 [0.738, 0.837]
FetMRIE	0.720±0.734	1.920±1.474	1.495±0.978	0.843 [0.778, 0.898]	0.558 [0.423, 0.695]	0.757 [0.718, 0.798]	0.821 [0.773, 0.870]

Table 6. Quantitative comparison of automated fetal brain biometry performance on the LFC dataset. Mean absolute error (MAE) ± standard deviation (mm) between automated measurements and expert ground truth are reported for Transverse Cerebellar Diameter (TCD), Vermis Anteroposterior Diameter (VAD), and Vermis Height (VH). The best result for each metric is highlighted in red.

Biometry (mm)	Original	BM4D	BME-X	ZS-N2N	Pixel2Pixel	FetMRIE
TCD	1.733±2.115	1.821±1.662	2.015±1.871	1.717±1.469	1.612±1.246	1.295±1.184
VAD	0.634±0.736	0.642±0.555	0.718±0.724	0.792±0.638	0.809±0.559	0.718±0.600
VH	1.193±0.803	0.986±0.674	1.049±0.761	1.026±0.773	1.196±0.548	0.932±0.932
Average	1.187±0.951	1.150±0.739	1.261±0.712	1.178±0.703	1.206±0.494	0.982±0.495

Table 7. Quantitative results of the ablation study. uMSE ($\times 10^{-3}$) and uPSNR (in dB) are reported as mean ± standard deviation across eight datasets. The full FetMRIE model is compared against two ablated variants: removing the diffusion-based structural restoration stage (w/o Stage II) and removing the adaptive state-matching mechanism during inference (w/o state matching). The best result for each metric is highlighted in red.

Methods	WCT-Normal-Test	WCT VM	WCT GMH	WCB	FeTA	LFC	CSH Normal	CSH GMH
uMSE ($\times 10^{-3}$)								
w/o stage II	0.753±1.108	0.646±0.307	19.68±23.28	2.200±0.788	3.503±5.229	1.525±1.380	1.018±0.714	0.823±0.537
w/o state matching	1.112±3.881	0.529±0.280	13.36±21.54	3.368±13.11	2.901±5.077	0.947±0.516	0.742±0.570	0.595±0.186
Full	0.587±0.788	0.435±0.199	0.950±1.320	1.542±0.633	0.512±0.235	0.813±0.434	0.489±0.258	0.515±0.163
uPSNR (dB)								
w/o stage II	31.95±2.08	32.35±1.98	22.50±8.47	31.08±4.27	28.27±5.51	29.36±3.07	30.90±2.95	31.71±2.76
w/o state matching	32.45±3.14	33.34±2.27	26.87±8.82	32.18±5.55	29.51±5.42	30.87±2.44	32.27±2.85	32.48±1.41
Full	33.02±3.46	34.06±2.00	33.05±4.48	33.01±3.95	33.37±2.09	31.49±2.30	33.68±2.29	33.12±1.46



$$L_{stage1} = |s_0 - x|_1 + \lambda_p L_{perceptual}(s_0, x) + \lambda_{KL} KL\left(q_\phi(z \mid x) \mid N(0, I)\right) \quad (1)$$

$$t^* = arg\min_{t \in [1,T]} \left|\sqrt{1-\overline{\alpha}_t} - \sigma_{res}\right| \quad (2)$$

$$L_{diff} = E_{t \sim U(1, T_{max}),\, \epsilon \sim N(0,I)} \left[|\epsilon_\theta\left(\sqrt{\overline{\alpha}_t}\, s_0 + \sqrt{1-\overline{\alpha}_t}\, \epsilon,\, t\right) - \epsilon|_2^2\right] \quad (3)$$

Figure 1. Overview of FetMRIE framework. (a) The proposed two-stage enhancement pipeline comprises VAE-based initial denoising (Stage I) and diffusion-based structural restoration (Stage II) guided by an adaptive state-matching strategy. (b) Original fetal brain MR and FetMRIE-enhanced images are compared across diverse scanners (1.5-T and 3.0-T), sequences (ss-FSE, TSE-SSH and B-TFE). (c-e) FetMRIE improves three clinically relevant downstream tasks, including (c) gestational age prediction, (d) pathological anomaly detection, and (e) biometry measurement.

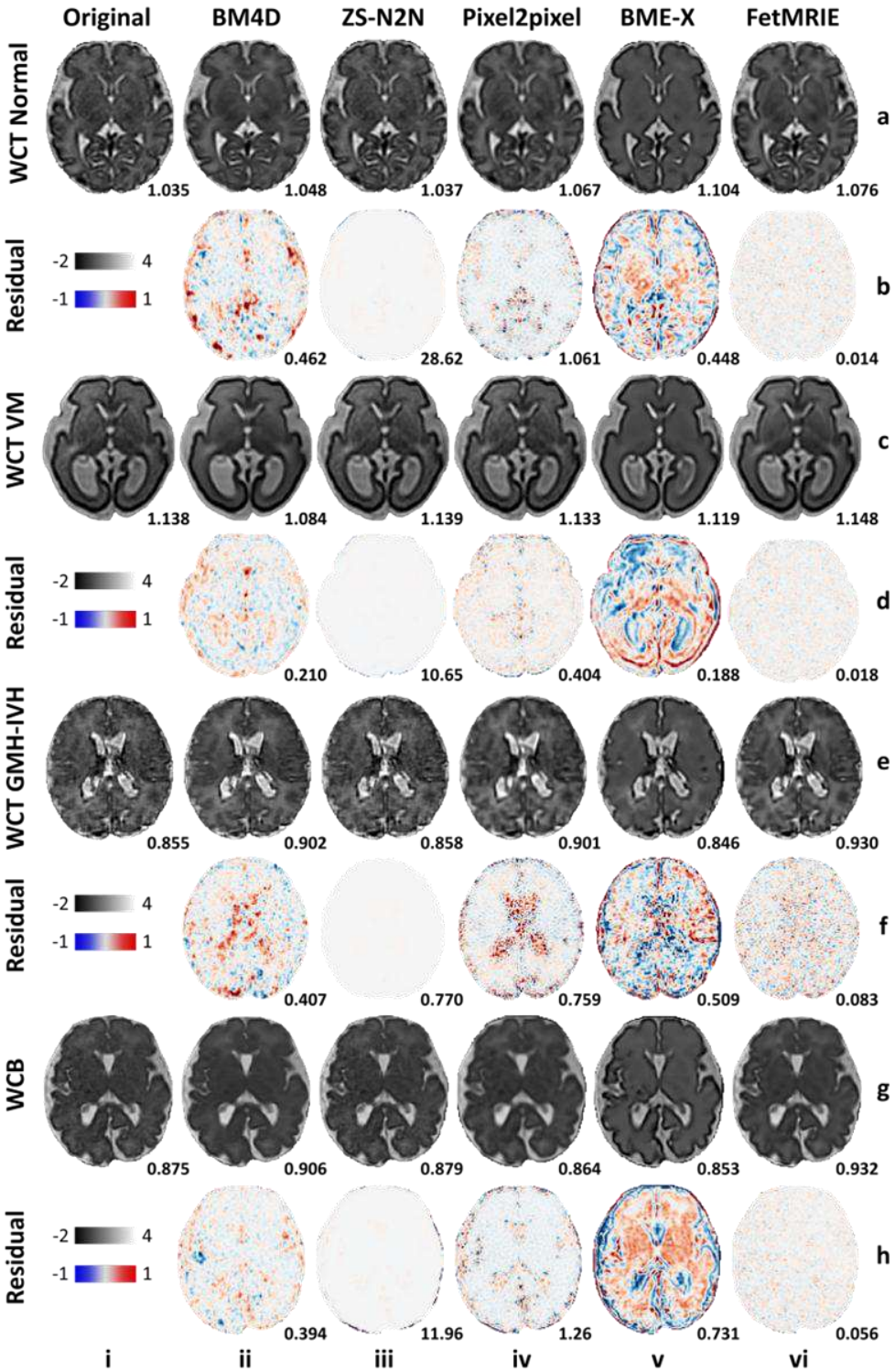


Figure 2. Qualitative comparison of enhancement results (internal validation). Exemplar axial slices display (i) the original noisy images and results from (ii) BM4D, (iii) ZS-N2N, (iv) Pixel2Pixel, (v) BME-X, and (vi) FetMRIE, from four internal validation datasets, including (a) WCT-Normal-Test, (c) WCT-VM, (e) WCT-GMH, and (g) WCB, with TCT listed below each image. (b, d, f, h) Maps of residuals between the enhanced output and original input are displayed, with CNC values listed below each map.

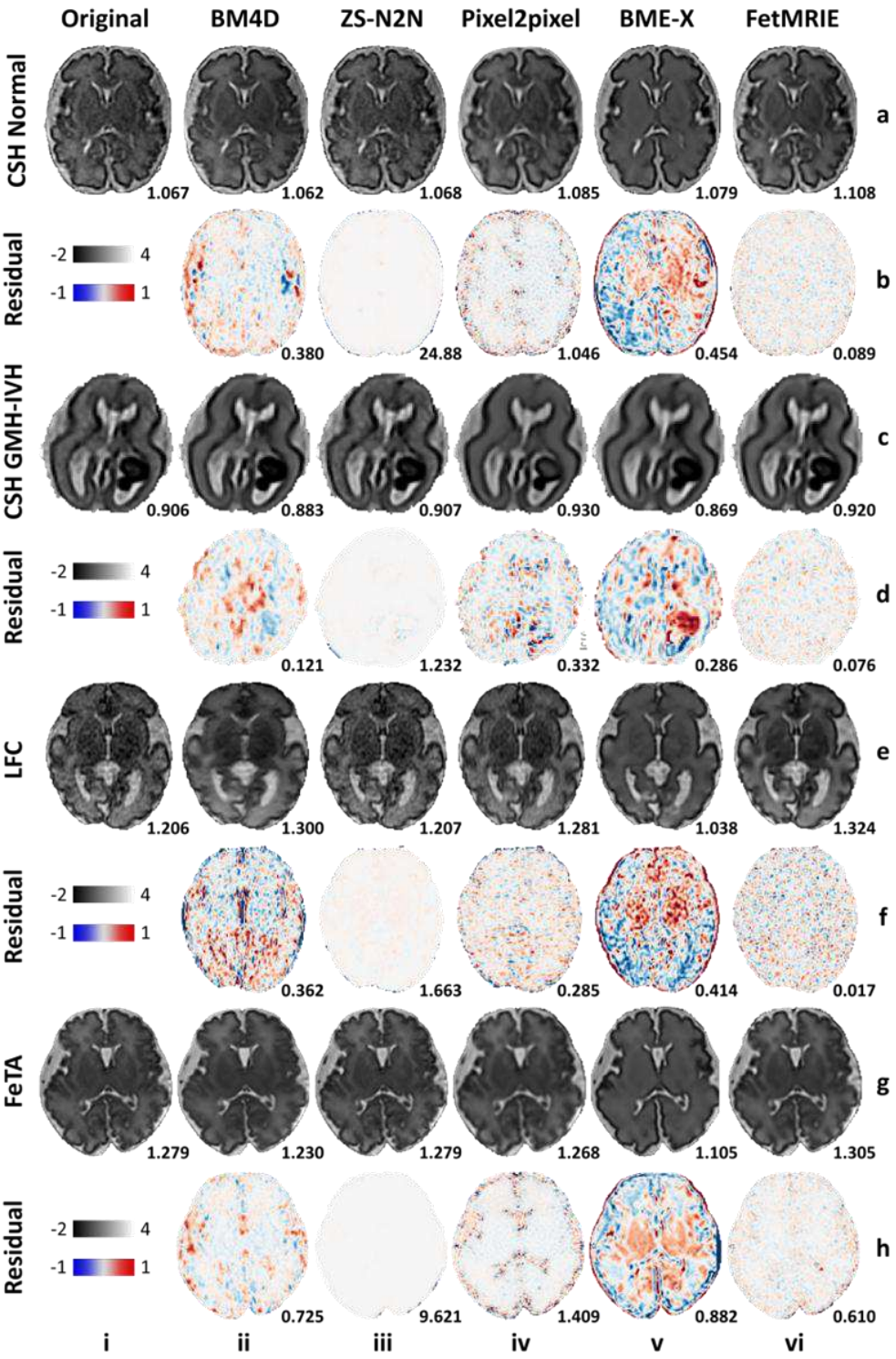


Figure 3. Qualitative comparison of enhancement results (external validation). Exemplar axial slices display (i) the original noisy images and results from (ii) BM4D, (iii) ZS-N2N, (iv) Pixel2Pixel, (v) BME-X, and (vi) FetMRIE, from four external validation datasets, including (a) CSH-Normal, (c) CSH-GMH-IVH, (e) LFC, and (g) FeTA, with TCT listed below each image. (b, d, f, h) Maps of residuals between the enhanced output and original input are displayed, with CNC values listed below each map

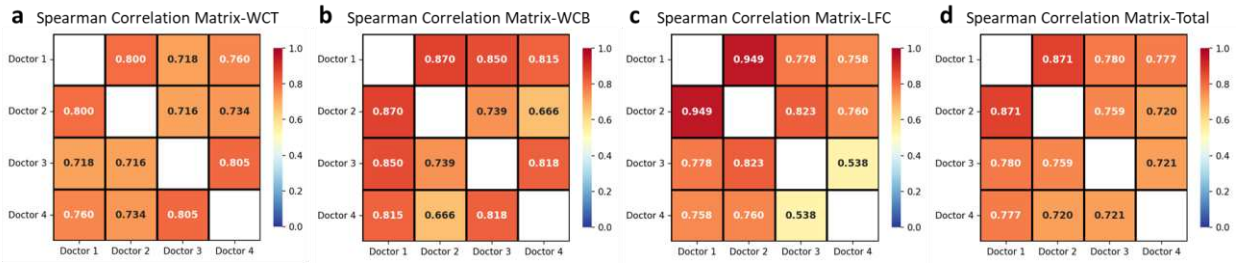


Figure 4. Inter-rater Pearson's correlation matrices for the reader assessment. The heatmaps display pairwise Pearson's correlation coefficients among the four participating radiologists on the (a) WCT-Normal-Test, (b) WCB, and (c) LFC datasets, as well as (d) the aggregate correlation matrix across all evaluated cases. Values closer to 1.0 (darker red) indicate stronger inter-rater consensus and higher evaluation reliability.

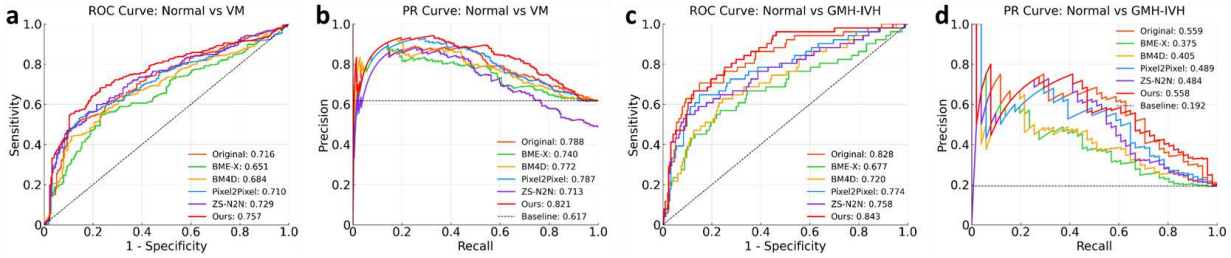


Figure 5. AUROC and AUPR curves for anomaly detection. Anomaly detection performance is shown for (a, b) Normal vs. VM and (c, d) Normal vs. GMH. (a, c) AUROC curves and (b, d) AUPR curves are shown, with dashed horizontal lines indicating the baseline prevalence of the positive class.

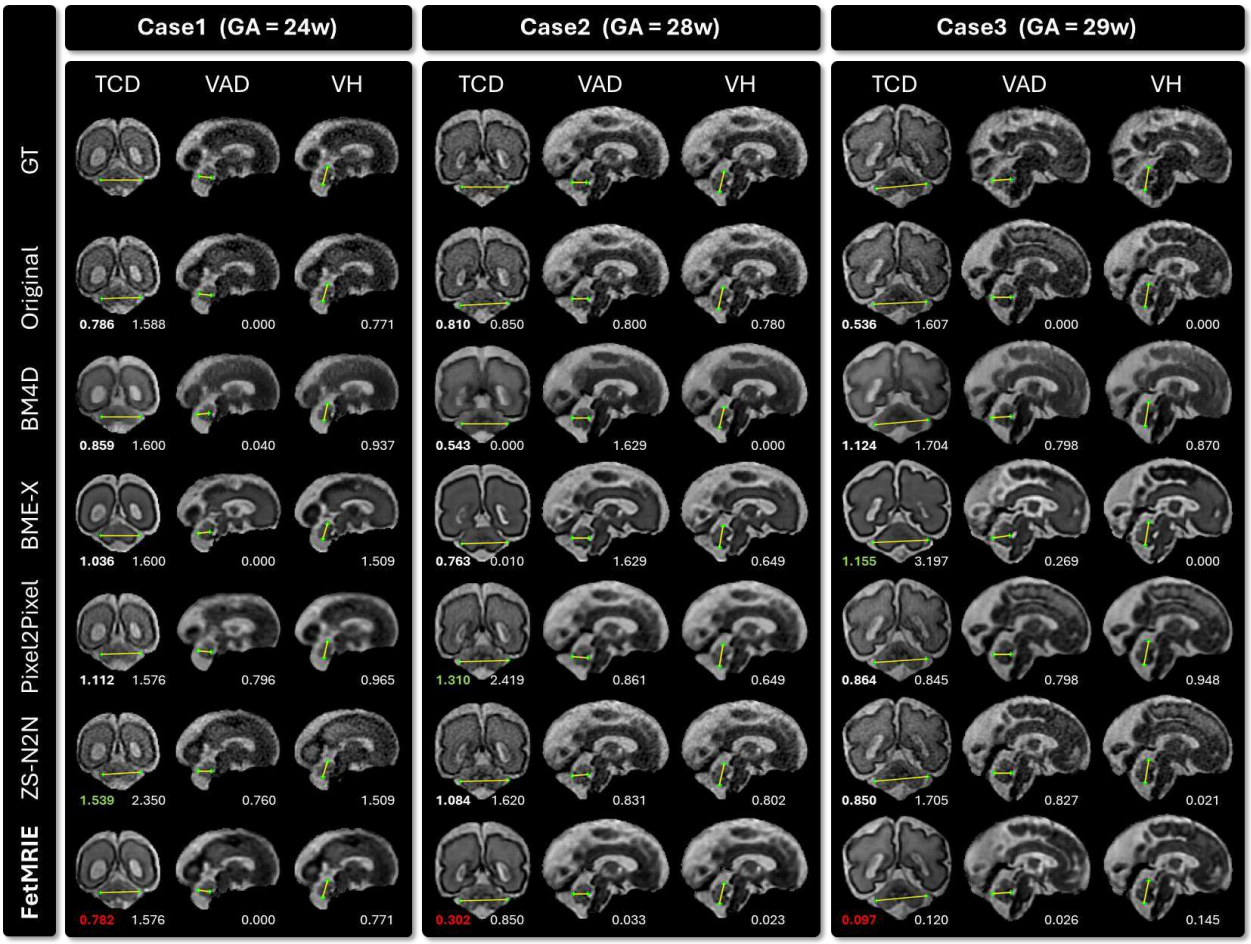


Figure 6. Qualitative comparison of automated biometry performance. Green dots and yellow lines indicate predicted landmarks and biometric lengths for transverse cerebellar diameter (TCD), vermis anteroposterior diameter (VAD), and vermis height (VH), respectively, for three representative cases at GAs of 24, 28, and 29 weeks, with absolute measurement errors (mm) relative to the ground truth listed at the bottom right of each image. The average mean absolute error (MAE) across the three measurements is shown at the bottom left in bold, with the best and worst values highlighted in red and green, respectively.

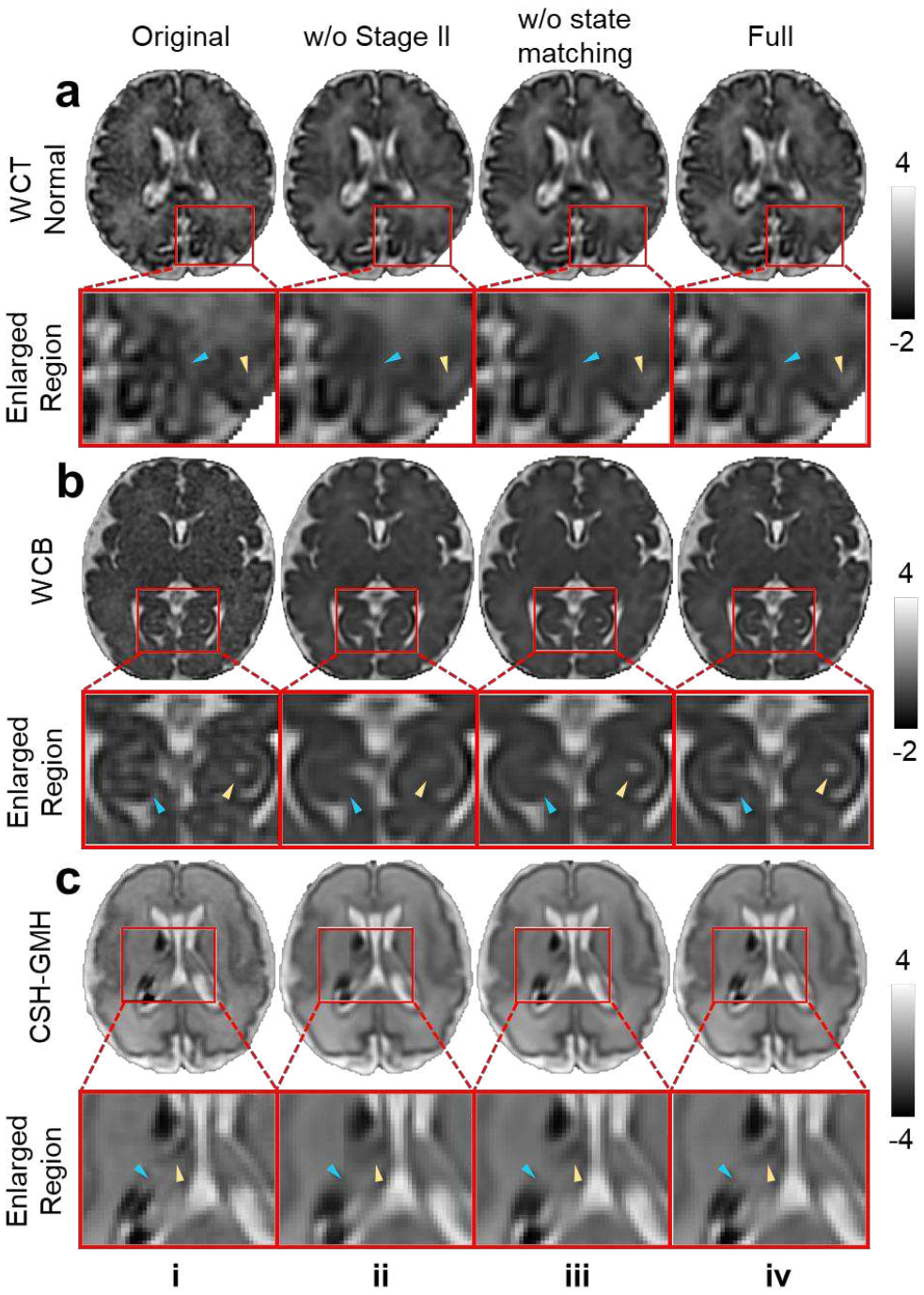


Figure 7. Qualitative results of the ablation study. Representative axial slices show (i) the original noisy images and results from the ablated variant (ii) without Stage II (VAE-only without diffusion-based structural restoration), (iii) without state matching (diffusion inference with a fixed schedule replacing the adaptive state-matching strategy), and (iv) the full FetMRIE model from (a) WCT-Normal-Test, (b) WCB, and (c) WCT-GMH datasets. Regions of interest are shown in enlarged views, with arrowheads highlighting regions where the ablated models exhibit structural blurring and loss of detail.