BrainDiff: Latent Diffusion for Unsupervised Correction of Motion Artifacts in Brain MRI Images

1st Raul Zapata
depto. Ingenieria Informatica y Ciencias de la Computacion
Universidad de Concepcion
Concepcion, Chile
rzapata2019@udec.cl

2nd Ricardo Flores
depto. Ingenieria Informatica y Ciencias de la Computacion
Universidad de Concepcion
Concepcion, Chile
riflores@udec.cl

***Abstract*—Patient motion during Magnetic Resonance Imaging (MRI) acquisition remains a critical challenge, often leading to artifacts that degrade diagnostic utility. Deep learning-based correction methods typically require large datasets of paired corrupted and motion-free images, which are practically impossible to acquire in clinical settings. To address this limitation, we propose *BrainDiff*, a novel framework for unsupervised motion correction that leverages Latent Diffusion Models (LDM) to synthesize training pairs from unpaired data. Our approach operates in a compressed latent space via a KL-Regularized VAE to ensure computational efficiency. We train a diffusion model exclusively on motion-corrupted scans to capture the statistical distribution of artifacts. By conditioning this model on clean anatomical scans through a reverse diffusion injection process, we generate a realistic synthetic paired dataset. Finally, a supervised Attention U-Net is trained on these synthetic pairs to perform artifact removal. Quantitative evaluation on the MR-ART dataset demonstrates that our generative pipeline successfully simulates realistic motion artifacts, allowing the correction network to reduce the Mean Relative Error (MRE) in corrupted images without requiring ground-truth clinical pairs.**

I. Introduction

Magnetic Resonance Imaging (MRI) is a powerful and non-invasive medical tool that enables the diagnosis of complex neurological and systemic diseases in humans. However, the acquisition process is highly sensitive to patient motion, as scans typically require long acquisition times and strict stillness. Even subtle head or body movements can introduce motion artifacts that distort the anatomical structures or blur the image, significantly reducing its diagnostic value. Depending on the severity of motion, the resulting images may become unusable, leading to longer examination times, repeated scans.

With the rapid advancement of generative modeling techniques, new possibilities have emerged for correction of motion artifacts directly from magnitude MRI data. These methods aim to reconstruct anatomically consistent and diagnostically reliable images without the need for re-acquisition.

II. literature review

The presence of motion artifacts in brain MRI remains a critical obstacle to high-fidelity imaging, undermining both qualitative assessments and quantitative analyses such as cortical thickness, volumetry, or brain-age predictions [1]. Early solutions focused on acquisition strategies and reconstruction methods that inherently resist or correct motion. PROPELLER samples k-space in overlapping strips and enables retrospective correction of rigid head motion [2]. Another approach uses variable-density spiral trajectories in MRI to mitigate motion artifacts by distributing motion effects more evenly across k-space [3].Techniques, such as optical tracking or navigator-based methods (e.g. vNav, PROMO), have demonstrated significant reductions in motion artifacts in both functional and structural MRI when the required hardware or navigators are available [4], [5].

In parallel, reconstruction-based methods estimate motion and correct it during image formation. For instance, motion-corrected compressed sensing techniques jointly recover the image and motion parameters by imposing sparsity, achieving robust results in dynamic and free-breathing cardiac MRI [6]. The concept of leveraging sparse representations to correct for motion (e.g. sparse MRI motion correction) provides a flexible, data-driven [7].

Simulating motion on magnitude MRI data (i.e. after magnitude reconstruction rather than in k-space) is another important line of work. Such simulations allow creation of synthetic corrupted/uncorrupted image pairs to benchmark correction algorithms and quantify the impact of motion on image quality and segmentation. highlight that even moderate motion can degrade segmentation of cortical structures [8].

The impact of motion on morphometric measures and downstream analyses is well-documented. Motion biases cortical thickness, regional volumes, and other structural metrics, which in turn can confound studies of aging, disease, or longitudinal change. Because of this, retrospective estimation of head motion in structural MRI has been proposed to adjust or exclude data in large cohorts [9]. In brain-age modeling studies, motion has been shown to systematically bias predictions, motivating deeper investigation of motion's confounding role [10].

More recently, deep learning approaches have dominated retrospective motion correction. Unsupervised disentanglement models such as MAUDGAN attempt to separate artifact and anatomical content so as to correct motion without paired training data [11]. On the other hand, diffusion-based models have emerged as a powerful generative paradigm for motion correction in MRI. MoCo-Diff introduces an adaptive condi-

tional prior in a diffusion network tailored for MRI motion correction [12]. DIMA (Diffusing Motion Artifacts) further extends this idea: it first learns the distribution of motion artifacts from unpaired data using diffusion, then synthesizes corrupted images from clean volumes to train a supervised correction network [13]. Variants and extensions (such as ResMoCoDiff) refine architecture for 3D MRI volumes [14].

Alongside these, the rise of foundation models in MRI offers a promising path. Large-scale pretraining on multimodal MRI data enables models that can perform denoising, super-resolution, harmonization, and motion correction in a unified manner. These foundation models, when fine-tuned or applied in cascade, consistently improve downstream tasks—including segmentation, registration, and diagnostic prediction—while reducing dependence on annotated data from specific sites [15], [16].

III. Method

We propose a framework for synthesizing paired datasets suitable for supervised learning from initially unpaired collections. The procedure utilizes two distinct cohorts: one comprising motion-degraded images and another consisting of artifact-free scans, which may originate from different subjects.

As illustrated in Fig. 1, we first train a generative Diffusion model exclusively on the motion-affected dataset to capture the statistical distribution of motion artifacts. Rather than generating random samples, we condition this model on the artifact-free images to inject realistic motion patterns onto them. This process yields a synthetic paired dataset where each clean image has a corresponding motion-corrupted counterpart. Subsequently, this dataset is employed to train a supervised artifact correction model, the performance of which is benchmarked against alternative simulation techniques and real-world paired data.

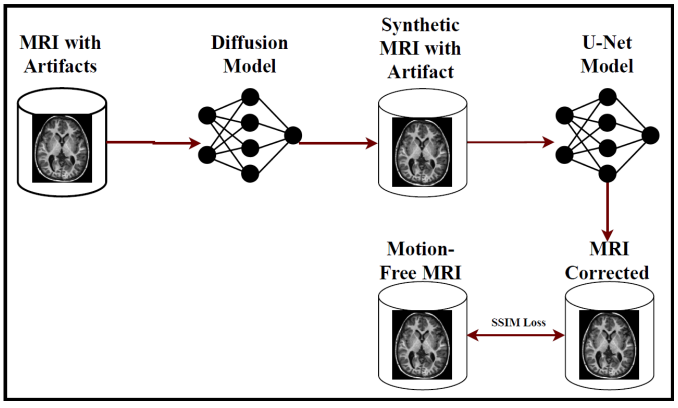

Fig. 1. Model framework.

A. Phase 1: Latent Space Compression (VAE)

To reduce computational complexity and focus on high-level semantic features, we trained a KL-Regularized Variational Autoencoder (AutoencoderKL). This model learns to compress an input image x into a lower-dimensional latent representation z and subsequently reconstruct it as $\hat{x}$.

The VAE was optimized using a composite loss function to ensure high-fidelity reconstruction:

$$\mathcal{L}_{VAE} = \mathcal{L}_{L1} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{perc}\mathcal{L}_{LPIPS} + \lambda_{adv}\mathcal{L}_{GAN} \quad (1)$$

Where $\mathcal{L}_{LPIPS}$ is the Learned Perceptual Image Patch Similarity loss, which ensures preservation of textural details, and $\mathcal{L}_{GAN}$ is a patch-based adversarial loss to enforce realism.

B. Phase 2: Learning Motion Priors via Latent Diffusion

We employed a Latent Diffusion Model (LDM) to capture the statistical distribution of motion artifacts. Crucially, this model was trained exclusively on the motion-corrupted subset of the dataset.

The model architecture is a U-Net with attention mechanisms. It functions as a Denoising Diffusion Probabilistic Model (DDPM), trained to reverse a Markov chain that gradually adds Gaussian noise to the latent representation of motion images. By training solely on corrupted data, the diffusion model learns a "motion prior," meaning it interprets any input noise as a potential motion artifact during the denoising process.

C. Phase 3: Synthetic Data Generation

To overcome the lack of paired ground-truth data (clean vs. motion), we implemented the Deep Image Motion Artifact injection strategy [13]. This process generates a synthetic paired dataset $\mathcal{D} = \{(x_{clean}, x_{artifact})\}$ as follows:

1) **Encoding:** A clean image x_{clean} is encoded into latent space z_0.
2) **Forward Diffusion:** Partial noise is added to z_0 up to a timestep t, where t is sampled uniformly from a range $[0.35T, 0.65T]$ (where $T = 1000$). This destroys fine details while preserving gross anatomy.
3) **Reverse Diffusion:** The noisy latent is denoised using the motion-trained LDM from Phase 2. Since the LDM only knows motion patterns, it "hallucinates" or injects motion artifacts onto the clean anatomy during reconstruction.
4) **Decoding:** The resulting latent is decoded back to pixel space, yielding $x_{artifact}$.

We generated 5 stochastic variants for each clean subject, resulting in a dataset of approximately 740 paired images.

D. Phase 4: Artifact Correction Network

The final restoration model is an Attention U-Net designed to map $x_{artifact} \rightarrow x_{clean}$. We employed a Residual Learning strategy: instead of predicting the clean image directly, the network predicts the artifact residual $\mathcal{R}$. The restored image is obtained by subtraction:

$$\hat{x}_{clean} = x_{artifact} - \text{Net}(x_{artifact}) \quad (2)$$

To train this network, we utilized a hybrid loss function:

- **Pixel-wise Loss:** L_1 distance between the restored image and the ground truth.

- **Contrastive Loss:** We utilized the frozen VAE encoder to extract latent features from the prediction, the target, and the input. We applied a contrastive objective (InfoNCE) to maximize the semantic similarity between the prediction and the clean target, while minimizing similarity to the motion-corrupted input.

IV. Experiments

A. Dataset

Our experiments were primarily conducted using the MR-ART dataset [30], a comprehensive collection of brain MRI scans designed for motion artifact analysis. This dataset includes T1-weighted scans from 148 subjects. For each subject, the dataset provides a "Standard" acquisition (motion-free) where the patient remained still, and "Head-motion" acquisitions where the patient performed purposeful movements to induce realistic artifacts.

From this 3D dataset, we extracted 2D slices from the sagittal plane. To ensure data consistency, all images were resized to a spatial resolution of 192×256 pixels and normalized to the $[0, 1]$ intensity range. The dataset was split into two distinct subsets to prevent data leakage:

- **Motion Prior Subset:** Comprising exclusively motion-affected images from 90% of the subjects. These were used solely to train the Latent Diffusion Model in Phase 2, ensuring the model learned the distribution of artifacts without seeing clean anatomy.
- **Correction Subset:** Comprising the standard (clean) images from the remaining subjects. These were used as the source for DIMA [13] to generate the synthetic paired dataset for training the correction network.

For the final evaluation, we utilized a held-out test set of real paired images from the MR-ART dataset (Standard vs. Head-motion) that were never seen by either the generator or the corrector during training.

B. Implementation Details

1) Latent Space Compression (VAE): In Phase 1, we trained a KL-regularized Autoencoder (AutoencoderKL) to compress the input space. The network architecture consists of an encoder and decoder with 2 residual blocks per level and channel configurations of $(128, 128, 256)$. The model compresses the $1 \times 192 \times 256$ input into a latent representation of $3 \times 48 \times 64$. **Hyperparameters:** The VAE was trained for 50 epochs with a batch size of 8. We used the Adam optimizer with a learning rate of $1e-4$. The loss weights were set as follows: $\lambda_{KL} = 1e-6$ for regularization, and $\lambda_{perc} = 1.0$ for the perceptual loss using a VGG-based LPIPS network.

2) Motion Diffusion Model: For Phase 2, we implemented a Denoising Diffusion Probabilistic Model (DDPM) utilizing a U-Net architecture (DiffusionModelUNet) operating in the latent space defined by the VAE. The network features 2 residual blocks per level, channel multipliers of $(1, 2, 4)$, and attention mechanisms at resolutions of 16 and 8. **Hyperparameters:** The diffusion process used a linear beta schedule with $\beta_{start} = 0.0015$ and $\beta_{end} = 0.0195$ over $T = 1000$ time-steps. The model was trained using the Adam optimizer with a learning rate of $2.5e-5$ and a batch size of 16. Training was conducted for 100 epochs until convergence of the Evidence Lower Bound (ELBO) loss.

3) Injection Strategy: To generate the synthetic training pairs for the corrector, we applied the DIMA [13] injection algorithm. Clean images were encoded to latent space, corrupted with noise levels sampled uniformly from $t \in [350, 650]$ (representing 35% to 65% signal destruction), and then reconstructed using the motion-trained diffusion model. We generated 5 stochastic variants for each clean slice, resulting in a synthetic dataset of approximately 740 paired images.

4) Artifact Correction Network: For Phase 4, we trained an Attention U-Net to perform artifact removal. Unlike standard denoising approaches, we employed a *Residual Learning* strategy where the network predicts the artifact component rather than the clean image directly. The architecture includes attention gates in the skip connections to focus processing on corrupted regions. **Hyperparameters:** The corrector was trained using a hybrid loss function combining L_1 loss and a Contrastive Loss (weight $\lambda = 0.1$) utilizing the frozen VAE as a feature extractor. We used the Adam optimizer with a learning rate of $1e-4$ and a batch size of 16. We implemented early stopping with a patience of 10 epochs, monitoring the validation loss to prevent overfitting, resulting in an optimal model at approximately epoch 15-20.

V. Results

A. Quantitative Evaluation

We evaluated the performance of each component of our pipeline using three quantitative metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Mean Relative Error (MRE). Table I summarizes the results across the three key stages of our framework.

TABLE I
Quantitative Results for Each Pipeline Stage

Model Stage	PSNR (dB)	SSIM	MRE
VAE (Reconstruction)	30.60	0.8707	0.0443
LDM (Artifact Injection)	21.46	0.6360	0.1337
Corrector (Restoration)	22.12	0.5724	0.1072

1) Latent Space Fidelity: The first row of Table I evaluates the AutoencoderKL (VAE). The high PSNR of 30.60 dB and SSIM of 0.8707 confirm that our compression strategy preserves the anatomical structure and fine details of the brain MRI with high fidelity. The low MRE (0.0443) indicates that the information loss due to latent compression is minimal, validating the use of the latent space for the subsequent diffusion tasks.

2) Synthetic Data Quality: The LDM row represents the images generated by the DIMA injection process (Phase 3). Here, a lower similarity to the original clean image is expected and desired, as it indicates the successful introduction of motion artifacts. The drop in SSIM to 0.6360 and the

increase in MRE to 0.1337 demonstrate that the diffusion model effectively corrupted the images with significant motion patterns, creating a challenging synthetic dataset for supervised training.

3) Restoration Performance: The final row displays the performance of the Attention U-Net Corrector on the test set. The model achieved a PSNR of 22.12 dB and an MRE of 0.1072. The reduction in Mean Relative Error compared to the motion-affected input (from 0.1337 in LDM down to 0.1072 in the Corrector) indicates that the model successfully removed a portion of the artifact signal and recovered anatomical information. However, the SSIM value of 0.5724 suggests that recovering high-frequency structural details from severe motion artifacts remains a challenging task, potentially due to the complexity of the synthetic artifacts generated during the training phase.

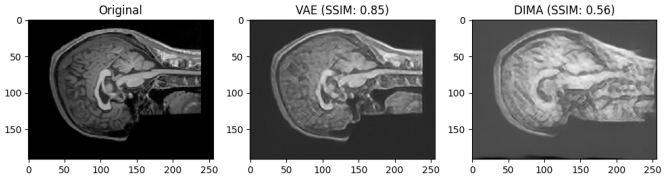

Fig. 2. Model framework.

VI. Conclusion

In this work, we presented a novel unsupervised framework for correcting motion artifacts in brain MRI, addressing the critical scarcity of paired ground-truth clinical data. By leveraging the generative capabilities of Latent Diffusion Models (LDM), we moved beyond traditional physics-based simulations, allowing the model to learn the complex, non-linear statistics of real motion artifacts directly from unpaired corrupted scans.

Our experimental results validate the efficacy of the proposed pipeline across its three stages. First, the KL-Regularized VAE demonstrated that high-resolution MRI data can be effectively compressed into a low-dimensional latent space (PSNR > 30 dB) without compromising anatomical fidelity. Second, the DIMA injection strategy proved successful in synthesizing a large-scale paired dataset, effectively transferring learned motion patterns onto clean anatomy. Finally, the supervised Attention U-Net trained on this synthetic data demonstrated a capability to reduce the relative error (MRE) in corrupted images.

However, the structural recovery (SSIM) of the final corrector highlights the inherent difficulty of the restoration task. While the model successfully denoises the image, recovering high-frequency fine details remains a challenge. This suggests that future iterations could benefit from integrating stronger adversarial losses or leveraging pre-trained foundation models to enhance textural realism.

Future work will focus on extending this architecture to fully 3D volumetric processing to capture through-plane motion effects and validating the clinical utility of the reconstructed images in downstream tasks such as segmentation and registration. Our findings suggest that latent generative models represent a promising direction for robust, data-driven medical image restoration.

References

[1] M. Zaitsev, J. Maclaren, and M. Herbst, "Motion artifacts in mri: A complex problem with many partial solutions," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 4, pp. 887–901, 2015.

[2] J. G. Pipe, "Motion correction with propeller mri: application to head motion and free-breathing cardiac imaging," *Magn Reson Med*, vol. 42, no. 5, pp. 963–969, 1999, pubMed.

[3] J.-R. Liao, J. M. Pauly, T. J. Brosnan, and N. J. Pelc, "Reduction of motion artifacts in cine mri using variable-density spiral trajectories," *Magn Reson Med*, vol. 37, no. 4, pp. 569–575, 1997, pubMed.

[4] N. Todd, O. Josephs, M. F. Callaghan, A. Lutti, and N. Weiskopf, "Prospective motion correction of 3d echo-planar imaging data for functional mri using optical tracking," *Neuroimage*, vol. 113, pp. 1–12, 2015, pubMed.

[5] N. White, C. Roddey, A. Shankaranarayanan, E. Han, D. Rettmann, J. Santos, J. Kuperman, and A. Dale, "Promo: real-time prospective motion correction in mri using image-based tracking," *Magn Reson Med*, vol. 63, no. 1, pp. 91–105, 2010, pubMed.

[6] M. Usman, D. Atkinson, F. Odille, C. Kolbitsch, G. Vaillant, T. Schaeffter, P. G. Batchelor, and C. Prieto, "Motion corrected compressed sensing for free-breathing dynamic cardiac mri," *Magn Reson Med*, vol. 70, no. 2, pp. 504–516, 2013, pubMed.

[7] Z. Yang, C. Zhang, and L. Xie, "Sparse mri for motion correction," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2013, pp. 962–965, arXiv.

[8] H. Olsson, J. M. Millward, L. Starke, and et al., "Simulating rigid head motion artifacts on brain magnitude mri data – outcome on image quality and segmentation of the cerebral cortex," *PLOS ONE*, Apr 16 2024, pLOS.

[9] D. Zacà and et al., "Method for retrospective estimation of natural head motion in structural mri," *J Magn Reson Imaging*, 2018, pubMed.

[10] R. Moqadam and et al., "Investigating the impact of motion in the scanner on brain age predictions," *Imaging Neuroscience / preprints*, 2023-2024, direct.mit.edu.

[11] M. Safari and et al., "Maudgan: Motion artifact unsupervised disentanglement generative adversarial network," 2023, preprint / medRxiv.

[12] F. Li and et al., "Moco-diff: Adaptive conditional prior on diffusion network for mri motion correction," in *MICCAI 2024*, 2024, papers.miccai.org.

[13] P. Angella, L. Balbi, F. Ferrando, and et al., "Dima: Diffusing motion artifacts for unsupervised correction in brain mri images," 2025, preprint arXiv.

[14] et al., "Res-mocodiff / variants: Recent articles and preprints on ddpm / denoising diffusion for mri motion correction," 2025, arXiv.

[15] Y. Sun and et al., "A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks," *Nature*, 2024/2025, nature / Commun. / Biomed Eng?

[16] et al., "Mri-core: A foundation model for magnetic resonance – examples of foundation models applied to segmentation and mr enhancement," Jun 2025, arXiv.