# Unpaired Multi-Site Brain MRI Harmonization with Image Style-Guided Latent Diffusion

Mengqi Wu[1,2], Minhui Yu[1,2], Weili Lin[1], Pew-Thian Yap[1], and Mingxia Liu[1*]

[1] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2] Lampe Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill and North Carolina State University, Chapel Hill, NC 27599, USA
*Corresponding author (mingxia_liu@med.unc.edu)

**Abstract.** Multi-site brain MRI heterogeneity caused by differences in scanner field strengths, acquisition protocols, and software versions poses a significant challenge for consistent analysis. Image-level harmonization, leveraging advanced learning methods, has attracted increasing attention. However, existing methods often rely on paired data (*e.g.*, human traveling phantoms) for training, which are not always available. Some methods perform MRI harmonization by transferring target-style features to source images but require explicitly learning disentangled image styles (*e.g.*, contrast) via encoder-decoder networks, which increases computational complexity. This paper presents an unpaired MRI harmonization (UMH) framework based on a new image style-guided diffusion model. UMH operates in two stages: (1) a *coarse harmonizer* that aligns multi-site MRIs to a unified domain via a conditional latent diffusion model while preserving anatomical content; and (2) a *fine harmonizer* that adapts coarsely harmonized images to a specific target using style embeddings derived from a pre-trained Contrastive Language-Image Pre-training (CLIP) encoder, which captures semantic style differences between the original MRIs and their coarsely-aligned counterparts, eliminating the need for paired data. By leveraging rich semantic style representations of CLIP, UMH avoids learning image styles explicitly, thereby reducing computation costs. We evaluate UMH on 4,123 MRIs from three distinct multi-site datasets, with results suggesting its superiority over several state-of-the-art (SOTA) methods across image-level comparison, downstream classification, and brain tissue segmentation tasks.

**Keywords:** Brain MRI Harmonization · Style Translation · Diffusion.

## 1 Introduction

Brain MRI analysis using learning-based methods, particularly deep learning, often suffers from limited data at individual acquisition sites, resulting in suboptimal performance and poor generalizability across sites. To address this, multi-site MRI pooling is increasingly used in neuroimaging studies to expand sample

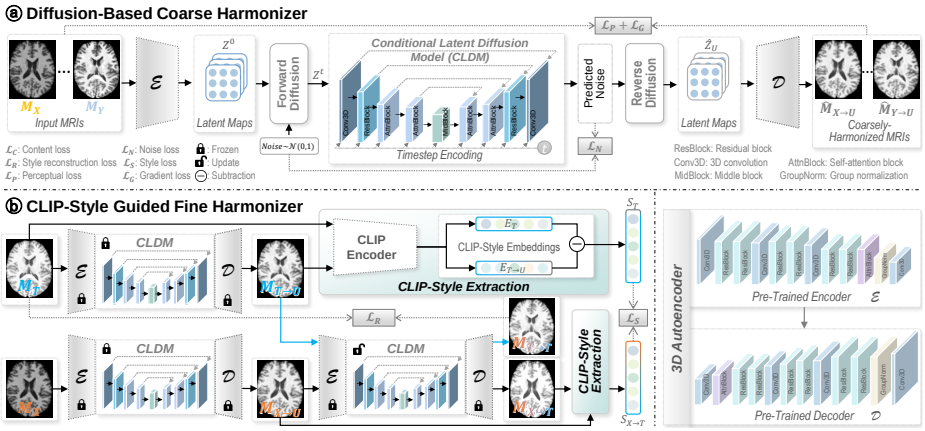


**Fig. 1.** Proposed UMH framework, containing (a) a diffusion-based coarse harmonizer (DCH) that aligns multi-site MRIs into a unified domain while preserving anatomical content, and (b) a CLIP-style guided fine harmonizer that aligns coarsely-harmonized MRIs to a target style using a disentangled CLIP-style loss. This dual-stage process achieves efficient MRI harmonization without explicit content-style disentanglement.

size, enhance cohort diversity, and improve statistical power [1,22,25]. However, variations in scanner vendors, scanning sequences, and field strength introduce site-specific non-biological style differences (*e.g.*, intensity, contrast, and signal-to-noise ratio), which can hinder model training [12,13,19,28].

Data harmonization techniques have been developed to reduce site effects in pre-extracted MRI features or directly normalize raw images. Existing feature-level harmonization methods, such as ComBat [11] and ComBat-GAM [21], utilize the Empirical Bayes framework that models and adjusts the site effect as multiplicative and additive errors in data batches. However, these methods heavily rely on the extracted features [1,4]. Existing image-level harmonization methods often use generative adversarial networks (GANs) for cross-domain image translation [6,16,18], which can be time-consuming and unstable to train. Some approaches use multiple encoder-decoder networks to learn disentangled image style and content representations [4,8,27,31], which is computationally heavy. In addition, they often require paired MRIs, such as traveling phantom data from the same subjects, which may not always be available in practice.

To this end, we propose an unpaired 3D brain MRI harmonization (UMH) framework with a conditional latent diffusion model, guided by semantic style embeddings derived from the contrastive language-image pre-training (CLIP) encoder [29]. As shown in Fig. 1, UMH operates in two stages. In the first stage, a conditional latent diffusion model (CLDM) is designed to coarsely align multi-site input MRIs into a unified domain while preserving anatomical content information. The second stage involves fine-tuning the coarse harmonizer to align the coarsely-harmonized MRIs to a specific target style guided by a CLIP-based style loss, which leverages the difference of CLIP-style embeddings

from the original MRIs and their coarsely-harmonized counterparts to model disentangled style information without requiring paired MRIs. This dual-stage process allows our UMH to leverage the stable training and generative power of diffusion models [20] and the rich semantic representations of CLIP, achieving efficient MRI style transfer without the need for separate encoders and decoders to learn content and style information explicitly. The UMH is trained and evaluated on three multi-site datasets with a total of 4,123 T1-weighted (T1w) MRIs through three tasks. Experimental results demonstrate the superiority of UMH over several state-of-the-art (SOTA) methods in aligning multi-site MRI styles while preserving critical biological and anatomical features.

## 2 Proposed Method

**Problem Formulation.** Our goal is to align the style (*e.g.*, intensity and contrast) of source MRIs with a target style while preserving their content (*i.e.*, anatomical structures). We utilize a two-stage approach. **(1) Coarse harmonization**: To reduce computational cost, we perform harmonization in a low-dimensional latent space using a 3D autoencoder. The encoder $\boldsymbol{E}$ compresses an MRI $M$ into a latent map $Z = \boldsymbol{E}(M) \in \mathbb{R}^{c \times h \times w \times d}$, while the decoder $\boldsymbol{D}$ reconstruct the MRI from the latent map $Z$. Here $c$, $w$, $h$, and $d$ stand for channel, width, height, and depth dimensions, respectively. To coarsely remove site-specific style variations while maintaining anatomical content, we train a conditional latent diffusion model (CLDM) $\boldsymbol{\Phi}$, which maps the latent representations of arbitrary MRIs to a unified latent domain $U$. The coarsely-harmonized MRIs are then reconstructed as: $M_U = \boldsymbol{D}(Z_U\{C, S_U\}) = \boldsymbol{D} \circ \boldsymbol{\Phi}(Z\{C, S\})$, where $C$ and $S$ represent content and style information, respectively. Two novel image-level constraints are designed to ensure image content preservation. **(2) Fine harmonization**: The CLDM is fine-tuned to align coarsely-harmonized MRIs to a specific target style. Given MR images from a source domain $X$ and a target domain $T$, we first map them to the unified domain $U$: for $i \in \{X, T\}$, $M_{i \to U} = \boldsymbol{D} \circ \boldsymbol{\Phi} \circ \boldsymbol{E}(M_i)$, generating coarsely-harmonized images. We then fine-tune CLDM to translate the coarsely-harmonized source latent map to match the target style: $Z_{X \to T}\{C_X, S_T\} = \boldsymbol{\Phi}(Z_U\{C_X, S_U\})$. This translation is guided by a disentangled CLIP-style loss, which captures style embedding differences between each original MRI and its coarsely-harmonized counterpart. Finally, the harmonized source image is reconstructed by decoding the translated latent map $Z_{X \to T}$: $M_{X \to T} = \boldsymbol{D}(Z_{X \to T}\{C, S_T\})$. This two-stage process allows unseen MRIs from a new site to be harmonized *by only fine-tuning the second stage.*

**Diffusion-based Coarse Harmonizer (DCH).** The DCH is trained to project multi-site MRIs into a coarsely-aligned unified latent domain by leveraging encoder $\boldsymbol{E}$ and CLDM (see Fig. 1). To begin with, we first transform an input MRI $M$ to a low-dimensional latent code $Z^0 = \boldsymbol{E}(M)$ through the pre-trained encoder $\boldsymbol{E}$. The CLDM $\boldsymbol{\Phi}$ is then trained to iteratively reconstruct $Z^0$ via forward diffusion process (FDP) and reverse-diffusion (RDP) operations, governed

by a Markov chain with a total of $T$ timesteps. During FDP, noise is gradually added into $Z^0$ to create a noisy latent map $Z^t$ at each timestep $t$:

$$Z^t = \sqrt{\bar{\alpha}_t} Z^0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

where $\epsilon$ is the sampled noise, $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$, $\alpha_t := 1-\beta_t$, and $\beta_t$ follows a predefined variance schedule. To suppress the influence of image style and provide a content condition, the original latent code $Z^0$ undergoes instance normalization (IN): $\mathrm{IN}(Z^0) = (Z^0 - \mu(Z^0))/\sigma(Z^0)$, where $\mu(\cdot)$ and $\sigma(\cdot)$ compute the mean and standard deviation across channels, effectively removing style information [15, 26]. The resulting $\mathrm{IN}(Z^0)$ can serve as the image content condition. The CLDM, implemented as a time-conditioned 3D U-Net, takes the noisy latent map $Z^t$ at timestep $t$ as input and $\mathrm{IN}(Z^0)$ as conditions to predict the noise $\epsilon_\theta$, which is compared against noise $\epsilon$ added during training via the *noise loss*:

$$\mathcal{L}_N = \|\epsilon - \epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))\|_2^2. \tag{2}$$

During RDP, the CLDM employs a DDIM [23] sampling strategy, which iteratively denoises $Z^t$ over $T_r$ timesteps ($t = T_r : 0$) to recover a coarsely-aligned image $Z_U$ during inference, through the following:

$$Z_U^{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{Z}_U + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))), \quad \hat{Z}_U = \frac{1}{\sqrt{\bar{\alpha}_t}}(Z^t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))), \tag{3}$$

where $\hat{Z}_U$ is the one-step estimate of $Z_U^0$. Typically, $Z_U^t$ is iteratively denoised until $Z_U^0$ is obtained. During training, we use this one-step estimate and decode it to a coarsely-aligned MRI $\hat{M}_U = \boldsymbol{D}(\hat{Z}_U)$ for efficiency in this work.

We further employ two novel *image-level constraints* to enforce MR image content preservation. This is achieved by incorporating a gradient loss $\mathcal{L}_G$ and a perceptual loss $\mathcal{L}_P$ in CLDM. Specifically, the gradient loss is designed to measure the difference between the gradient maps of $\hat{M}_U$ and $M$:

$$G(M) = \frac{1}{3}(\nabla_h M + \nabla_w M + \nabla_d M), \quad \mathcal{L}_G = \|G(\hat{M}_U) - G(M)\|_2^2, \tag{4}$$

where $\nabla$ denotes the gradient operation. The perceptual loss compares feature maps extracted from a pre-trained MedicalNet ResNet-50 [7]:

$$\mathcal{L}_P = \sum\nolimits_l \lambda_l \left\| \psi_l(M) - \psi_l(\hat{M}_U) \right\|_1, \tag{5}$$

where $\psi_l$ denote feature maps extracted from the $l$-th layer of the ResNet-50 and $\lambda_l$ controls the contribution of each layer. These losses defined at the image level ensure content and semantic consistency during training, effectively guiding CLDM to generate style-agnostic reconstruction while preserving anatomical integrity. The hybrid loss for training our DCH is defined as $\mathcal{L}_C = \mathcal{L}_N + \mathcal{L}_G + \mathcal{L}_P$.

**CLIP-Style Guided Fine Harmonizer.** In this stage, we fine-tune the DCH to align a coarsely-aligned source MRI $M_{X\to U}$ into a translated image $M_{X\to T}$ with a style of a target site $T$. We leverage a pre-trained CLIP encoder [29] to implicitly extract style embeddings and develop a *hybrid disentangled CLIP-style loss*, eliminating the need for paired data or explicit style definitions.

Given a source MRI $M_X$ and an unpaired target MRI $M_T$, we first map them to the coarsely-aligned unified domain through FDP defined in Eq. (1) and RDP in Eq. (3) while fixing the weight of the trained DCH:

$$M_{X\to U} = \boldsymbol{D} \circ \boldsymbol{\Phi} \circ \boldsymbol{E}(M_X), \quad M_{T\to U} = \boldsymbol{D} \circ \boldsymbol{\Phi} \circ \boldsymbol{E}(M_T). \tag{6}$$

The fine harmonizer is then trained to adapt $M_{X\to U}$ to match the style of $M_T$ through other FDP and RDP processes. Unlike FDP in DCH, where random noise is added to $Z^0$ through Eq. (1), we iteratively add learned noise from the coarse harmonizer over $T_f$ iterations ($t = 0 : T_f$) to get the noisy latent $Z^{T_f}$:

$$Z_X^0 = Z_{X\to U} = \boldsymbol{E}(M_{X\to U}), \quad Z_X^{t+1} = \sqrt{\bar{\alpha}_{t+1}}\hat{Z}_X^0 + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))), \tag{7}$$

where $\hat{Z}_X^0$ is the one-step estimate similar to $\hat{Z}_U$, see Eq. (3). RDP is then used to iteratively denoise $Z_X^{T_f}$ for $t = T_r : 0$ times to yield the final translated feature:

$$Z_{X\to T}^{T_r} = Z_X^{T_f}, \quad Z_{X\to T}^{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{Z}_{X\to T}^0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))), \tag{8}$$

where $\hat{Z}_{X\to T}^0$ is the one-step estimate. After $T_r$ iterative timesteps, we decode the translated latent map $Z_{X\to T}^0$ to obtain the translated MRI: $M_{X\to T} = \boldsymbol{D}(Z_{X\to T}^0)$.

To align source MRIs with target style, we extract embeddings for both target and coarsely-harmonized source images via a pre-trained CLIP encoder $\Psi$:

$$S_T = \Psi(M_T) - \Psi(M_{T\to U}), \quad S_{X\to T} = \Psi(M_{X\to T}) - \Psi(M_{X\to U}), \tag{9}$$

Since $M_{T\to U}$ is the coarsely-harmonized unified version of the target MRI $M_T$ (with the same content), the difference between their CLIP-space embeddings (*i.e.*, $S_T$) captures the target style information. Similarly, $S_{X\to T}$ captures the style of the harmonized source image. We define the *style translation loss* as:

$$\mathcal{L}_S = \|S_T - S_{X\to T}\|_1 + (1 - (S_T \cdot S_{X\to T})/(\|S_T\|\|S_{X\to T}\|)), \tag{10}$$

where the 1st term is the $l_1$ distance in the CLIP-embedding space and the 2nd term quantifies the directional discrepancy between two style embeddings.

To ensure style consistency, we further design a *style reconstruction loss* by minimizing style embeddings of each target MRI and its harmonized counterpart: $\mathcal{L}_R = \|S_T - S_{T\to T}\|_1$. The hybrid disentangled CLIP-style loss is defined as:

$$\mathcal{L}_F = \mathcal{L}_S + \mathcal{L}_R. \tag{11}$$

By leveraging CLIP's semantic-rich embeddings, our fine harmonizer effectively translates source MRIs to the target style *without requiring explicit image style and content disentanglement learning*, ensuring that anatomical content remains unchanged. The two-stage training strategy used in the proposed UMH allows the coarse harmonizer to be employed universally, requiring only the fine-tuning stage for harmonizing unseen MRI data, providing superior flexibility.

**Implementation.** We use the 3D AutoencoderKL from MONAI [5], with three groups of upsampling/downsampling 3D convolutional layers with residual blocks ({32, 64, 128} channels). CLDM is implemented as a time-conditioned 3D U-Net

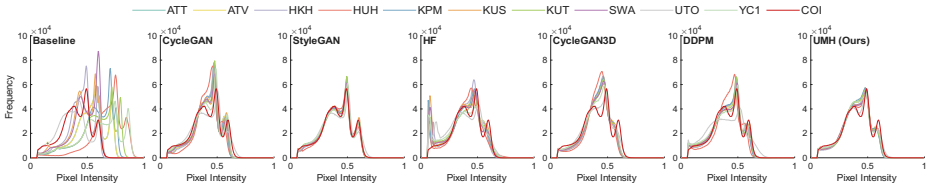


**Fig. 2.** Histograms of 22 SRPBS test MRIs across 11 sites, with COI as the target.

**Table 1.** Comparison between source site MRIs and corresponding target site (COI) MRIs with matching subjects in the SRPBS test set (2 subjects across 11 sites).

| Method | SSIM ↑ | PSNR ↑ | PCC ↑ | WD ↓ |
|---|---|---|---|---|
| Baseline | 0.879 ± 0.033 | 22.020 ± 4.282 | 0.979 ± 0.008 | 0.041 ± 0.026 |
| CycleGAN [6] | 0.868 ± 0.026 | 21.410 ± 1.652 | 0.973 ± 0.009 | 0.028 ± 0.011 |
| StyleGAN [16] | 0.893 ± 0.028 | 23.498 ± 1.756 | 0.975 ± 0.009 | **0.005 ± 0.002** |
| HF [2] | 0.890 ± 0.024 | 23.411 ± 1.427 | 0.976 ± 0.011 | 0.016 ± 0.006 |
| CycleGAN3D [30] | **0.937 ± 0.025** | 26.906 ± 1.771 | 0.988 ± 0.006 | 0.013 ± 0.004 |
| DDPM [10] | 0.869 ± 0.124 | 30.021 ± 10.027 | 0.968 ± 0.032 | 0.015 ± 0.009 |
| UMH (Ours) | 0.933 ± 0.097 | **31.263 ± 2.025** | **0.989 ± 0.005** | **0.005 ± 0.003** |

with a symmetric architecture: 2 upsampling/downsampling layers, 2 residual blocks, 4 self-attention blocks, and a middle block ({32, 64, 64, 64} channels). Both models are trained using the Adam optimizer with default settings and an initial learning rate of $1 \times 10^{-4}$. The variance scheduler $\beta$ is empirically set to increase linearly from 0.0015 to 0.0195. We empirically apply $T = 1,000$ noise steps for the first coarse harmonization stage and choose $T_f = 25, T_r = 30$ for the second fine harmonization stage through hyperparameter grid search.

## 3 Experiments

**Materials.** Three datasets are utilized: (1) OpenBHB [9], with T1w MRIs from 3,984 healthy subjects across 58 sites; (2) SRPBS [24], with T1w MRIs from 9 traveling subjects across 11 sites; and (3) DWI-THP [17] with T1w MRIs from 5 subjects scanned across 8 sites. The OpenBHB dataset is divided into a training set (3,227 MRIs) and a validation set (757 MRIs) for training the 3D autoencoder and CLDM. The SRPBS and DWI-THP datasets are used for fine-tuning and evaluation, with data splits detailed in the experimental section.

**Competing Methods.** We compare UMH with five SOTA image-level MRI harmonization methods: CycleGAN [6], StyleGAN [16], CycleGAN-3D [30], Harmonizing Flow (HF) [2], and DDPM [10]. We ensured consistent training data and hyperparameters across all methods for a fair comparison.

**Task 1: Histogram & Voxel-Level Comparison.** This experiment evaluates harmonization outcomes using SRPBS with ground-truth traveling phantom data. We train on 77 MRIs (7 subjects across 11 sites) and test on 22 MRIs (2 subjects across 11 sites). Each method harmonizes MRIs from 10 source sites

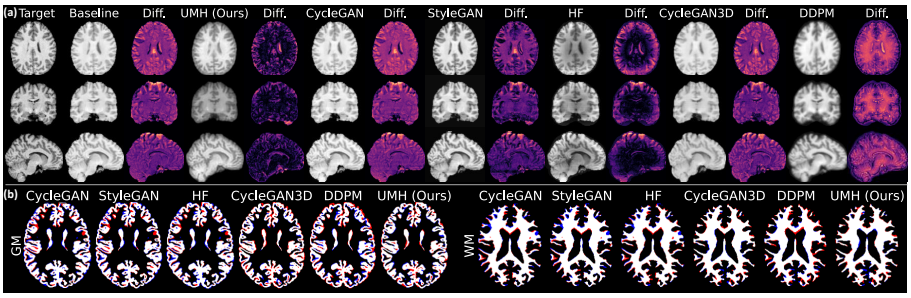


**Fig. 3.** (a) harmonization results and difference (Diff.) and (b) segmentation maps with *White*: accurate segmentation; *Red*: under-segmentation; *Blue*: over-segmentation.

**Table 2.** Results of AP and DSC metrics on GM and WM segmentation in SRPBS.

| Method | Anatomical Preservation (AP) ↑ | | | Dice Similarity Coefficient (DSC) ↑ | | |
|---|---|---|---|---|---|---|
| | GM | WM | Mean | GM | WM | Mean |
| CycleGAN [6] | $0.986 \pm 0.017$ | $\mathbf{0.994 \pm 0.004}$ | $0.990 \pm 0.009$ | $0.880 \pm 0.046$ | $0.920 \pm 0.030$ | $0.900 \pm 0.038$ |
| StyleGAN [16] | $0.992 \pm 0.006$ | $\mathbf{0.994 \pm 0.005}$ | $\mathbf{0.993 \pm 0.004}$ | $0.886 \pm 0.043$ | $0.924 \pm 0.029$ | $0.905 \pm 0.036$ |
| HF [2] | $0.985 \pm 0.016$ | $0.990 \pm 0.009$ | $0.987 \pm 0.010$ | $0.879 \pm 0.046$ | $0.918 \pm 0.029$ | $0.899 \pm 0.038$ |
| CycleGAN3D [30] | $0.988 \pm 0.019$ | $0.992 \pm 0.006$ | $0.990 \pm 0.010$ | $0.886 \pm 0.055$ | $\mathbf{0.932 \pm 0.032}$ | $0.909 \pm 0.043$ |
| DDPM [10] | $0.992 \pm 0.005$ | $0.989 \pm 0.009$ | $0.990 \pm 0.007$ | $0.829 \pm 0.168$ | $0.847 \pm 0.154$ | $0.838 \pm 0.161$ |
| UMH (Ours) | $\mathbf{0.993 \pm 0.010}$ | $0.992 \pm 0.006$ | $\mathbf{0.993 \pm 0.006}$ | $\mathbf{0.898 \pm 0.045}$ | $0.926 \pm 0.030$ | $\mathbf{0.912 \pm 0.038}$ |

to a target site, selected as COI due to its highest mean peak-signal-to-noise ratio (PSNR). Performance is assessed via histogram comparisons and voxel-level metrics: PSNR, structural similarity index (SSIM), Wasserstein distance (WD), and Pearson correlation coefficient (PCC) with raw source MRIs as Baseline. The results in Fig. 2 show significant site-wise intensity variations in raw MRIs. Our method effectively aligns the source histograms to the target's, performing comparably to StarGAN without explicit style learning. Quantitative results in Table 1 suggest that our UMH yields the highest PSNR and PCC, indicating superior voxel-level agreement, and the lowest WD, confirming effective style alignment. It also achieves the second-best SSIM, demonstrating strong anatomical preservation. Sample visualization in Fig. 3 (a) shows that UMH harmonized MRI is closer to the target. These results highlight UMH's ability to harmonize MRIs while maintaining high image quality and anatomical fidelity.

**Task 2: Brain Tissue Segmentation.** We further evaluate anatomical preservation via a brain tissue segmentation task. Each method is trained on 24 MRIs (3 subjects across 8 sites) and harmonizes 16 MRIs from 2 additional subjects in the DWI-THP dataset with site CCF as the target. We use FreeSurfer [3] to generate gray matter (GM) and white matter (WM) segmentation maps for original and harmonized MRIs. Segmentation quality is assessed using the Anatomical Preservation (AP) score [19], and Dice Similarity Coefficient (DSC). The results in Table 2 show that our method achieves the highest performance for GM and mean scores and the second-highest for WM. Figure 3 (b) further shows that UMH yields fewer segmentation errors in the WM and GM tissue boundaries.

**Table 3.** Site classification and age prediction results on harmonized OpenBHB MRIs.

| Method | Site Classification | | | | Age Prediction | |
|---|---|---|---|---|---|---|
| | Balanced Accuracy (BAC) ↓ | F1 ↓ | Precision (PRE) ↓ | Recall ↓ | MAE ↓ | MSE ↓ |
| Baseline | $0.343 \pm 0.024$ | $0.663 \pm 0.023$ | $0.757 \pm 0.018$ | $0.732 \pm 0.019$ | $5.295 \pm 0.260$ | $47.446 \pm 1.405$ |
| CycleGAN [6] | $0.425 \pm 0.016$ | $0.695 \pm 0.027$ | $0.770 \pm 0.030$ | $0.739 \pm 0.020$ | $6.625 \pm 0.264$ | $78.951 \pm 10.541$ |
| StyleGAN [16] | $0.258 \pm 0.022$ | $0.593 \pm 0.012$ | $0.662 \pm 0.015$ | $0.651 \pm 0.015$ | $7.314 \pm 0.494$ | $85.716 \pm 12.905$ |
| HF [2] | $0.342 \pm 0.011$ | $0.665 \pm 0.020$ | $0.736 \pm 0.021$ | $0.723 \pm 0.021$ | $5.835 \pm 0.221$ | $57.009 \pm 3.849$ |
| CycleGAN3D [30] | $0.324 \pm 0.029$ | $0.656 \pm 0.019$ | $0.751 \pm 0.019$ | $0.723 \pm 0.017$ | $5.901 \pm 0.360$ | $\mathbf{33.348 \pm 9.671}$ |
| DDPM [10] | $0.166 \pm 0.016$ | $0.560 \pm 0.013$ | $\mathbf{0.545 \pm 0.007}$ | $0.535 \pm 0.020$ | $5.333 \pm 0.258$ | $48.548 \pm 6.796$ |
| UMH (Ours) | $\mathbf{0.129 \pm 0.013}$ | $\mathbf{0.510 \pm 0.035}$ | $0.625 \pm 0.024$ | $\mathbf{0.522 \pm 0.042}$ | $\mathbf{5.240 \pm 0.141}$ | $52.022 \pm 5.052$ |

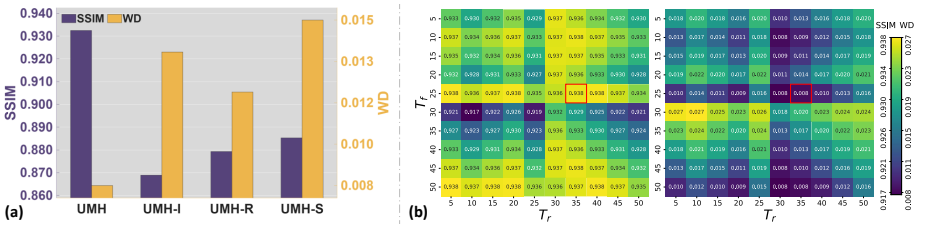


**Fig. 4.** Results of (a) UMH and its variants, and (b) UMH with different $T_f$ and $T_r$.

**Task 3: Site Classification & Brain Age Prediction.** We evaluate UMH in reducing site-related variations while preserving anatomical information through site classification and brain age prediction. Each method is trained on OpenBHB training data and applied to harmonize the validation data with Site 17 selected as the target site. We extract features from these harmonized MRIs using ResNet18 [14] with its final layer removed. A logistic regression, trained on 70% of these deep features and tested on 30%, performs multi-class site classification, and a ridge regressor predicts brain age. Both tasks are repeated 5 times for random data partition, with mean±standard deviation results reported in Table 3. Lower site classification results indicate better removal of site-related variations, while lower age prediction error suggests superior anatomical feature preservation. Table 3 shows that UMH yields the lowest site classification performance in BAC, F1, and Recall, effectively removing site-related features and maintaining faithful anatomical integrity with slightly lower mean absolute error (MAE).

**Ablation Study.** We validate three key components by comparing UMH with its three degraded variants: UMH-I (no image-level constraints during coarse harmonizer training), UMH-R (without style reconstruction loss), and UMH-S (without style translation loss). Results in Fig. 4 (a) show that UMH-I achieves the lowest SSIM, indicating our image-level constraints help preserve structural details. Additionally, UMH-S yields the highest WD, reflecting poor style alignment without style translation loss. UMH-R's suboptimal performance further underscores the importance of style reconstruction loss during fine-tuning.

**Parameter Analysis.** We conduct a hyperparameter grid search on $T_f$ (forward diffusion steps) and $T_r$ (reverse diffusion steps), evaluating volume-level metrics on the SRPBS test set. Results in Fig. 4 (b) show that our model performs better with larger $T_r$ than $T_f$, with peak performance when $T_f = 25$ and $T_r = 30$.

**Computational Cost Comparison.** The proposed UMH employs a lightweight latent-diffusion U-Net and a 3D autoencoder with 3.0M and 3.3M trainable parameters, respectively, totaling 6.3M—smaller than CycleGAN (28.3M), StyleGAN (161.3M), CycleGAN3D (22.6M), and DDPM (10.3M), and comparable to HF (5.7M). Training UMH on the SRPBS cohort takes about 6.5 hours (3.0h for Stage 1 on OpenBHB; 2.5h for Stage 2 fine-tuning), which is faster than CycleGAN (16.2h), StyleGAN (10.5h), HF (13.4h), and CycleGAN3D (12.4h), and comparable to DDPM (5.5h). All models were trained on an H100 GPU.

## 4 Conclusion

We propose an unpaired MRI harmonization framework using a latent diffusion model, which aligns MRIs into a unified latent space and translates them into the target style via disentangled CLIP-based style losses. Leveraging the semantic-rich CLIP embeddings, our method enables effective volume-level harmonization without paired MRIs or auxiliary style encoders. Extensive experiments across three multi-site datasets demonstrate UMH's superiority in style alignment, site variation removal, and anatomical preservation.

**Acknowledgments.** This study was supported in part by NIH grants (Nos. AG073297, AG082938, EB035160, and NS13484).

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

## References

1. An, L., Chen, J., Chen, P., Zhang, C., He, T., Chen, C., Zhou, J.H., Yeo, B.T., of Aging, L.S., Initiative, A.D.N., et al.: Goal-specific brain MRI harmonization. Neuroimage **263**, 119570 (2022)
2. Beizaee, F., Desrosiers, C., Lodygensky, G.A., Dolz, J.: Harmonizing Flows: Unsupervised MR harmonization based on normalizing flows. In: International Conference on Information Processing in Medical Imaging. pp. 347–359. Springer (2023)
3. Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E.: Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proceedings of the National Academy of Sciences **120**(9), e2216399120 (2023)
4. Cackowski, S., Barbier, E.L., Dojat, M., Christen, T.: Imunity: A generalizable VAE-GAN solution for multicenter MR image harmonization. Medical Image Analysis **88**, 102799 (2023)
5. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
6. Chang, X., Cai, X., Dan, Y., Song, Y., Lu, Q., Yang, G., Nie, S.: Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. Physics in Medicine & Biology **67**(14), 145004 (2022)
7. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)

8. Dewey, B.E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E.M., Newsome, S., Oh, J., Calabresi, P.A., Prince, J.L.: A disentangled latent space for cross-site MRI harmonization. In: Medical Image Computing and Computer-Assisted Intervention. pp. 720–729. Springer (2020)
9. Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E.: OpenBHB: A large-scale multi-site brain MRI data-set for age prediction and debiasing. NeuroImage **263**, 119637 (2022)
10. Durrer, A., Wolleb, J., Bieder, F., Sinnecker, T., Weigel, M., Sandkühler, R., Granziera, C., Yaldizli, Ö., Cattin, P.C.: Diffusion models for contrast harmonization of magnetic resonance images. arXiv preprint arXiv:2303.08189 (2023)
11. Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al.: Harmonization of cortical thickness measurements across scanners and sites. NeuroImage **167**, 104–120 (2018)
12. Gadewar, S.P., Zhu, A.H., Gari, I.B., Somu, S., Thomopoulos, S.I., Thompson, P.M., Nir, T.M., Jahanshad, N.: Synthesizing study-specific controls using generative models on open access datasets for harmonized multi-study analyses. arXiv preprint arXiv:2403.00093 (2024)
13. Glocker, B., Robinson, R., Castro, D.C., Dou, Q., Konukoglu, E.: Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. arXiv preprint arXiv:1910.04597 (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
16. Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N.: Style transfer using generative adversarial networks for multi-site MRI harmonization. In: Medical Image Computing and Computer Assisted Intervention, Part III 24. pp. 313–322. Springer (2021)
17. Magnotta, V.A., Matsui, J.T., Liu, D., Johnson, H.J., Long, J.D., Bolster Jr, B.D., Mueller, B.A., Lim, K., Mori, S., Helmer, K.G., et al.: Multicenter reliability of diffusion tensor imaging. Brain Connectivity **2**(6), 345–355 (2012)
18. Modanwal, G., Vellal, A., Buda, M., Mazurowski, M.A.: MRI image harmonization using cycle-consistent generative adversarial network. In: Computer-Aided Diagnosis. vol. 11314, pp. 259–264. SPIE (2020)
19. Parida, A., Jiang, Z., Packer, R.J., Avery, R.A., Anwar, S.M., Linguraru, M.G.: Quantitative metrics for benchmarking medical image harmonization. In: IEEE ISBI. pp. 1–5. IEEE (2024)
20. Pfaff, L., Wagner, F., Vysotskaya, N., Thies, M., Maul, N., Mei, S., Wuerfl, T., Maier, A.: No-new-denoiser: A critical analysis of diffusion models for medical image denoising. In: MICCAI. pp. 568–578. Springer (2024)
21. Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., et al.: Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage **208**, 116450 (2020)
22. Schnack, H.G., van Haren, N.E., Brouwer, R.M., van Baal, G.C.M., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T.D., Huttunen, M., Lepage, C., et al.: Mapping reliability in multicenter MRI: Voxel-based morphometry and cortical thickness. Human Brain Mapping **31**(12), 1967–1982 (2010)

23. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
24. Tanaka, S., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., Imamizu, H.: A multi-site, multi-disorder resting-state magnetic resonance image database. Scientific Data **8**(1), 227 (2021)
25. Tofts, P., Collins, D.: Multicentre imaging measurements for oncology and in the brain. The British Journal of Radiology **84**, S213–S226 (2011)
26. Ulyanov, D.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
27. Wu, M., Zhang, L., Yap, P.T., Zhu, H., Liu, M.: Disentangled latent energy-based style translation: An image-level structural MRI harmonization framework. Neural Networks **184**, 107039 (2025)
28. Yu, M., Guan, H., Fang, Y., Yue, L., Liu, M.: Domain-prior-induced structural MRI adaptation for clinical progression prediction of subjective cognitive decline. In: MICCAI. pp. 24–33. Springer (2022)
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024)
30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)
31. Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A.: Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. NeuroImage **243**, 118569 (2021)