SPATIAL-DIFFERENTIAL GUIDANCE AND STRUCTURE ANCHORING FOR SELECTIVE MILD MOTION ARTIFACT CORRECTION IN HEAD CT

Anonymous authors
Paper under double-blind review

ABSTRACT

Head CT scans are critical for diagnosing acute neurological conditions, yet involuntary patient motion during scanning frequently introduces motion artifacts that degrade image quality. A particularly challenging scenario is mild motion artifacts in stationary CT, where artifacts are axially sparse but spatially entangled with normal anatomical structures. Consequently, correction requires both precise artifact localization and preservation of anatomical consistency. To address these challenges, we propose a spatial-differential and structure-anchoring framework for head CT motion artifact correction. The framework introduces an explicit gradient-based discriminative response field (DRF) as a spatial prior, which is incorporated through a spatial-differential attention (SDA) module to guide the correction network toward artifact-corrupted regions. Furthermore, we propose a semantic structure anchoring (SSA) strategy that aligns intermediate representations with artifact-free anatomical features from a pretrained 3D visual encoder, encouraging structural consistency during restoration. Extensive experiments on simulated datasets with varying artifact severity, a public dataset, and a real-world dataset demonstrate our framework's effectiveness. The proposed method achieves improved restoration quality on mild motion artifacts, reducing MAE by 22.7% and achieving 37.62dB PSNR, while maintaining strong performance across diverse evaluation settings.

1 INTRODUCTION

Head computed tomography (CT) scans are critical for diagnosing acute neurological conditions such as intracerebral hemorrhage and ischemic stroke (Campbell et al., 2019; Greenberg et al., 2022). However, involuntary patient motion during scanning introduces motion artifacts into the reconstructed images, severely compromising image quality and diagnostic reliability. While severe motion artifacts produce conspicuous degradation that is readily identified by radiologists, prompting a repeat scan, mild motion artifacts are more insidious and clinically hazardous. These artifacts often manifest as streak-like or blurred structures superimposed on normal anatomy, which can mimic or obscure subtle lesions, leading to missed or misdiagnoses cases (Boas et al., 2012; Barrett & Keat, 2004). Therefore, accurate correction of mild motion artifacts is of significant clinical value for avoiding unnecessary repeat scans and preserving the diagnostic utility of head CT scans.

However, correcting mild motion artifacts presents two major challenges, as illustrated in Figure 1. On the one hand, these artifacts are axially sparse but spatially entangled with normal anatomical structures, necessitating fine-grained localization to distinguish artifact-corrupted regions from those that should be preserved. Existing methods mainly learn a direct mapping from corrupted to clean images or employ generative models for restoration. However, these approaches generally lack explicit artifact localization and may perform unnecessary corrections on unaffected anatomical regions.

To address these limitations, we propose a novel spatial-differential and structure-anchoring framework for mild motion artifact correction. The core idea is to guide the restoration process with fine-grained artifact localization while anchoring it to artifact-free anatomical representations to preserve structural fidelity. Specifically, we derive an explicit gradient-based discriminative response

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

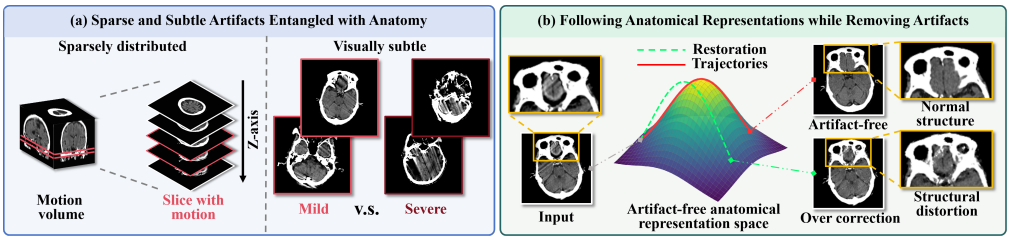

Figure 1: Two major challenges in head CT motion artifact correction. (a) Mild motion artifacts are axially sparse and visually subtle, entangling with anatomy. (b) The correction process should follow the anatomical representations while removing artifacts to avoid over-correction.

field (DRF) as a spatial prior, which is injected into the correction network via a spatial-differential attention (SDA) module to direct the network's representational capacity toward artifact-corrupted regions while leaving normal tissues largely untouched. To prevent over-correction, we further propose a semantic structure anchoring (SSA) strategy that aligns intermediate features of the correction network with high-quality artifact-free representations derived from a pre-trained visual encoder, anchoring the restoration trajectory to artifact-free anatomical feature representations. Extensive experiments across multiple datasets validate the effectiveness of our framework, demonstrating its superior performance and robust generalization.

The main contributions of this paper are summarized as follows:

- We formulate mild head CT motion artifact correction as a spatially selective restoration problem and propose a framework that explicitly addresses artifact localization and anatomical preservation.
- We design the DRF mechanism and SDA module to translate artifact discriminative knowledge into a spatial prior, guiding the network to focus correction on artifact-corrupted regions.
- We further introduce the SSA strategy to anchor the restoration trajectory to the artifact-free anatomical representations, preventing over-correction and structural distortions throughout the process.
- Extensive experiments on datasets with varying artifact severity, a public dataset, and a real-world dataset validate the effectiveness of our framework across diverse artifact severities and evaluation datasets.

2 Related Work

Motion artifact correction can be approached from either projection-domain or image-domain data (Maier et al., 2023; Sun et al., 2016; Bruder et al., 2016; Ouadah et al., 2017; Zhang et al., 2020). However, raw projection data is rarely accessible in routine clinical CT workflows, which has shifted research emphasis toward image-domain methods in recent years. Current image-domain motion artifact correction has advanced along two primary technical paths. The first employs CNNs to learn direct mappings from corrupted images to clean ones. Ko et al. (2021) introduced a channel attention mechanism into residual networks, achieving motion compensation by adaptively amplifying or attenuating residual features. Su et al. (2022) further proposed a U-Net (Ronneberger et al., 2015) architecture with dilated ResNet (He et al., 2016) blocks, which preserves high resolution while leveraging a large receptive field to capture long-range streak artifacts, and improves restoration quality through explicit residual prediction. Al-masni et al. (2022) proposed a stacked U-Net with self-supervised priors to preserve fine-grained details. The second path leverages generative models, which exploit distribution modeling capabilities to achieve higher-quality restoration. Chen et al. (2025) proposed HM-EDM, adapting the Elucidated Diffusion Model (EDM) (Karras et al., 2022) to portable head CT with histogram equalization to bridge the intensity gap between skull and brain tissue, achieving effective correction for severe motion artifacts in portable scanners. However, the aforementioned methods lack explicit spatial localization capabilities for artifact regions and tend

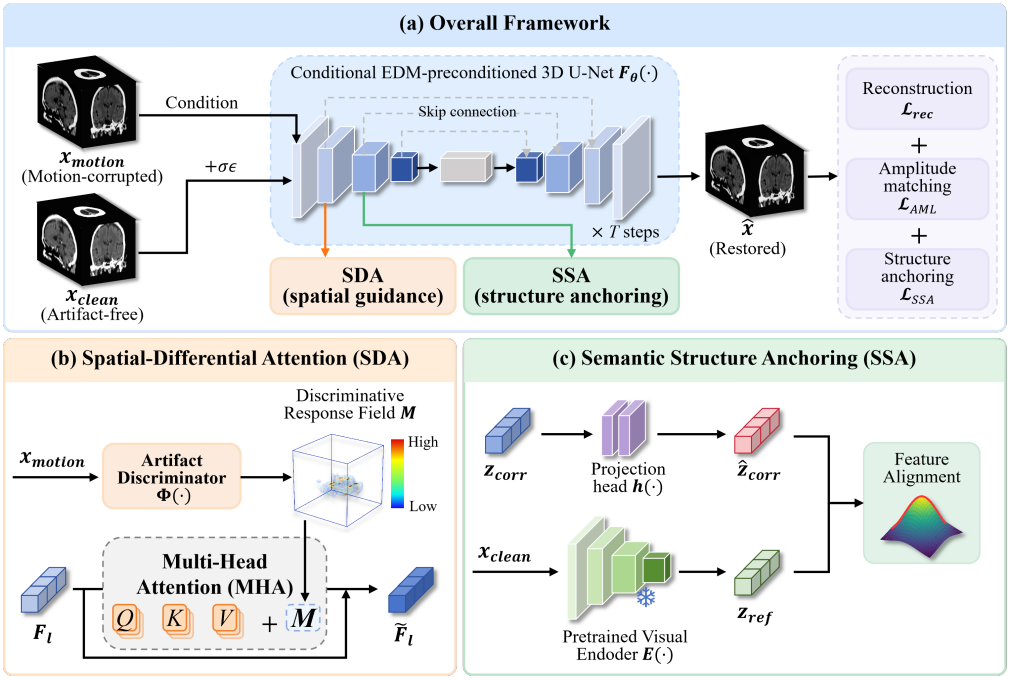

Figure 2: Architecture of the proposed framework. (a) Overall Framework: The conditional EDM-preconditioned 3D U-Net takes the noisy target x_σ and motion-corrupted volume x_{motion} as inputs, while SDA provides spatial guidance and SSA enforces anatomical feature alignment. (b) Spatial-Differential Attention (SDA) Module: the discriminative response field M is injected as a spatial bias into attention layers. (c) Semantic Structure Anchoring (SSA) Module: intermediate restoration features are aligned with reference features extracted from a frozen pretrained 3D visual encoder through $\mathcal{L}_{SSA}$.

to perform homogeneous restoration across the entire image or volume, making them unable to accurately correct mild motion artifacts.

Moreover, the absence of structural preservation constraints causes these methods to struggle to maintain the integrity of normal anatomical structures during artifact removal, leading to over-correction. Existing structural preservation strategies broadly fall into two categories. One prevalent approach employs perceptual losses (Gatys et al., 2016; Johnson et al., 2016) using pre-trained networks (e.g., VGG (Simonyan & Zisserman, 2015)) to enforce feature-space similarity between generated and target images. Another line of work resorts to explicit semantic supervision, where anatomical priors from segmentation networks regularize specific organs or tissues (Dalca et al., 2018; Zhang et al., 2018). Nevertheless, these methods face challenges in mild motion artifact scenarios. Artifacts are embedded within normal anatomical structures with subtle visual differences and insufficient contrast. Consequently, constraints based solely on global image-wise similarity or discrete pixel-space anatomical priors struggle to provide effective supervision. Recently, representation alignment strategies (Yu et al., 2025; Zheng et al., 2025; Singh et al., 2025) have provided a new avenue for preserving anatomical structure consistency. By aligning intermediate features of the generative model with high-quality self-supervised representations extracted from a pre-trained visual encoder, this approach enables the generation process to implicitly inherit rich spatial and semantic information, thereby achieving semantic structural consistency in the feature representation space. Inspired by this, we propose a semantic structure anchoring strategy that anchors the intermediate features of the correction process to the artifact-free anatomical representations. Simultaneously employing an explicit spatial-differential prior to guide the correction network to focus on artifact regions, thereby achieving accurate artifact removal while preserving normal structures.

SDA	SSA	AML	Brain Region				Overall Foreground			
			MAE↓	RMSE↓	SSIM↑	PSNR↑	MAE↓	RMSE↓	SSIM↑	PSNR↑
			24.50 ± 19.08	67.94 ± 39.85	0.8755 ± 0.0633	34.06 ± 4.41	65.45 ± 31.86	131.80 ± 53.60	0.8898 ± 0.0556	27.43 ± 3.32
			23.43 ± 18.97	64.14 ± 43.77	0.8843 ± 0.0603	34.84 ± 4.85	63.55 ± 33.35	133.38 ± 58.06	0.8943 ± 0.0583	27.75 ± 3.52
		✓	23.20 ± 16.34	60.16 ± 34.03	0.8862 ± 0.0642	35.07 ± 4.35	63.38 ± 27.64	132.94 ± 45.87	0.8898 ± 0.0560	27.57 ± 3.03
	✓		22.33 ± 18.19	62.05 ± 39.74	0.8937 ± 0.0612	35.05 ± 4.79	62.77 ± 32.92	132.53 ± 55.77	0.8971 ± 0.0574	27.81 ± 3.57
✓			21.75 ± 16.20	60.64 ± 36.50	0.8903 ± 0.0593	35.21 ± 4.76	60.92 ± 30.27	128.53 ± 50.65	0.8986 ± 0.0605	28.04 ± 3.53
	✓	✓	20.62 ± 18.36	51.82 ± 40.06	0.8945 ± 0.0594	37.06 ± 5.46	55.94 ± 32.44	115.26 ± 54.40	0.9112 ± 0.0563	29.11 ± 3.70
✓		✓	18.61 ± 15.36	51.57 ± 41.49	0.9062 ± 0.0529	37.28 ± 5.70	52.18 ± 31.25	111.42 ± 57.65	0.9174 ± 0.0610	29.55 ± 4.01
✓	✓		19.35 ± 16.81	50.10 ± 38.25	0.8988 ± 0.0585	37.31 ± 5.37	52.82 ± 30.91	109.85 ± 53.37	0.9169 ± 0.0543	29.58 ± 3.84
✓	✓	✓	$\mathbf{18.12 \pm 14.78}$	$\mathbf{48.70 \pm 36.53}$	$\mathbf{0.9065 \pm 0.0538}$	$\mathbf{37.62 \pm 5.49}$	$\mathbf{50.42 \pm 28.24}$	$\mathbf{106.99 \pm 50.88}$	$\mathbf{0.9211 \pm 0.0536}$	$\mathbf{29.82 \pm 3.90}$

Table 1: Ablation study on the Mild Dataset. SDA, SSA, and AML denote the spatial-differential attention, semantic structure anchoring, and spectral amplitude matching loss, respectively. All metrics are computed on motion-corrupted slices only and reported as mean $\pm$ std. **Bold** indicates the best result in each column.

3 METHODOLOGY

3.1 OVERVIEW

As illustrated in Figure 2, the proposed framework is built upon a conditional EDM restoration process with two complementary mechanisms: an explicit spatial-differential prior for artifact localization and a semantic structure anchoring strategy for anatomical preservation.

Given a motion-corrupted volume x_{motion} and noisy target x_σ, the conditional EDM network progressively restores the artifact-free volume. The key challenge in mild motion artifact correction is that artifacts are spatially sparse along the axial direction and highly entangled with normal anatomical structures. Therefore, uniform restoration may either leave residual artifacts or introduce unnecessary anatomical distortion.

To address this challenge, we introduce three synergistic components. First, a gradient-based Discriminative Response Field (DRF) is derived from an artifact discriminative function to provide an explicit spatial prior indicating artifact-sensitive regions. Second, the Spatial-Differential Attention (SDA) module incorporates the DRF into the restoration network, guiding feature aggregation toward corrupted regions while preserving unaffected anatomy. Third, the Semantic Structure Anchoring (SSA) module aligns intermediate restoration features with artifact-free anatomical representations extracted from a pretrained 3D visual encoder, constraining the restoration trajectory toward anatomically consistent solutions.

3.2 EXPLICIT SPATIAL-DIFFERENTIAL PRIOR

Mild motion artifacts are often distributed sparsely along the axial direction and visually intertwined with normal anatomy. To enable selective correction, we introduce a gradient-based Discriminative Response Field (DRF) that converts artifact-discriminative knowledge into an explicit spatial prior.

Given an input volume $x \in \mathbb{R}^{H \times W \times D}$, we employ a differentiable artifact discriminative function $\Phi(\cdot)$ that operates on axial slices. For the i-th slice x_i, the artifact response is defined as:

$$s = \sum_{i=1}^{D} \Phi(x_i), \tag{1}$$

where $\Phi(\cdot)$ outputs the confidence of motion artifact presence.

To obtain fine-grained spatial information, we calculate the gradient of the artifact response with respect to the input volume:

$$g = \frac{\partial s}{\partial x} \in \mathbb{R}^{H \times W \times D}. \tag{2}$$

The gradient magnitude reflects the contribution of each voxel to artifact discrimination. We normalize the gradient response using min-max normalization:

$$M = \frac{|g| - \min(|g|)}{\max(|g|) - \min(|g|) + \epsilon} \in [0,1]^{H \times W \times D}. \tag{3}$$

The resulting DRF M provides an explicit spatial prior indicating regions that are more likely to contain motion-related degradation.

3.3 Spatial-Differential Attention Injection

The DRF provides spatial guidance for artifact localization. To incorporate this prior into the restoration process, we propose a Spatial-Differential Attention (SDA) module that introduces the DRF as a spatial bias into the self-attention operation.

Let $F_l \in \mathbb{R}^{C_l \times H_l \times W_l \times D_l}$ denote the feature map at layer l. The DRF is downsampled to the corresponding spatial resolution:

$$M_l \in [0,1]^{H_l \times W_l \times D_l}. \tag{4}$$

After flattening the spatial dimensions into $N_l = H_l W_l D_l$ tokens, the DRF response is represented as:

$$m_l \in [0,1]^{N_l}. \tag{5}$$

For the h-th attention head, the attention weight between query token i and key token j is defined as:

$$a_{ij}^{(h)} = \mathrm{softmax}_j \left(\frac{(q_i^{(h)})^T k_j^{(h)}}{\sqrt{d}} + m_{l,j} \right), \tag{6}$$

where $q_i^{(h)}$, $k_j^{(h)}$, and d denote the query, key, and feature dimension, respectively.

The attention output is obtained through standard value aggregation:

$$\tilde{F}_l = F_l + \mathrm{MHA}_{\mathrm{SDA}}(F_l, M_l). \tag{7}$$

By introducing the DRF as an explicit spatial bias, SDA enables the restoration network to preferentially aggregate information from artifact-sensitive regions while maintaining the original representations of normal tissues.

3.4 Semantic Structure Anchoring

Mild motion artifacts are tightly coupled with anatomical structures, creating a risk of overcorrection. To preserve anatomical consistency, we introduce Semantic Structure Anchoring (SSA), which constrains intermediate restoration representations toward artifact-free anatomical representations.

Let $E(\cdot)$ denote a frozen pretrained 3D visual encoder. Given an artifact-free target volume x_{clean}, the reference anatomical representation is extracted as:

$$z_{\mathrm{ref}} = E(x_{\mathrm{clean}}). \tag{8}$$

Meanwhile, an intermediate feature map z_{corr} is extracted from the correction network. Since the two feature spaces have different dimensions, a trainable projection head $h(\cdot)$ is introduced:

$$\hat{z}_{\text{corr}} = h(z_{\text{corr}}). \tag{9}$$

The semantic alignment loss is calculated using voxel-wise cosine similarity:

$$L_{\text{SSA}} = 1 - \frac{1}{N}\sum_{n=1}^{N} \frac{\hat{z}_{\text{corr}}^{(n)} \cdot z_{\text{ref}}^{(n)}}{\|\hat{z}_{\text{corr}}^{(n)}\|_2 \|z_{\text{ref}}^{(n)}\|_2}. \tag{10}$$

By aligning restoration features with artifact-free anatomical representations, SSA constrains the restoration trajectory toward anatomically consistent solutions and reduces structural distortion caused by excessive artifact removal.

3.5 Loss Functions

The correction network is trained within the EDM framework. Given paired motion-corrupted and artifact-free volumes $(x_{\text{motion}}, x_{\text{clean}})$, Gaussian noise is added to the clean target:

$$x_\sigma = x_{\text{clean}} + \sigma\epsilon, \tag{11}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and σ is sampled from the EDM noise distribution.

The denoising prediction is formulated as:

$$\hat{x}_0 = c_{\text{skip}}(\sigma)x_\sigma + c_{\text{out}}(\sigma)F_\theta(c_{\text{in}}(\sigma)x_\sigma, c_{\text{noise}}(\sigma); x_{\text{motion}}, M), \tag{12}$$

where F_θ denotes the conditional denoising network and M represents the DRF spatial prior.

The EDM reconstruction objective is:

$$L_{\text{rec}} = \mathbb{E}_{\sigma,\epsilon}\left[w(\sigma)\|\hat{x}_0 - x_{\text{clean}}\|_2^2\right]. \tag{13}$$

To recover frequency components suppressed by motion artifacts, we introduce an amplitude matching loss:

$$L_{\text{AML}} = \mathbb{E}\left[\||\mathcal{F}(\hat{x}_0)| - |\mathcal{F}(x_{\text{clean}})|\|_2^2\right], \tag{14}$$

where $\mathcal{F}(\cdot)$ denotes the 3D Fourier transform.

Finally, the overall optimization objective is:

$$L_{\text{total}} = \lambda_{\text{rec}}L_{\text{rec}} + \lambda_{\text{AML}}L_{\text{AML}} + \lambda_{\text{SSA}}L_{\text{SSA}}. \tag{15}$$

4 Experiments

4.1 Datasets and Preprocessing

The in-house dataset comprises multi-center non-contrast head CT scans. All scans were visually inspected by radiologists and categorized into 360 artifact-free volumes and 360 real motion-artifact volumes. All in-house volumes were resampled to a uniform spatial resolution of 1.0×1.0×2.5 mm^3. Each volume was then padded to 80 slices, and each slice was resized from 512×512 to 256×256 pixels. Finally, intensities were clipped to [-1000, 2000]HU and min–max normalized to [-1, 1]. Motion artifacts were simulated following Chen et al. (2024) Chen et al. (2024) using rigid motion trajectories and CT projection-reconstruction simulation. Based on this pipeline, we constructed four datasets:

Mild Dataset. Mild artifacts were simulated on the 360 artifact-free volumes from the in-house dataset. For each volume, the proportion of motion-corrupted slices was randomly sampled from

Method	Mild Dataset				Severe Dataset				Public Dataset			
	MAE↓	RMSE↓	SSIM↑	PSNR↑	MAE↓	RMSE↓	SSIM↑	PSNR↑	MAE↓	RMSE↓	SSIM↑	PSNR↑
Brain Region ([-100, 100]HU)												
Motion images	24.50	67.94	0.8755	34.06	83.62	202.62	0.7907	23.73	14.14	44.42	0.9191	37.41
Pix2Pix	24.14	67.46	0.8778	34.14	73.78	155.77	0.8198	26.12	24.58	54.95	0.8994	35.24
cDDPM	58.22	106.48	0.8295	29.71	53.63	148.33	0.8516	26.58	20.43	34.32	0.9388	39.20
CNN-Att	27.75	76.70	**0.9172**	32.64	83.64	207.35	0.7958	23.51	16.97	57.66	**0.9470**	34.82
CNN-Res	32.06	74.78	0.8338	32.93	69.98	157.93	0.7660	28.43	26.79	61.11	0.8756	34.15
EDM	23.43	64.14	0.8843	34.84	54.91	158.51	0.8464	25.92	13.22	28.92	0.9348	41.10
Rectified Flow	25.55	61.28	0.8764	35.13	53.10	151.00	0.8478	26.44	12.48	28.81	0.9247	41.16
DDBM	24.98	67.86	0.8793	33.95	83.27	199.69	0.7914	23.86	14.19	43.28	0.9253	37.61
MeanFlows	27.47	72.71	0.8848	33.34	70.70	177.99	0.8173	24.89	14.51	39.09	0.9273	38.77
HM-EDM	25.66	74.47	0.8791	33.24	61.15	172.36	0.8350	25.26	**9.50**	29.32	0.9466	41.36
Ours	**18.12**	**48.70**	0.9065	**37.62**	**38.20**	**110.91**	**0.8901**	**35.69**	9.96	**25.56**	0.9382	**42.46**
Overall Foreground ([-100, +∞]HU)												
Motion images	65.45	131.80	0.8898	27.43	183.93	339.76	0.6821	19.10	43.63	100.05	0.9239	29.99
Pix2Pix	65.59	136.95	0.8901	27.42	154.54	281.19	0.7249	20.76	55.43	109.27	0.9030	29.07
cDDPM	104.15	181.71	0.8485	24.80	150.58	303.32	0.7369	20.14	38.76	72.96	0.9530	32.60
CNN-Att	75.55	164.08	0.8881	25.52	184.82	344.22	0.6824	18.97	51.46	129.89	0.9202	27.47
CNN-Res	75.73	150.93	0.8649	26.48	148.63	271.26	0.7463	23.07	64.46	132.26	0.8824	27.30
EDM	63.55	133.38	0.8943	27.75	145.75	298.49	0.7396	20.32	32.20	69.02	0.9510	33.28
Rectified Flow	61.84	123.94	0.8989	28.36	135.04	274.05	0.7476	21.03	30.84	65.98	0.9513	33.63
DDBM	66.68	138.95	0.8883	27.20	183.91	338.89	0.6831	19.12	42.51	97.93	0.9267	30.12
MeanFlows	71.41	147.82	0.8779	26.67	162.23	309.72	0.7104	19.90	37.20	83.85	0.9352	31.62
HM-EDM	70.06	148.50	0.8773	26.67	163.00	325.12	0.7160	19.58	29.59	69.68	0.9544	33.17
Ours	**50.42**	**106.99**	**0.9211**	**29.82**	**97.53**	**204.83**	**0.8194**	**30.35**	**25.85**	**57.79**	**0.9605**	**34.89**

Table 2: Comparison with state-of-the-art methods in the brain region ([-100, 100]HU) and the overall foreground ([-100, $+\infty$]HU) on motion-corrupted slices of the Mild, Severe, and Public Datasets. Mean values are reported. **Bold** with red shading indicates the best mean in each column; blue shading indicates the second best.

the range [10%, 30%], with rotation amplitudes of 1–2° and translation amplitudes of 1–2 mm, and the number of directional changes per volume was at most 2. The simulated mild artifacts closely approximate the distribution of the real-world motion artifact dataset, with Fréchet Inception Distance (FID) (Heusel et al., 2017) reaching 5.31.

Severe Dataset. Severe artifacts were simulated on the same 360 artifact-free volumes, with rotation amplitudes of 5–10°, translation amplitudes of 3–6 mm, and 40–60% of slices corrupted per volume, with all other settings identical to those in the Mild Dataset.

Public Dataset. To evaluate the robustness of the proposed framework on an independent public cohort, we additionally constructed a simulated dataset from the SynthRAD2023 challenge dataset. Specifically, only the brain subset was used, containing 360 CT volumes. The same preprocessing and motion simulation pipeline as the Mild Dataset was applied, while the original spatial resolution of $1 \times 1 \times 1$ mm^3 was preserved and each volume was resized to $256 \times 256 \times 128$ voxels.

Real-world Dataset. The real motion-corrupted volumes from the in-house dataset were used exclusively for real-world evaluation. Since paired artifact-free references are unavailable, the corrected volumes were evaluated using distribution-based metrics by comparing them with artifact-free volumes.

Mild, Severe, and Public Datasets were each split into training and testing sets at a 4:1 ratio without patient overlap, while the Real-world Dataset was used exclusively for testing.

4.2 Comparative Methods

We compare our methods with two groups of baselines. The first group comprises six representative image translation and generative models spanning the major generative paradigms: Pix2Pix (Isola et al., 2017), cDDPM (Ho et al., 2020), EDM (Karras et al., 2022), DDBM (Zhou et al., 2024), Rectified Flow (Liu et al., 2023), MeanFlows (Geng et al., 2025). The second group comprises three

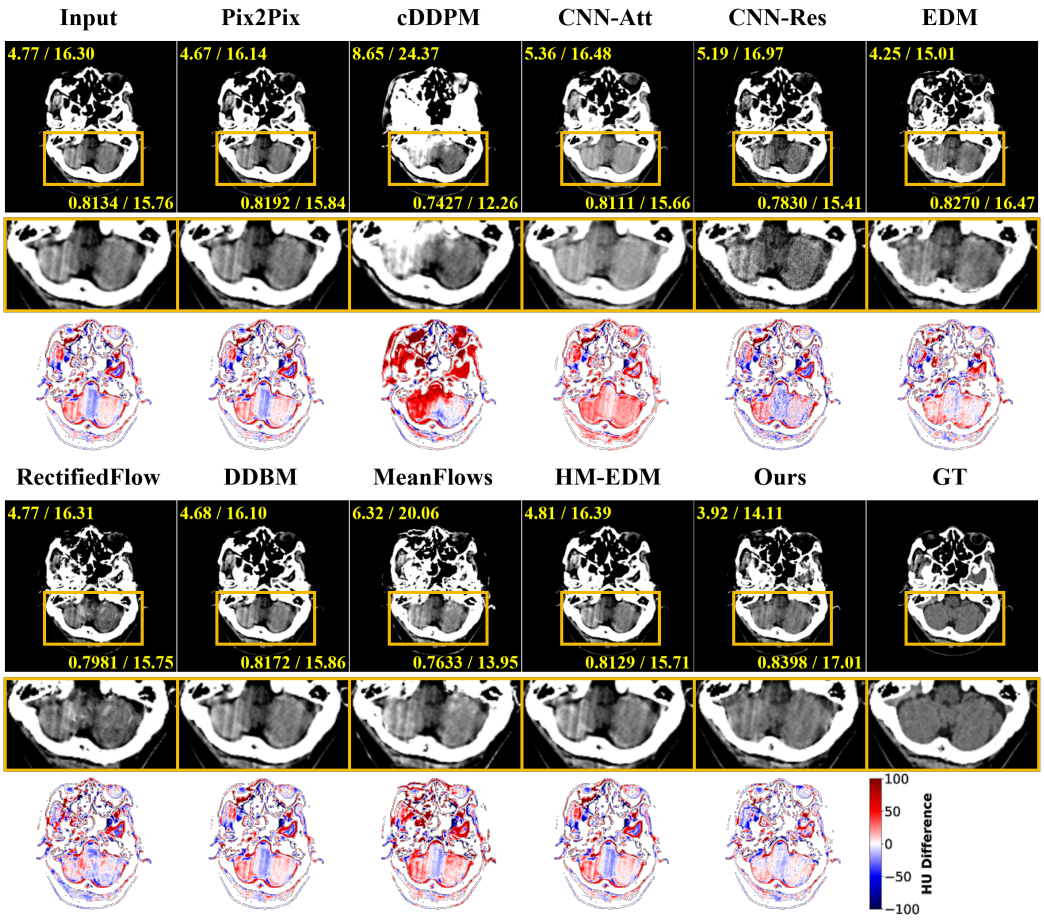

Figure 3: Visualization comparisons on the Mild Dataset. The display window is [0, 100]HU. For each subfigure, the top-left and bottom-right yellow numbers indicate MAE / RMSE and SSIM / PSNR, respectively.

methods dedicated to head CT motion artifact correction: CNN-Att (Ko et al., 2021), CNN-Res (Su et al., 2022), and HM-EDM (Chen et al., 2025).

4.3 Evaluation Metrics

We use distinct metrics for different datasets. For the Mild, Severe, and Public Datasets, where ground-truth artifact-free volumes are available, we report results using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Structural Similarity Index (SSIM) (Wang et al., 2004), and Peak Signal-to-Noise Ratio (PSNR). For the Real-world Dataset, we employ FID and Kernel Inception Distance (KID) (Binkowski et al., 2018) as distribution-based metrics. For FID and KID evaluation, axial CT slices were rendered using a fixed brain window and treated as grayscale images for feature extraction. All compared methods were evaluated under the same preprocessing protocol.

4.4 Implementation Details

All models are implemented in PyTorch and trained for 150 epochs with the Adam optimizer (lr = 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.99$). The batch size is 4. A StepLR scheduler is applied with a step size of 40 epochs and a decay factor of 0.95. The loss weights λ_{AML} and λ_{SSA} are empirically set to 10 and 0.05, respectively. For representation alignment, we employ a frozen 3DINO-ViT (Xu et al., 2025) encoder for $E(\cdot)$ and a trainable 3D convolutional projection head for $h(\cdot)$. The artifact

discriminative function $\Phi(\cdot)$ is implemented as a ResNet50, pretrained on motion-corrupted data with the same patient split. For a fair comparison, all diffusion- and flow-based baselines share the same 3D U-Net backbone, a unified 50 deterministic sampling steps, and are trained under identical settings.

4.5 Results

4.5.1 Ablation Study

We conduct ablation experiments on the Mild Dataset to dissect the contribution of each component in our framework: the spatially-differential attention injection (SDA), the semantic structure anchoring (SSA), and the spectral amplitude matching loss (AML). Starting from the EDM baseline, we progressively add each component and evaluate all variants under identical training and inference settings. Quantitative results are reported in Table 1. Each component alone yields consistent improvements over the uncorrected inputs, among which SDA brings the largest gain, reducing the MAE from 23.43HU to 21.75HU in the brain region of motion-affected slices.

4.5.2 Comparison on Simulated Dataset

Table 2 reports the quantitative comparisons on motion-corrupted slices of the Mild, Severe, and Public Datasets, evaluated in the brain region and the overall foreground. On the Mild Dataset, several comparative methods bring marginal or even negative gains over the uncorrected inputs. In contrast, our framework achieves the best results across most metrics in both regions. Compared with the second-best results on the Mild Dataset, it reduces the brain-region MAE from 23.43HU to 18.12HU (-22.7%) and improves the PSNR from 35.13dB to 37.62dB (+2.49dB). On the Severe Dataset with larger motion degradation, our framework also achieves strong restoration performance, obtaining 38.20HU MAE and 35.69dB PSNR in the brain region. Additional qualitative comparisons under severe artifacts are provided in the supplementary material. On the Public Dataset, our framework achieves strong performance on an independent public cohort with different acquisition characteristics. Specifically, it obtains the best RMSE (25.56HU) and PSNR (42.46dB) in the brain region, and achieves the best results across all four metrics in the overall foreground. These results demonstrate the robustness of the proposed framework across diverse artifact severities and evaluation settings.

4.5.3 Comparison on Real-world Dataset

Figure 4 reports the FID and KID scores on the Real-world Dataset, measuring the distributional discrepancy between the corrected volumes and real artifact-free volumes. Our framework achieves the best FID (8.40) and KID (3.54×10^{-3}), outperforming the second-best results by 14.9% and 15.1%, respectively.

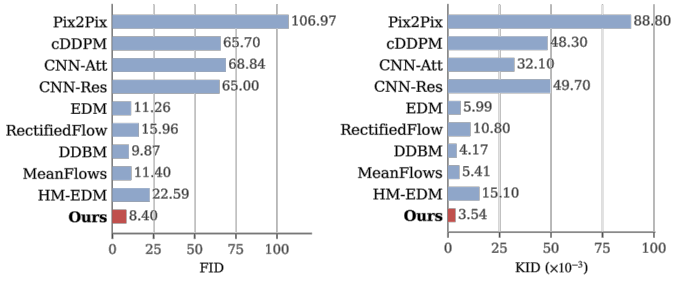

Figure 4: FID and KID scores on the Real-world Dataset. Lower values indicate a closer distribution to real artifact-free volumes.

5 Conclusion

We propose a spatial-differential and structure-anchoring framework for mild head CT motion artifact correction. DRF provides explicit artifact localization, SDA enables spatially selective restoration, and SSA preserves anatomical consistency through feature alignment. Experiments across simulated, public, and real-world datasets demonstrate improved restoration quality over existing methods.

ETHICS STATEMENT

The in-house dataset was retrospectively collected from three township hospitals with institutional approval; all scans were de-identified prior to use. This work does not involve any intervention with patients. The proposed method is intended as an assistive tool for improving the diagnostic utility of motion-corrupted head CT scans, and corrected images should always be reviewed by qualified radiologists.

REPRODUCIBILITY STATEMENT

All datasets, preprocessing steps, motion simulation parameters, training configurations, and evaluation protocols are described in detail in the Experiments section. The simulation pipeline follows the publicly available method of Chen et al. (2024), and the Public Dataset is constructed from the openly accessible SynthRAD2023 challenge data.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing of the manuscript. We have not used generative AI tools for research ideation, experimental design, data analysis, or figure generation. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Mohammed A. Al-masni, Seul Lee, Jaeuk Yi, Sewook Kim, Sung-Min Gho, Young Hun Choi, and Dong-Hyun Kim. Stacked u-nets with self-assisted priors towards robust correction of rigid motion artifact in brain MRI. *NeuroImage*, 259:119411, 2022. doi: 10.1016/J.NEUROIMAGE.2022.119411. URL https://doi.org/10.1016/j.neuroimage.2022.119411.

Julia F Barrett and Nicholas Keat. Artifacts in ct: recognition and avoidance. *Radiographics*, 24(6): 1679–1691, 2004.

Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

F Edward Boas, Dominik Fleischmann, et al. Ct artifacts: causes and reduction techniques. *Imaging Med*, 4(2):229–240, 2012.

H Bruder, C Rohkohl, K Stierstorfer, and T Flohr. Compensation of skull motion and breathing motion in ct using data-based and image-based metrics, respectively. In *Medical Imaging 2016: Physics of Medical Imaging*, volume 9783, pp. 348–359. SPIE, 2016.

Bruce CV Campbell, Deidre A De Silva, Malcolm R Macleod, Shelagh B Coutts, Lee H Schwamm, Stephen M Davis, and Geoffrey A Donnan. Ischaemic stroke. *Nature reviews Disease primers*, 5 (1):70, 2019.

Zhennong Chen, Quanzheng Li, and Dufan Wu. Estimate and compensate head motion in non-contrast head ct scans using partial angle reconstruction and deep learning. *Medical physics*, 51 (5):3309–3321, 2024.

Zhennong Chen, Siyeop Yoon, Quirin D. Strotzer, Rehab Naeem Khalid, Matthew Tivnan, Quanzheng Li, Rajiv Gupta, and Dufan Wu. Portable head CT motion artifact correction via diffusion-based generative model. *Comput. Medical Imaging Graph.*, 119:102478, 2025. doi: 10.1016/J.COMPMEDIMAG.2024.102478. URL https://doi.org/10.1016/j.compmedimag.2024.102478.

Adrian V. Dalca, John V. Guttag, and Mert R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp.

9290–9299. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00968. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Dalca_Anatomical_Priors_in_CVPR_2018_paper.html.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2414–2423. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.265. URL https://doi.org/10.1109/CVPR.2016.265.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/6d13e085b79d454da5910e4ca82a3d9d-Abstract-Conference.html.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Steven M Greenberg, Wendy C Ziai, Charlotte Cordonnier, Dar Dowlatshahi, Brandon Francis, Joshua N Goldstein, J Claude Hemphill III, Ronda Johnson, Kiffon M Keigher, William J Mack, et al. 2022 guideline for the management of patients with spontaneous intracerebral hemorrhage: a guideline from the american heart association/american stroke association. *Stroke*, 53(7):e282–e361, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5967–5976. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.632. URL https://doi.org/10.1109/CVPR.2017.632.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pp. 694–711. Springer, 2016. doi: 10.1007/978-3-319-46475-6_43. URL https://doi.org/10.1007/978-3-319-46475-6_43.

Patricia M Johnson and Maria Drangova. Conditional generative adversarial network for 3d rigid-body motion correction in mri. *Magnetic resonance in medicine*, 82(3):901–910, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Youngjun Ko, Seunghyuk Moon, Jongduk Baek, and Hyunjung Shim. Rigid and non-rigid motion artifact reduction in x-ray CT using attention module. *Medical Image Anal.*, 67:101883, 2021. doi: 10.1016/J.MEDIA.2020.101883. URL https://doi.org/10.1016/j.media.2020.101883.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Joscha Maier, TJ Herbst, Stefan Sawall, Marcel Arheit, Pascal Paysan, and Marc Kachelrieß. Projection-based motion correction for shifted-detector cone-beam ct. In *2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD)*, pp. 1–1. IEEE, 2023.

Gyutaek Oh, Sukyoung Jung, Jeong Eun Lee, and Jong Chul Ye. Annealed score-based diffusion model for MR motion artifact reduction. *IEEE Trans. Computational Imaging*, 10:43–53, 2024. doi: 10.1109/TCI.2023.3347917. URL https://doi.org/10.1109/TCI.2023.3347917.

Sarah Ouadah, Matthew Jacobson, Joseph Webster Stayman, Tina Ehtiati, C Weiss, and Jeffrey H Siewerdsen. Correction of patient motion in cone-beam ct using 3d–2d registration. *Physics in medicine & biology*, 62(23):8813–8831, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd international conference on learning representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *CoRR*, abs/2512.10794, 2025. doi: 10.48550/ARXIV.2512.10794. URL https://doi.org/10.48550/arXiv.2512.10794.

Bin Su, Yuting Wen, Yanyan Liu, Shu Liao, Jianwei Fu, Guotao Quan, and Zhenlin Li. A deep learning method for eliminating head motion artifacts in computed tomography. *Medical Physics*, 49(1):411–419, 2022.

Tao Sun, Jung-Ha Kim, Roger Fulton, and Johan Nuyts. An iterative projection-based motion estimation and compensation scheme for head x-ray ct. *Medical physics*, 43(10):5705–5716, 2016.

Adrian Thummerer, Erik van der Bijl, Arthur Jr Galapon, Joost J. C. Verhoeff, Johannes A. Langendijk, Stefan Both, Cornelis A. T. van den Berg, and Matteo Maspero. Synthrad2023 grand challenge dataset: generating synthetic CT for radiotherapy. *CoRR*, abs/2303.16320, 2023. doi: 10.48550/ARXIV.2303.16320. URL https://doi.org/10.48550/arXiv.2303.16320.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G. Krishnan, Anne L. Martel, and Maged Goubran. A generalizable 3d framework and model for self-supervised learning in medical imaging. *npj Digit. Medicine*, 8(1), 2025. doi: 10.1038/S41746-025-02035-W. URL https://doi.org/10.1038/s41746-025-02035-w.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Qiyang Zhang, Zhanli Hu, Changhui Jiang, Hairong Zheng, Yongshuai Ge, and Dong Liang. Artifact removal using a hybrid-domain convolutional neural network for limited-angle computed tomography imaging. *Physics in Medicine & Biology*, 65(15):155010, 2020.

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 9242–9251. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00963. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Translating_and_Segmenting_CVPR_2018_paper.html.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *CoRR*, abs/2510.11690, 2025. doi: 10.48550/ARXIV.2510.11690. URL https://doi.org/10.48550/arXiv.2510.11690.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FKksTayvGo.