MODEFOLD: MODAL-DEGRADATION JOINT UNFOLDING FOR ALL-IN-ONE MEDICAL IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

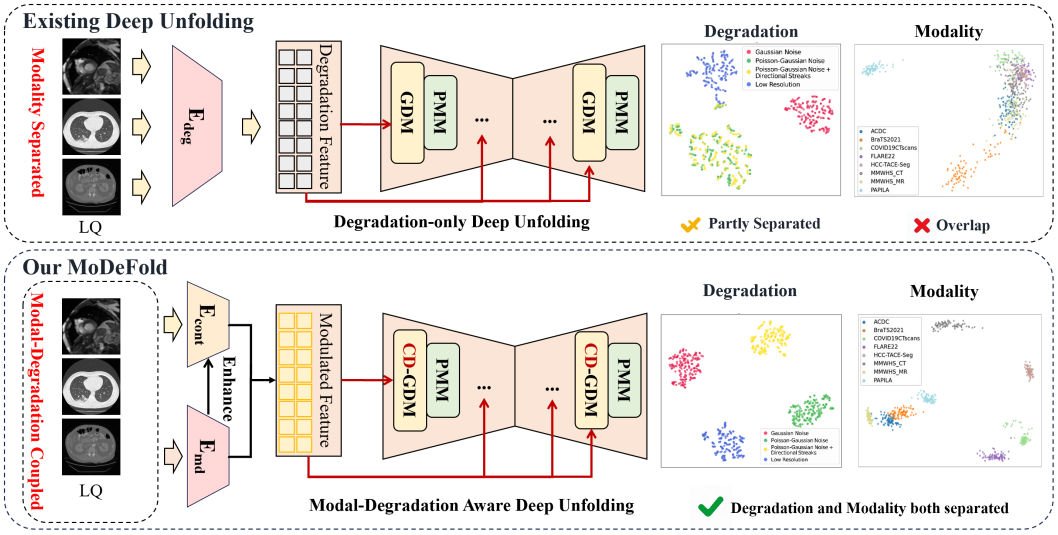

Figure 1: **Motivation of MoDeFold.** All-in-one restoration must preserve modality-dependent anatomy while removing heterogeneous degradations. The right-hand projections are colored by degradation or dataset; the eight datasets and four degradation labels are detailed in Appendix D. Degradation-oriented unfolding shows partial degradation separation and dataset overlap (top). MoDeFold enhances clean-code-aligned content features with a modal-degradation embedding and fuses both to guide shared unfolding, producing clearer grouping (bottom).

ABSTRACT

All-in-one medical image restoration (MedIR) seeks a single model for heterogeneous, modality-coupled degradations, including k-space undersampling, low-dose noise, and limited-photon degradation. Existing Deep Unfolding Networks (DUNs) do not fully account for the interaction between modality-dependent content and heterogeneous degradations within a shared restoration pipeline. Across modalities, differences in tissue contrast and appearance are further mixed with corruption, making reliable content representations difficult to obtain. We propose **Mo**dal-**De**gradation un**Fold**ing (MoDeFold), a unified unfolding framework that reconciles divergent modality and degradation distributions for all-in-one MedIR. MoDeFold models modality-dependent content and modal-degradation embedding as complementary priors for a shared unfolding process. Our core insight is to use the modal-degradation prior to enhance modality-specific content features, then combine the enhanced features with the modal-degradation prior to jointly control the unfolding process. A multi-scale vector-quantized autoencoder supplies content features supervised by clean-code references, a Modal-Degradation Embedder trained with contrastive learning provides the joint prior, and embedding-conditioned modulation and fusion connect the two. Experiments demonstrate that MoDeFold outperforms baselines in all-in-one restoration, achieving an improvement of 1.126 dB in average PSNR, with consistent gains under single-task setting.

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.

1 INTRODUCTION

All-in-one medical image restoration (MedIR) (Yang et al., 2024b) aims to reconstruct high-quality (HQ) images from low-quality (LQ) observations across modalities such as magnetic resonance imaging (MRI) (Wang et al., 2023b), computed tomography (CT) (Li et al., 2025), and positron emission tomography (PET) (Jang et al., 2023). Improving image quality is important for reliable interpretation and clinical diagnosis. A common formulation is

$$y = \Phi x + n, \tag{1}$$

where x is the clean image, y is the degraded observation, Φ denotes an imaging-related degradation operator, and n represents measurement uncertainty. Both the operator and the statistics of the uncertainty vary with the imaging and acquisition setting.

Single-task MedIR methods can achieve strong performance by specializing to a particular degradation. Clinical workflows, however, involve heterogeneous settings, including k-space undersampling in MRI, low-dose noise in CT, and limited-photon degradation in PET. An all-in-one model must restore these observations using a shared network. Existing approaches range from model-based methods (Chen et al., 2020; Zhang et al., 2020b; Mou et al., 2022) to deep learning methods (Ronneberger et al., 2015; Wang et al., 2023a; Yang et al., 2024a) and Deep Unfolding Networks (DUNs) (Ma et al., 2019; Zeng et al., 2025). DUNs turn iterative solvers, such as Proximal Gradient Descent (PGD) (Beck & Teboulle, 2009), into trainable stages that combine an observation-driven update with a learned image prior. For the linear model in Eq. (1), a PGD step takes the form

$$z^{(k)} = x^{(k-1)} - \rho \Phi^T \big(\Phi x^{(k-1)} - y\big), \tag{2}$$

$$x^{(k)} = \arg\min_x \frac{1}{2}\|x - z^{(k)}\|_2^2 + \rho\lambda\Psi(x), \tag{3}$$

where k denotes the iteration, ρ is the step size, and Ψ is an image-prior regularizer (Chang et al., 2020). The Gradient Descent Module (GDM) implements the observation-driven update in Eq. (2), while the Proximal Mapping Module (PMM) refines its intermediate estimate $z^{(k)}$.

The difficulty in extending this structure to all-in-one MedIR is that *both modules become coupled to the degradation*. The GDM depends explicitly on Φ: changing the sampling pattern or corruption process changes the residual and the correction required at each stage. The PMM is coupled to the same process through its input. Because $z^{(k)}$ is produced by the GDM, the artifacts and uncertainty remaining in $z^{(k)}$ also depend on the degradation. When the two modules are trained together, the PMM can therefore learn to remove the residual patterns produced by particular GDM updates rather than learning a prior that transfers equally well across settings. Sharing an unfolding architecture alone does not remove this coupling; a change in degradation can affect both the direction of the observation-driven update and the type of refinement needed afterward.

Medical data add a second source of ambiguity. Imaging modality and degradation are often correlated through their acquisition processes, while clean MRI, CT, and PET images differ in contrast, intensity statistics, and anatomical appearance. Consequently, features extracted from an LQ observation mix cues about the underlying content with cues about how it was corrupted. A condition learned mainly from degradation may provide insufficient guidance about which structures should be preserved, whereas content features extracted directly from the LQ image may retain artifacts as if they were anatomy. In a shared DUN, these ambiguous cues can lead to unsuitable updates across tasks, leaving residual corruption or suppressing subtle structures. Conventional operator-specific unfolding (Yang et al., 2016; Zhang & Ghanem, 2018; Zhang et al., 2020a) and more flexible learned or degradation-conditioned updates (Mou et al., 2022; Zeng et al., 2025) do not by themselves resolve how clean, modality-dependent content should guide the degradation-sensitive modules.

To address this problem, we propose **Mo**dal-**De**gradation un**Fold**ing (MoDeFold), a unified framework for all-in-one MedIR. **Our core insight is to explicitly learn both modality and degradation through dual contrastive supervision, and use this joint representation to modulate the structural content prior for restoration guidance.** As illustrated in Fig. 1, a Modal-Degradation Embedder applies modality-based and degradation-based contrastive objectives to a shared embedding, capturing imaging characteristics and corruption cues. This embedding modulates modality-specific content features learned by a multi-scale vector-quantized autoencoder (VQVAE) under clean-code

supervision. The enhanced content features and joint embedding are then fused to guide unfolding updates, connecting modality–degradation discrimination with structure-aware restoration.

The contributions of this paper are summarized as follows:

- We propose MoDeFold, a unified unfolding framework in which content and modal-degradation priors jointly guide all-in-one MedIR.
- We learn complementary priors using a multi-scale VQVAE with clean-code alignment and a Modal-Degradation Embedder trained with contrastive learning.
- We use the modal-degradation prior to enhance modality-specific content features before fusing the two representations to control the shared unfolding process.
- Experiments under single-task and all-in-one protocols demonstrate strong restoration performance, and ablations verify the contribution of each component.

2 Related Works

2.1 Deep Unfolding Networks for Image Restoration

Deep unfolding networks turn iterative reconstruction algorithms into trainable networks. Early methods such as ADMM-Net (Yang et al., 2016) and ISTA-Net (Zhang & Ghanem, 2018) retain the structure of the underlying solver while learning reconstruction transforms and regularization. USRNet (Zhang et al., 2020a) extends this approach to super-resolution with different blur kernels, scale factors, and noise levels. Incorporating the forward model gives these networks a direct connection to the measurement process, although applying them to a new imaging setting requires a suitable operator specification. Subsequent work has made unfolding less dependent on a prescribed operator. DGUNet (Mou et al., 2022) learns gradient estimation and proximal mapping for a range of restoration tasks. VLU-Net (Zeng et al., 2025) further uses CLIP-derived degradation representations to guide all-in-one unfolding. Deep LoRA-Unfolding Networks (LoRun) (Wang et al., 2026b) focus on differences between unfolding stages, sharing a pretrained denoiser and adapting its PMMs with stage-specific low-rank modules. This reduces parameter duplication while allowing different denoising behavior along the reconstruction trajectory. These advances make unfolding more flexible, but medical restoration also involves changes in the clean-image distribution across modalities. MoDeFold uses the modal-degradation embedding to modulate modality-specific content features learned against clean quantized references, then fuses both representations to control unfolding.

2.2 All-in-One Medical Image Restoration

Medical restoration methods such as SwinMR (Huang et al., 2022a), SDAUT (Huang et al., 2022b), CTformer (Wang et al., 2023a), DASMamba (Chan et al., 2025), A-DPS (Aali et al., 2025), DiFusion (Wu et al., 2025), Corruption2Self (Tu et al., 2025) and Restore-RWKV (Yang et al., 2026) improve image fidelity and anatomical detail in task-specific restoration settings. However, effective architectures for individual tasks do not by themselves provide joint modeling of modality-dependent content and degradation in an all-in-one model. Such modeling must distinguish changes in anatomical appearance from corruption while accommodating different degradations within the same modality. DRMC (Yang et al., 2023) explores dynamic routing for multi-center PET synthesis, allowing a generalist model to accommodate data from different centers. AMIR (Yang et al., 2024a) extends task-adaptive processing to MRI, CT, and PET through spatial and channel routing. TAT (Yang et al., 2025) further introduces task-specific weight generation and adaptive loss balancing to reduce interference and uneven optimization across tasks. Together, these methods show the value of adapting shared computation to heterogeneous medical data. Their main emphasis is on where and how tasks share processing, rather than learning content representations against explicit clean-image references. DiffCode (Chen et al., 2025) and VQPET (Chen et al., 2026) takes a different approach by constructing a task-adaptive codebook bank of HQ features. It uses latent diffusion to refine the retrieval of these priors and compensate for information lost in degraded images. This brings clean-content guidance into all-in-one MedIR and is closely related to our motivation. MoDeFold instead uses the modal-degradation embedding to adjust content queries before quantization and aligns their representations with paired clean codes. The enhanced content features then combine with the modal-degradation embedding to jointly control the unfolding process.

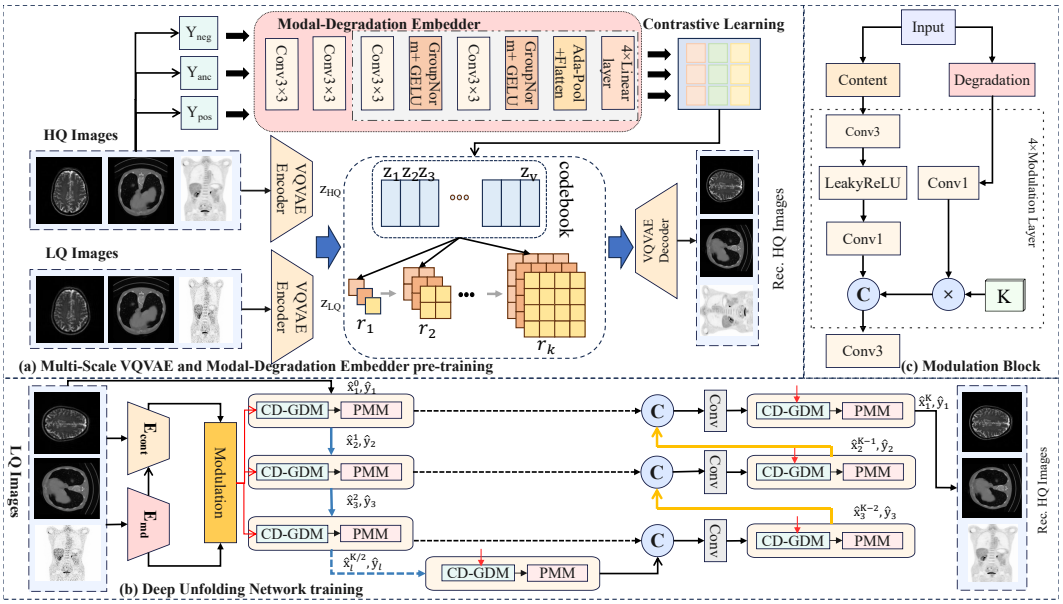

Figure 2: Overview of MoDeFold. (a) Modal-Degradation Embedder pretraining and embedding-conditioned modulation of modality-specific VQVAE features. (b) Hierarchical unfolding controlled by the joint prior. (c) Fusion of the enhanced content features and modal-degradation prior.

3 METHODS

3.1 UNIFIED MODAL-DEGRADATION UNFOLDING FRAMEWORK

Building on the observation model in Eq.(1), all-in-one MedIR must account for both modality-dependent content and heterogeneous corruption. We use the modal-degradation prior to modulate modality-specific content features, then fuse the two representations to condition the shared unfolding process,

$$g = E_{\text{md}}(y), \qquad c = E_{\text{cont}}(y; m, g), \qquad p = \mathcal{M}(c, g). \tag{4}$$

Here, y is the degraded observation and m its modality label. E_{md} produces the joint modal-degradation embedding g, while E_{cont} extracts spatial content features c conditioned on m and g and supervised by clean-code references. The fusion module $\mathcal{M}$ combines both into the unfolding condition p.

As shown in Fig. 2(c), fusion $\mathcal{M}$ uses g to generate channel-wise scales and shifts that modulate spatial content features, then transforms and aggregates them into p. This makes unfolding control sensitive to both the corruption to remove and the anatomical structures to preserve, rather than conditioning on either prior alone. With p fixed for each observation, we express restoration as

$$\hat{x} = \arg\min_{x} \frac{1}{2}\|\Phi(x; p) - y\|_2^2 + \lambda\Psi(x; p). \tag{5}$$

where the first term is a learned observation-fidelity surrogate rather than an explicit modality-specific physical forward model, and Ψ represents the image prior. Following the PGD structure in Eqs.(2–3), we learn the conditional updates

$$z^{(k)} = x^{(k-1)} - \rho_k \mathcal{B}_k\big(\Phi_k(x^{(k-1)}; p) - y\big), x^{(k)} = \mathcal{P}_k(z^{(k)}). \tag{6}$$

Here, $\mathcal{B}_k$ maps the observation-feature residual into an estimate-space correction, scaled by the learned step size ρ_k. CD-GDM performs this correction under p, and the PMM $\mathcal{P}_k$ refines $z^{(k)}$. Since $\mathcal{B}_k$ is learned independently of Φ_k, the update is PGD-inspired. Its MDTA implementation and stage details are given in Appendix C.

3.2 MODAL-DEGRADATION EMBEDDER

Medical observations vary in both modality-dependent appearance and degradation characteristics: CT images may contain low-dose noise or streak artifacts, while CT depicts bone with high attenuation and MRI provides richer soft-tissue contrast. We therefore introduce a Modal-Degradation Embedder (E_{md}) to learn a joint representation for restoration conditioning through modality–degradation dual contrastive learning. Given a mixed-modality mini-batch $\{(y_i, m_i, d_i)\}_{i=1}^{B}$, where m_i and d_i denote the source modality and degradation label, we define two positive sets for anchor y_i: $P_m(i) = \{j \neq i : m_j = m_i\}$ and $P_d(i) = \{j \neq i : d_j = d_i\}$. The modality objective associates same-modality observations across available degradations, while the degradation objective associates same-degradation observations across available modalities. Samples with a different corresponding label serve as negatives. Both objectives act on the shared token $g_i = E_{\mathrm{md}}(y_i)$, normalized as $v_i = g_i/\|g_i\|_2$. For each relation $r \in \{m, d\}$, we optimize

$$\mathcal{L}_{\mathrm{con}}^{r} = -\frac{1}{|I_r|}\sum_{i\in I_r}\frac{1}{|P_r(i)|}\sum_{j\in P_r(i)}\log\frac{\exp\left(v_i^{\top}v_j/\tau\right)}{\sum_{a=1,a\neq i}^{B}\exp\left(v_i^{\top}v_a/\tau\right)}, \tag{7}$$

where $I_r = \{i : |P_r(i)| > 0\}$ contains anchors with at least one positive, and τ is the temperature. The denominator includes all non-anchor samples; the loss is set to zero if I_r is empty. An auxiliary degradation-classification loss further reinforces category discrimination. The embedder is pretrained using a weighted combination of the modality contrastive, degradation contrastive, and auxiliary classification losses, with implementation details provided in Appendix A.

For restoration, we retain the continuous token $g = E_{\mathrm{md}}(y)$ rather than a predicted class, using it for content modulation and fusion to provide image-dependent unfolding guidance.

3.3 MODAL-DEGRADATION-CONDITIONED CONTENT PRIORS

Clean medical images differ in tissue contrast and appearance across modalities, while degradation further obscures their structural cues. The multi-scale VQVAE (E_{cont}) addresses these two sources of variation separately. A lightweight modality-specific residual adapter maps encoder features into a shared content space. We then apply embedding-conditioned modulation to enhance these features before they query the pretrained multi-scale codebook (van den Oord et al., 2017; Qu et al., 2025):

$$h_m = E(y) + A_m(E(y)), \qquad c = (1 + \gamma(g)) \odot h_m + \beta(g). \tag{8}$$

Here, A_m accommodates modality-level statistics, while the channel-wise affine coefficients $\gamma(g)$ and $\beta(g)$ adapt the query to the observed corruption. This avoids treating every modality–degradation combination as an independent content domain.

The shared codebook $\mathcal{Z} = \{Z_1, \ldots, Z_V\}$ contains D-dimensional latent embeddings. At scale k, the latent residual is quantized on an $n_k \times n_k$ grid, yielding indices $r_k \in \{1, \ldots, V\}^{n_k^2}$ and retrieved features $F_k = \mathcal{Z}[r_k] \in \mathbb{R}^{n_k^2 \times D}$. The sequence $(F_1, \ldots, F_K)$ captures coarse spatial organization followed by finer residual variations in boundaries and texture. Reshaping, spatial alignment, and residual accumulation produce the quantized feature map $Q(c)$, rather than a global code vector.

We supervise the modulated features using the paired clean image. Let h_m^x denote its feature from the same modality adapter without embedding-conditioned modulation. The quantized maps are aligned through

$$\mathcal{L}_{\mathrm{align}} = \|Q(c) - \mathrm{sg}[Q(h_m^x)]\|_1. \tag{9}$$

Stop-gradient prevents the alignment loss from pulling the clean target toward the degraded view. Together with reconstruction and VQ objectives, this alignment encourages the conditioned feature to recover clean content using the multi-scale code references.

At inference, the continuous feature c carries clean-code supervision into restoration and is fused with g to form $p = \mathcal{M}(c, g)$ (Fig. 2(c)). Thus, clean-content alignment and joint embedding learning shape the same unfolding condition. Implementation details are provided in Appendices A, B, and C.

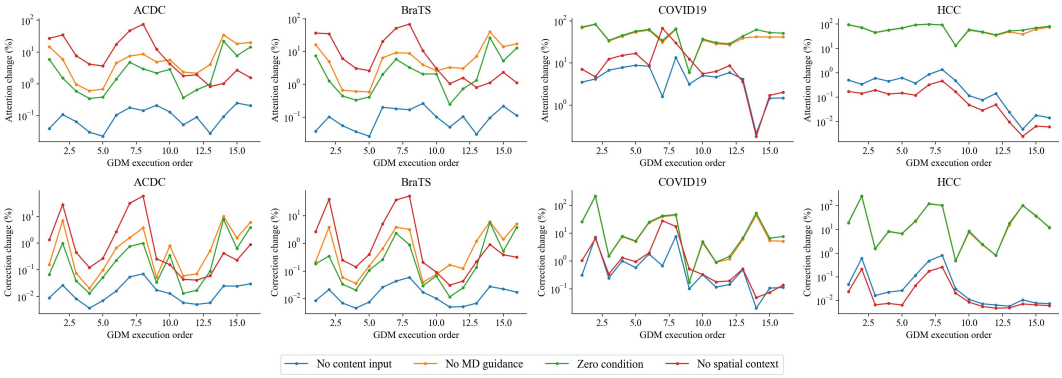

Figure 3: **Prior-guided unfolding.** Content and modal-degradation guidance influence both channel attention and unfolding corrections, with effects that vary across layers and examples.

Effect of prior conditioning on unfolding. Fig. 3 compares four interventions: removing content input, disabling modal-degradation guidance, zeroing the fused condition, and removing spatial context (detailed in Appendix C). With layer inputs fixed, the horizontal axis indexes 16 GDMs, and the vertical axis reports relative changes (%) from full fusion in attention (top) and corrections (bottom). For example, for COVID19, removing modal-degradation guidance changes attention and corrections by 43.56% and 26.42% on average, versus 5.03% and 1.31% without content input. Both priors therefore control attention and the resulting updates, with stronger modal-degradation influence in this example. These changes measure conditioning sensitivity, not restoration gains.

4 Experiments

4.1 Datasets

Our main benchmark evaluates restoration across modalities and degradation conditions using ACDC (cardiac MRI) (Bernard et al., 2018), BraTS2021 (brain MRI) (Baid et al., 2021), COVID19CTscans (thoracic CT) (Ma et al., 2021), and HCC-TACE-Seg (abdominal CT) (Moawad et al., 2021), distributed through IMed-361M Cheng et al. (2025). MRI inputs undergo Fourier-domain truncation combined with bicubic resampling. CT inputs receive Poisson–Gaussian noise, with stronger noise and additional directional streaks for HCC-TACE-Seg. For prior pretraining only, we additionally include FLARE22 (Ma & Wang, 2022), MMWHS-MR and MMWHS-CT (Zhuang & Shen, 2016), and PAPILA (Kovacs et al., 2020), yielding eight datasets for learning E_{cont} and E_{md}. The training/validation/test ratio is 7:1:2. Preparation and degradation simulation are detailed in Appendix D. We also evaluate on the datasets adopted by AMIR Yang et al. (2024a), covering MRI super-resolution (LLC), CT denoising (McCollough et al., 2017), and PET synthesis.

4.2 Experimental Settings

We train a separate all-in-one model on each benchmark. For one-by-one training, an independent model is trained for each dataset on our benchmark, and for each of the three tasks on AMIR. On our benchmark, E_{md} and E_{cont} are pretrained for 50 epochs each on the eight-dataset collection using modality–degradation contrastive learning and clean-code-aligned content learning, respectively. This stage covers the four degradation categories illustrated in Fig. 1: Gaussian noise, Poisson–Gaussian noise, Poisson–Gaussian noise with directional streaks, and low resolution. Subsequent end-to-end unfolding training uses only ACDC, BraTS2021, COVID19CTscans and HCC-TACE-Seg with three degradation settings–MRI low resolution, CT low-dose noise, and CT low-dose noise with streaks, for up to 300,000 updates with AdamW.

On our benchmark, all-in-one baselines are Restormer (Zamir et al., 2022), Restore-RWKV (Yang et al., 2026), DASMamba (Chan et al., 2025), AirNet (Li et al., 2022), PromptIR (Potlapalli et al., 2023), AMIR (Yang et al., 2024a), VLU-Net (Zeng et al., 2025), TAT (Yang et al., 2025), and R2R (Wang et al., 2026a). The one-by-one comparison uses Restormer and Restore-RWKV for both

Table 1: All-in-one restoration on our benchmark. Original Paradigm denotes the setting for which each method was originally designed. Best and second-best results are bolded and underlined.

Original Paradigm	Method	Param.	Reference	ACDC		COVID19CTscans		HCC-TACE-Seg		BraTS2021		Average	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Task-Specific	Restormer	26M	CVPR 2022	24.6970	0.6919	30.6683	0.8714	40.5314	0.9599	28.5701	0.8837	31.1167	0.8517
	Restore-RWKV	28M	JBHI 2026	26.4272	0.7773	31.3859	0.9132	40.1901	0.9214	29.1601	0.8892	31.7908	0.8753
	DASMamba	27M	MICCAI 2025	26.3650	0.7798	32.3760	0.9161	40.9213	0.9528	29.1747	0.8891	32.2093	0.8845
All-in-One	AirNet	9M	CVPR 2022	24.2286	0.6703	35.4095	0.9640	36.9141	0.9571	28.6947	0.8823	31.3117	0.8684
	PromptIR	33M	NeurIPS 2023	24.3568	0.6634	35.1940	0.8755	38.5289	0.9495	29.1270	0.8755	31.8017	0.8410
	AMIR	24M	MICCAI 2024	24.6732	0.6921	30.0948	0.8492	40.4505	0.9642	27.5235	0.8746	30.6855	0.8450
	VLU-Net	35M	CVPR 2025	24.3075	0.6921	**38.1786**	0.9421	40.6964	0.9529	29.1109	0.8800	33.0734	0.8668
	TAT	42M	MICCAI 2025	25.6111	0.6897	33.8687	0.8718	41.2228	0.9683	28.6242	0.8838	32.3317	0.8534
	R2R	30M	CVPR 2026	26.2389	0.7008	34.8370	0.9499	43.1277	0.9673	30.5578	0.9058	33.6904	0.8810
	MoDeFold	49M	Ours	**27.6824**	**0.7831**	36.0758	**0.9700**	**44.3118**	**0.9829**	**31.1950**	**0.9112**	**34.8163**	**0.9118**

Table 2: All-in-One restoration on the AMIR benchmark. Original Paradigm denotes the setting for which each method was originally designed. Best and second-best results are bolded and underlined.

Original Paradigm	Method	Param.	Reference	MRI Super-Resolution			CT Denoising			PET Synthesis			Average		
				PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Task-Specific	Restormer	26M	CVPR 2022	31.72	0.9362	30.05	33.61	0.9177	8.53	37.14	0.9473	0.09	34.16	0.9337	12.89
	Spach Transformer	13M	TMI 2023	31.18	0.9290	31.83	33.47	0.9155	8.67	37.05	0.9445	0.09	33.90	0.9297	13.53
	DRMC	0.5M	MICCAI 2023	29.55	0.9032	38.17	33.28	0.9153	8.87	36.19	0.9376	0.10	33.01	0.9187	15.71
	Restore-RWKV	28M	JBHI 2026	32.10	0.9408	28.95	33.74	0.9189	8.41	37.32	0.9479	0.09	34.39	0.9359	12.48
All-in-One	AirNet	9M	CVPR 2022	31.09	0.9316	33.11	33.24	0.9142	8.98	37.14	0.9442	**0.085**	33.82	0.9300	14.06
	PromptIR	33M	NeurIPS 2023	30.51	0.9244	31.84	33.33	0.9159	8.81	36.62	0.9406	0.092	33.49	0.9270	13.58
	AMIR	24M	MICCAI 2024	32.03	0.9396	29.10	33.70	0.9182	8.45	37.12	0.9475	0.088	34.28	0.9351	12.55
	MoCE-IR	25M	CVPR 2025	31.38	0.9309	31.25	33.58	0.9172	8.56	37.15	0.9354	0.086	34.04	0.9278	13.30
	VLU-Net	35M	CVPR 2025	31.18	0.9289	32.85	33.35	0.9162	8.63	37.33	0.9471	**0.085**	33.95	0.9307	13.86
	TAT	42M	MICCAI 2025	32.10	0.9402	28.91	33.80	0.9192	8.36	37.28	0.9480	0.086	34.39	0.9358	12.45
	MoDeFold	49M	Ours	**32.14**	**0.9441**	**26.05**	**33.81**	**0.9196**	**8.27**	**37.38**	**0.9481**	**0.085**	**34.44**	**0.9373**	**11.47**

modalities, DASMamba and MSM (Park et al., 2026) for MRI, and RISE-MAR (Ma et al., 2025) and FIND-Net (Tasharofi et al., 2025) for CT. On the AMIR benchmark, the all-in-one comparison includes Restormer, Spach Transformer (Jang et al., 2023), DRMC (Yang et al., 2023), Restore-RWKV, AirNet, PromptIR, AMIR, MoCE-IR (Zamfir et al., 2025), VLU-Net, and TAT. Its one-by-one baselines are DAGAN (Yang et al., 2017), SwinMR (Huang et al., 2022a), SDAUT (Huang et al., 2022b), MambaIR (Guo et al., 2024), and PGASR (Luo et al., 2025) for MRI; EDCNN (Chen et al., 2017), DML_CT (Jung et al., 2022), CTformer (Wang et al., 2023a), DenoMamba (Öztürk et al., 2024), and MRED-Net (Liu et al., 2026) for CT; and DCNN (Chen et al., 2017), CycleWGAN (Ni et al., 2022), ARGAN (Luo et al., 2022), DRMC, and VLU-Net for PET.

4.3 All-in-One Restoration

Tables 1 and 2 report all-in-one restoration on our main benchmark and the AMIR benchmark, respectively. On our benchmark, MoDeFold achieves average PSNR and SSIM of 34.8163 dB and 0.9118, exceeding R2R by 3.34% and DASMamba by 3.09%, respectively, relative to the strongest competing averages. Its PSNR exceeds the best competing results on HCC-TACE-Seg and BraTS2021 by 2.75% and 2.09%, respectively. ACDC SSIM also exceeds DASMamba by 0.42%, although COVID19CTscans PSNR remains 5.51% below VLU-Net. On the AMIR benchmark, MoDeFold achieves the best average PSNR, SSIM, and RMSE among baselines, raising PSNR and SSIM over TAT by 0.15% and 0.16% and reducing RMSE by 7.87%.

Fig. 4 compares MRI and CT absolute error maps against HQ references, with darker regions indicating smaller errors. MoDeFold reduces residual errors while retaining anatomical structure. Fig. 5 provides AMIR comparisons for MRI super-resolution, CT denoising, and PET synthesis.

4.4 One-by-One Restoration

MoDeFold achieves the highest PSNR and SSIM on all four datasets under dataset-specific training (Table 3); AMIR results appear in Appendix Table 5. PSNR gains over the strongest baselines are 2.02% on ACDC, 3.74% on COVID19CTscans, 1.92% on HCC-TACE-Seg, and 1.60% on BraTS2021. Specialization improves PSNR over all-in-one training on both CT datasets and BraTS2021, but reduces it by 3.64% on ACDC, indicating dataset-dependent benefits.

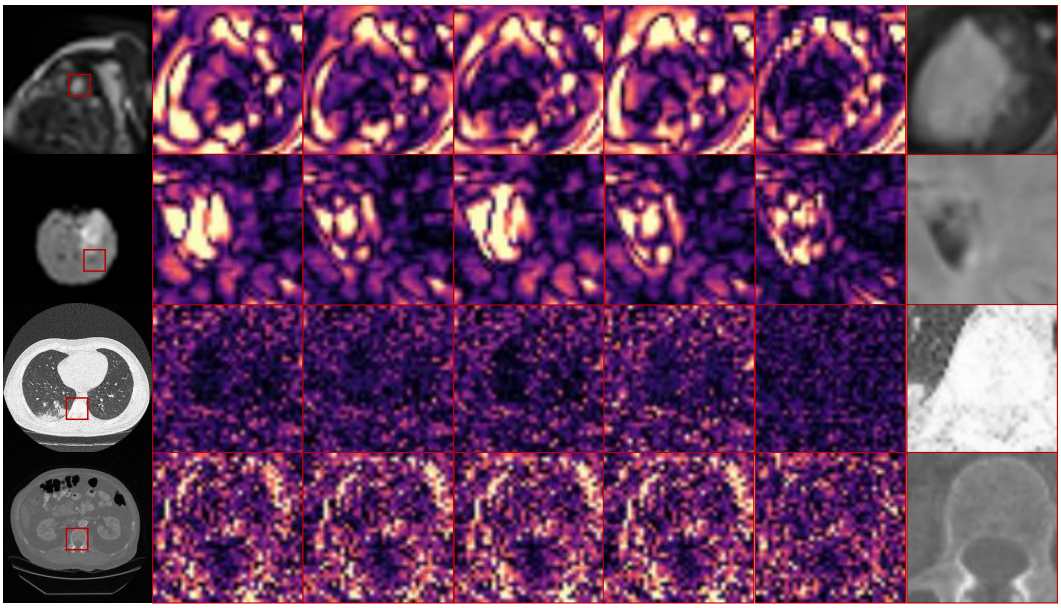

Figure 4: Visual comparison on our benchmark. MRI results and CT results as absolute error maps, where darker colors indicate smaller errors.

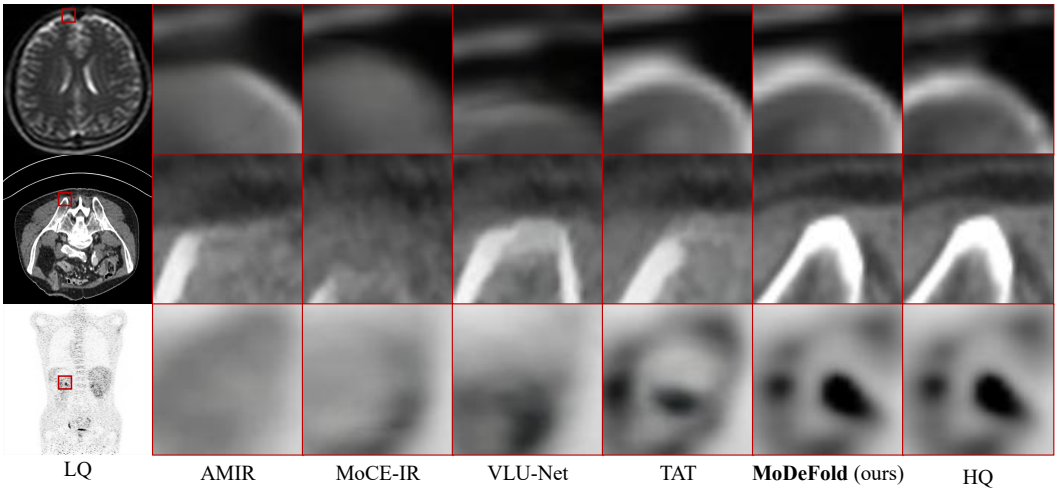

Figure 5: All-in-one visual comparison on the AMIR benchmark. Representative test examples from MRI super-resolution, CT denoising, and PET synthesis are compared with their HQ references.

4.5 Ablation Study

We evaluate the contribution of prior guidance and content representation learning on our benchmark in Table 4. The unfolding backbone and frequency decomposition remain unchanged in all variants.

w/o Content Guidance zeros the content input to fusion while retaining modal-degradation guidance and VQVAE supervision. *w/o Modal-Degradation Guidance* disables token-based content modulation and zeros the embedding supplied to fusion. *w/o Both Priors* supplies a zero unfolding condition. *Late Conditioning Only* bypasses pre-quantization content modulation but retains the token in downstream fusion. *w/o Contrastive Learning* removes supervised contrastive pretraining while retaining degradation classification. *w/o Clean-Code Alignment* removes paired quantized-feature alignment while retaining reconstruction and VQ objectives.

Table 3: One-by-one restoration on our benchmark. A separate model is trained for each dataset and degradation setting. Best and second-best results are bolded and underlined, respectively.

Method	ACDC		Method	COVID19CTscans		Method	HCC-TACE-Seg		Method	BraTS2021	
	PSNR↑	SSIM↑		PSNR↑	SSIM↑		PSNR↑	SSIM↑		PSNR↑	SSIM↑
Restormer	24.3689	0.6810	Restormer	32.9750	0.9624	Restormer	41.3564	0.9708	Restormer	29.7004	0.8752
Restore-RWKV	24.2654	0.6795	Restore-RWKV	33.0018	0.9263	Restore-RWKV	41.0371	0.9650	Restore-RWKV	28.8889	0.8853
DASMamba	24.3992	0.6831	RISE-MAR	32.8688	0.9261	RISE-MAR	41.7291	0.9621	DASMamba	28.9329	0.8868
MSM	<u>26.1487</u>	<u>0.7235</u>	FIND-Net	<u>34.8435</u>	<u>0.9687</u>	FIND-Net	<u>43.9742</u>	<u>0.9795</u>	MSM	<u>30.9981</u>	<u>0.9124</u>
MoDeFold	**26.6756**	**0.7556**	**MoDeFold**	**36.1481**	**0.9703**	**MoDeFold**	**44.8193**	**0.9842**	**MoDeFold**	**31.4941**	**0.9141**

Table 4: Ablation of prior guidance and representation learning on our benchmark. All variants use all-in-one restoration. Best results in each column are bolded.

Variant	ACDC		COVID19CTscans		HCC-TACE-Seg		BraTS2021		Average	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MoDeFold (Full)	**27.6824**	**0.7831**	**36.0758**	**0.9700**	**44.3118**	**0.9829**	31.1950	0.9112	**34.8163**	**0.9118**
w/o Content Guidance	25.7342	0.7221	35.1387	0.9442	43.7953	0.9632	30.1264	0.8997	33.6987	0.8823
w/o Modal-Degradation Guidance	25.4072	0.7037	34.3776	0.9223	43.4895	0.9529	30.0018	0.8796	33.3190	0.8646
w/o Both Priors	24.4578	0.6799	29.7971	0.8409	41.0467	0.9511	29.0614	0.8872	31.0908	0.8398
Late Conditioning Only	25.5412	0.7013	32.1482	0.8967	42.1242	0.9582	29.5127	0.8994	32.3316	0.8639
w/o Contrastive Learning	26.3154	0.7154	34.9742	0.9381	43.8119	0.9701	31.4176	0.9134	34.1298	0.8843
w/o Clean-Code Alignment	26.3775	0.7413	34.8765	0.9372	43.6614	0.9802	**31.4772**	**0.9136**	34.0982	0.8931

Prior guidance. Removing both priors reduces average PSNR by 10.70% and SSIM by 7.90% relative to the full model. Removing content or modal-degradation guidance alone lowers average PSNR by 3.21% and 4.30%, respectively. Both sources therefore contribute to the shared restoration condition. The larger loss without modal-degradation guidance is consistent with the need to adapt correction to different corruption patterns, while the loss without content guidance shows that modal-degradation embeddings alone do not provide sufficient structural guidance.

Embedding-conditioned content modulation. The clearest evidence for our core insight comes from *Late Conditioning Only*. Although the joint embedding still reaches the fusion block, average PSNR drops by 7.14% and SSIM by 5.25% relative to the full model. The decrease is particularly pronounced on COVID19CTscans, where PSNR decreases by 10.89%. This supports modulating modality-specific content features with the modal-degradation prior before fusing the two representations for joint unfolding control. Removing contrastive learning reduces average PSNR and SSIM by 1.97% and 3.02%; removing clean-code alignment reduces them by 2.06% and 2.05%. These results support joint embedding organization and clean-reference alignment. On BraTS2021, however, the retained variants without contrastive learning or clean-code alignment exceed the updated full model, so these benefits are not uniform across datasets.

5 Conclusion

We presented MoDeFold, a unified unfolding framework that jointly accounts for modality-dependent anatomical content and heterogeneous degradations in medical image restoration. Our core insight is to use the modal-degradation prior to enhance modality-specific content features, then combine the enhanced features with the modal-degradation prior to jointly control the unfolding process. The Modal-Degradation Embedder learns a joint embedding through contrastive learning and uses it to modulate content features supervised by clean quantized references. Content–embedding fusion combines the enhanced features with this token to condition the shared unfolding updates. Experiments on our benchmark and the AMIR benchmark evaluate this strategy under all-in-one and one-by-one settings, while ablations support the complementary roles of the priors and their fusion. These findings highlight joint modeling of content, modality, and degradation as a means of adapting deep unfolding to heterogeneous medical observations.

AI USE STATEMENT

In this work, we used generative AI tools for drafting parts of the manuscript, editing the paper to improve readability and language, polishing the text, translating content, creating or modifying scientific figures or images, and formatting references. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, or interpreting results. We have reviewed all AI-assisted work. Specifically, all AI-generated text was checked for factual accuracy and potential plagiarism through a manual literature survey; all AI-assisted figures were verified and corrected by the authors to ensure they accurately represent the underlying data and methodology; and all AI-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alex Dimakis, and Jonathan I. Tamir. Ambient diffusion posterior sampling: Solving inverse problems with diffusion models trained on corrupted data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederic Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. Imag.*, 37(11):2514–2525, 2018.

Simon C. K. Chan, Lulin Shi, Bingxin Huang, and Terence T. W. Wong. Directional adaptive shuffle-based visual state-space models for medical image restoration. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025 - 28th International Conference, Daejeon, South Korea, September 23-27, 2025, Proceedings, Part XIII*, volume 15972 of *Lecture Notes in Computer Science*, pp. 160–170. Springer, 2025.

Yi Chang, Luxin Yan, Xi-Le Zhao, Houzhang Fang, Zhijun Zhang, and Sheng Zhong. Weighted low-rank tensor recovery for hyperspectral image restoration. *IEEE Transactions on Cybernetics*, 50(11):4558–4572, 2020. doi: 10.1109/TCYB.2020.2983102.

Dongdong Chen, Mike E. Davies, and Mohammad Golbabaee. Compressive mr fingerprinting reconstruction with neural proximal gradient iterations. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 13–22, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59713-9.

Haowei Chen, Zhiwen Yang, Haotian Hou, Hui Zhang, Bingzheng Wei, Gang Zhou, and Yan Xu. All-in-one medical image restoration with latent diffusion-enhanced vector-quantized codebook prior. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025 - 28th International Conference, Daejeon, South Korea, September 23-27, 2025, Proceedings, Part XVI*, volume 15975 of *Lecture Notes in Computer Science*, pp. 67–77. Springer, 2025.

Haowei Chen, Zhiwen Yang, Yang Zhou, Xiaoqian Zhang, Hui Zhang, Dan Zhao, Bingzheng Wei, Gang Zhou, and Yan Xu. VQPET: leveraging vector-quantized codebook prior for PET image synthesis. *IEEE Trans. Medical Imaging*, 45(5):2542–2560, 2026.

Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022.

Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Chen, Jingwen Li, Yanzhou Su, Min Zhu, and Junjun He. Interactive medical image segmentation: A benchmark dataset and baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20841–20851. Computer Vision Foundation / IEEE, 2025.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, 2024.

Jiahao Huang, Yingying Fang, Yinzhe Wu, Huanjun Wu, Zhifan Gao, Yang Li, Javier Del Ser, Jun Xia, and Guang Yang. Swin transformer for fast mri. *Neurocomputing*, 493:281–304, 2022a.

Jiahao Huang, Xiaodan Xing, Zhifan Gao, and Guang Yang. Swin deformable attention u-net transformer (sdaut) for explainable fast mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 538–548. Springer, 2022b.

Se-In Jang, Tinsu Pan, Ye Li, Pedram Heidari, Junyu Chen, Quanzheng Li, and Kuang Gong. Spach transformer: spatial and channel-wise transformer based on local and global self-attentions for pet image denoising. *IEEE Transactions on Medical Imaging*, 2023.

Chanyong Jung, Joonhyung Lee, Sunkyoung You, and Jong Chul Ye. Patch-wise deep metric learning for unsupervised low-dose CT denoising. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VI*, volume 13436 of *Lecture Notes in Computer Science*, pp. 634–643. Springer, 2022.

Laszlo Kovacs, A Dadya, Andras Domonkos, et al. Papila: Dataset with fundus images for glaucoma analysis and diagnosis. *Scientific Data*, 7(1):1–9, 2020.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, 2022.

Linxuan Li, Wenjia Wei, Luyao Yang, Wenwen Zhang, Jiashu Dong, Yahua Liu, Hongshi Huang, and Wei Zhao. Ct-mamba: A hybrid convolutional state space model for low-dose ct denoising. *Computerized Medical Imaging and Graphics*, pp. 102595, 2025. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2025.102595. URL https://www.sciencedirect.com/science/article/pii/S0895611125001041.

Yuhang Liu, Huazhong Shu, Qiang Chi, Yue Zhang, Zidong Liu, Fuzhi Wu, Xin Chen, Lei Wang, Yi Liu, Pengcheng Zhang, and Zhiguo Gui. Mred-net: A pure tokens-to-token visual mamba-based residual encoder-decoder network for low-dose CT denoising. *Expert Syst. Appl.*, 301: 130305, 2026. doi: 10.1016/J.ESWA.2025.130305. URL https://doi.org/10.1016/j.eswa.2025.130305.

MultiMedia LLC. Ixi dataset. URL https://brain-development.org/ixi-dataset/. Accessed: 2024-01-15.

Yanmei Luo, Luping Zhou, Bo Zhan, Yuchen Fei, Jiliu Zhou, Yan Wang, and Dinggang Shen. Adaptive rectification based adversarial network with spectrum constraint for high-quality pet image synthesis. *Medical Image Analysis*, 77:102335, 2022.

Yiwen Luo, Xiaoying Tang, and Yixuan Yuan. Probabilistic prior-guided anatomical alignment for MRI super-resolution. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025 - 28th International Conference, Daejeon, South Korea, September 23-27, 2025, Proceedings, Part IV*, volume 15963 of *Lecture Notes in Computer Science*, pp. 425–434. Springer, 2025.

Chenglong Ma, Zilong Li, Yuanlin Li, Jing Han, Junping Zhang, Yi Zhang, Jiannan Liu, and Hongming Shan. Radiologist-in-the-loop self-training for generalizable CT metal artifact reduction. *IEEE Trans. Medical Imaging*, 44(6):2504–2514, 2025.

Jiawei Ma, Xiao-Yang Liu, Zheng Shou, and Xin Yuan. Deep tensor ADMM-net for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Jun Ma and Bo Wang (eds.). *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation - MICCAI 2022 Challenge*, volume 13816, 2022.

Jun Ma, Cheng Ge, Yixin Wang, Xingle An, Jiantao Gao, Ziqi Yu, Minqing Zhang, Xin Liu, Xueyuan Deng, Shucheng Cao, et al. Towards efficient covid-19 ct annotation: A benchmark for lung and infection segmentation. *Medical Physics*, 48(10):5629–5648, 2021.

Cynthia H McCollough, Adam C Bartley, Rickey E Carter, Baiyu Chen, Tammy A Drees, Phillip Edwards, David R Holmes III, Alice E Huang, Farhana Khan, Shuai Leng, et al. Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. *Medical Physics*, 44(10):e339–e352, 2017.

A. W. Moawad, D. Fuentes, A. Morshid, et al. Multimodality annotated hcc cases with and without advanced imaging segmentation, 2021.

Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *CVPR*, 2022.

Zhangkai Ni, Wenhan Yang, Hanli Wang, Shiqi Wang, Lin Ma, and Sam Kwong. Cycle-interactive generative adversarial network for robust unsupervised low-light enhancement. In *ACM Multimedia*, pp. 1484–1492, 2022.

Şaban Öztürk, Oğuz Can Duran, and Tolga Cukur. Denomamba: A fused state-space model for low-dose ct denoising. *arXiv preprint arXiv:2409.13094*, 2024.

Chicago Y. Park, Shirin Shoushtari, Hongyu An, and Ulugbek S. Kamilov. Measurement score-based diffusion model. In *International Conference on Learning Representations*, 2026.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *NeurIPS*, 2023.

Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. *The International Conference on Machine Learning (ICML)*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Farid Tasharofi, Fuxin Fan, Melika Qahqaie, Mareike Thies, and Andreas K. Maier. Find-net - fourier-integrated network with dictionary kernels for metal artifact reduction. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025 - 28th International Conference, Daejeon, South Korea, September 23-27, 2025, Proceedings, Part XIII*, volume 15972 of *Lecture Notes in Computer Science*, pp. 192–201. Springer, 2025.

Jiachen Tu, Yaokun Shi, and Fan Lam. Score-based self-supervised MRI denoising. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Chenxu Wang, Kai Zhang, and Jian Yang. Retrieve-to-restore: Efficient all-in-one image restoration with a retrieval-based degradation bank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1277–1287, June 2026a.

Dayang Wang, Fenglei Fan, Zhan Wu, Rui Liu, Fei Wang, and Hengyong Yu. Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising. *Physics in Medicine & Biology*, 68(6):065012, 2023a.

Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M. Jorge Cardoso, and Razvan Marinescu. Inversesr: 3d brain mri super-resolution using a latent diffusion model. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 438–447, Cham, 2023b. Springer Nature Switzerland. ISBN 978-3-031-43999-5.

Xiangming Wang, Haijin Zeng, Benteng Sun, Jiezhang Cao, Kai Zhang, Qiangqiang Shen, and Yongyong Chen. Deep lora-unfolding networks for image restoration. *IEEE Trans. Image Process.*, 35:1858–1869, 2026b.

Chenxu Wu, Qingpeng Kong, Zihang Jiang, and S. Kevin Zhou. Self-supervised diffusion MRI denoising via iterative and stable refinement. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017.

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing MRI. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 10–18, 2016.

Zhiwen Yang, Yang Zhou, Hui Zhang, Bingzheng Wei, Yubo Fan, and Yan Xu. Drmc: A generalist model with dynamic routing for multi-center pet image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36–46. Springer, 2023.

Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Yi, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. All-in-one medical image restoration via task-adaptive routing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 67–77. Springer, 2024a.

Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Zhou, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Region attention transformer for medical image restoration. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part VII*, volume 15007 of *Lecture Notes in Computer Science*, pp. 603–613. Springer, 2024b.

Zhiwen Yang, Jiaju Zhang, Yang Yi, Jian Liang, Bingzheng Wei, and Yan Xu. TAT: task-adaptive transformer for all-in-one medical image restoration. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025 - 28th International Conference, Daejeon, South Korea, September 23-27,*

2025, Proceedings, Part XVI, volume 15975 of *Lecture Notes in Computer Science*, pp. 565–575. Springer, 2025.

Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-rwkv: Efficient and effective medical image restoration with RWKV. *IEEE J. Biomed. Health Informatics*, 30(1):513–526, 2026.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12753–12763, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022.

Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, and Jie Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1828–1837. Computer Vision Foundation / IEEE Computer Society, 2018.

Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 3214–3223. Computer Vision Foundation / IEEE, 2020a.

Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3217–3226, 2020b.

Xiahai Zhuang and Juan Shen. Multi-modality whole heart segmentation: an unaligned challenge for medical image segmentation. *Scientific Reports*, 6(1):1–10, 2016.

A Modal-Degradation Embedder Implementation

The Modal-Degradation Embedder E_{md} has a convolutional stem and two downsampling layers with channel widths 24, 48, and 96, followed by 16 residual blocks. Global average pooling and an MLP with widths $96 \to 1024 \to 1024 \to 1024 \to 256$ produce the joint image-level token g. Restoration consumes this continuous token rather than a predicted class.

Both contrastive objectives in Eq. 7 act on the same normalized token, with temperature 0.07 and unit weights. Modality positives share m_i, whereas degradation positives share d_i; different corresponding labels define negatives. Thus, a pair may be positive for one objective and negative for the other. All non-anchor samples enter each denominator, and anchors without positives are excluded from that objective's average; an empty valid-anchor set gives zero loss. Modality labels identify the imaging modality, not the dataset. Auxiliary degradation and modality classifiers are trained with cross-entropy, each with unit weight in the current implementation. Their accuracies are reported separately and do not directly measure clustering of the token space. The pretrained embedder is frozen during content adaptation and unfolding training.

The manuscript's main training protocol pretrains E_{md} for 50 epochs on the eight-dataset, four-degradation collection described in Appendix D. The shorter diagnostic run used for Fig. 3 is specified separately below.

B Multi-Scale VQVAE Implementation

The multi-scale VQVAE retains a pretrained encoder, decoder, and a shared 4096-entry codebook with 32-dimensional latent features. Multi-scale residual quantization represents coarse structure and successive residual details. The same vocabulary is used across modalities rather than constructing an independent codebook for every task.

After the encoder projection, a known modality selects a residual adapter consisting of two bias-free 1×1 convolutions with a 16-channel bottleneck. Its learnable residual scale is initialized to zero. This produces the modality-adapted spatial feature h_m, into which the degradation token is injected before multi-scale quantization.

The frozen embedder processes the same degraded image to obtain $g \in \mathbb{R}^{256}$. Layer normalization and a linear projection to 64 dimensions produce two 32-dimensional vectors, $\gamma(g)$ and $\beta(g)$. Broadcasting them spatially gives $c = (1+\gamma(g)) \odot h_m + \beta(g)$ in Eq. 8. The token adjusts codebook queries rather than becoming an extra codebook entry. Zero-initialized projection weights and biases make the initial modulation an identity mapping. The adapter handles modality-level appearance, while token conditioning provides image-dependent adjustment.

The degraded and clean paths share the VQVAE and modality adapters, but only the degraded path receives token modulation. Their accumulated quantized representations are aligned using Eq. 9, with the clean target detached from the alignment gradient. Both reconstruction paths use the clean image as the target. Reconstruction, VQ consistency, clean-code alignment, and an auxiliary adversarial objective supervise adaptation. In the main protocol, this stage runs for 50 epochs on the same eight-dataset collection, optimizing the adapters and token-conditioned affine layers while freezing the shared encoder, decoder, codebook, and embedder.

Token modulation remains active during unfolding training and inference. The degraded observation is encoded, modality-adapted, and modulated by g to produce the continuous content feature c, which is passed to the prior fusion module rather than quantized and decoded into an intermediate image. The adapters and token-conditioned affine layers continue to be optimized during unfolding training, while the shared VQVAE backbone and codebook remain frozen. The degradation token therefore participates in two successive steps. It first modulates modality-specific content features within the multi-scale VQVAE, with clean quantized references supervising the enhanced representation. The content–degradation fusion block then combines c with the same token g to generate p, through which both representations jointly control unfolding.

C UNFOLDING IMPLEMENTATION

A 1×1 projection maps the spatial content feature into the fusion space. Four residual modulation blocks generate channel-wise scales and shifts from the normalized degradation token and transform the conditioned content using a depthwise 3×3 convolution, GELU, and a pointwise convolution. Global pooling and an MLP produce the shared 256-dimensional condition p. This condition combines the enhanced modality-specific content features and the degradation prior and is reused throughout reconstruction.

At resolution k, let $x^{(k-1)}$ and y_k denote the estimate and observation features with C_k channels. CD-GDM constructs the prompt

$$P_k = \mathrm{Broadcast}(\mathrm{MLP}_k(p)) + W_k^{\mathrm{ctx}} * y_k, \tag{10}$$

where W_k^{ctx} is a bias-free 1×1 convolution. The MLP maps the 256-dimensional condition through a hidden width $\max(256, C_k)$ and GELU to C_k channels. Thus, the prompt retains the feature channel width rather than passing through a narrow channel bottleneck. The broadcast term supplies global prior conditioning, while the projected observation features supply spatial context.

Queries and values come from the estimate feature, and keys come from P_k, using pointwise and depthwise convolutions. Within each of two heads, queries and keys are normalized over spatial positions. Their dot products, multiplied by a learned per-head scale and passed through row-wise softmax, produce a channel-attention matrix of size $(C_k/2) \times (C_k/2)$. Weighted values are projected to predict the observation feature. This is channel attention, not a matrix of pixel-to-pixel probabilities, and the learned mapping does not select from physical degradation matrices.

Subtracting the observation feature produces the data residual. A Restormer-style multi-DConv head transposed attention (MDTA) module (Zamir et al., 2022) maps this residual back into the estimation space. Its output, scaled by the learned step size ρ_k, is subtracted from the current estimate to obtain $z^{(k)}$. The back-mapping $\mathcal{B}_k$ is independently learned. For a nonlinear forward mapping, an exact fidelity gradient would involve its Jacobian transpose; we do not impose this constraint. The stage is therefore a learned PGD-inspired update rather than an exact gradient step for Eq. 5.

As illustrated in Fig. 6, the PMM refines the corrected feature $z^{(k)}$ using NAFNet-based residual blocks (Chen et al., 2022) to approximate the proximal map. Each block combines a local convolutional path with SimpleGate and simplified channel attention (SCA), followed by a gated channel-mixing path. SimpleGate multiplies two channel groups element-wise, while SCA reweights channels using globally pooled features. Both paths use residual connections. Prior guidance reaches the PMM through the corrected feature, without a separate prompt input. Algorithm 1 summarizes the conditional update.

A Haar transform separates the input into LL, containing low-pass responses along both axes, and LH, HL, and HH, containing a high-pass response along one or both axes. The three high-frequency bands are concatenated. Two independently parameterized hierarchical unfolding networks process LL and the concatenated high-frequency bands, respectively, before inverse Haar reconstruction. The $\times 2$ annotation in Fig. 6 denotes the two separately parameterized NAFNet-based proximal networks for these low- and high-frequency branches, not two successive NAFBlocks within a stage. Both branches receive the same condition p but share neither GDM nor PMM parameters.

Unfolding training uses AdamW for up to 300,000 updates, optimizing the unfolding network, prior fusion, and content adapters while freezing the degradation extractor and shared VQVAE backbone. Evaluation uses complete grayscale slices resized to 256×256, without random test-time crops. PSNR and SSIM are computed per image on normalized $[0, 1]$ outputs in FP32, then averaged within each dataset. SSIM uses an 11×11 Gaussian window with standard deviation 1.5 and padding 5. Overall benchmark averages are unweighted dataset means. AMIR additionally reports RMSE following its evaluation convention.

Here, y_k is the feature representation of y at the current resolution. Queries and keys are normalized over spatial positions, and softmax acts over key channels within each attention head. The projection steps include the pointwise and depthwise convolutions described above. The pseudocode describes a local stage; resolution changes and encoder–decoder skip connections follow the hierarchical network in Fig. 2.

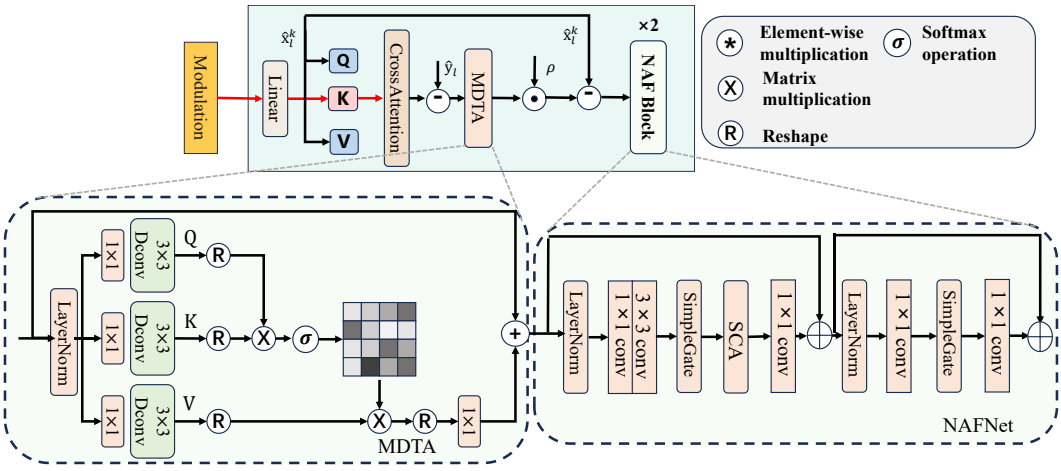

Figure 6: Detailed components in each unfolding stage. Prior-conditioned cross-attention predicts the observation feature, and MDTA back-projects the residual for a learned correction with step size ρ_k. A NAFNet-based PMM then approximates the proximal map. The $\times 2$ annotation denotes separate low- and high-frequency proximal networks with no shared parameters.

Algorithm 1 Dual-prior conditioning and one unfolding-stage update

Require: Observation y, modality m, current estimate feature $x^{(k-1)}$, corresponding observation feature y_k

$g \leftarrow E_{\text{md}}(y)$; $c \leftarrow E_{\text{cont}}(y; m, g)$
$p \leftarrow \mathcal{M}(c, g)$ {Computed once and reused across stages}
$P_k \leftarrow \text{Broadcast}(\text{MLP}_k(p)) + W_k^{\text{ctx}} * y_k$
$Q_k, V_k \leftarrow \text{Project}_{QV,k}(x^{(k-1)})$
$K_k \leftarrow \text{Project}_{K,k}(P_k)$
$A_k \leftarrow \text{Softmax}\big(t_k \ \text{Norm}(Q_k)\,\text{Norm}(K_k)^\top\big)$
$\widehat{y}_k \leftarrow \text{Project}_{out,k}(A_k V_k)$
$r_k \leftarrow \widehat{y}_k - y_k$
$z^{(k)} \leftarrow x^{(k-1)} - \rho_k \ \text{MDTA}_k(r_k)$
$x^{(k)} \leftarrow \text{NAFBlocks}_k(z^{(k)})$
return $x^{(k)}$

Attention-conditioning diagnostic. Fig. 3 uses a separate four-dataset run configured for 10 epochs each of embedder pretraining, content adaptation, and unfolding training; the figure uses the saved third unfolding epoch, not the final checkpoint of the main protocol. One test example per dataset is selected nearest the median full-model PSNR within a fixed eight-example subset, without selecting for intervention effects. Each low- and high-frequency branch contains eight GDMs, giving 16 modules in execution order.

At each GDM, the estimate and observation inputs from the full-model forward pass are held fixed. Four interventions are evaluated without retraining: (i) *No content input*: use $\mathcal{M}(0, g)$; (ii) *No MD guidance*: bypass token modulation in the content encoder and use $\mathcal{M}(c_{\text{raw}}, 0)$; (iii) *Zero condition*: set $p = 0$; and (iv) *No spatial context*: omit $W_k^{\text{ctx}} * y_k$ in Eq. 10 while retaining full prior conditioning. Here, c_{raw} retains modality adaptation but excludes token modulation. Zeroing an input does not remove learned biases; in particular, $p = 0$ still retains the prompt MLP's biases and observation-context branch. These interventions isolate local responses rather than propagating an altered trajectory through all subsequent layers.

For attention A_k and correction $u_k = \rho_k \mathcal{B}_k(\widehat{y}_k - y_k)$, the plotted relative change is

$$\Delta_k(T) = 100 \frac{\|T_k^{\text{int}} - T_k^{\text{full}}\|_F}{\|T_k^{\text{full}}\|_F + \epsilon}, \qquad T \in \{A, u\}. \tag{11}$$

Attention metrics use all heads. The horizontal axis follows GDM execution order; the vertical axis uses a symmetric logarithmic scale. Percentages quoted in the main text average the 16 module-wise changes for the illustrated example, not a dataset-wide population. Larger values indicate stronger conditioning sensitivity, not better restoration or greater performance loss. The spatial-context intervention is also not a retrained comparison with the earlier narrow-projection architecture.

D Datasets and Degradation Simulation

For the four-dataset quantitative benchmark reported in Tables 1–4, the training/validation/test ratio is 7:1:2. Degraded observations are generated offline and fixed across methods.

We use the two-dimensional images distributed through IMed-361M (Cheng et al., 2025), including ACDC cardiac MRI (Bernard et al., 2018), BraTS2021 brain MRI (Baid et al., 2021), COVID19CTscans thoracic CT (Ma et al., 2021), and HCC-TACE-Seg abdominal CT (Moawad et al., 2021). Reference slices serve as HQ targets. These data cover modality-dependent anatomy as well as different corruption conditions within CT.

For ACDC and BraTS2021, a two-dimensional FFT is applied to each reference slice. We retain the central 12.5% of coefficients along each spatial axis, zero the remaining coefficients, and invert the transform. Separately, the reference is downsampled to 12.5% of its width and height and bicubically upsampled to its original size. The Fourier reconstruction and resampled image are combined with weights 0.25 and 0.75 and clipped to $[0, 1]$. Both MRI datasets use this setting. This is an image-based resolution simulation rather than reconstruction from acquired raw k-space.

Normalized reference intensities are scaled by 3.0 and mapped to transmission values through an exponential attenuation model. Poisson counts with photon-budget parameter I_0 are converted back through a negative logarithm and rescaled, followed by additive Gaussian noise. COVID19CTscans uses $I_0 = 1800$ and Gaussian standard deviation 0.035. HCC-TACE-Seg uses $I_0 = 1500$ and standard deviation 0.04, with ten randomly oriented, spatially localized sinusoidal streak components. The combined streak pattern is normalized and scaled to amplitude 0.06, and the final observation is clipped to $[0, 1]$. This is an image-domain approximation, not a projection-domain CT acquisition model.

We follow AMIR (Yang et al., 2024a) for MRI super-resolution, CT denoising, and PET synthesis using IXI MRI (LLC), AAPM-Mayo CT (McCollough et al., 2017), and paired low-/full-dose PET data. Its task definitions and preprocessing are retained; the synthetic corruptions above apply to our benchmark, not to AMIR.

Prior pretraining uses eight datasets: ACDC, BraTS2021, COVID19CTscans, FLARE22, HCC-TACE-Seg, MMWHS_CT, MMWHS_MR, and PAPILA. Both E_{md} and E_{cont} use this collection, covering Gaussian noise, Poisson–Gaussian noise, Poisson–Gaussian noise with directional streaks, and low resolution. End-to-end unfolding and the four-dataset quantitative benchmark use only ACDC, BraTS2021, COVID19CTscans, and HCC-TACE-Seg under the three settings specified above; both MRI datasets share the low-resolution setting. The additional four datasets are used only for prior pretraining, not end-to-end restoration training.

Fig. 1 visualizes the eight-dataset collection. Colors in panels titled "Modality" denote dataset identity, not eight distinct imaging modalities; degradation colors denote the four corruption categories. Dataset-colored separation should therefore not be interpreted as independent evidence of modality–degradation disentanglement.

E One-by-One Results on the AMIR Benchmark

Table 5 reports task-specific results for MRI super-resolution, CT denoising, and PET synthesis.

MoDeFold achieves the highest PSNR and lowest RMSE across all three tasks, exceeding the strongest competing PSNR by 0.21, 0.18, and 0.35 dB for MRI, CT, and PET, respectively. It also achieves the highest SSIM on MRI and PET. On CT, its SSIM of 0.919 is slightly below MRED-Net's 0.920, despite lower reconstruction error. The largest PSNR gain occurs in PET synthesis, while the MRI results show improvements in both intensity accuracy and structural similarity. These

Table 5: One-by-one restoration on the AMIR benchmark. Per-task baselines and measurements are retained from the original comparison.

Method	MRI Super-Resolution			Method	CT Denoising			Method	PET Synthesis		
	PSNR↑	SSIM↑	RMSE↓		PSNR↑	SSIM↑	RMSE↓		PSNR↑	SSIM↑	RMSE↓
DAGAN	30.55	0.919	34.087	EDCNN	33.19	0.911	8.943	DCNN	36.27	0.924	0.095
SwinMR	30.93	0.925	32.734	DML_CT	33.22	0.912	8.903	CycleWGAN	36.62	0.929	0.091
SDAUT	30.96	0.926	32.593	CTformer	33.25	0.913	8.897	ARGAN	36.73	0.941	0.090
MambaIR	31.77	0.937	29.837	DenoMamba	33.53	0.915	8.612	DRMC	36.00	0.935	0.100
PGASR	32.03	0.94	26.32	MRED-Net	33.62	0.92	8.42	VLU-Net	37.12	0.940	0.088
MoDeFold	32.24	0.945	25.864	**MoDeFold**	33.80	0.919	8.360	**MoDeFold**	37.47	0.947	0.086

results complement the all-in-one evaluation, showing that the dual-prior unfolding framework remains effective when trained separately for individual restoration tasks.