EasyChair Preprint

№ 16019

MetaDB: Metadata-Guided Diffusion Bridge Model for High-Fidelity Medical Image Synthesis

Yanjun Chi, Keqiang Wang, Wei Huang, Wei Xu, Jiaen Liang and Jun Yu

May 5, 2026

MetaDB: Metadata-Guided Diffusion Bridge Model for High-Fidelity Medical Image Synthesis

Yanjun Chi
University of Science and Technology of China
Hefei, China
yiChi@mail.ustc.edu.cn

Keqiang Wang
Ping An Technology Co., Ltd
Shanghai, China
wangkeqiang265@pingan.com.cn

Wei Huang
Unisound AI Technology Co., Ltd
Beijing, China
huangwei@unisound.com

Wei Xu
The First Affiliated Hospital, Division of Life Sciences and Medicine, University of Science and Technology of China
Hefei, China
xw199807@163.com

Jiaen Liang
Unisound AI Technology Co., Ltd
Beijing, China
liangjiaen@unisound.com

Jun Yu*
University of Science and Technology of China
Hefei, China
harryjun@ustc.edu.cn

Abstract

Medical image synthesis is pivotal in modern clinical workflows, addressing the issue of missing imaging modalities. While diffusion-based models have shown promise, existing approaches often neglect the rich clinical metadata, leading to synthesized images that lack semantic fidelity and fail to maintain strict consistency with the target modality. To address these challenges, we propose a metadata-guided diffusion bridge model, termed MetaDB, a novel framework that leverages textual clinical priors to steer the source-to-target translation process. Our method introduces two key innovations to ensure high-fidelity synthesis. First, we design a text-guided adaptive normalization layer, which dynamically modulates the feature statistics of the diffusion backbone using encoded clinical metadata. This mechanism explicitly aligns the synthesized features with the target modality's attributes, ensuring semantic consistency throughout the generation process. Second, to prevent semantic degradation during the iterative denoising steps, we propose a semantics reconstruction network. This auxiliary module imposes a constraint that forces the network to preserve deep semantic representations, further reinforcing the semantic consistency between the generated output and the target description. Extensive experiments on multiple medical imaging datasets demonstrate that our approach achieves state-of-the-art performance in terms of quantitative metrics and visual quality, generating images that are both anatomically accurate and semantically faithful to clinical protocols.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Medical Image Synthesis; Diffusion Bridge Model; Clinical Metadata

ACM Reference Format:
Yanjun Chi, Keqiang Wang, Wei Huang, Wei Xu, Jiaen Liang, and Jun Yu. 2026. MetaDB: Metadata-Guided Diffusion Bridge Model for High-Fidelity Medical Image Synthesis. In *International Conference on Multimedia Retrieval (ICMR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3805622.3810728

*Corresponding author.

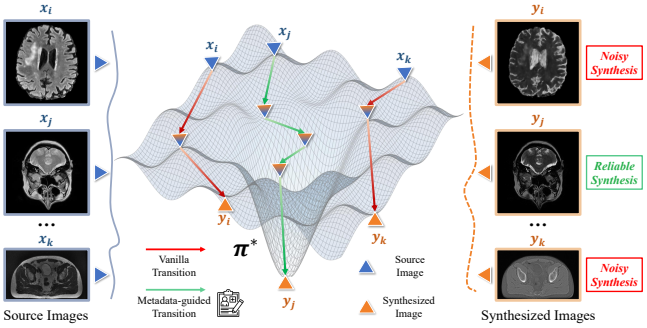

Figure 1: Comparison between vanilla and our proposed metadata-guided synthesis paradigms. Unlike the vanilla transition which follows an unconstrained path resulting in unreliable samples, our metadata-guided transition actively corrects the transition pathway using clinical metadata. This ensures that the synthesized images accurately converge to the underlying target distribution π^*, yielding reliable and high-fidelity synthesis.

1 Introduction

Medical image synthesis is pivotal in modern clinical workflows, addressing the critical issue of missing imaging modalities [4, 13, 18, 33]. While multi-modal data (e.g., MRI and CT) provide complementary anatomical and functional information, acquiring a complete set is often precluded by cost, radiation exposure, and patient compliance [37, 41, 42]. This data scarcity significantly impairs downstream tasks such as radiotherapy planning, lesion characterization, and joint medical image segmentation [4, 7]. Consequently, learning-based image translation has emerged as a dominant solution to map available source modalities to target ones in a data-driven manner [6, 8, 24–30, 35, 38, 39, 43–45].

Among existing paradigms, Generative Adversarial Networks (GANs) have been extensively explored but often suffer from training instability and error accumulation, which can compromise anatomical correctness [1, 34]. Recently, Denoising Diffusion Models (DDMs) offered improved stability via probabilistic modeling [20]. However, standard DDMs are primarily designed for noise-to-image generation. The influence of the source modality often dilutes during the reverse process, leading to a misalignment between denoising and translation objectives [32].

To mitigate this, Denoising Diffusion Bridge Models (DDBMs) have been introduced to establish direct source-target diffusion pathways, preserving better anatomical structure [51, 52]. Despite their promise, current DDBMs typically rely on pixel-level constraints, neglecting rich clinical metadata. This oversight leads to two critical limitations: 1) the inability to disentangle modality-specific styles without explicit clinical priors, and 2) semantic degradation, where intermediate states drift away from the target clinical attributes during the iterative process.

To address these challenges, we propose a metadata-guided diffusion bridge model, termed MetaDB. As illustrated in Fig. 1, MetaDB explicitly incorporates clinical metadata to guide the diffusion bridge transition toward the target modality. Specifically, we design a text-guided adaptive normalization layer to dynamically modulate feature statistics, ensuring continuous alignment with the target modality's attributes. Furthermore, to combat semantic degradation, we introduce a semantics reconstruction network that imposes a latent constraint, forcing the network to preserve deep semantic representations faithful to the clinical description.

Our key contributions are summarized as follows:

- We propose MetaDB, a novel framework that integrates clinical metadata into the diffusion bridge paradigm to ensure strict semantic adherence to target modalities.
- We introduce two core designs: a TA-Norm layer for dynamic feature modulation and a Semantics Reconstruction Network to prevent semantic degradation during denoising.
- Extensive experiments on multi-contrast MRI and MRI-to-CT tasks demonstrate that MetaDB significantly outperforms state-of-the-art methods in both quantitative metrics and visual fidelity.

2 Related Work

2.1 Medical Image Synthesis

Generative Adversarial Networks (GANs) initially dominated the field, utilizing paired learning like Pix2Pix [5] or cycle-consistency mechanisms like CycleGAN [48] for cross-modal translation. Subsequent variants further enhanced performance through attention mechanisms [46] and feature disentanglement strategies [3], though training instability and mode collapse remain persistent challenges. Recently, Denoising Diffusion Models (DDMs) have emerged as a robust alternative, offering superior training stability and distribution coverage compared to adversarial approaches, albeit often at the cost of slower inference speeds and potential structural degradation during the noise-to-image generation process.

2.2 Diffusion Bridges

To address the structural limitations of standard DDMs, Denoising Diffusion Bridge Models (DDBMs) [15, 51, 52] were introduced to establish direct generative trajectories between source and target domains, effectively bypassing the information loss associated with pure Gaussian noise. In medical imaging, notable advancements include DBIM [51], which employs implicit sampling to accelerate the translation process, and DualDB [40], which integrates dual-domain alignments to preserve high-frequency anatomical fidelity. Other variants have also demonstrated efficacy in tasks such as fundus image enhancement [14]. However, despite their success in structural alignment, these methods predominantly rely on pixel-level constraints, often neglecting the explicit semantic guidance offered by clinical metadata. Advanced optimization strategies, such as bilevel and constraint learning, alongside robust degradation modeling, have also been extensively explored to address complex constraints in other broad vision tasks like low-light imaging, underwater restoration, and generalized transfer attacks [9–12, 16, 17, 21–23, 47, 49].

3 Methods

3.1 Overview

MetaDB utilizes a diffusion bridge framework to progressively transform a source-modality image y into a target-modality image x_0. Unlike standard diffusion models that degrade data to Gaussian noise, our forward process bridges the target image x_0 and the clean source image y (denoted as x_T) through intermediate noisy states x_t. The forward transition probability is defined as:

$$q(x_t|x_0, y) = \mathcal{N}\left(x_t; (1 - m_t)x_0 + m_t y, \delta_t^2 I\right), \tag{1}$$

where m_t controls the interpolation between domains and δ_t^2 represents the noise variance at each timestep $t \in [0, T]$.

The reverse process aims to reconstruct the target image x_0 from x_T conditioned on the source image y. At each timestep t, the text-guided generator G_θ predicts an estimate of the target image $\hat{x}_0$:

$$\hat{x}_0 = G_\theta(x_t, t, y, \mathbf{E}_{text}). \tag{2}$$

Here, $\mathbf{E}_{text}$ represents the clinical text embeddings. Once the initial estimate $\hat{x}_0$ is obtained, the subsequent state x_{t-1} is sampled from the posterior distribution using $\hat{x}_0$ as a surrogate for the ground truth, ensuring a stable trajectory towards the target modality.

3.2 Text-Guided Image Synthesis

To ensure the synthesized images are not only structurally coherent but also semantically faithful to clinical protocols, we integrate explicit textual guidance into the generation process. This is achieved through two core components: the Text-guided Adaptive Normalization (TA-Norm) layer and the Semantics Reconstruction Network.

3.2.1 Text-guided Adaptive Normalization. Standard normalization layers in diffusion models often fail to explicitly disentangle the modality-specific styles of the source and target domains. Motivated by the normalization design in UniSyn [36], we incorporate the TA-Norm layer into the encoder and decoder blocks of G_θ, as illustrated in Fig. 3.

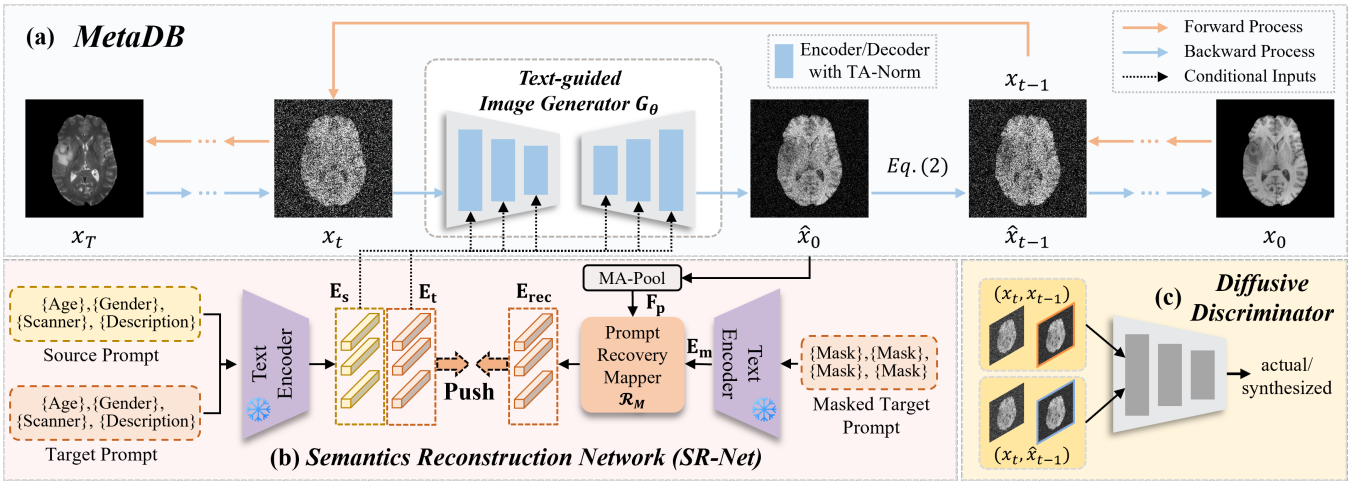

Figure 2: Schematic of MetaDB. (a) The diffusion bridge transitions target x_0 to source $x_T = y$ (forward) and reconstructs x_0 via generator G_θ (reverse). (b) The semantics reconstruction network (SR-Net) recovers masked text embeddings from global image features via mapper $\mathcal{R}_M$ to enforce consistency. (c) The Diffusive Discriminator distinguishes real transition pairs (x_t, x_{t-1}) from synthesized ones $(x_t, \hat{x}_{t-1})$.

This module leverages clinical metadata to dynamically modulate feature statistics. Let $\mathbf{E}_s$ and $\mathbf{E}_t$ denote the text embeddings of the source and target prompts, extracted by a pre-trained text encoder. We employ learnable mapping networks, composed of linear layers followed by sigmoid activation functions, to project these embeddings into affine transformation parameters. Specifically, the source embedding produces normalization parameters $(\boldsymbol{\alpha}_s, \boldsymbol{\beta}_s)$, while the target embedding produces modulation parameters $(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$.

For an input feature map $\mathbf{F}_i$, TA-Norm first "normalizes" it to strip away source-specific attributes using $(\boldsymbol{\alpha}_s, \boldsymbol{\beta}_s)$, and then "modulates" it to inject target-specific attributes using $(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$. The operation is formulated as:

$$\mathbf{F}_o = \left(\frac{\mathbf{F}_i - \boldsymbol{\beta}_s}{\boldsymbol{\alpha}_s}\right) \cdot \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t. \tag{3}$$

This dual-step process explicitly aligns the feature statistics with the target clinical description while preserving the underlying anatomical structure.

3.2.2 Semantics Reconstruction Network. During the iterative reverse process, the semantic consistency of the estimated $\hat{x}_0$ may degrade. To counter this, we introduce a Semantics Reconstruction Network (Fig. 2(b)) that imposes a semantic consistency constraint.

We employ a mask-then-recover strategy. Specifically, we mask key tokens in the target text prompt (e.g., masking the modality token "T1") to obtain a masked prompt embedding $\mathbf{E}_{mask}$. Simultaneously, we extract global visual features from the generator's predicted image $\hat{x}_0$ using a multi-scale pooling (MA-Pool) module, which combines global average pooling and global max pooling:

$$\mathbf{F}_p = \text{Concat}(\text{AvgPool}(\hat{x}_0), \text{MaxPool}(\hat{x}_0)). \tag{4}$$

These visual features $\mathbf{F}_p$ contain the semantic information present in the current image estimate. A Prompt Recovery Mapper $\mathcal{R}_M$ then takes both $\mathbf{E}_{mask}$ and $\mathbf{F}_p$ as input to reconstruct the complete target text embedding $\mathbf{E}_{rec}$. Specifically, $\mathcal{R}_M$ is implemented as a lightweight adapter module, consisting of concatenated projection layers followed by a multi-layer perceptron (MLP) to effectively fuse the multi-modal features and project them back into the text embedding space.

$$\mathbf{E}_{rec} = \mathcal{R}_M(\mathbf{E}_{mask}, \mathbf{F}_p). \tag{5}$$

By forcing the reconstructed embedding $\mathbf{E}_{rec}$ to match the original unmasked target embedding $\mathbf{E}_t$, we enforce that the synthesized

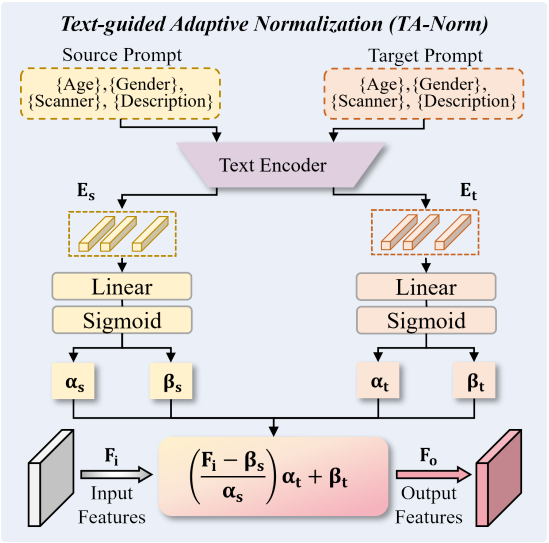

Figure 3: Illustration of the Text-guided Adaptive Normalization (TA-Norm). It normalizes input features using source parameters $(\boldsymbol{\alpha}_s, \boldsymbol{\beta}_s)$ and modulates them with target parameters $(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$ to achieve semantic alignment.

Table 1: Descriptions of experimental datasets.

Dataset	IXI	BRATS	MR-CT
Modality	T1/T2/PD	T1/T2/FLAIR	T1/T2/CT
Image Number	12000	16500	4500
Train/Val/Test	25/5/10	25/10/20	9/2/4
Regions	EU	NA/EU/Asia	EU

image $\hat{x}_0$ must contain the necessary semantic information to fill in the missing gaps in the text prompt.

3.3 Optimization Objective

The optimization objective of MetaDB comprises three components: reconstruction loss for anatomical fidelity, semantic consistency loss for textual alignment, and adversarial diffusive loss for perceptual realism.

To ensure strict anatomical fidelity, we minimize the pixel-level L_1 distance between the predicted image $\hat{x}_0$ and the ground truth x_0:

$$\mathcal{L}_{rec} = \mathbb{E}_{x_0,y,t}\left[\|x_0 - \hat{x}_0\|_1\right]. \tag{6}$$

To prevent semantic degradation, we introduce a semantic consistency loss via the Semantics Reconstruction Network. This term maximizes the cosine similarity $\mathcal{H}$ between the original target text embedding $\mathbf{E}_t$ and the one reconstructed by $\mathcal{R}_M$ using the masked embedding $\mathbf{E}_{mask}$ and visual features $\mathbf{F}_p$:

$$\mathcal{L}_{sem} = \mathbb{E}\left[\mathcal{H}\left(\mathbf{E}_t, \mathcal{R}_M(\mathbf{E}_{mask}, \mathbf{F}_p)\right)\right]. \tag{7}$$

To enhance trajectory authenticity, we employ an adversarial diffusive loss where a discriminator D_θ distinguishes synthesized transition pairs $(\hat{x}_{t-1}, x_t)$ from real ones. The generator optimization is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}_{x_t,\hat{x}_{t-1}}\left[-\log(D_\theta(x_t, \hat{x}_{t-1}))\right]. \tag{8}$$

The final objective is a weighted sum of these terms, where λ_1 and λ_2 balance the adversarial and semantic contributions:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_1\mathcal{L}_{adv} + \lambda_2\mathcal{L}_{sem}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Dataset. We evaluated MetaDB on three diverse datasets: IXI[1], BRATS [2], and Pelvic MRI-CT [19], covering various modalities (T1, T2, PD, FLAIR, CT) and anatomical regions (see Table 1). Data was partitioned into training, validation, and test sets using a strict subject-level split to prevent data leakage. We selected 100 axial cross-sections per subject, applying standard preprocessing including spatial alignment, intensity normalization to $[0, 1]$, and resizing to 256×256.

Implementation Details. Implemented in PyTorch on NVIDIA RTX 4090 GPUs, our framework consists of a text-guided generator G_θ, a semantics reconstruction network, and a diffusive discriminator D_θ. G_θ utilizes a U-Net backbone [31] with channel multipliers $[1, 1, 2, 2, 4, 4]$ and our proposed TA-Norm. We employ a frozen pretrained BiomedCLIP [50] as the text encoder to preserve semantic alignment. The discriminator D_θ adopts a patch-based architecture with temporal embedding integration.

We train the model for 50 epochs (batch size 16) using the Adam optimizer ($\beta = [0.5, 0.9]$), with learning rates of 1.6×10^{-4} for G_θ and 1.0×10^{-4} for D_θ. Loss weights are set to $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$. The diffusion process is configured with $N = 10$ discrete steps using a linear noise schedule ($\beta_{start} = 0.1, \beta_{end} = 3.0$). During inference, we apply a self-consistency recursion depth of $R = 2$ to enhance generation quality.

Evaluation Metrics. Quantitative performance is evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Statistical significance is verified via non-parametric Wilcoxon signed-rank tests ($p < 0.05$).

4.2 Experimental Results

We compare our proposed MetaDB with several state-of-the-art medical image synthesis methods. These include representative GAN-based approaches such as Pix2Pix [5] and DTF-Net [3], as well as recent diffusion-based frameworks like SynDiff [20], DDBM [52], DBIM [51], and DualDB [40].

4.2.1 Qualitative Comparison. The visual comparisons are presented in Fig. 4. MetaDB consistently generates images with superior realism and anatomical fidelity compared to competing methods. In MRI translation tasks (e.g., T1→T2), GAN-based methods [3, 5] often suffer from structural blurring and mode collapse, while standard diffusion bridges like DDBM [52] exhibit slight texture distortions due to the lack of semantic constraints. In contrast, MetaDB produces sharper tissue interfaces and more accurate contrast phenotypes, benefiting from the dynamic modulation of TA-Norm. For pathological cases, competitors often fail to clearly delineate tumor boundaries or misrepresent heterogeneous lesion textures. MetaDB, however, accurately reconstructs the tumor core and edema regions, demonstrating the effectiveness of the Semantics Reconstruction Network in preserving high-level diagnostic features. In the challenging MRI→CT task, methods like Pix2Pix [5] and SynDiff [20] frequently introduce streaking artifacts or intensity inhomogeneity in bone structures. Our approach excels at rendering cortical bone with distinct boundaries and uniform density, ensuring high clinical interpretability.

4.2.2 Quantitative Results. Table 2 reports the quantitative comparison results across the three datasets. As observed, MetaDB consistently achieves the highest PSNR and SSIM scores, establishing a new state-of-the-art performance. Notably, MetaDB significantly outperforms the recent strong baseline, DualDB, across all translation tasks. These substantial gains are primarily attributed to the integration of our text-guided adaptive normalization (TA-Norm) and semantics reconstruction network. Unlike DualDB [40] and DBIM [51], which rely heavily on pixel-level or gradient-based constraints, MetaDB leverages explicit clinical textual priors to steer the generation process. This semantic guidance is particularly critical in cross-modal mappings with complex contrast variations (e.g., MRI→CT), where it effectively resolves intensity ambiguities that pure visual-based methods struggle to handle. Non-parametric Wilcoxon signed-rank tests confirm that these improvements are

[1] http://brain-development.org/ixi-dataset/

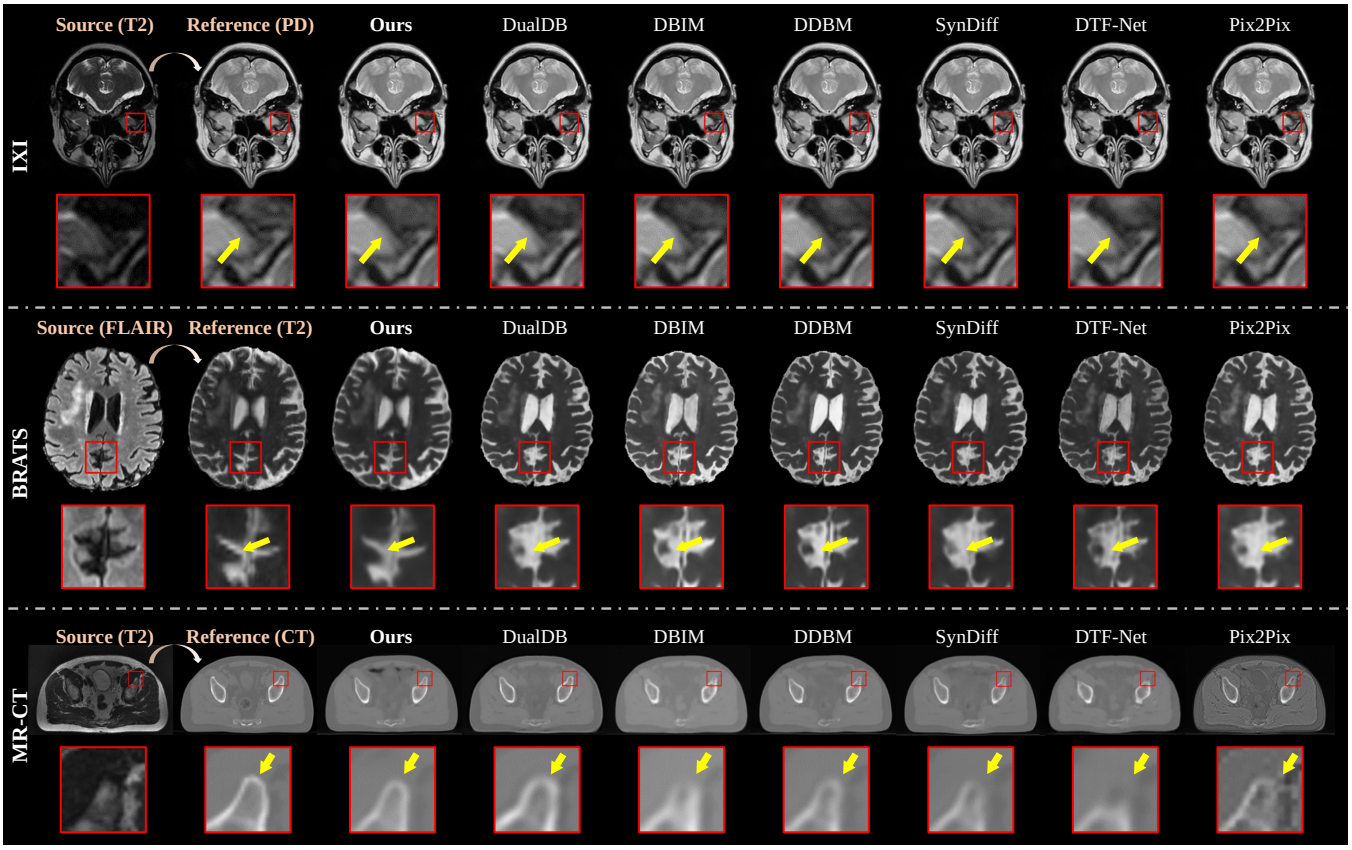

Figure 4: Qualitative comparison of MetaDB with state-of-the-art methods across three medical datasets. Zoom-in views below each result highlight anatomical consistency and synthesis accuracy.

statistically significant ($p < 0.05$), validating the necessity of integrating clinical metadata for reliable medical synthesis.

4.2.3 Downstream Segmentation Validation. High-fidelity medical synthesis must not only look realistic but also preserve the underlying pathological semantics essential for diagnosis. To verify this, we conducted a downstream segmentation experiment on the BRATS dataset. We utilized a pre-trained segmentation network to generate prediction masks from the synthesized images produced by MetaDB and competing methods. As visualized in Fig. 6, the segmentation maps include three tumor sub-regions: necrosis (red), enhancing tumor (blue), and edema/invasion (green). Existing GAN-based methods often fail to preserve the precise boundaries of necrotic cores or misclassify edema regions due to texture blurring. Similarly, standard diffusion baselines occasionally suffer from semantic drift, leading to fragmented tumor masks. In contrast, MetaDB generates images that yield segmentation results highly consistent with the ground truth. This indicates that our proposed Semantics Reconstruction Network effectively preserves complex pathological structures, ensuring that the synthesized images retain their clinical diagnostic value.

4.2.4 Robustness Analysis. We further tested the models by corrupting the source input images with Gaussian noise ($\sigma = 10$) (Bottom row of Fig. 5). Pure pixel-based translation methods typically amplify input noise or lose structural details during the denoising attempts. MetaDB, leveraging the generative prior of the diffusion backbone and the semantic guidance of TA-Norm, demonstrates superior noise resilience. It successfully recovers clean anatomical structures from the noisy source, highlighting its potential for enhancing low-quality clinical scans.

4.3 Ablation Study

To validate the effectiveness of the proposed framework, we conducted comprehensive ablation studies analyzing the individual contributions of our core modules and the impact of different textual conditioning strategies.

4.3.1 Effect of TA-Norm. To verify the superiority of our proposed Text-guided Adaptive Normalization (TA-Norm) over conventional conditioning mechanisms, we compared it against the standard Cross-Attention mechanism widely used in latent diffusion models. As presented in Table 4, replacing Cross-Attention with TA-Norm yields consistent improvements across all datasets. Standard cross-attention tends to inject semantic information spatially, which is highly effective for object generation but less suitable for medical

Table 2: Quantitative comparison of text-guided cross-modal synthesis. Text guidance is injected into all baselines for fairness. Results are reported as mean ± std. † denotes statistically significant improvement over all competitors ($p < 0.05$).

Method	PD→T2		T2→PD		T2→T1		T1→T2	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Pix2Pix [5]	29.02±1.85	92.35±1.52	30.59±1.48	92.98±1.58	27.26±1.36	93.48±1.55	27.01±1.86	92.01±1.74
DTF-Net [3]	31.98±1.58	95.81±1.29	32.63±1.14	95.26±1.17	28.78±1.02	95.11±1.31	29.23±1.74	94.61±1.33
SynDiff [20]	32.23±1.05	96.01±1.01	32.88±1.22	95.36±1.05	28.48±1.01	94.91±1.29	29.28±1.24	94.76±1.27
DDBM [52]	32.53±1.27	96.46±1.08	33.58±1.34	95.81±1.00	30.03±1.31	95.66±1.17	29.68±1.38	95.11±1.29
DBIM [51]	32.73±1.25	96.51±1.05	33.71±1.33	95.96±0.98	30.29±1.29	95.91±1.15	29.93±1.36	95.34±1.28
DualDB [40]	34.12±1.23	97.15±0.95	34.89±1.48	96.58±0.93	31.96±1.55	96.48±1.18	30.87±1.49	95.58±1.25
MetaDB (*ours*)	**34.42±1.17**†	**97.48±0.90**†	**35.21±1.41**†	**96.93±0.88**†	**32.19±1.47**†	**96.82±1.12**†	**31.08±1.42**†	**95.90±1.19**†

(a) Quantitative comparison on the IXI dataset.

Method	FLAIR→T2		T2→FLAIR		T2→T1		T1→T2	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Pix2Pix [5]	24.78±1.98	86.52±3.70	27.06±1.60	87.37±3.34	27.90±1.23	92.29±2.35	26.72±2.02	91.73±2.79
DTF-Net [3]	25.58±1.79	89.47±3.18	27.71±1.57	89.00±2.93	28.35±1.17	93.43±2.04	26.79±1.87	91.82±2.83
SynDiff [20]	25.75±1.70	91.37±3.06	27.21±1.68	89.50±2.72	27.96±1.34	92.58±2.01	26.40±2.04	91.62±3.24
DDBM [52]	26.05±1.76	90.43±2.99	27.98±1.63	89.77±2.77	28.30±1.58	93.67±2.08	26.75±1.50	92.17±2.48
DBIM [51]	26.30±1.72	90.68±2.93	28.05±1.60	90.22±2.70	28.42±1.54	93.96±2.03	26.99±1.46	92.09±2.43
DualDB [40]	27.88±1.97	92.12±2.71	29.04±1.80	90.88±2.60	29.45±1.58	94.78±2.05	28.16±2.15	92.95±2.80
MetaDB (*ours*)	**28.02±1.87**†	**92.44±2.57**†	**29.18±1.71**†	**91.23±2.47**†	**29.59±1.52**†	**95.10±1.94**†	**28.32±2.04**†	**93.27±2.66**†

(b) Quantitative comparison on the BRATS dataset.

Method	MRI(T2)→CT		CT→MRI(T2)		MRI(T1)→CT		CT→MRI(T1)	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Pix2Pix [5]	26.23±1.50	87.15±2.64	23.26±1.65	79.27±2.91	24.76±2.60	87.48±7.00	21.72±2.86	79.50±7.70
DTF-Net [3]	26.62±1.44	87.73±2.58	23.64±1.59	79.75±2.84	25.24±2.53	88.10±6.91	22.22±2.78	80.12±7.60
SynDiff [20]	27.33±1.86	92.23±1.93	24.20±2.05	84.17±2.12	26.74±2.41	91.57±4.79	23.70±2.65	83.59±5.26
DDBM [52]	26.99±1.91	90.54±2.48	23.87±2.10	83.56±2.73	27.76±4.45	93.02±5.05	24.74±4.89	85.04±5.56
DBIM [51]	27.27±1.88	90.87±2.45	24.14±2.07	83.89±2.70	28.10±4.39	93.40±5.01	25.06±4.83	85.42±5.51
DualDB [40]	28.92±2.08	93.68±1.70	25.63±2.29	85.38±1.87	28.88±3.35	93.52±5.25	24.96±3.69	85.85±5.78
MetaDB (*ours*)	**29.22±1.98**†	**94.31±1.62**†	**25.86±2.18**†	**86.09±1.78**†	**29.18±3.18**†	**94.15±4.99**†	**25.20±3.51**†	**86.47±5.49**†

(c) Quantitative comparison on the Pelvic MR-CT dataset.

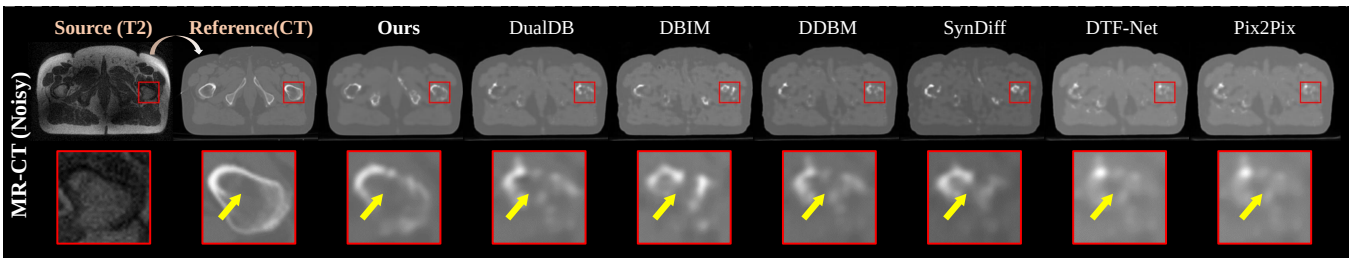

Figure 5: Qualitative robustness comparison under two challenging scenarios: out-of-distribution (OOD) testing on the in-house dataset (top) and Gaussian noise corruption with $\sigma = 10$ (bottom).

modality translation where global intensity modulation is paramount. In contrast, TA-Norm directly modulates the affine parameters of the feature maps using clinical priors. This allows for a global alignment of feature statistics with the target modality's attributes, ensuring that the synthesized image adheres to the specific contrast protocols defined in the metadata.

4.3.2 Effect of SR-Net. The Semantics Reconstruction Network (SR-Net) is crucial for mitigating semantic degradation. As evidenced in Table 3, removing SR-Net leads to a notable performance drop, particularly in the challenging MRI→CT task. Without this constraint, intermediate states tend to drift from the textual description during reverse diffusion. By enforcing the recovery of masked prompts from visual features, SR-Net effectively locks the generation trajectory to the target clinical protocols, ensuring high-level semantic fidelity.

4.3.3 Effect of Textual Prompt Content. We analyzed specific metadata contributions on the IXI and Pelvic datasets (Table 5). The baseline without text guidance yields the lowest performance, validating its necessity. Among attributes, *Description* (modality/imaging parameters) provides the most significant gain, serving as the primary

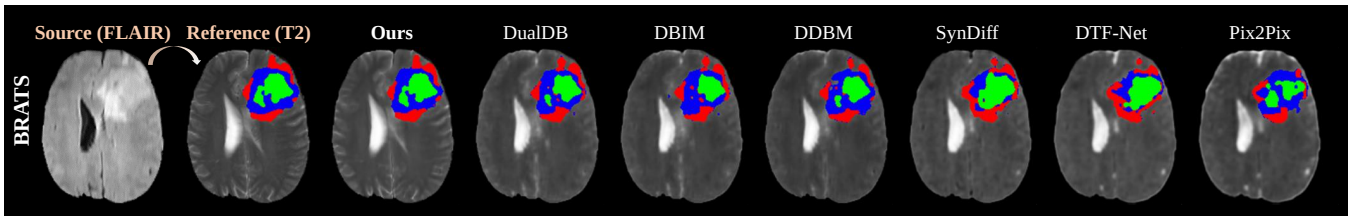

Figure 6: Visual comparison of segmentation results on images synthesized by different methods, with necrosis, enhancing tumor, and edema/invasion annotated in red, blue, and green, respectively.

Table 3: Ablation study of the key components in MetaDB. TA-Norm denotes the text-guided adaptive normalization, and SR-Net denotes the semantics reconstruction network.

TA-Norm	SR-Net	T2→PD		FLAIR→T2		MRI(T2)→CT	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
-	-	32.53 ±1.27	96.46 ±1.08	26.05 ±1.76	90.43 ±2.99	26.99 ±1.91	90.54 ±2.48
-	✓	33.85 ±1.35	96.82 ±1.02	27.15 ±1.82	91.20 ±2.80	27.95 ±1.95	92.10 ±2.15
✓	-	34.50 ±1.38	97.10 ±0.95	27.60 ±1.85	91.95 ±2.65	28.60 ±1.92	93.50 ±1.85
✓	✓	**35.21 ±1.41**	**97.48 ±0.90**	**28.02 ±1.87**	**92.44 ±2.57**	**29.22 ±1.98**	**94.31 ±1.62**

Table 4: Ablation study on text-guided conditioning components. Comparison of conventional cross-attention versus our proposed text-guided adaptive normalization (TA-Norm).

Method	T2→PD		FLAIR→T2		MRI(T2)→CT	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/ Cross-Attention	34.67 ±1.48	96.38 ±0.93	27.65 ±1.97	91.89 ±2.71	28.67 ±2.08	93.36 ±1.70
w/ TA-Norm (*ours*)	**35.21 ±1.41**	**97.48 ±0.90**	**28.02 ±1.87**	**92.44 ±2.57**	**29.22 ±1.98**	**94.31 ±1.62**

driver for synthesis. *Scanner* information aids domain adaptation, while demographic attributes (*Age*, *Gender*) offer marginal individual improvements. However, the full MetaDB integrating all attributes achieves optimal results, demonstrating that detailed patient metadata provides complementary guidance for refining subtle anatomical features.

4.3.4 Effect of Prompt Content. To verify the necessity of the dual-prompt mechanism in our TA-Norm layer, we evaluated the model under four settings: 1) without any text prompt, 2) with only the source prompt (for feature stripping), 3) with only the target prompt (for style injection), and 4) with both prompts (MetaDB). As shown in Table 6, the complete absence of prompts yields the lowest performance (e.g., 32.53 dB on T2→PD), confirming that text guidance is essential for the diffusion bridge. Using only the source prompt provides marginal improvement, as it aids in removing source-specific characteristics but lacks the target modality's guidance. Conversely, using only the target prompt brings a significant boost (34.15 dB), as injecting target style is critical for synthesis. However, the best performance is achieved only when both prompts are utilized (35.21 dB). This validates the design of TA-Norm, which requires a "strip-then-inject" process: effectively removing source information before injecting target attributes ensures the most accurate modality translation.

Table 5: Ablation study on textual attribute contributions in MetaDB. Results for the FLAIR→T2 task are omitted due to reliance on modality-only text.

Age	Gender	Scanner	Description	T2→PD		MRI(T2)→CT	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑
				33.15 ±1.68	95.85 ±1.18	27.18 ±2.25	92.75 ±1.85
✓				33.28 ±1.65	95.96 ±1.15	27.25 ±2.22	92.88 ±1.83
	✓			33.30 ±1.63	96.02 ±1.14	27.30 ±2.20	92.92 ±1.81
		✓		33.75 ±1.55	96.45 ±1.08	27.85 ±2.12	93.35 ±1.75
			✓	34.65 ±1.48	97.05 ±1.02	28.70 ±2.05	93.95 ±1.68
✓	✓	✓	✓	**35.21 ±1.41**	**97.48 ±0.90**	**29.22 ±1.98**	**94.31 ±1.62**

Table 6: Ablation study on the source and target prompt configurations in TA-Norm. Both prompts are required for the optimal "strip-then-inject" feature modulation.

Configuration	T2→PD		MRI(T2)→CT	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
None (w/o Text)	32.53 ± 1.27	96.46 ± 1.08	26.99 ± 1.91	90.54 ± 2.48
Source Prompt Only	33.02 ± 1.35	96.65 ± 1.05	27.45 ± 2.05	91.20 ± 2.30
Target Prompt Only	34.15 ± 1.30	97.05 ± 0.98	28.50 ± 1.95	93.15 ± 1.90
Both (MetaDB)	**35.21 ± 1.41**	**97.48 ± 0.90**	**29.22 ± 1.98**	**94.31 ± 1.62**

5 Conclusion

In this paper, we proposed MetaDB, a novel diffusion bridge framework that leverages clinical metadata to guide high-fidelity medical image synthesis. By integrating text-guided adaptive normalization for dynamic feature modulation and a semantics reconstruction network for consistency enforcement, our method effectively resolves the semantic ambiguity prevalent in existing approaches. Extensive experiments on three diverse datasets demonstrate that MetaDB significantly outperforms state-of-the-art methods in both anatomical accuracy and clinical faithfulness. Future work will explore extending this paradigm to broader multi-modal downstream tasks.

Acknowledgments

This work was supported by the Natural Science Foundation of China (62276242), Hefei Municipal Natural Science Foundation (HZR2431), CAAI-MindSpore Open Fund, developed on OpenI Community.

References

[1] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. 2020. MedGAN: Medical image translation using GANs. *Computerized medical imaging and graphics* 79 (2020), 101684.

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

[3] Zengyang Che, Zheng Zhang, Yaping Wu, and Meiyun Wang. 2025. Disentangle and Then Fuse: A Cross-Modal Network for Synthesizing Gadolinium-Enhanced Brain MR Images. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).

[4] Nicolas Cordier, Hervé Delingette, Matthieu Lê, and Nicholas Ayache. 2016. Extended modality propagation: image synthesis of pathological cases. *IEEE transactions on medical imaging* 35, 12 (2016), 2598–2608.

[5] Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. 2019. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE transactions on medical imaging* 38, 10 (2019), 2375–2388.

[6] Xin Di, Long Peng, Peizhe Xia, Wenbo Li, Renjing Pei, Yang Cao, Yang Wang, and Zheng-Jun Zha. 2025. Qmambabsr: Burst image super-resolution with query state space model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23080–23090.

[7] Xin Fan, Xiaolin Wang, Jiaxin Gao, Jia Wang, Zhongxuan Luo, and Risheng Liu. 2024. Bi-level Learning of Task-Specific Decoders for Joint Registration and One-Shot Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11574–11583.

[8] ZhanFeng Feng, Long Peng, Xin Di, Yong Guo, Wenbo Li, Yulun Zhang, Renjing Pei, Yang Wang, Yang Cao, and Zheng-Jun Zha. [n. d.]. PMQ-VE: Progressive Multi-Frame Quantization for Video Enhancement. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[9] Jiaxin Gao, Xiaokun Liu, Risheng Liu, and Xin Fan. 2023. Learning adaptive hyper-guidance via proxy-based bilevel optimization for image enhancement. *The Visual Computer* 39, 4 (2023), 1471–1484.

[10] Jiaxin Gao and Yaohua Liu. 2024. Enhancing Images with Coupled Low-Resolution and Ultra-Dark Degradations: A Tri-level Learning Framework. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 8642–8651.

[11] Jiaxin Gao, Yaohua Liu, Ziyu Yue, Xin Fan, and Risheng Liu. 2024. Collaborative brightening and amplification of low-light imagery via bi-level adversarial learning. *Pattern Recognition* 154 (2024), 110558.

[12] Jiaxin Gao, Ziyu Yue, Yaohua Liu, Sihan Xie, Xin Fan, and Risheng Liu. 2024. A dual-stream-modulated learning framework for illuminating and super-resolving ultra-dark images. *IEEE transactions on neural networks and learning systems* 36, 4 (2024), 7500–7513.

[13] Yawen Huang, Ling Shao, and Alejandro F Frangi. 2017. Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6070–6079.

[14] Sehui Kim, Hyungjin Chung, Se Hie Park, Eui-Sang Chung, Kayoung Yi, and Jong Chul Ye. 2024. Fundus image enhancement through direct diffusion bridges. *IEEE Journal of Biomedical and Health Informatics* (2024).

[15] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. 2023. I2SB: image-to-image Schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*. 22042–22062.

[16] Risheng Liu, Jiaxin Gao, Xuan Liu, and Xin Fan. 2024. Learning with constraint learning: New perspective, solution strategy and various applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 7 (2024), 5026–5043.

[17] Yaohua Liu, Jiaxin Gao, Xuan Liu, Xianghao Jiao, Xin Fan, and Risheng Liu. 2024. Advancing generalized transfer attack with initialization derived bilevel optimization and dynamic sequence truncation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 1137–1145.

[18] Shaocong Mo, Ming Cai, Lanfen Lin, Ruofeng Tong, Qingqing Chen, Fang Wang, Hongjie Hu, Yutaro Iwamoto, Xian-Hua Han, and Yen-Wei Chen. 2021. Mutual information-based graph co-attention networks for multimodal prior-guided magnetic resonance imaging segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 5 (2021), 2512–2526.

[19] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, et al. 2018. MR and CT data with multiobserver delineations of organs in the pelvic area—Part of the Gold Atlas project. *Medical physics* 45, 3 (2018), 1295–1300.

[20] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. 2023. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging* 42, 12 (2023), 3524–3539.

[21] Yuwen Pan, Rui Sun, Wangkai Li, and Tianzhu Zhang. 2025. Exploring weather-aware aggregation and adaptation for semantic segmentation under adverse conditions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13952–13962.

[22] Yuwen Pan, Rui Sun, Yuan Wang, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. 2024. Purify Then Guide: A Bi-Directional Bridge Network for Open-Vocabulary Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[23] Yuwen Pan, Rui Sun, Yuan Wang, Tianzhu Zhang, and Yongdong Zhang. 2024. Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2031–2040.

[24] Long Peng, Yang Cao, Renjing Pei, Wenbo Li, Jiaming Guo, Xueyang Fu, Yang Wang, and Zheng-Jun Zha. 2024. Efficient real-world image super-resolution via adaptive directional gradient convolution. *arXiv preprint arXiv:2405.07023* (2024).

[25] Long Peng, Yang Cao, Yuejin Sun, and Yang Wang. 2024. Lightweight adaptive feature de-drifting for compressed image classification. *IEEE Transactions on Multimedia* 26 (2024), 6424–6436.

[26] Long Peng, Xin Di, Zhanfeng Feng, Wenbo Li, Renjing Pei, Yang Wang, Xueyang Fu, Yang Cao, and Zheng-Jun Zha. 2025. Directing mamba to complex textures: An efficient texture-aware state space model for image restoration. *arXiv preprint arXiv:2501.16583* (2025).

[27] Long Peng, Wenbo Li, Jiaming Guo, Xin Di, Haoze Sun, Yong Li, Renjing Pei, Yang Wang, Yang Cao, and Zheng-Jun Zha. 2024. Unveiling hidden details: A raw data-enhanced paradigm for real-world super-resolution. *arXiv preprint arXiv:2411.10798* (2024).

[28] Long Peng, Wenbo Li, Renjing Pei, Jingjing Ren, Jiaqi Xu, Yang Wang, Yang Cao, and Zheng-Jun Zha. [n. d.]. Towards Realistic Data Generation for Real-World Super-Resolution. In *The Thirteenth International Conference on Learning Representations*.

[29] Long Peng, Yang Wang, Xin Di, Xueyang Fu, Yang Cao, Zheng-Jun Zha, et al. 2025. Boosting image de-raining via central-surrounding synergistic convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6470–6478.

[30] Long Peng, Anran Wu, Wenbo Li, Peizhe Xia, Xueyuan Dai, Xinjie Zhang, Xin Di, Haoze Sun, Renjing Pei, Yang Wang, et al. 2025. Pixel to gaussian: Ultra-fast continuous super-resolution with 2d gaussian modeling. *arXiv preprint arXiv:2503.06617* (2025).

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[32] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. 2022. Solving Inverse Problems in Medical Imaging with Score-Based Generative Models. In *International Conference on Learning Representations*.

[33] Gijs Van Tulder and Marleen de Bruijne. 2015. Why does synthesized data improve multi-sequence classification?. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 531–538.

[34] Guanhua Wang, Enhao Gong, Suchandrima Banerjee, Dann Martin, Elizabeth Tong, Jay Choi, Huijun Chen, Max Wintermark, John M Pauly, and Greg Zaharchuk. 2020. Synthesize high-quality multi-contrast magnetic resonance imaging from multi-echo acquisition using multi-task deep generative model. *IEEE transactions on medical imaging* 39, 10 (2020), 3089–3099.

[35] Yang Wang, Long Peng, Liang Li, Yang Cao, and Zheng-Jun Zha. 2023. Decoupling-and-aggregating for image exposure correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18115–18124.

[36] Yulin Wang, Honglin Xiong, Kaicong Sun, Jiameng Liu, Xin Lin, Ziyi Chen, Yuanzhe He, Qian Wang, and Dinggang Shen. 2025. Unisyn: A generative foundation model for universal medical image synthesis across mri, ct and pet. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 673–682.

[37] ZhenZhou Wang. 2020. Automatic localization and segmentation of the ventricles in magnetic resonance images. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 2 (2020), 621–631.

[38] Zhijie Wang, Aiping Liu, Jinbao Wei, Qingguo Xie, Kongqiao Wang, and Xun Chen. 2024. DDC-Net: Dual-Domain Cascaded Network with POCS Prior for Fast MRI Reconstruction. *IEEE Sensors Journal* (2024).

[39] Zhijie Wang, Jinbao Wei, Gang Yang, Aiping Liu, Wei Wei, Bensheng Qiu, and Xun Chen. 2025. Dual-Domain Self-Consistency-Enhanced Deep Unfolding

Network for accelerated MRI reconstruction. *Computer Methods and Programs in Biomedicine* (2025), 108995.

[40] Jinbao Wei, Yuhang Chen, Zhijie Wang, Gang Yang, Shimin Tao, Jian Gao, Aiping Liu, and Xun Chen. 2025. Rethinking Diffusion Bridge Model with Dual Alignments for Medical Image Synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 1052–1061.

[41] Jinbao Wei, Zhijie Wang, Kongqiao Wang, Li Guo, Xueyang Fu, Ji Liu, and Xun Chen. 2023. Accurate MRI Reconstruction via Multi-Domain Recurrent Networks.. In *IJCAI*. 1524–1532.

[42] Jinbao Wei, Gang Yang, Zhijie Wang, Yu Liu, Aiping Liu, and Xun Chen. 2024. Misalignment-Resistant Deep Unfolding Network for multi-modal MRI super-resolution and reconstruction. *Knowledge-Based Systems* 296 (2024), 111866.

[43] Jinbao Wei, Gang Yang, Zhijie Wang, Shimin Tao, Aiping Liu, and Xun Chen. 2025. Degradation-Aware Prompted Transformer for Unified Medical Image Restoration. *IEEE Transactions on Image Processing* 34 (2025), 8583–8598.

[44] Jinbao Wei, Gang Yang, Wei Wei, Aiping Liu, and Xun Chen. 2025. Multi-contrast mri arbitrary-scale super-resolution via dynamic implicit network. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).

[45] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. 2025. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870* (2025).

[46] Qian Xie, Yusong Lin, Meiyun Wang, and Yaping Wu. 2024. Synthesis of gadolinium-enhanced glioma images on multisequence magnetic resonance images using contrastive learning. *Medical Physics* 51, 7 (2024), 4888–4897.

[47] Sihan Xie, Peiming Li, Jiaxin Gao, Ziyu Yue, Xin Fan, and Risheng Liu. 2024. Breaking the water dilemma: Transmission-guided bilevel adaptive learning for underwater imagery. *Neurocomputing* 596 (2024), 127909.

[48] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. 2018. Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 174–182.

[49] Ziyu Yue, Jiaxin Gao, and Zhixun Su. 2024. Unveiling details in the dark: Simultaneous brightening and zooming for low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6899–6907.

[50] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).

[51] Kaiwen Zheng, Guande He, Jianfei Chen, Fan Bao, and Jun Zhu. 2025. Diffusion bridge implicit models. In *The Thirteenth International Conference on Learning Representations*.

[52] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. 2024. Denoising Diffusion Bridge Models. In *The Twelfth International Conference on Learning Representations*.