TAIR: Text-Guided Adaptive Prompt Refinement for Coarse-to-Fine All-In-One Medical Image Restoration

Jiaqi Cui[1], Haocheng Xiong[1], Jize Han[2], Bo Liu[1], Xi Wu[3], Yan Wang[1(✉)]

[1] School of Computer Science, Sichuan University, Chengdu, China
wangyanscu@hotmail.com
[2] China Mobile (Suzhou) Software Technology Company Limited, Suzhou, China
[3] School of Computer Science, Chengdu University of Information Technology, China

Abstract. All-in-one medical image restoration (MedIR) aims to recover high-quality images from diverse degradations. However, existing methods predominantly rely on visual features and adopt a one-step mapping from low-quality inputs to high-quality outputs, which limits their ability to identify and adapt to heterogeneous degradations. To tackle these limitations, we propose TAIR, a novel **T**ext-guided **A**dapt**I**ve prompt **R**efinement framework for coarse-to-fine all-in-one MedIR. TAIR leverages dual-level textual prompts to guide restoration, where the task-level prompt specifies the restoration objective, indicating what to restore; and the instruction-level prompt directs how to restore via a coarse-to-fine strategy, with a coarse instruction recovering global structures and a fine instruction refining textural details. Based on this design, TAIR further constructs two composite prompts, i.e., the coarse composite prompt (CCP), which integrates degradation-related visual cues with task-level prompt and coarse instruction to recover global structures in the coarse phase; and the composite fine prompt (CFP), which combines task-level prompt with fine instruction and the coarsely restored image for detailed refinement in the fine phase. During restoration, an adaptive multi-prompt interaction (AMPI) module dynamically weights the contributions of each component in CCP and CFP for effective guidance. In addition, a text-driven quality-aware loss aligns restored images with high-quality textual descriptions while suppressing low-quality associations, enhancing overall fidelity and semantic consistency. Experiments across MRI, CT, and PET demonstrate that TAIR achieves competitive performance in single-task and all-in-one MedIR. Code is available at https://github.com/gluucose/TAIR.

Keywords: All-In-One, Medical Image Restoration, Multi-modal Learning.

1 Introduction

Medical images play a fundamental role in modern clinical practice [1, 2]. However, medical imaging is often degraded due to factors such as reduced radiation exposure time, insufficient radiation intensity, or patient-related factors, resulting in low-quality (LQ) images with substantially diminished diagnostic value [3-5]. Facing this challenge, medical image restoration (MedIR) has emerged as a key technique for trans-

forming high-quality (HQ) images from LQ inputs. Over the past few decades, MedIR methods have achieved notable success across various tasks, including MRI super-resolution [5, 6], CT denoising [7-8], and PET synthesis [9-11].

Although effective, these single-task models are inherently limited in broader applicability, as they are tailored to specific types of degradation and often experience substantial performance drops when applied to other MedIR tasks [12, 13]. In clinic, maintaining separate models for each task is inefficient and burdensome, and often infeasible on resource-constrained medical devices. These challenges highlight the urgent need for a single universal model capable of simultaneously handling multiple MedIR tasks. Recently, all-in-one image restoration methods [14, 15] have gained prominence in natural image processing by addressing various restoration tasks with a single model. For example, Li *et al.* [16] designed prompt-in-prompt which encodes degradation into learnable prompt embeddings for single-step restoration, without decoupling what to restore from how to restore. However, substantial modality gaps and varied degradation patterns in medical images hinder the direct application of these models to MedIR, leading to limited exploration of all-in-one solutions in the medical domain. To bridge this gap, Yang *et al.* [17] presented AMIR, a pioneering all-in-one MedIR method that employs a task routing strategy to allocate different tasks to distinct network paths for effective restoration. Following this, Yang *et al.* [18] designed a task-adaptive transformer that mitigates task interference and imbalance via adaptive weighting and loss balancing. Chen *et al.* [19] further developed a task-adaptive codebook based on latent diffusion to encode task-specific priors, boosting all-in-one performance.

Despite recent advances, existing MedIR approaches solely rely on visual features to infer task-relevant information. This visual-only paradigm is limited in two aspects. On the one hand, the visual features learned from LQ inputs may fail to capture sufficient context or even introduce ambiguous cues, making it difficult to accurately identify the corresponding restoration task among multiple possible tasks. On the other hand, current methods typically adopt a one-step restoration process that directly maps LQ images to high-quality outputs, which struggle to handle heterogeneous degradations in medical images and often produce suboptimal results. In contrast, text, serving as an easily accessible medium to encapsulate the semantic information, offers an explicit and intuitive way to convey the task objective and restoration process, complementing visual representations and improving adaptability across tasks. However, text-guided all-in-one MedIR remains unexplored.

To address the above limitations, in this paper, we propose TAIR, a **T**ext-guided **A**dapt**I**ve prompt **R**efinement framework for coarse-to-fine all-in-one MedIR. The core of TAIR lies in utilizing textual information to explicitly guide both the task objective (what to restore) and the restoration process (how to restore). Specifically, TAIR introduces dual-level textual prompts: (1) the task-level prompt specifies the restoration objective (MRI super-resolution, *etc.*), providing explicit task intent; and (2) the instruction-level prompt guides the restoration process through complementary instructions, with a coarse instruction recovering global structures and a fine instruction refining textural details. As one-step LQ-to-HQ mappings often fail to capture degraded details while ensuring global consistency, TAIR adopts a coarse-to-fine strategy based on the instruction-level prompt to progressively enhance the restoration quality. In the coarse

phase, the model involves the task-level prompt, the coarse instruction, and a degradation-related prompt derived from the low-quality input, forming a coarse composite prompt (CCP). An adaptive multi-prompt interaction (AMPI) module weights its contributions during restoration via weighted cross-attention. The coarse phase produces an intermediate result (i.e., coarsely restored image) with improved global structures, serving as input for the subsequent phase. In the fine phase, a degradation-free prompt extracted from the intermediate result, together with the fine instruction and task-level prompt, forms a composite fine prompt (CFP). Injected via AMPI, the CFP enables refinement of texture details. In addition, a text-driven quality-aware loss aligns restored images with high-quality textual descriptions while distancing them from low- and medium-quality descriptions, further improving overall visual fidelity.

The contributions of this paper can be summarized as follows. (1) We are the first to leverage textual guidance for all-in-one MedIR. By introducing dual-level textual prompts and restoring in a coarse-to-fine manner, the model not only understands what to restore by recognizing the task-level cues, but also how to restore according to the provided instructions. (2) We design an AMPI module to weigh prompt contributions and emphasize the most relevant guidance for more controlled restoration. In addition, a text-driven quality-aware loss drives restored images closer to high-quality textual descriptions, enhancing overall fidelity. (3) Extensive experiments show that our proposed TAIR achieves competitive performance in both single-task and all-in-one MedIR tasks across MRI, CT, and PET modalities.

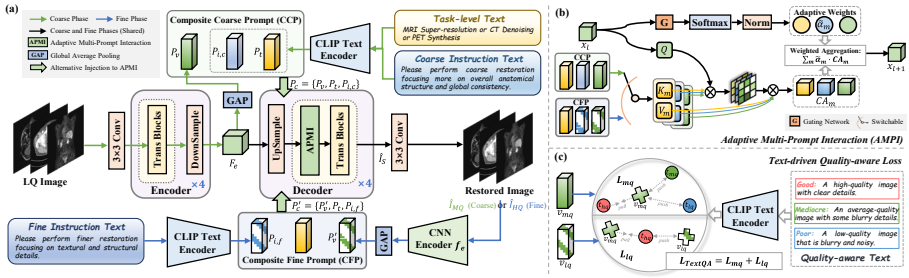

Fig. 1. (a) Overview of the proposed TAIR model. (b) The AMPI module that dynamically adjusts prompt contributions during restoration. (c) The text-driven quality-aware loss that leverages quality-aware textual prompts to further enhance the overall fidelity of restored images.

2 Methodology

2.1 Overview

The overall framework of TAIR is illustrated in Fig. 1 (a). Given a LQ input $I_{LQ} \in \mathbb{R}^{H\times W}$, a 3×3 convolutional layer first extracts shallow features I_S, which are then encoded into latent representations $F_e \in \mathbb{R}^{h\times w\times c}$ through a 4-level transformer-based encoder. Different from previous one-phase approaches that directly map LQ inputs to HQ outputs, TAIR performs an end-to-end coarse-to-fine restoration, which breaks down a challenging restoration process into more manageable sub-processes that sequentially focus on the recovery of global structures and fine texture details, thereby

enabling more stable and effective learning. Particularly, in the coarse phase, the coarse composite prompt (CCP) P_c is incorporated into the decoder via the adaptive multi-prompt interaction (AMPI) module, producing the restored shallow features $\hat{I}_S$. Symmetrically, a 3×3 convolutional layer then transforms $\hat{I}_S$ into a residual image $\hat{I}_R$, which is added to the original LQ image I_{LQ} to obtain the coarsely restored intermediate result $\hat{I}_{MQ} = \hat{I}_S + \hat{I}_R$. In the fine phase, $\hat{I}_{MQ}$ contributes to forming the composite fine prompt (CFP) P_c', which is injected into the decoder via AMPI to refine textural details, yielding the final predicted HQ images $\hat{I}_{HQ}$.

In TAIR, both coarse and fine restoration phases are supervised by L1 loss as:

$$\mathcal{L}_{C2F} = ||\hat{I}_{MQ} - I_{HQ}||_1(\text{Coarse Phase}) + ||\hat{I}_{HQ} - I_{HQ}||_1(\text{Fine Phase}). \quad (1)$$

In addition, a text-driven quality-aware loss $\mathcal{L}_{TextQA}$ guides restored images toward high-quality textual descriptions while suppressing low-quality associations, thereby improving the overall fidelity and consistency (detailed in Sec. 2.4). Therefore, the total objective function is formulated as a λ-weighted sum of the above two losses:

$$\mathcal{L}_{total} = \mathcal{L}_{C2F} + \lambda\mathcal{L}_{TextQA}. \quad (2)$$

2.2 Composite Coarse-to-Fine Prompt

To address the diverse degradations in all-in-one MedIR, we exploit the complementary strengths of textual and visual information by designing a composite coarse prompt (CCP) and a composite fine prompt (CFP). These prompts deliver sequential task-aware guidance for coarse-to-fine restoration, enabling the model to perceive what to restore (i.e., the specific task associated with the degraded input), and how to restore (i.e., first recovering global anatomical structures, then refining texture details).

Composite Coarse Prompt (CCP). The CCP guides the model in learning the degraded spatial context while perceiving global structures during the coarse restoration phase. Overall, CCP comprises dual-level textual prompts (i.e., task-level prompt and coarse instruction prompt), together with a degradation-related visual prompt, providing semantic and instructive guidance. Specifically, we first design the task-level text T_t that specifies the restoration task with the content such as "MRI Super-resolution", and coarse instruction text $T_{i,c}$ that offers phase-specific guidance, i.e., "Please perform coarse restoration focusing more on overall anatomical structure and global consistency." Then, these texts are processed through a frozen CLIP text encoder [20] to produce the corresponding prompts P_t and $P_{i,c} \in \mathbb{R}^{1\times d}$, where d signifies the dimension of the encoded prompts. In parallel, the latent visual features F_e are flattened, projected and pooled to generate the degradation-related visual prompt $P_v \in \mathbb{R}^{1\times d}$, which captures global degradation patterns. By concatenating all prompts, the CCP is formed as $P_c = \{P_v, P_t, P_{i,c}\}$. Injected into the transformer decoder via the AMPI, CCP interacts with visual features in the main decoding branch to produce the initial coarse restoration $\hat{I}_{MQ}$, exhibiting recovered structures and mitigated major degradations.

Composite Fine Prompt (CFP). Following the coarse restoration phase, the coarsely restored image $\hat{I}_{MQ}$is obtained under the guidance of the CCP. Given that $\hat{I}_{MQ}$ approximates a clean HQ image, it inherently involves degradation-free information that provides high-fidelity visual cues. Accordingly, the CFP P_c' exploits these cues from $\hat{I}_{MQ}$

to guide fine-grained enhancement during the fine restoration phase. Overall, the CFP comprises a task-level prompt, a fine instruction prompt, and a degradation-free visual prompt, which together enable precise and detail-oriented restoration. To obtain the prompt representation, $\hat{I}_{MQ}$ is encoded by a lightweight convolutional encoder f_e, followed by a linear projection to produce a compact degradation-free visual prompt $P_v' \in \mathbb{R}^{1\times d}$, which substitutes the degradation-related visual prompt P_v in the CCP. Meanwhile, the coarse instruction prompt $P_{i,c}$ is replaced by a fine instruction prompt $P_{i,f}$ with text content of "Please perform finer restoration focusing on textural and structural details.", whereas the task-level prompt P_t remains unchanged. Finally, the CFP is represented as $P_c' = \{P_v', P_t, P_{i,f}\}$. Injected into the transformer decoder through the AMPI module using the same mechanism as the CCP, the CFP further emphasizes local structural and textural details while preserving global anatomical consistency.

2.3 Adaptive Multi-Prompt Interaction (AMPI)

As shown in Fig. 1(b), we develop an AMPI module to effectively exploit multiple prompts in CCP and CFP for versatile image restoration. Each prompt conveys distinct semantic or structural information, whose importance to the restoration varies across images and tasks. As a result, static fusion strategies often fail to fully harness these heterogeneous cues. AMPI addresses this limitation by dynamically learning the relative impact of each prompt through its interactions with visual features in the decoder, allowing the model to emphasize the most important prompt guidance according to the specific degradation type and image content. Taking CCP as an example, let the visual features at the l-th decoder block be $x_l \in \mathbb{R}^{h_l\times w_l\times d_l}$, they are flattened into n_l ($n_l = h_l \times w_l$) patches to produce visual tokens $x_l' \in \mathbb{R}^{n_l\times d_l}$. Given the CCP prompt set as $P_c = \{P_v, P_t, P_{i,c}\}$, AMPI performs independent cross-attention between x_l' and each prompt component $m \in P_c$ as follows:

$$CA_m = \mathrm{Softmax}(QK_m^{\mathrm{T}}/\sqrt{d}) \cdot V_m, \tag{3}$$

where the query Q is a linear projection of x_l', and the key K_m and value V_m are sourced from the prompt component $m \in \{P_v, P_t, P_{i,c}\}$. This design enables each output $CA_m$to capture context-aware representations by aligning the visual query with the corresponding prompt-specific key-value space, achieving task-adaptive feature modulation that reflects both degradation characteristics and restoration intent. Then, a lightweight gating network f_θ predicts adaptive weights $\alpha_m = \mathrm{Softmax}(f_\theta(m))$, which are subsequently normalized across all prompt components to produce $\tilde{\alpha}_m$, dynamically weighing the contribution of each prompt. Note that, unlike previous prompt-based restoration methods [16], which inject prompts through static prompt–feature interactions, AMPI adaptively reweights heterogeneous prompts for each image and task, thereby prioritizing the most relevant guidance. The final representation for the next (l+1-th) decoder block is obtained via weighted aggregation of CA_m as:

$$x_{l+1} = \sum_{m\in\{P_v,P_t,P_{i,c}\}} \tilde{\alpha}_m \cdot CA_m. \tag{4}$$

A similar approach is applied to CFP. By adaptively integrating prompts into the restoration process, this module improves model capacity to generalize across diverse degradation types and restoration tasks.

2.4 Text-driven Quality-aware Loss

We devise the text-driven quality-aware loss to direct the restoration toward perceptually high-quality results by aligning restored visual features with textual quality descriptions. Using the CLIP text encoder, as shown in Fig. 1(c), we define three textual descriptions (i.e., good, mediocre, and poor), which correspond to three visual quality levels (i.e., high, medium, and low). Latent textual representations are extracted as t_{hq}, t_{mq} and t_{lq}. For visual features, we extract representations v_{mq} and v_{lq} for the restored and degraded images, respectively, where v_{mq} is obtained from the restored image $\hat{I}_{HQ}$ via a lightweight network f_e, and v_{lq} is derived by linearly projecting P_v. The loss contrasts visual and textual representations through two complementary contrastive objectives. For restored images, we encourage v_{mq} to be closer to the "good" text and farther from the lower-quality ones ("mediocre" and "poor") as:

$$\mathcal{L}_{mq} = -log \frac{\exp(sim(v_{mq}, t_{hq})/\tau)}{\sum_{k\in\{lq,mq,hq\}} \exp(sim(v_{mq}, t_k)/\tau)}. \quad (5)$$

For degraded images v_{lq}, as degraded visual features are unstable under distortions, the "good" semantic is used as a stable anchor to push away LQ visual features as:

$$\mathcal{L}_{lq} = -log \frac{\exp(sim(t_{hq}, v_{mq})/\tau)}{\sum_{k\in\{lq,mq\}} \exp(sim(t_{hq}, v_k)/\tau)}. \quad (6)$$

The final objective is $\mathcal{L}_{TextQA} = \mathcal{L}_{mq} + \mathcal{L}_{lq}$. This formulation semantically aligns restored images with high-quality textual concepts while effectively distancing them from degraded ones, thereby enhancing the overall perceptual quality.

3 Experiment

3.1 Dataset and Implementation

Dataset. (1) MRI Super-Resolution. We use the IXI MRI dataset [21] with 578 HQ T2 MRI scans. For each scan, the central 100 slices sized 256×256 are extracted, and LQ images are generated by 4× k-space downsampling. The dataset is split into 405/59/114 for training/validation/testing. **(2) CT Denoising.** From the NIH AAPM-Mayo Low-Dose CT dataset [22], paired HQ and LQ CT images of size 512×512 from 10 patients are used, with 8/1/1 patients for training/validation/testing. **(3) PET Synthesis.** A clinical dataset of 159 PET scans acquired using the PolarStar m660 system is involved. For each scan, we extract 2D slices sized 192×400. LQ images are generated by 12× list-mode dose reduction, and reconstructed using the OSEM method [23]. The dataset is split into 120/10/29 for training/validation/testing.

Implementation. For network architecture, the encoders and decoders utilize [3, 4, 4, 6] transformer-based blocks at each level. The input channel C is specified as 48. The weighting coefficient λ in Eq. (2) is set as 0.001. During training, we utilize patches cropped at 128×128 with a batch size of 8. In the all-in-one setting, our model is optimized using the Adam optimizer for 2e5 iterations, starting with an initial learning rate of 2e-4, gradually decreasing to 1e-6 using cosine annealing.

Table 1. Single task medical image restoration results.

Methods	MRI Super-resolution			Methods	CT Denoising			Methods	PET Synthesis		
	PSNR↑	SSIM↑	RMSE↓		PSNR↑	SSIM↑	RMSE↓		PSNR↑	SSIM↑	RMSE↓
SwinMR [24]	30.93	0.9253	32.7339	RDCNN [3]	33.19	0.9113	8.9427	DCNN [13]	36.27	0.9243	0.0954
F-UNet [6]	31.26	0.9314	31.5675	Eformer [26]	33.35	0.9175	8.8030	CWGAN [11]	36.62	0.9290	0.0910
MambaIR [25]	31.77	0.9369	29.8372	CTformer [8]	33.25	0.9134	8.8974	ARGAN [9]	36.73	0.9406	0.0902
Restormer [26]	31.85	0.9378	29.7005	DMamba [27]	33.53	0.9149	8.6115	DRMC [10]	36.00	0.9352	0.0998
TAIR	**32.03**	**0.9390**	**29.3218**	**TAIR**	**33.79**	**0.9194**	**8.3645**	**TAIR**	**37.36**	**0.9488**	**0.0867**

Table 2. All-in-one medical image restoration results.

Methods	MRI Super-resolution			CT Denoising			PET Synthesis			Average		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
ARGAN [9]	30.08	0.9083	35.7999	32.92	0.9111	9.2110	36.75	0.9389	0.0907	33.25	0.9194	15.0339
DMamba [28]	30.32	0.9091	34.6972	33.18	0.9115	8.9512	36.81	0.9367	0.0895	33.44	0.9191	14.5793
Restormer [26]	31.72	0.9362	30.0549	33.61	0.9177	8.5329	37.14	0.9473	0.0872	34.15	0.9337	12.8917
SFformer [29]	30.83	0.9164	33.1332	33.41	0.9160	8.7310	36.80	0.9369	0.0899	33.68	0.9231	13.9847
AirNet [14]	31.39	0.9316	31.1131	33.62	0.9176	8.5226	37.17	0.9451	0.0864	34.06	0.9314	13.2410
AdaIR [15]	31.35	0.9295	31.2693	33.58	0.9168	8.5654	37.25	0.9458	0.0857	34.06	0.9307	13.3068
AMIR [17]	32.03	0.9396	29.0988	33.70	0.9182	8.4520	37.12	0.9475	0.0876	34.28	0.9351	12.5461
DiffCode [19]	31.92	0.9376	29.3838	33.73	0.9193	8.4219	37.21	0.9476	0.0868	34.29	0.9348	12.6308
TAT [18]	32.10	0.9402	28.9145	33.80	0.9192	8.3642	37.28	0.9480	0.0856	34.39	0.9358	12.4548
TAIR	**32.11**	**0.9411**	**28.9045**	**33.88**	**0.9215**	**8.2177**	**37.33**	**0.9483**	**0.0855**	**34.44**	**0.9370**	**12.4026**

3.2 Comparative Experiment

For single-task MedIR, we compare TAIR against four leading methods for each task. For all-in-one MedIR, we compare TAIR with nine methods, including three task-specific leading methods (ARGAN [9], DMamba [28], and Restormer [26]), one general restoration method (SFformer [29]), and five leading all-in-one methods (AirNet [14], AdaIR [15], AMIR [17], DiffCode [19], and TAT [18]).

Single-Task Medical Image Restoration. Table 1 demonstrates that TAIR outperforms all single-task baselines. Although not specifically designed for single-task MedIR, its strong performance confirms the effectiveness and generalizability of the proposed text-guided adaptive prompt refinement strategies across individual tasks.

All-in-One Medical Image Restoration. Table 2 illustrates that TAIR surpasses all comparison methods, including advanced models such as DiffCode and TAT. With

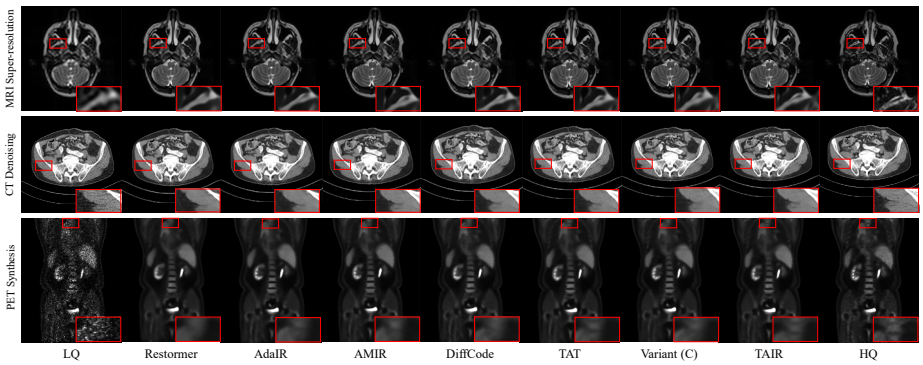

Fig. 2. Visual comparisons on all-in-one image restoration.

Table 3. Generalization of all-in-one models to out-of-distribution degradations.

Methods	PET Synthesis (20×)			Sparse-view CT Reconstruction			Average		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
TAT	38.79	0.8777	0.0163	25.22	0.7613	45.7313	32.01	0.8195	22.8738
TAIR	39.56	0.9148	0.0145	26.13	0.7960	41.9725	32.85	0.8554	20.9935

Table 4. Ablation study results of TAIR.

Methods		MRI Super-resolution			CT Denoising			PET Synthesis		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
TAIR		**32.11**	**0.9411**	**28.9045**	**33.88**	**0.9215**	**8.2177**	**37.33**	**0.9483**	**0.0855**
CCP	(A) w/o textual prompts	31.77	0.9368	29.8629	33.70	0.9173	8.4736	37.24	0.9475	0.0861
	(B) w/o degradation-related prompt	31.86	0.9372	29.5745	33.71	0.9187	8.4417	37.28	0.9475	0.0857
	(C) w/o all (remove coarse phase)	31.60	0.9349	30.4312	33.60	0.9162	8.5481	37.18	0.9452	0.0865
CFP	(D) w/o textual prompts	31.82	0.9371	29.7853	33.69	0.9181	8.4610	37.26	0.9476	0.0859
	(E) w/o degradation-free prompt	31.63	0.9355	30.2617	33.70	0.9182	8.4572	37.26	0.9472	0.0860
Others	(F) w/o AMPI	31.82	0.9374	29.7438	34.13	0.9350	7.8619	37.18	0.9469	0.0866
	(G) w/o $\mathcal{L}_{TextQA}$	31.74	0.9361	29.9843	33.70	0.9187	8.4558	37.25	0.9450	0.0858

only 24.4 M parameters and an end-to-end training pipeline, in contrast to TAT with 41.69 M parameters and DiffCode with a complex two-stage diffusion process, TAIR still achieves superior performance, demonstrating efficiency and effectiveness. This improvement arises from its smart integration of textual prompts, which involve task objectives and restoration strategies to complement visual representations, enhancing the adaptability across tasks. Moreover, TAIR (0.082/0.061/0.109 on MRI/CT/PET) achieves better LPIPS than second-best TAT (0.084/0.063/0.123), confirming its perceptual gains. The visual comparisons in Fig. 2 further demonstrate the superior ability of TAIR to restore both global structures and fine details.

Generalization to Out-of-Distribution Degradations. To further evaluate the generalization of TAIR, we directly apply the all-in-one pretrained model to two challenging scenarios: (1) an unseen degradation level of PET with a 20× low dose from another hospital, and (2) an unseen degradation type of sparse-view CT. In this experiment, we reuse training prompts, replacing only task-level text (e.g., "sparse-view CT reconstruction" for unseen degradation type). As presented in Table 3, TAIR consistently achieves superior and robust performance to unseen degradations.

3.3 Ablation Study

The ablation study in Table 4 validates the effectiveness of the key components, i.e., CCP, CFP, AMPI, and the loss $\mathcal{L}_{\text{TextQA}}$. Specifically, removing either textual or degradation-related visual prompts in CCP and CFP leads to a performance decline, highlighting their complementary roles in guiding accurate restoration. Notably, variant (C), which omits the entire coarse phase, experiences a substantial performance drop. Fig. 2 also illustrates the structural distortions in the obtained images, further validating the potential of the coarse-to-fine strategy. In addition, variant (F), which substitutes AMPI with standard cross-attention, yields reduced overall performance, verifying the benefits of using adaptive weights to modulate prompt contributions during restoration. Moreover, the removal of $\mathcal{L}_{\text{TextQA}}$ leads to a performance drop, suggesting the importance of quality-aware semantics in enhancing overall fidelity. More importantly, unlike conventional learnable embeddings, our textual guidance decouples what to re-

store from how to restore for phase-specific guidance, which also enables generalization (as shown in Table 3) to unseen tasks by simply rewriting the prompt.

4 Conclusion

In this paper, we propose TAIR, a text-guided adaptive prompt refinement framework for all-in-one MedIR. By leveraging textual prompts, TAIR effectively guides the model to learn the restoration objectives and perform progressive coarse-to-fine refinement, ensuring the quality of both global structures and fine details. The AMPI module then dynamically modulates multiple prompt contributions, emphasizing task-relevant guidance for controlled restoration. In addition, a text-driven quality-aware loss enhances fidelity by aligning restored images with high-quality textual descriptions. Extensive experiments demonstrate our feasibility and superiority.

Acknowledgments. This work is supported by National Natural Science Foundation of China (NSFC 62371325), Sichuan Science and Technology Program 2025NSFJQ0050, Sichuan Science and Technology Program 2024ZDZX0018.

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

References

1. Azizi, S., Culp, L., Freyberg, J., et al.: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat. Biomed. Eng. 7(6), 756-779 (2023)
2. Chen, W.: Clinical applications of PET in brain tumors. J. Nucl. Med. 48(9), 1468–1481 (2007)
3. Chen, H., Zhang, Y., Kalra, M.K., et al: Low-dose ct with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging 36(12), pp. 2524–2535 (2017)
4. Spuhler, K., Serrano-Sosa, M., Cattell, R., et al.: Full-count PET recovery from low-count image using a dilated convolutional neural network. Med. Phys. 47(10), 4928–4938 (2020)
5. Huang, J., Fang, Y., Wu, Y., et al.: Swin transformer for fast mri. Neurocomputing 493, 281–304 (2022)
6. Sun, H., Li, Y., Li, Z., et al.: Fourier convolution block with global receptive field for mri reconstruction. Med. Image Anal. 99, 103349 (2025)
7. Liang, T., Jin, Y., Li, Y., Wang, T.: Edcnn: Edge enhancement-based densely connected network with compound loss for low-dose ct denoising. In: IEEE International conference on signal processing (ICSP). vol. 1, pp. 193–198. IEEE (2020)
8. Wang, D., Fan, F., Wu, Z., et al.: Ctformer: convolutionfree token2token dilated vision transformer for low-dose ct denoising. Phys. Med. Biol. 68(6), 065012 (2023)
9. Cui, J., Zeng, X., Zeng, P., et al.: Mcad: Multi-modal conditioned adversarial diffusion model for high-quality pet image reconstruction. In: Linguraru, M.G., et al. (eds.) MICCAI 2024, vol. 15007, pp. 467–477. Springer, Cham (2024)
10. Yang, Z., Zhou, Y., Zhang, H., et al.: Drmc: A generalist model with dynamic routing for multi-center pet image synthesis. In: Greenspan, H., et al. (eds.) MICCAI 2023, vol. 14222, pp. 36–46. Springer, Cham (2023)

11. Zhou, L., Schaefferkoetter, J.D., Tham, I.W., et al.: Supervised learning with cyclegan for low-dose fdg pet image denoising. Med. Image Anal. 65, 101770 (2020)
12. Mu, X., Zhang, F., Li, M., et al.: Fibroblast activation imaging in rheumatoid arthritis: evaluating disease activity and treatment response using [18F] FAPI PET/CT. Eur. Radiol. 35, 6104–6114 (2025)
13. Chan, C., Zhou, J., Yang, L., et al.: Noise adaptive deep convolutional neural network for whole-body pet denoising. In: IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC). pp. 1–4. IEEE (2018)
14. Li, B., Liu, X., Hu, P., et al.: 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)
15. Y. Cui, S.W. Zamir, S. Khan, et al.: Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In: Proceedings of the International Conference on Learning Representations (2024)
16. Li, Z., Lei, Y., Ma, C., et al.: Prompt-in-prompt learning for universal image restoration. Available at SSRN 6115113 (2023)
17. Yang, Z., Chen, H., Qian, Z., et al.: All-in-one medical image restoration via task-adaptive routing. In: Linguraru, M.G. (eds.) MICCAI 2024, vol. 15007, pp. 67–77. Springer, Cham (2024)
18. Yang, Z., Zhang, J., Yi, Y., et al.: TAT: Task-Adaptive Transformer for All-in-One Medical Image Restoration. In: Gee, J.C. (eds.) MICCAI 2025, vol. 15975, pp. 565-575. Springer, Cham (2025)
19. Chen, H., Yang, Z., Hou, H., et al.: All-in-One Medical Image Restoration with Latent Diffusion-Enhanced Vector-Quantized Codebook Prior. In: Gee, J.C. (eds.) MICCAI 2025, vol. 15975, pp. 67-77. Springer, Cham (2025)
20. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, vol. 139, pp. 8748-8763 (2021)
21. Zhao, X., Zhang, Y., Zhang, T., Zou, X.: Channel splitting network for single mr image super-resolution. IEEE Trans. Image Process. 28(11), 5649–5662 (2019)
22. McCollough, C.H., Bartley, A.C., Carter, R.E., et al.: Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. Med. Phys. 44(10), e339–e352 (2017)
23. Hudson, H.M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging 13(4), 601–609 (1994)
24. Huang, J., Fang, Y., Wu, Y., et al.: Swin transformer for fast mri. Neurocomputing 493, 281–304 (2022)
25. Guo, H., Li, J., Dai, T., et al.: Mambair: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision. pp. 222–241. Springer (2025)
26. Zamir, S.W., Arora, A., Khan, S., et al.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
27. Luthra, A., Sulakhe, H., Mittal, T., et al.: Eformer: Edge enhancement based transformer for medical image denoising. arXiv preprint, arXiv:2109.08044 (2021)
28. Öztürk, Ş., Duran, O.C., Çukur, T.: Denomamba: A fused state-space model for low-dose ct denoising. arXiv preprint, arXiv:2409.13094 (2024)
29. Jiang, X., Zhang, X., Gao, N. and Deng, Y. When fast fourier transform meets transformer for image restoration. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds) ECCV 2024, vol. 15103, pp. 381-402. Springer, Cham (2024)