CT-DegradBench: A Physics-Informed Benchmark for CT Degradation Detection and Severity Estimation

Yousra Nabila Taifour[1], Marouane Tliba[1], Zuheng Ming[1], Marie Luong[1], Nour Aburaed[2], Aladine Chetouani[1], Gorkem Durak[3], Alessandro Bruno[4], Faouzi Alaya Cheikh[5], Habib Zaidi[6], Ulas Bagci[3], Azeddine Beghdadi[1]
[1]Université Sorbonne Paris Nord, [2]University of Dubai
[3]Northwestern University, [4]IULM University
[5]Norwegian University of Science and Technology, [6]University of Geneva
taifouryousra2017@gmail.com

Abstract

Computed tomography (CT) images are frequently degraded by acquisition artifacts, including noise, blur, streaking, aliasing, and metal artifacts. Yet CT enhancement is still largely evaluated using image quality metrics with limited perceptual and clinical validity, while existing datasets remain focused on isolated restoration tasks, hindering unified benchmarking across diverse degradation types. We present ***CT-DegradBench****, a dataset and benchmark for CT degradation detection and severity estimation under controlled single- and mixed-artifact settings. CT-DegradBench enables systematic evaluation across multiple degradation families and severity levels within a common experimental framework. We further propose* ***SeSpeCT*** *(****Se****mantic–****Spe****ctral* ***CT*** *degradation estimation), a framework that combines semantic priors from medical vision–language models with complementary frequency-domain cues for artifact analysis. SeSpeCT constructs a training-free semantic quality axis in the multimodal embedding space using radiology-informed text prompts, without task-specific fine-tuning, and combines it with spectral features that capture degradation-specific frequency patterns. The resulting representation enables joint prediction of artifact type and severity. Experimental results show that SeSpeCT consistently outperforms the evaluated baselines under both single- and mixed-degradation settings. The framework is available at https://github.com/yousranb/CT-DEGRADBENCH.*

1. Introduction

Computed tomography (CT) is a core medical imaging modality that provides high-resolution anatomical information essential for clinical diagnosis, treatment planning, and image-guided intervention [4, 14]. In routine practice, however, CT images are often degraded by artifacts introduced during acquisition and reconstruction, including noise, blur, streaking, aliasing, and metal artifacts [4, 11, 19, 28]. These degradations arise from diverse sources, such as dose reduction, sparse-view sampling, metallic implants, patient motion, and reconstruction parameter choices, and they frequently occur at different severity levels or in combination. Such effects compromise visual interpretation and negatively impact downstream automated analysis [18].

A large body of work has been devoted to CT image restoration, with deep learning methods now dominating low-dose denoising, sparse-view reconstruction, metal artifact reduction, and related enhancement tasks [4, 7, 11, 16, 17, 22, 23, 28]. Yet despite this progress, evaluation still depends heavily on image quality assessment (IQA) metrics whose perceptual and clinical reliability across diverse degradation conditions remains limited [4, 10, 11, 26]. At the same time, most existing datasets are designed for isolated restoration settings, such as low-dose denoising, sparse-view reconstruction, or metal artifact reduction, rather than for unified analysis across multiple degradation types and severity levels. As a result, they provide limited support for systematic benchmarking of degradation detection, severity estimation, and IQA reliability under controlled conditions.

This exposes an important gap: while CT restoration has been widely studied, there is still no unified benchmark for analyzing how different degradations and severity levels affect CT image quality, nor for assessing whether commonly used metrics and modern representation models remain sensitive and reliable across such conditions. This problem is especially relevant in practice, where degradations rarely appear in a single canonical form and often co-occur in compound scenarios.

To address this gap, we introduce **CT-DegradBench**, a dataset and benchmark for CT artifact detection and severity estimation under controlled single- and mixed-degradation settings. CT-DegradBench is built from paired reference and degraded CT images generated using physics-informed degradation models and covers five common acquisition-related degradation types across calibrated severity levels. It also includes radiologist-informed mixed-artifact settings to approximate clinically plausible compound degradations.

Using CT-DegradBench, we first conduct a systematic benchmark of classical and deep learning-based IQA metrics commonly used in CT enhancement evaluation, with a focus on degradation sensitivity and severity monotonicity. We then extend the benchmark to modern multimodal models by studying whether medical vision-language models (VLMs) encode quality-aware representations that correlate with degradation type and severity.

Building on this analysis, we propose **SeSpeCT** (**Se**mantic-**Spe**ctral **CT** degradation estimation), a framework for joint artifact detection and severity estimation. SeSpeCT constructs a training-free semantic quality axis in the multimodal embedding space of medical VLMs using radiology-informed text prompts, without task-specific fine-tuning, and combines this semantic signal with complementary spectral features that capture degradation-specific frequency structure. Together, these components provide a robust representation for CT degradation analysis under both isolated and mixed-artifact conditions. Our contributions are summarized as follows:

- We introduce **CT-DegradBench**, a controlled benchmark for CT degradation detection and severity estimation, covering five common degradation types, calibrated severity levels, and radiologist-informed mixed-artifact settings.
- We provide the first **systematic evaluation of CT degradation sensitivity** across classical IQA metrics, learned perceptual metrics, and medical vision-language models, analyzing their behavior under diverse single- and mixed-artifact conditions.
- We propose **SeSpeCT**, a semantic-spectral framework that combines a training-free prompt-derived quality axis from medical vision-language models with frequency-domain features for joint artifact type and severity prediction.

2. CT-DegradBench: Dataset and Benchmark

Existing CT datasets are largely designed for task-specific restoration problems, such as denoising or metal artifact reduction, and therefore do not support unified benchmarking across degradation types, severity levels, and mixed-artifact settings. We address this limitation with **CT-DegradBench**, a controlled multi-degradation CT dataset and benchmark for artifact detection and severity estimation. CT-DegradBench generates paired *reference–degraded* CT slices with explicit degradation labels, calibrated severity levels, and clinically plausible artifact mixtures, enabling systematic evaluation under a common experimental framework.

2.1. Problem setting and dataset overview

Let $x^{\mathrm{ref}} \in \mathbb{R}^{H\times W}$ denote a reference CT image. We generate a degraded counterpart x^{deg} by applying one or more degradation operators $\mathcal{D}_k(\cdot)$ with known severity:

$$x^{\mathrm{deg}} = \mathcal{D}_K \circ \cdots \circ \mathcal{D}_2 \circ \mathcal{D}_1 \left(x^{\mathrm{ref}}\right), \tag{1}$$

where each operator corresponds to a degradation type, namely noise, loss of sharpness, streaking, aliasing, or metal artifacts, and is associated with one of four severity levels (L0–L3), where L0 denotes the lowest severity level and L3 the highest. Each generated sample is accompanied by structured metadata specifying the degradation type(s), severity level(s), mixture composition, and generation parameters. This formulation enables controlled construction of both isolated and mixed degradations, making CT-DegradBench suitable for standardized benchmarking of degradation detection, severity estimation, and quality-aware representation learning.

2.2. CT imaging background

Computed tomography reconstructs cross-sectional images from X-ray projections acquired at multiple view angles. In clinical CT, reconstructed images are commonly represented in Hounsfield Units (HU), which encode attenuation relative to water. Given a reference CT slice x^{ref} in HU, we convert it to linear attenuation coefficients as

$$\mu = \mu_{\mathrm{w}} \left(1 + \frac{x^{\mathrm{ref}}}{1000}\right), \tag{2}$$

where μ_{w} denotes the attenuation coefficient of water [9].

The acquisition process is modeled in the projection domain. An ideal sinogram s is obtained by applying the parallel-beam Radon transform

$$s = \mathcal{R}(\mu). \tag{3}$$

Image reconstruction is then performed using filtered backprojection (FBP) [2]:

$$\hat{\mu} = \mathcal{R}^{-1}_{\mathrm{FBP}}(s), \qquad \hat{x} = 1000 \left(\frac{\hat{\mu}}{\mu_{\mathrm{w}}} - 1\right), \tag{4}$$

where $\hat{x}$ is the reconstructed CT image in HU.

This formulation provides a physics-grounded basis for degradation simulation. In CT-DegradBench, degradations are introduced prior to reconstruction, either directly in the projection domain or by modifying the attenuation map in the image domain (e.g., for metal artifacts), and are subsequently propagated to the image domain through filtered backprojection.

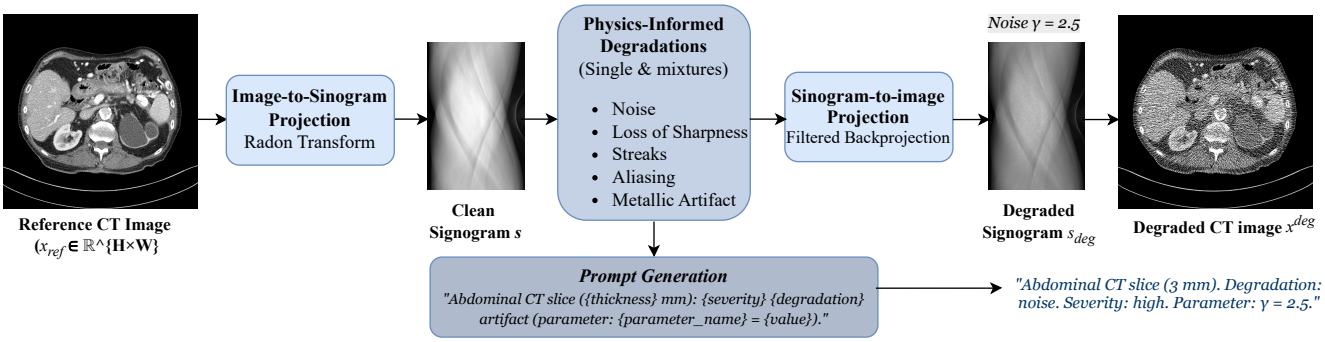

Figure 1. **CT-DegradBench generation pipeline.** A reference CT image is forward-projected to the sinogram domain, where physics-informed degradations are applied individually or as realistic mixtures with controlled severity. The degraded sinogram is reconstructed via filtered backprojection to obtain the final degraded CT image, while structured metadata are generated in parallel for prompt construction.

2.3. Degradation modeling and severity levels

To enable controlled benchmarking, we model five common CT degradations, noise, loss of sharpness, streaking, aliasing, and metal artifacts, each at four calibrated severity levels (L0–L3).

(1) Noise (mixed Poisson–Gaussian, severity γ). CT noise arises mainly from photon counting statistics and electronic detector noise in low-dose CT imaging. We simulate it in the projection domain using the mixed Poisson–Gaussian model of [29]. Given a clean sinogram s, photon counts are sampled as

$$K \sim \text{Poisson}\left(\alpha I_0 e^{-s}\right), \tag{5}$$

where I_0 is the incident photon intensity and α controls the effective dose level. Detector noise is modeled by

$$K' = K + \mathcal{N}(0, \sigma^2), \tag{6}$$

and the noisy sinogram is obtained via

$$s_{\text{noisy}} = -\log\left(\frac{K' + \delta}{\alpha I_0}\right), \tag{7}$$

where δ stabilizes the logarithm.

To generate controlled degradation levels for benchmarking, we introduce a residual noise scaling mechanism that enables calibrated control of noise severity while preserving noise characteristics consistent with the underlying CT acquisition physics. Specifically, the residual noise is scaled as

$$s^{\gamma}_{\text{noisy}} = s + \gamma\left(s_{\text{noisy}} - s\right), \qquad \gamma \in \{1,\ 2,\ 2.5,\ 4\}. \tag{8}$$

The degraded sinogram is then reconstructed using FBP (Eq. 4).

(2) Loss of sharpness / blur (severity σ). The dominant resolution loss in CT arises from detector aperture and detector response, which primarily affect spatial resolution along the detector direction rather than across projection angles. Accordingly, we blur the sinogram along the detector axis with a 1D Gaussian kernel:

$$s_{\text{blur}} = G_\sigma * s, \quad \sigma \in \{0.8, 1, 1.5, 2.5\}, \tag{9}$$

where $*$ denotes convolution and G_σ is applied along the detector dimension. The blurred sinogram is then reconstructed using FBP.

(3) Streaks (severity ΔL). In practice, streak artifacts often co-occur with noise and other degradations. To isolate their effect, we introduce structured outliers in the sinogram that mimic directional streaks observed in clinical CT images:

$$s_{\text{streak}} = s + \Delta L \cdot M, \qquad \Delta L \in \{0.25,\ 0.5,\ 1,\ 2\}, \tag{10}$$

where M is a binary mask selecting the affected regions. Reconstructing s_{streak} yields the streak-corrupted image.

(4) Aliasing artifacts (severity: number of views). Aliasing artifacts from sparse-view CT [7] are simulated by subsampling projection angles. Let Θ denote the full set of acquisition angles and $\Theta_N \subset \Theta$ a subset of N views. We form a sparse sinogram

$$s_N = \mathcal{R}_{\Theta_N}(\mu), \qquad N \in \{180,\ 90,\ 60,\ 45\}, \tag{11}$$

and reconstruct it using FBP. Fewer views produce stronger aliasing artifacts.

(5) Metal artifacts. Metal artifacts are simulated by inserting a metal region into the attenuation map, following [8]. Given a binary metal mask $m(\mathbf{x}) \in \{0, 1\}$ and metal attenuation coefficient μ_{metal}, we define

$$\mu_{\text{metal}}(\mathbf{x}) = (1 - m(\mathbf{x}))\,\mu(\mathbf{x}) + m(\mathbf{x})\,\mu_{\text{metal}}. \tag{12}$$

We then forward-project and reconstruct using Eq. 4, yielding artifacts similar to those observed around metallic implants. Severity is controlled by increasing the size of the metal region. Because metal artifacts are spatially localized, we also provide bounding-box annotations derived from the mask $m(\mathbf{x})$.

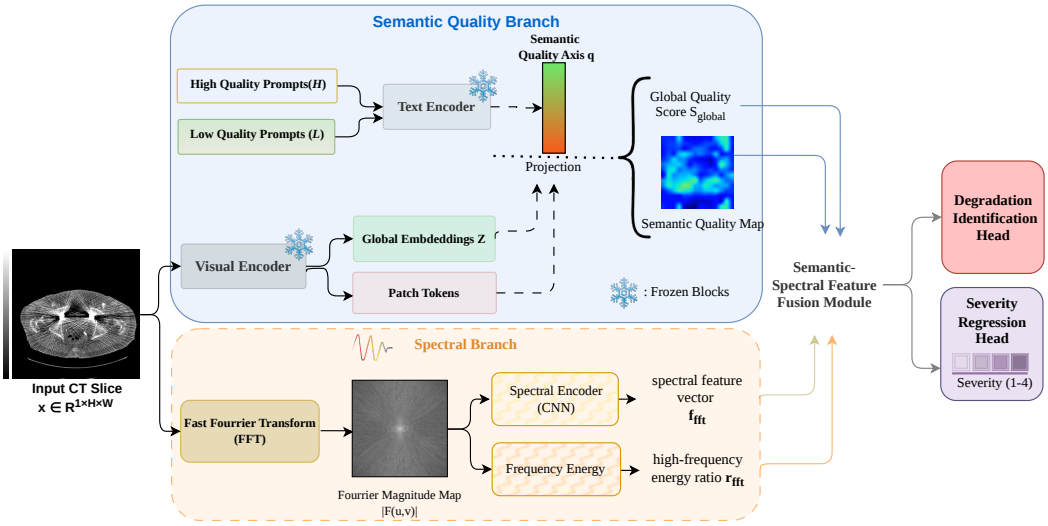

Figure 2. **Overview of the proposed SeSpeCT framework.** The model combines a semantic quality branch derived from a medical vision–language model with frequency-domain descriptors extracted from the Fourier spectrum. The fused representation is used to jointly predict degradation type and severity.

Mixed degradations. In clinical CT, multiple artifacts often co-occur due to interacting acquisition conditions, reconstruction processes, and patient-specific factors. To reflect this setting, CT-DegradBench includes mixed degradations in addition to isolated artifacts. We define five representative mixture configurations: *Blur + Noise*, *Streaks + Noise*, *Metal + Noise*, *Aliasing + Noise*, and *Metal + Blur + Noise*. These combinations capture common interactions between acquisition noise and other degradation mechanisms. Degradations are applied sequentially in an order that approximates their manifestation in reconstructed CT images; the corresponding clinical scenarios are summarized in the supplementary material (Table 8).

Each mixture is assigned a global severity level controlling the overall degradation strength. Component severities are sampled from neighboring levels around this global level to maintain comparable artifact magnitudes and avoid unrealistic combinations. We define the final mixture severity as the maximum severity among its constituent degradations.

2.4. Metadata Description

To support reproducible benchmarking, each CT-DegradBench sample is associated with structured metadata describing the degradation type(s), mixture order, severity level, and generation parameters. For metal artifacts, localization annotations are additionally provided. A structured natural-language prompt describing the degradation configuration is included to benchmark Vision–Language Models (VLMs), as illustrated in Figure 1.

3. Proposed Method

We propose **SeSpeCT**, a framework for CT degradation detection and severity estimation that combines semantic quality cues from a medical vision–language model with complementary Fourier-domain features. As shown in Fig. 2, the method consists of three parts: (1) a **Semantic Quality Branch** that derives quality-aware descriptors from a prompt-defined semantic axis, (2) a **Spectral Branch** that captures degradation-sensitive frequency patterns, and (3) a **fusion and prediction module** that jointly predicts degradation type and severity.

3.1. Semantic Quality Branch

Prompt-derived semantic quality axis. Let $\mathcal{H}$ and $\mathcal{L}$ denote two sets of radiology-informed prompts describing high-quality and degraded CT images, respectively. Using the frozen BioMedCLIP text encoder $f_{\text{text}}(\cdot)$ [31], we compute normalized prompt embeddings

$$e_h = \frac{f_{\text{text}}(h)}{\|f_{\text{text}}(h)\|}, \quad h \in \mathcal{H}, \qquad e_l = \frac{f_{\text{text}}(l)}{\|f_{\text{text}}(l)\|}, \quad l \in \mathcal{L}. \tag{13}$$

We then compute the corresponding prompt prototypes

$$\mu_H = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} e_h, \qquad \mu_L = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} e_l, \tag{14}$$

and define the semantic quality axis as

$$q = \frac{\mu_H - \mu_L}{\|\mu_H - \mu_L\|}. \tag{15}$$

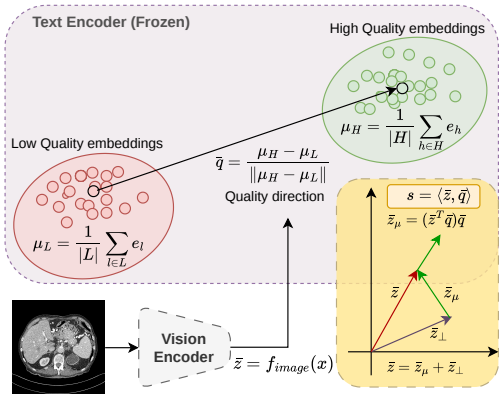

Figure 3. Semantic quality axis in the Vision-Language Model embedding space. High- and low-quality prompts produce embeddings $\bar{\mu}_H$ and $\bar{\mu}_L$, whose normalized difference defines the quality direction $\bar{q}$. Image embeddings $\bar{z}$ are projected onto this axis to obtain a quality score $s_{\text{global}} = \bar{z}^\top \bar{q}$.

As illustrated in Fig. 3, q defines a direction in the shared vision–language embedding space that captures the transition from degraded to high-quality CT appearance.

Global and local semantic descriptors. Given an input CT slice x, the frozen BioMedCLIP image encoder produces a normalized global embedding

$$z = \frac{f_{\text{image}}(x)}{\|f_{\text{image}}(x)\|}. \tag{16}$$

Projecting z onto the semantic quality axis yields a global semantic quality score

$$s_{\text{global}} = z^\top q. \tag{17}$$

To capture localized degradations, we additionally project Vision Transformer patch embeddings $T = \{t_1, \ldots, t_N\}$ onto the same axis:

$$s_i = \hat{t}_i^\top q, \qquad \hat{t}_i = \frac{t_i}{\|t_i\|}. \tag{18}$$

These scores form a semantic quality map over the image, as illustrated in Fig. 2. We summarize them using a pooling operator $\phi(\cdot)$ to obtain a fixed-length local descriptor $f_{\text{loc}} = \phi(s_1, \ldots, s_N)$. The final semantic representation is $f_{\text{sem}} = [\, s_{\text{global}} \| f_{\text{loc}} \,]$. In practice, $\phi(\cdot)$ is implemented using simple statistics over patch responses, such as their mean, maximum, and standard deviation.

3.2. Spectral Branch

Several CT degradations are more clearly expressed in the frequency domain than in the image domain. In particular, aliasing and streak artifacts produce structured spectral distortions, while noise alters the high-frequency energy distribution (See Fig. 4). We therefore complement the semantic branch with frequency-domain features.

Given an input image x, we compute its discrete Fourier transform $F(u, v) = \mathcal{F}(x)(u, v)$ and form the log-magnitude spectrum $M(u, v) = \log(1 + |F(u, v)|)$. This spectrum is then processed by a lightweight convolutional encoder $g_{\text{fft}}(\cdot)$ to obtain a spectral feature vector h_{fft}.

To capture the relative concentration of high-frequency content, we also compute a high-frequency energy ratio

$$r_{\text{fft}} = \frac{\sum_{(u,v)\in\Omega_h} |F(u, v)|^2}{\sum_{u,v} |F(u, v)|^2}, \tag{19}$$

where Ω_h denotes a predefined high-frequency region. The final spectral descriptor is formed by concatenating the learned spectral feature and the energy ratio, i.e., $f_{\text{spec}} = [\, h_{\text{fft}} \| r_{\text{fft}} \,]$.

3.3. Fusion and Multi-Task Prediction

The semantic descriptor f_{sem} and spectral descriptor f_{spec} provide complementary information: the former captures quality-aware semantic structure, while the latter captures degradation-specific frequency patterns. We concatenate them and project them into a shared embedding space using a multilayer perceptron:

$$z_f = \text{MLP}\big([\, f_{\text{sem}} \| f_{\text{spec}} \,]\big), \qquad z_f \in \mathbb{R}^d. \tag{20}$$

From the fused embedding z_f, we predict both degradation category and severity. The classification head outputs

$$o = W_{\text{cls}} z_f + b_{\text{cls}}, \tag{21}$$

where $o \in \mathbb{R}^C$ are the logits over C degradation classes. The predicted class is

$$\hat{y} = \arg\max_c \text{softmax}(o)_c. \tag{22}$$

The severity head predicts a scalar severity score

$$\hat{s} = w_{\text{reg}}^\top z_f + b_{\text{reg}}. \tag{23}$$

3.4. Training Objective

We train the model with a multi-task objective that combines degradation classification, severity regression, ordinal ranking, and supervised contrastive learning:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{24}$$

Here, $\mathcal{L}_{\text{cls}}$ is the standard cross-entropy loss for degradation classification, and $\mathcal{L}_{\text{reg}} = \text{SmoothL1}(\hat{s}, s)$ is the regression loss for severity prediction, where s denotes the ground-truth severity label.

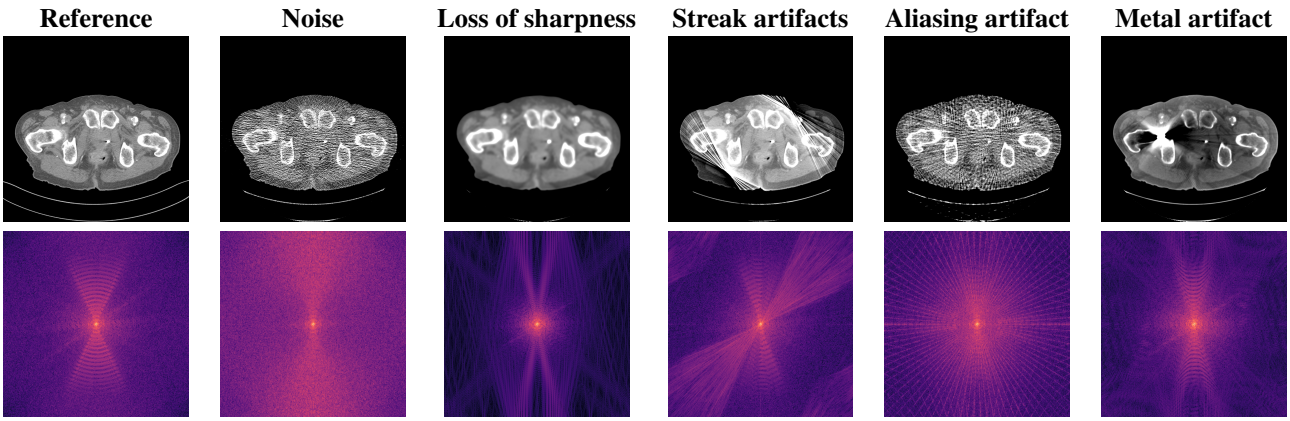

Figure 4. Visualization of mid-severity degradations in the image domain (top) and Fourier domain (bottom). The distinct spectral patterns motivate the use of frequency-domain features as complementary cues for degradation detection and severity estimation.

To preserve the ordinal structure of severity levels, we introduce a pairwise ranking loss over sample pairs (i, j) such that $s_i > s_j$:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max\big(0, m-(\hat{s}_i-\hat{s}_j)\big), \mathcal{P} = \{(i,j) \mid s_i > s_j\}. \tag{25}$$

To further structure the fused embedding space, we apply a supervised contrastive loss on z_f, where samples sharing the same degradation class and severity level are treated as positives:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{V}|} \sum_{i\in\mathcal{V}} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(z_{f,i}^{\top} z_{f,p}/\tau)}{\sum_{a\neq i} \exp(z_{f,i}^{\top} z_{f,a}/\tau)}. \tag{26}$$

We set $\lambda_{\text{rank}} = 0.3$ and $\lambda_{\text{con}} = 0.05$. The classification and regression terms provide the main supervision, while the ranking and contrastive terms improve ordinal consistency and embedding structure.

4. Experiments and Results

4.1. Experimental Setup

Dataset and protocol. We construct CT-DegradBench from the Mayo 2016 Low-Dose CT Grand Challenge dataset [13], which provides high-quality clinical CT scans together with acquisition parameters required for physics-informed degradation simulation. We select the high-dose CT acquisitions (200 mAs, 3 mm slice thickness, sharp reconstruction kernel) as reference images and generate degraded counterparts using the degradation pipeline described in Section 2. In total, the benchmark contains 2,378 reference slices of size 512×512 from 10 patients, from which we generate 9,512 degraded samples spanning multiple degradation types and severity levels.

Unless otherwise stated, benchmark analyses of IQA metrics and frozen VLM representations are conducted on the full dataset to study their behavior across degradation types and severity levels. For evaluation of the proposed method, we additionally report results on a held-out test split comprising 30% of the dataset in order to assess generalization to unseen samples.

Benchmarked IQA metrics and VLMs. We first evaluate widely used full-reference CT image quality assessment (IQA) metrics, including PSNR, SSIM [24], VIF [20], LPIPS [30], and DISTS [5], which are commonly used to assess CT enhancement methods [4, 10, 11]. Their sensitivity to degradation severity is quantified using both **Spearman** rank correlation and **Pearson** correlation, together with severity-wise metric trends. We also benchmark representative vision–language models, namely OpenCLIP [3], MedCLIP [25], BioMedCLIP [31], and Merlin [1]. For each model, we measure the cosine embedding drift between reference and degraded CT images and report its Spearman and Pearson correlation with degradation severity.

Finally, we evaluate the proposed method, **SeSpeCT**, on two tasks: degradation classification and severity estimation. Classification performance is reported using Accuracy and F1 score, while severity prediction is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Quadratic Weighted Kappa (QWK), and Spearman/Pearson correlation with ground-truth severity.

4.2. Quantitative Results

4.2.1. Classical IQA metrics

Table 1 and Table 2 summarize the correlation between degradation severity and widely used full-reference IQA metrics on the full dataset and the held-out test set, respectively. Detailed severity-wise statistics for the full

Single distortions (ρ / r)						Mixtures (ρ / r)					
Setting	**PSNR ↑**	**SSIM ↑**	**VIF ↑**	**LPIPS ↓**	**DISTS ↓**	**Setting**	**PSNR ↑**	**SSIM ↑**	**VIF ↑**	**LPIPS ↓**	**DISTS ↓**
S1_noise	-0.6636/-0.6691	-0.7095/-0.7087	-0.6712/-0.6396	0.6463/0.6475	0.6979/0.6756	M1_b+n	-0.4894/-0.4924	-0.5429/-0.5420	-0.6273/-0.6655	0.5330/0.5360	0.5670/0.5940
S2_blur	-0.2837/-0.2490	-0.7936/-0.7747	-0.8926/-0.8953	0.8301/0.8231	0.9107/0.9181	M2_s+n	-0.6129/-0.6150	-0.6284/-0.6283	-0.5475/-0.5781	0.5533/0.5544	0.5539/0.5743
S3_streak	-0.6883/-0.6729	-0.9345/-0.8719	-0.7093/-0.6921	0.9447/0.9405	0.9068/0.8984	M3_m+n	-0.8055/-0.8024	-0.8074/-0.8114	-0.7244/-0.7641	0.7001/0.7191	0.7250/0.7666
S4_aliasing	-0.5356/-0.5272	-0.9518/-0.9521	-0.9548/-0.9379	0.9600/0.9495	0.9548/0.9303	M4_a+n	-0.5437/-0.5499	-0.6877/-0.6834	-0.6203/-0.6583	0.6052/0.6102	0.5986/0.6209
S5_metal	-0.6413/-0.6340	-0.5317/-0.4794	-0.2461/-0.2490	0.0431/0.0497	0.2929/0.2861	M5_m+b+n	-0.8062/-0.7941	-0.8139/-0.8127	-0.7941/-0.8246	0.7478/0.7620	0.7702/0.8043
Mean	-0.5625/-0.5504	**-0.7842/-0.7574**	-0.6948/-0.6828	0.6848/0.6821	0.7526/0.7417	**Mean**	-0.6515/-0.6508	**-0.6961/-0.6956**	-0.6627/-0.6981	0.6279/0.6363	0.6429/0.6720

Entries are reported as ρ/r (Spearman/Pearson). Positive values indicate the metric increases with severity; negative values indicate it decreases with severity.
Setting keywords: S1-S5 = single degradations; M1-M5 = mixtures; L0-L3 = severity levels (increasing with L).
Mixture shorthand: b = blur, n = noise, s = streaks, a = aliasing, m = metal.

Table 1. Spearman rank correlation (ρ) and Pearson correlation (r) results between objective quality metrics and degradation severity levels. (Full Dataset)

Single distortions (ρ / r)						Mixtures (ρ / r)					
Setting	**PSNR ↑**	**SSIM ↑**	**VIF ↑**	**LPIPS ↓**	**DISTS ↓**	**Setting**	**PSNR ↑**	**SSIM ↑**	**VIF ↑**	**LPIPS ↓**	**DISTS ↓**
S1_noise	-0.6644/-0.6743	-0.7470/-0.7411	-0.6873/-0.6644	0.7120/0.7175	0.7321/0.7255	M1_b+n	-0.5221/-0.5254	-0.6095/-0.5978	-0.7084/-0.6673	0.6319/0.6169	0.6726/0.6362
S2_blur	-0.3108/-0.2307	-0.8631/-0.8600	-0.9126/-0.9209	0.8335/0.8316	0.9144/0.9202	M2_s+n	-0.6245/-0.6287	-0.6538/-0.6613	-0.6057/-0.5824	0.6095/0.6129	0.5982/0.5916
S3_streak	-0.7252/-0.6910	-0.9449/-0.8697	-0.7453/-0.7256	0.9427/0.9366	0.9271/0.9182	M3_m+n	-0.8035/-0.8054	-0.8258/-0.8300	-0.7592/-0.7223	0.7661/0.7518	0.7656/0.7376
S4_aliasing	-0.4856/-0.5009	-0.9432/-0.9386	-0.9660/-0.9541	0.9632/0.9501	0.9415/0.9209	M4_a+n	-0.5438/-0.5497	-0.7243/-0.7093	-0.6638/-0.6343	0.6852/0.6774	0.6441/0.6267
S5_metal	-0.6261/-0.6401	-0.4833/-0.4559	-0.2847/-0.2660	0.0460/0.0489	0.3169/0.3318	M5_m+b+n	-0.8161/-0.8058	-0.8362/-0.8378	-0.8282/-0.7998	0.8136/0.7955	0.8362/0.7993
Mean	-0.5624/-0.5474	**-0.7963/-0.7731**	-0.7192/-0.7062	0.6995/0.6969	0.7664/0.7633	**Mean**	-0.6628/-0.6630	**-0.7299/-0.7272**	-0.7131/-0.6812	0.7013/0.6909	0.7033/0.6783

Entries are reported as ρ/r (Spearman/Pearson). Positive values indicate the metric increases with severity; negative values indicate it decreases with severity.
Setting keywords: S1-S5 = single degradations; M1-M5 = mixtures.

Table 2. Spearman rank correlation (ρ) and Pearson correlation (r) results between objective quality metrics and degradation severity levels (Test-Set; Patients: L506, L192, L310).

Single distortions (ρ / r)						Mixtures (ρ / r)					
Setting	**OpenCLIP [3]**	**MedCLIP [25]**	**BioMedCLIP [31]**	**Merlin [1]**	**Ours**	**Setting**	**OpenCLIP [3]**	**MedCLIP [25]**	**BioMedCLIP [31]**	**Merlin [1]**	**Ours**
S1_noise	0.611/0.538	0.345/0.276	0.682/0.582	0.621/0.484	**0.798/0.791**	M1_b+n	0.543/0.493	0.236/0.216	0.663/0.582	0.437/0.291	**0.709/0.685**
S2_blur	0.649/0.596	0.533/0.486	0.885/0.796	0.169/0.191	**0.887/0.869**	M2_s+n	0.485/0.420	0.301/0.283	0.657/0.615	0.495/0.363	**0.668/0.657**
S3_streak	0.769/0.617	0.595/0.551	0.893/0.858	0.399/0.418	**0.926/0.894**	M3_m+n	0.653/0.655	0.228/0.215	0.542/0.518	0.679/0.605	**0.745/0.737**
S4_aliasing	0.839/0.759	0.738/0.640	0.912/0.855	0.604/0.543	**0.854/0.846**	M4_a+n	0.595/0.551	0.255/0.206	0.692/0.600	0.438/0.361	**0.668/0.658**
S5_metal	0.114/0.073	0.037/0.041	0.114/0.108	0.107/0.090	**0.435/0.422**	M5_m+b+n	0.691/0.697	0.227/0.217	0.606/0.585	0.677/0.596	**0.758/0.756**
Mean	0.635/0.588	0.330/0.313	0.665/0.610	0.452/0.394	**0.780/0.764**	**Mean**	0.635/0.588	0.330/0.313	0.665/0.610	0.452/0.394	**0.710/0.699**

Entries are Spearman/Pearson correlations.
Positive values indicate increasing score with increasing severity.

Table 3. Correlation between predicted severity and ground-truth degradation level for frozen VLM baselines and the proposed method under single and mixed degradation settings. (Test Set; patients : L506, L192, L310)

dataset are provided in the supplementary material (Table 10). Overall, the results show that these metrics are not equally sensitive to all CT degradations. PSNR is most responsive to noise, whereas perceptual and structural metrics such as VIF and DISTS better capture blur severity. Structured degradations, including streaking and aliasing, exhibit clearer monotonic trends across most metrics. In contrast, metal artifacts yield substantially weaker correlations, indicating that global quality metrics are less sensitive to spatially localized degradations. This is particularly important for CT enhancement evaluation, where commonly used metrics may fail to reflect restoration quality when degradations are localized rather than globally distributed.

4.2.2. VLM embedding sensitivity to degradation

Table 3 reports the correlation between degradation severity and cosine embedding drift in frozen vision–language models on the test dataset. Detailed severity-wise drift statistics for the full dataset are provided in the supplementary material (Tables 11 and 12). Across degradations, OpenCLIP and BioMedCLIP show the strongest monotonic relationship between embedding drift and severity. In particular, blur, streaking, and aliasing produce consistent representation shifts that increase with degradation strength.

By contrast, MedCLIP and Merlin show weaker severity correlations for several degradations, especially for metal artifacts and mixtures. More broadly, the VLM trends are consistent with those observed for classical IQA metrics: degradations that exhibit stronger monotonicity in structural or perceptual measures also tend to induce larger shifts in multimodal embedding space.

4.2.3. Performance of SeSpeCT

Table 3 reports severity correlation on the held-out test set for SeSpeCT and frozen VLM baselines. In contrast to Table 11, which analyzes representation drift on the full

Model	B_{sem}	B_{FFT}	L_{reg}	L_{rank}	L_{con}	Degradation classification Acc↑	Degradation classification F1↑	Severity estimation MAE↓	Severity estimation RMSE↓	Severity estimation QWK↑
w/o B_{sem}	✗	✓	✓	✓	✓	0.151	0.061	1.494	1.858	0.000
w/o B_{FFT}	✓	✗	✓	✓	✓	0.698	0.684	0.696	0.855	0.683
w/o L_{reg}	✓	✓	✗	✓	✓	0.599	0.579	1.492	1.783	0.030
w/o L_{rank}	✓	✓	✓	✗	✓	0.609	0.578	1.332	1.547	0.383
w/o L_{con}	✓	✓	✓	✓	✗	0.695	0.686	0.966	1.173	0.571
Proposed	✓	✓	✓	✓	✓	**0.783**	**0.776**	**0.632**	**0.794**	**0.738**

B_{sem}: semantic quality branch. B_{FFT}: frequency branch. L_{reg}: regression loss. L_{rank}: ranking loss. L_{con}: contrastive loss. ↑ higher is better; ↓ lower is better.

Table 4. **Ablation study of SeSpeCT.** We remove one branch or loss term at a time and report degradation classification (Accuracy, F1) and severity estimation (MAE, RMSE, QWK) performance.

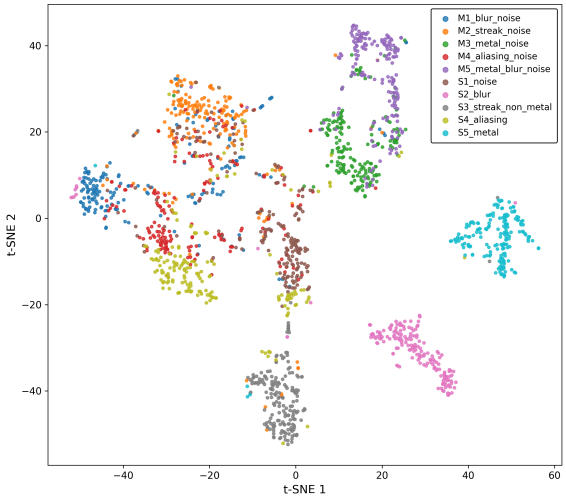

Figure 5. **t-SNE visualization colored by degradation type.** Samples form distinct clusters corresponding to different degradation categories, indicating that the learned representation captures degradation-specific characteristics.

dataset, this experiment evaluates predictive severity estimation on unseen samples. SeSpeCT is trained for joint degradation analysis using the semantic–spectral representation described in Section 3.

SeSpeCT achieves the highest correlation across both single and mixed settings, reaching a mean Spearman correlation of 0.780 for single degradations and 0.710 for mixtures, outperforming OpenCLIP (0.635) and BioMedCLIP (0.665). Gains are particularly pronounced for challenging cases such as metal artifacts, where correlations improve from $\rho < 0.12$ to $\rho = 0.435$.

These results indicate that combining semantic and spectral features improves sensitivity to both localized and structured degradations. Furthermore, Figure 5 shows well-separated clusters across degradation types, confirming that the learned representation captures degradation-specific characteristics. Overall, SeSpeCT provides a more reliable and monotonic representation of degradation severity than frozen VLM features.

4.2.4. Ablation study

Table 4 presents an ablation study of SeSpeCT. Removing the semantic branch causes the largest drop in performance, reducing classification accuracy from 0.783 to 0.151 and collapsing severity estimation performance, which highlights the importance of the prompt-derived semantic quality representation. Removing the spectral branch also leads to a substantial degradation, confirming that Fourier-domain cues provide complementary information beyond the semantic branch alone. Among the loss terms, the regression loss is essential for meaningful severity prediction, while the ranking loss improves ordinal consistency across severity levels. The supervised contrastive term further improves both classification and severity estimation by structuring the fused embedding space.

5. Conclusion

This work addressed CT degradation analysis under a unified benchmark and method setting. We introduced **CT-DegradBench**, a physics-informed benchmark for controlled evaluation of CT degradation type and severity under both isolated and mixed settings. Our experiments showed that commonly used IQA metrics and frozen vision–language model embeddings do not consistently reflect degradation severity, particularly for localized and compound artifacts. We therefore proposed **SeSpeCT**, a semantic–spectral framework that combines a prompt-derived semantic quality axis from medical vision–language models with frequency-domain descriptors for joint degradation detection and severity estimation. Together, CT-DegradBench and SeSpeCT will support more reliable quality assessment and evaluation protocols for future CT restoration and analysis pipelines.

Acknowledgments

This work was partially supported by the NIH grants R01-HL171376 and U01-CA268808.

References

[1] Louis Blankemeier et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, 2024. 6, 7, 5

[2] David J. Brenner and Eric J. Hall. Computed tomography—an increasing source of radiation exposure. *New England Journal of Medicine*, 357(22):2277–2284, 2007. 2

[3] Mehdi Cherti et al. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 6, 7, 5

[4] Clement David-Olawade et al. Ai-driven advances in low-dose imaging and enhancement—a review. *Diagnostics*, 15 (6):689, 2025. 1, 6

[5] Keyan Ding et al. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020. 6

[6] Nir Goren, Thomas Dowrick, James Avery, and David Holder. Uclh stroke eit dataset - radiology data (ct), 2017. 4

[7] Bing Guan et al. Generative modeling in sinogram domain for sparse-view ct reconstruction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 8(2):195–207, 2023. 1, 3

[8] Eri Haneda et al. Aapm ct metal artifact reduction grand challenge. *Medical Physics*, 52(10):e70050, 2025. 3, 4

[9] J. H. Hubbell and S. M. Seltzer. Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients, 2004. Version 1.4, National Institute of Standards and Technology, Gaithersburg, MD. 2

[10] Wonjin Kim et al. A systematic review of deep learning-based denoising for low-dose computed tomography from a perceptual quality perspective. *Biomedical Engineering Letters*, 14(6):1153–1173, 2024. 1, 6

[11] Yiming Lei et al. Ct image denoising and deblurring with deep learning: current status and perspectives. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 8(2): 153–172, 2023. 1, 6

[12] Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1):109, 2021. 4

[13] Cynthia McCollough. Tu-fg-207a-04: overview of the low dose ct grand challenge. *Medical physics*, 43(6Part35):3759–3760, 2016. 6, 4

[14] Cynthia H McCollough, Masahiro Jinzaki, and Hatem Alkadhi. Advances in ct technology and clinical applications—introductory editorial, 2025. 1

[15] Taylor R Moen, Baiyu Chen, David R Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G Fletcher, and Cynthia H McCollough. Low-dose ct image and projection dataset. *Medical physics*, 48(2):902–911, 2021. 4

[16] Taifour Yousra Nabila, Azeddine Beghdadi, Marie Luong, Zuheng Ming, Habib Zaidi, and Faouzi Alaya Cheikh. D-perceptct: Deep perceptual enhancement for low-dose ct images. In *European Workshop on Visual Information Processing (EUVIP)*, pages 1–6, 2025. 1

[17] Rabeya Tus Sadia et al. Ct image denoising methods for image quality improvement and radiation dose reduction. *Journal of Applied Clinical Medical Physics*, 25(2):e14270, 2024. 1

[18] Yazdan Salimi, Zahra Mansouri, Chang Sun, Amirhossein Sanaat, Mohammadhossein Yazdanpanah, Hossein Shooli, René Nkoulou, Sana Boudabbous, and Habib Zaidi. Deep learning-based segmentation of ultra-low-dose ct images using an optimized nnu-net model. *La Radiologia Medica*, pages 1–17, 2025. 1

[19] Mark Selles et al. Advances in metal artifact reduction in ct images: A review of traditional and novel metal artifact reduction techniques. *European Journal of Radiology*, 170: 111276, 2024. 1

[20] Hamid R. Sheikh and Alan C. Bovik. Image information and visual quality. *IEEE TIP*, 15(2):430–444, 2006. 6

[21] Emil Y Sidky and Xiaochuan Pan. Report on the aapm deep-learning sparse-view ct grand challenge. *Medical physics*, 49 (8):4935–4943, 2022. 4

[22] Farid Tasharofi et al. Find-net: Fourier-integrated network with dictionary kernels for metal artifact reduction. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 192–201, 2025. 1

[23] Tao Wang, Wenjun Xia, Jingfeng Lu, and Yi Zhang. A review of deep learning ct reconstruction from incomplete projection data. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 8(2):138–152, 2023. 1

[24] Zhou Wang et al. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[25] Zifeng Wang et al. Medclip: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 3876, 2022. 6, 7, 5

[26] Siyi Xun, Qiaoyu Li, Xiaohong Liu, et al. Charting the path forward: Ct image quality assessment - an in-depth review. *Journal of King Saud University Computer and Information Sciences*, 37(5):92, 2025. 1

[27] Xin Yi and Paul Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31(5):655–669, 2018. 4

[28] Sungho Yun et al. Tmaa-net: Tensor-domain multi-planal anti-aliasing network for sparse-view ct image reconstruction. *Physics in Medicine & Biology*, 69(22):225012, 2024. 1

[29] Dong Zeng et al. A simple low-dose x-ray ct simulation from high-dose scan. *IEEE Transactions on Nuclear Science*, 62 (5):2226–2233, 2015. 3

[30] Richard Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[31] Sheng Zhang et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. 4, 6, 7, 5