

Research on a γ -Photon Flow-Field Image Denoising Algorithm Based on the SFGD Model

JianTang Liu

Nanjing University of Aeronautics and Astronautics

YuLu Song

Nanjing University of Aeronautics and Astronautics

LiYing Hou

Nanjing University of Aeronautics and Astronautics

Hui Xiao

xiaohui@nuaa.edu.cn

Nanjing University of Aeronautics and Astronautics

Article

Keywords: photon imaging, SFGD model, SCGM, TAFM, image denoising

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10701211/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Research on a γ -Photon Flow-Field Image Denoising Algorithm Based on the SFGD Model

Liu JianTang¹, Song YuLu¹, Hou LiYing^{2,3} and Xiao Hui^{2,3,*}

1 College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, Nanjing 211106, China

2 College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, Nanjing 211106, China

3 Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen 518110, China

* Correspondence: xiaohui@nuaa.edu.cn

Abstract

High-energy γ photons have strong penetration and are electrically neutral, making γ -photon tomographic imaging suitable for non-contact monitoring of industrial flow fields. However, limited photon counts, detection noise, and insufficient projection data can blur boundaries and disrupt local structures. Existing diffusion-based denoising approaches may fail to preserve structural priors adequately or adapt frequency information across timesteps, leading to contour distortion and over-smoothing. We therefore propose a Structure-Frequency Guided Diffusion (SFGD) model for γ -photon flow-field image denoising. Its Structure-Cross Guidance Module (SCGM) uses a multi-scale structure stream and cross-attention to reinforce edge and local morphological information, while its Time-Adaptive Frequency Modulation (TAFM) module adjusts frequency-domain enhancement according to the diffusion timestep to balance noise suppression and detail recovery. On the independent test dataset, SFGD achieved a peak signal-to-noise ratio (PSNR) of 28.325 dB and a structural similarity index (SSIM) of 0.926, exceeding the baseline conditional diffusion model by 5.499 dB and 0.153, respectively. These results suggest that combining persistent structural guidance with timestep-adaptive frequency modulation improves structural preservation and detail restoration in γ -photon flow-field images.

Keywords: photon imaging; SFGD model; SCGM; TAFM; image denoising

Introduction

When a positron produced by the β^+ decay of a radionuclide annihilates with a surrounding electron, a pair of γ photons is emitted in opposite directions, each with an energy of 511 keV. The trajectories of each photon pair define a line of response (LOR), and each coincidence event defines an LOR^[1]. In practical reconstruction, coincidence events acquired by γ -photon imaging can either be stored event by event as list-mode data or binned according to LOR positions to form sinogram projection data. This study adopts the latter procedure: coincidence events are grouped by LOR to form sinograms, a system matrix is used to describe the contribution of each pixel in a slice to each projection line, and γ -photon flow-field slice images are then obtained using an iterative image reconstruction algorithm^[2]. Because γ photons have high penetration capability and are electrically

neutral, the detection results are relatively insensitive to temperature and electromagnetic fields, making this technique suitable for non-contact flow-field monitoring in complex industrial environments^[3]. Nevertheless, source activity, acquisition time, detection, and reconstruction constraints often leave the sinogram projection data incomplete. Effective structures in the reconstructed images are therefore susceptible to noise and artifacts, which reduce boundary discernibility and impair local structural continuity, ultimately affecting subsequent physical analysis^[4]. Image denoising is thus a critical preprocessing step for γ -photon flow-field images.

To address this problem, researchers have introduced various deep-learning denoising methods into flow-field image processing. Discriminative convolutional neural network methods, such as DnCNN and RIDNet^{[5][6]}, learn a deterministic end-to-end mapping from noisy images to clean images. Although they have achieved substantial progress in general image-denoising tasks, a single forward pass cannot adequately represent the uncertainty of the denoising process when confronted with the complex mixed noise in γ -photon flow-field images. Their outputs therefore tend to be oversmoothed, with pronounced loss of flow-field details. Denoising diffusion probabilistic models (DDPMs) use a progressive, iterative probabilistic denoising process to model the complex relationship between noise and image structure more fully, showing greater potential for image restoration. SR3, proposed by Saharia et al., directly concatenates a degraded image as a condition and extends diffusion models to conditional image-restoration tasks^[7]. Subsequent studies have continued to improve conditional fusion and frequency-domain processing, promoting diffusion-model applications in complex-noise scenarios such as medical and remote-sensing imaging.

Despite these advances, directly applying conditional diffusion models to γ -photon flow-field images remains challenging. Limited coincidence counts and incomplete projection data weaken vortex boundaries and other local structures. In a conventional conditional diffusion model, structural information from the degraded condition image is injected mainly by channel concatenation and can be diluted during repeated downsampling. Moreover, fixed frequency processing does not reflect the changing noise level across reverse-diffusion timesteps: aggressive high-frequency enhancement may amplify noise early in sampling, whereas insufficient enhancement may oversmooth edges later. To address these limitations, we develop a structure- and frequency-guided diffusion model (SFGD).

The main contributions are as follows:

(1) A Structure-Cross Guidance Module (SCGM) is introduced to maintain structural conditioning throughout the reverse-denoising network. A multi-scale structure stream extracts morphological priors from the degraded condition image, while window-based cross-attention and redesigned skip connections propagate these priors from the encoder to the decoder. This design strengthens the preservation of principal flow-field contours, boundary locations, and local structural continuity.

(2) A Time-Adaptive Frequency Modulation (TAFM) module is designed to account for the variation in noise level across reverse-diffusion timesteps. Initialized using power spectral density statistics from the training data, TAFM combines the diffusion timestep with the cumulative signal-retention coefficient to regulate the range and gain of frequency-domain modulation. It therefore emphasizes noise suppression during the early sampling stages and progressively enhances edges and local details as the underlying structure is recovered.

(3) A unified optimization and evaluation framework is established for the proposed model.

Noise-prediction, structural-consistency, and frequency-consistency objectives are integrated through uncertainty-based adaptive weighting to constrain restoration in the noise, spatial, and frequency domains. Comparisons with five representative denoising models and controlled ablation experiments are conducted using the same phantom-level training, validation, and independent test split, allowing the individual and complementary effects of SCGM and TAFM to be quantified.

Related work

(1) CNN-based image denoising

Convolutional neural networks (CNNs) progressively extract edges, textures, and high-level semantic features from images through hierarchical combinations of convolutional layers, pooling layers, and activation functions^[8]. A convolutional layer uses learnable kernels to perform weighted operations over local regions. As network depth increases, the receptive field expands, enabling image structures to be represented at different scales. Owing to their strong feature-extraction and nonlinear-mapping capabilities, CNNs have become an important technical foundation for image denoising^[9].

CNN-based image denoising research has primarily focused on residual learning, feature enhancement, and computational efficiency. DnCNN predicts a noise residual and removes it from the input image, reducing the difficulty of directly learning a clean-image mapping and advancing the application of deep CNNs to image denoising^[5]. Building on this idea, RIDNet uses hierarchical residual connections and feature attention to improve the modeling of complex noise and interchannel dependencies^[6]. BRDNet expands feature representation and the receptive field through parallel subnetworks, batch renormalization, and dilated convolution^[10]. MWDCNN combines dynamic convolution with wavelet transforms to jointly exploit spatial structures and frequency information^[11]. NAFNet simplifies nonlinear activations and network structure to reduce computation and parameter count while maintaining restoration performance^[12]. Collectively, these methods have improved CNN denoising in terms of feature representation, detail restoration, and model efficiency.

Nevertheless, these methods still learn an essentially deterministic end-to-end mapping from noisy images to clean images. Limited-count statistics, insufficient projection data, and iterative reconstruction make the noise in γ -photon flow-field images particularly complex^{[1][4]}. Under these conditions, a single forward pass cannot adequately describe uncertainty in the denoising process and may oversmooth local structures such as vortex edges while suppressing noise. A generative modeling framework with iterative inference is therefore needed to characterize the relationship between complex noise and image structure more fully, which motivates the use of denoising diffusion probabilistic models.

(2) Denoising diffusion probabilistic models

Denoising diffusion probabilistic models (DDPMs) establish a probabilistic generative process through progressive noise addition and iterative denoising, and can learn the conditional distribution from degraded images to clean images under the guidance of conditional information^[13]. Unlike the single deterministic mapping used by CNNs, a diffusion model decomposes a complex image-restoration process into multiple progressive denoising steps, providing a generative modeling framework for image restoration in complex-noise environments.

Sohl-Dickstein et al. introduced diffusion probabilistic models in 2015. Their method

progressively destroys the data distribution through a forward process and then learns a reverse process to restore its structure, although early models had low sampling efficiency^[14]. Ho et al. subsequently proposed DDPM, simplifying the training objective through a forward process that gradually injects Gaussian noise and a reverse process that predicts the added noise. This formulation substantially improved generation quality and training stability^[15]. To address the large number of sampling steps required by DDPM, Song et al. proposed DDIM, which reduces sampling steps through a non-Markovian process without changing the training objective^[16]. For conditional image restoration, SR3 takes a degraded image as a condition and feeds it into the network together with the noisy image at the current timestep, demonstrating the effectiveness of conditional diffusion models for image-restoration tasks^[7]. DDRM further uses pretrained diffusion models to provide a unified solution to linear inverse problems such as super-resolution, deblurring, and image inpainting^[17]. These studies have advanced diffusion models from unconditional generation toward conditional image restoration.

A diffusion model comprises a forward noising process and a reverse denoising process. In the forward process, Gaussian noise is gradually added to a clean image x_0 according to a predefined noise schedule. Let β_t denote the noise variance at timestep t , and define:

$$\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s \quad (1)$$

The noisy image at any timestep can then be expressed directly as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon, \quad \varepsilon : N(0, I) \quad (2)$$

where x_t is the noisy image at timestep t , ε is standard Gaussian noise, and $\bar{\alpha}_t$ denotes the cumulative retention of the original image signal. As the timestep increases, $\bar{\alpha}_t$ gradually decreases and the noise component in the image correspondingly increases. During the reverse process, the model predicts the noise from the noisy image x_t at timestep t , the degraded condition image I_{blur} , and timestep t , and progressively samples a clean image.

The denoising network in a diffusion model is typically implemented using a U-Net architecture^[18]. U-Net consists of a symmetric encoder-decoder structure and skip connections. The encoder progressively downsamples feature maps to extract multi-scale semantic features, whereas the decoder gradually restores spatial resolution through upsampling. Skip connections concatenate features from each encoder level with the corresponding decoder level, preventing low-level details from being lost during downsampling. Because the model must perceive both global structures and local details at different timesteps t , the multi-scale feature-fusion mechanism of U-Net naturally suits diffusion-based denoising. In addition, timestep embedding maps the scalar t to a high-dimensional vector through sinusoidal positional encoding and injects it into every U-Net level, allowing the model to adapt its output to the current denoising stage^[15]. However, a standard conditional diffusion model mainly relies on spatial-domain convolution and multi-scale feature propagation, so its treatment of different frequency components is generally implicit. In γ -photon flow-field images, high-frequency regions contain both random noise and physical details such as vortex edges, and the requirements for high-frequency suppression and preservation vary across diffusion timesteps. It is therefore necessary to analyze image-energy distributions in the frequency domain and clarify how different frequency components relate to noise and structural details.

(3) Fourier transform

The Fourier transform is a fundamental tool in signal and image processing. Its central idea is to decompose a signal in the time or spatial domain into a superposition of sinusoidal components at different frequencies. For a discrete image signal, the two-dimensional discrete Fourier transform maps a spatial-domain image $f(x, y)$ to a frequency-domain representation $F(u, v)$, defined as:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cdot e^{-i2\pi(\frac{ux}{M} + \frac{vy}{N})} \quad (3)$$

where M and N denote the image height and width, respectively; x and y are spatial coordinates; and u and v are frequency coordinates.

In the frequency domain, low-frequency components mainly represent overall brightness, contours, and large-scale structures produced by gradual grayscale variations, whereas high-frequency components mainly correspond to edges, textures, and regions with abrupt grayscale changes. For γ -photon flow-field images, low-frequency information describes the overall flow distribution and principal contours. High-frequency information contains not only random noise but also physically meaningful fine structures, including turbulent structures and vortex edges. Therefore, applying the same processing strategy to the full frequency range, or simply suppressing all high-frequency components, can remove important structures together with noise and lead to blurred details and excessive smoothing^[19].

To process different frequency components selectively, researchers have progressively introduced frequency-domain transforms into deep networks. MWCNN and SDWNet use wavelet decomposition for downsampling and parallel reconstruction, respectively, to reduce information loss during sampling and restore high-frequency details^{[20][21]}. FcaNet and frequency-aware dynamic high-pass filtering further improve the perception of specific frequency components through the discrete cosine transform or dynamic filtering^{[22][23]}. Building on these approaches, WINNet, SFNet, and FFTformer extend fixed frequency decomposition to learnable or content-adaptive frequency selection, enabling networks to preserve, suppress, or enhance frequency components according to the input features^{[24][25][26]}. These studies demonstrate that frequency-domain information provides a complementary representation for structure preservation in image denoising beyond spatial-domain features.

However, existing frequency-domain methods are regulated mainly by fixed transforms or current input content and generally do not explicitly incorporate the diffusion timestep and its corresponding noise level. At the beginning of reverse sampling, noise dominates the image, and excessive high-frequency enhancement may amplify that noise. As sampling proceeds and the principal structure gradually recovers, the need to preserve high-frequency details such as vortex edges increases. Frequency processing in diffusion denoising should therefore not use a fixed strategy; it should instead adapt dynamically to the noise level at different timesteps.

Methods

(1) Overview of SFGD

γ -photon flow-field images are reconstructed from a limited number of coincidence events through data reorganization and iterative reconstruction. They are jointly affected by statistical fluctuations in coincidence events, noise introduced during detection, and reconstruction artifacts caused by

insufficient projection data, and therefore often exhibit blurred boundaries and discontinuous local structures. When a basic conditional diffusion model is used for denoising, the condition image is usually introduced through channel concatenation. Its structural prior can progressively weaken during multilayer feature propagation, making it difficult to maintain constraints on the contour and spatial position of the principal flow structure. In addition, the reverse sampling process applies a uniform strategy to the entire frequency range without distinguishing differences in noise levels among timesteps. Consequently, turbulent structures, vortex edges, and other fine textures may be oversmoothed together with the noise, resulting in structural distortion.

To address these problems, this paper proposes a structure-frequency guided diffusion model for γ -photon flow-field image denoising. The overall SFGD framework is built on the theory of conditional diffusion probabilistic models. It introduces a degraded image as conditional information at every timestep of the reverse denoising process, thereby learning an end-to-end mapping from noisy flow-field images to clean ones. The forward process follows the standard diffusion procedure, progressively injecting Gaussian noise into a clean image until it degenerates into a standard normal distribution. In the reverse process, the model receives the current noisy image and the degraded condition image at each timestep t . Unlike conventional channel concatenation, a structure-guidance mechanism adaptively fuses the structural prior from the degraded condition image into the diffusion branch to predict the noise applied at the current step. By optimizing single-step noise prediction at arbitrary timesteps during training, the model learns to handle different noise levels and can progressively restore clean flow-field images during inference. To address structural distortion caused by uniform full-band processing, SFGD introduces two key modifications to the U-Net backbone of a conventional conditional diffusion model. On the encoder side, SCGM extracts structural priors from the degraded condition image through a multi-scale structure stream and adaptively fuses them into the feature representation of the diffusion branch through layer-wise windowed cross-attention. The original U-Net skip connections are also redesigned to prevent structural information from being diluted during transmission from the encoder to the decoder. On the decoder side, TAFM initializes its frequency-domain modulation parameters using power spectral density statistics from the training set. It then combines the diffusion timestep t and the corresponding cumulative signal-retention coefficient $\bar{\alpha}_t$ to apply time-varying frequency-domain modulation to the upsampled decoder feature maps. These parameters are further optimized adaptively during end-to-end training, enabling the range and strength of high-frequency enhancement to vary across diffusion stages and thereby dynamically balancing noise suppression against detail preservation. The complete SFGD model is optimized end to end using a joint loss comprising noise-prediction, structural-consistency, and frequency-consistency losses. A task-uncertainty-based adaptive weighting strategy combines the three terms, jointly supervising the restoration of flow-field structural details in the noise, image, and frequency domains. The SFGD architecture is shown in Fig. 1.

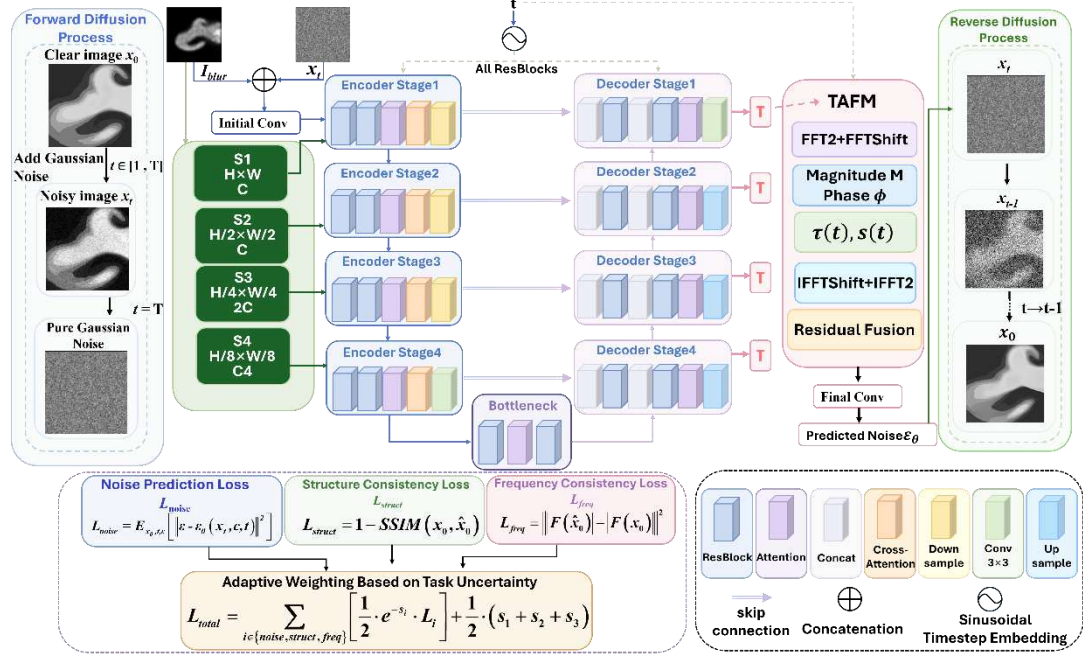


Fig. 1. Overall architecture of the SFGD model

(2) Structure-cross guidance module (SCGM)

Although degraded γ -photon flow-field images lose high-frequency details, they retain low-frequency structural information such as the overall flow contour and boundary orientation. This structural prior is important for guiding the diffusion model to recover the correct flow-field topology during reverse sampling. A standard conditional diffusion model, however, merely concatenates the condition image with the noisy image before feeding them into the network. This early-fusion strategy progressively dilutes structural information after multiple convolutional downsampling stages, making clear structural constraints difficult to preserve in deep features. Accordingly, this paper proposes the Structure-Cross Guidance Module (SCGM). Through the coordinated design of multi-scale structure-stream extraction, layer-wise windowed cross-attention fusion, and redesigned skip connections, SCGM enables the structural prior in the degraded condition image to guide noise-branch feature learning continuously and explicitly at every U-Net encoder level. The SCGM architecture is shown in Fig. 2.

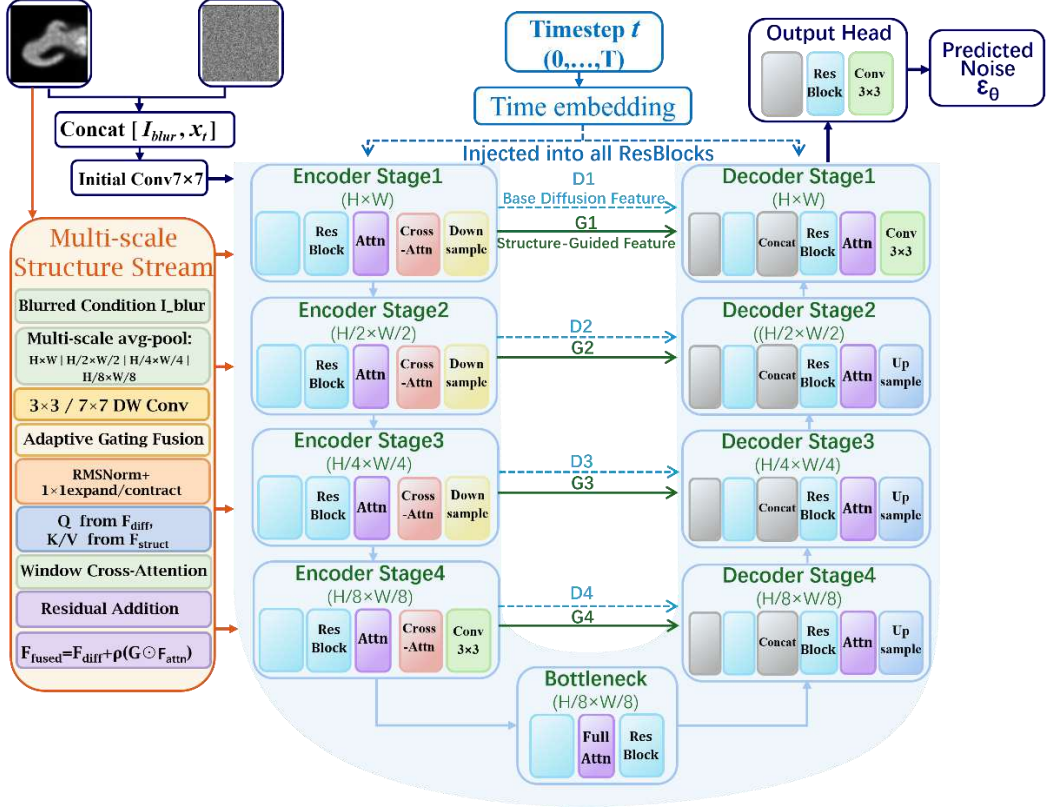


Fig. 2. Architecture of SCGM

Multi-scale structure stream

The multi-scale structure stream extracts a set of structural feature priors from the degraded condition image, with spatial resolutions aligned to those of the diffusion-branch feature maps. Specifically, the input degraded image is first downsampled according to the current network level to match the spatial resolution of the corresponding diffusion-branch encoder feature map exactly, thereby avoiding interpolation loss during cross-scale fusion. The downsampled image is then projected into an initial feature representation through a basic convolutional layer.

To capture the complex geometric characteristics of the flow field comprehensively, the structure stream constructs two parallel feature-extraction paths with different receptive fields. Motivated by the comparison of depthwise convolution kernels with different sizes reported by Liu et al.^[27], the two branches employ 3×3 and 7×7 depthwise separable convolutions, respectively, to provide complementary receptive fields. The 3×3 branch focuses on local edge textures and fine contours, whereas the 7×7 branch captures structural orientations and topological cues over a larger region. A spatially gated fusion mechanism then concatenates the features to generate spatially adaptive weights and dynamically combines the two paths. This design allows the network to adjust its relative dependence on local details and global contours according to the structural complexity at each spatial location, producing a more expressive fused representation.

The fused features are subsequently passed through a lightweight residual refinement block. The block first applies RMSNorm, expands the number of channels fourfold using a 1×1 convolution, applies GELU activation, projects the features back to the original channel dimension using another 1×1 convolution, and performs RMSNorm again^[28]. The normalized result is then added residually to the features obtained from the initial 3×3 convolutional projection. A final 3×3

convolution produces the structural feature F_{struct} . The detailed structure is shown in Fig. 3.

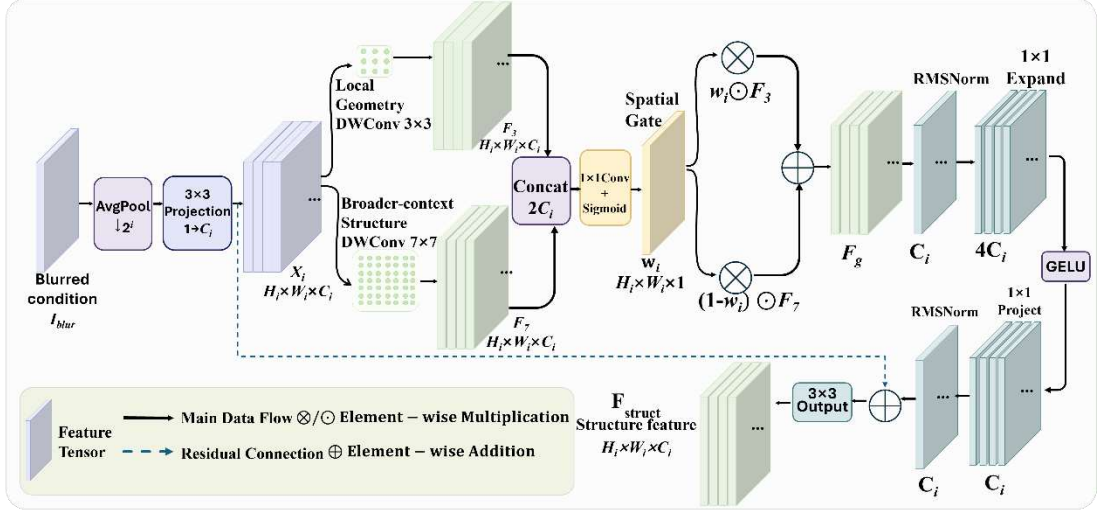


Fig. 3. Multi-scale structure stream

The structural features obtained at each level have the same spatial resolution as their corresponding diffusion features and can therefore be used directly in the subsequent windowed cross-attention fusion.

Windowed cross-attention fusion

Convolutional neural networks perform well in extracting basic image features through local receptive fields, but they have limitations in modeling content relationships among features. Attention mechanisms address this limitation by computing similarities among positions in the feature space and dynamically assigning weights to emphasize important information while suppressing redundant responses^[29]. This content-adaptive relational modeling allows a network to focus on regions with salient semantic or structural characteristics in complex scenes, providing effective representations for subsequent feature fusion and reconstruction.

After the multi-scale structure stream extracts structural features that are strictly aligned with the diffusion branch, conventional channel concatenation or elementwise addition alone may still cause the constraining effect of the structural prior to decay as network depth increases, preventing persistent structural guidance in deep features. This paper therefore designs a layer-wise cross-attention fusion mechanism that enables diffusion-branch features to actively retrieve and aggregate relevant information from the structural features, thereby supporting dynamic interactions between deep feature representations.

Specifically, at each level of the U-Net encoder, lightweight 1×1 convolutions first project the noisy diffusion-branch feature map into a query matrix Q and the structure-stream feature map F_{struct} at the current level into a key matrix K and a value matrix V . This projection improves feature discrimination while reducing computational cost. Following local-window attention^[30], Q , K , and V are then partitioned into nonoverlapping 4×4 windows, and cross-attention is computed independently within each window. Fixed window partitioning is used without shifted-window operations. The detailed structure is shown in Fig. 4.

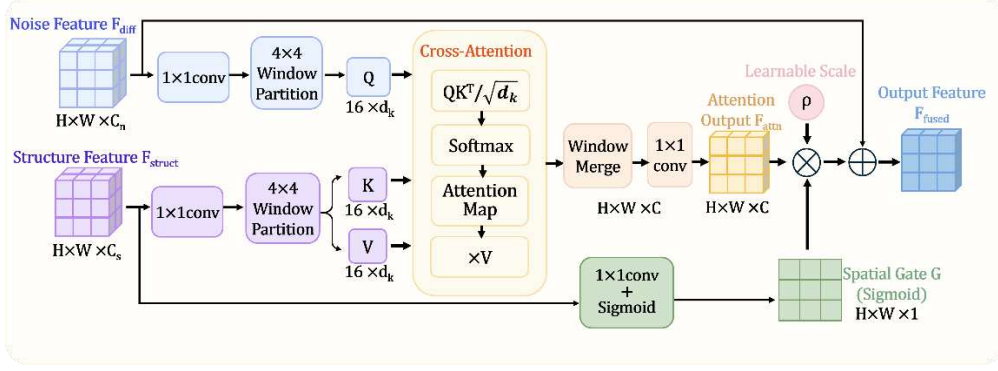


Fig. 4. Architecture of the Cross-Attention block

Let Q_n , K_n , and V_n denote the query, key, and value in the n th window, respectively. Because each window contains 16 spatial positions, all three have dimensions of $16 \times d_k$. The network computes scaled dot-product similarities between Q_n and K_n to construct an intra-window spatial affinity matrix. This matrix evaluates the dependence of each current noisy feature on every structural-prior position within the same window and generates attention weights used to aggregate V_n . The core operation is expressed as:

$$F_{attn}^n = \text{Soft max} \left(\frac{Q^n (K^n)^T}{\sqrt{d_k}} \right) V^n \quad (4)$$

where d_k is the channel dimension of the latent space and scales the attention scores to stabilize gradients during training. The equation shows that the output reorganizes the structural prior V according to the retrieval requirements of the noisy feature Q , allowing each noisy-feature position to receive spatially relevant structural guidance from within its window. After attention has been computed for all windows, the window outputs are merged at their original spatial positions and passed through a 1×1 convolution to obtain the attention-aggregated feature F_{attn} . To stabilize the initial stage of optimization and prevent insufficiently learned structural guidance from disrupting the original noise-prediction backbone, a learnable scalar scaling coefficient ρ is introduced into the residual fusion path at each encoder level. Each coefficient is initialized to 0, so the cross-attention module initially approximates the original diffusion branch. This zero-initialized residual scaling follows the residual-gating principle of ReZero^[31]. As optimization proceeds, ρ is updated adaptively through backpropagation to control the overall influence of the structure-guidance branch on the diffusion features. No explicit range constraint is imposed on ρ ; it is therefore a feature-scaling coefficient rather than a probability or injection ratio bounded between 0 and 1. The final fused feature is computed as:

$$F_{fused} = F_{diff} + \rho \cdot (G \odot F_{attn}) \quad (5)$$

where F_{diff} is the original noisy feature, F_{attn} is the attention-aggregated structural feature, G is the spatial gating weight generated by the sigmoid function, and $\odot$ denotes elementwise multiplication. This progressive and adaptive guidance strategy allows the network to adjust its dependence on local details and global contours according to the structural complexity of each spatial location.

Because this cross-attention mechanism operates independently at multiple resolution levels of

the U-Net encoder, shallow levels exchange local structural details such as edge textures. As feature resolution decreases, the same 4×4 window corresponds to a larger region in the original image, enabling deeper levels to fuse structural semantics over a broader spatial range. In this way, the structural prior guides the reverse denoising process at multiple levels.

Redesigned skip connections

Standard U-Net skip connections directly concatenate encoder feature maps with the corresponding decoder levels. In a conditional diffusion model, however, encoder features contain a mixture of noise information and structure-guidance information. If these features are transmitted without distinction, the decoder may have difficulty separating noise residuals from structural priors during upsampling, and the structural information may be overwritten by reconstruction in the noise branch. The skip connections are therefore redesigned as shown in Fig. 5.

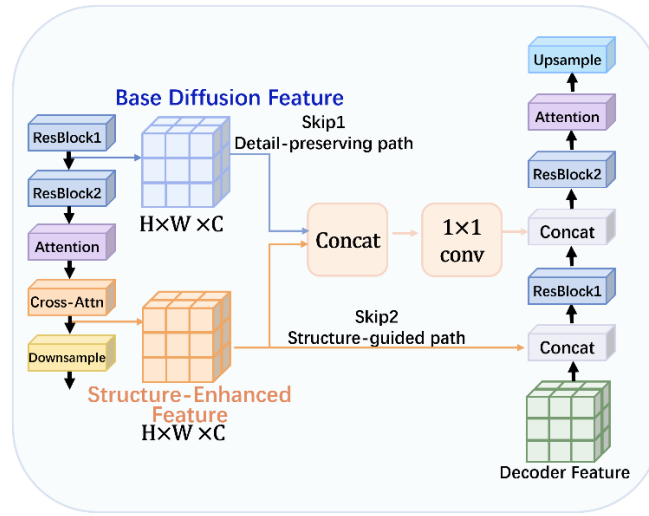


Fig. 5. Redesigned skip connections

Feature transmission from each encoder level is explicitly divided into two independent paths. The first retains the original noise-branch features from the early feature-extraction stage, preventing the loss of basic textures and shallow details. The second transmits the structure-enhanced feature F_{fused} produced by cross-attention fusion. At the corresponding decoder level, the network receives and concatenates these two feature streams. A lightweight 1×1 convolution then adjusts and reduces the channel dimension of the concatenated high-dimensional features and integrates them directly into the decoder feature space.

This design allows structure-enhanced features to reach the decoder through an independent path and to be fused with the basic diffusion features during decoding, reducing the risk that the structural prior will be diluted or overwritten during multilayer propagation.

(3) Time-adaptive frequency modulation (TAFM)

After SCGM guidance, the noise branch contains multi-scale structural priors. However, the recovery of high-frequency details, such as turbulent structures and vortex edges, still depends on how the decoder processes high-frequency information. Existing frequency-enhancement methods can adjust different frequency components according to the input content, but their modulation parameters generally do not explicitly depend on the diffusion timestep and therefore cannot readily adapt to the continuously changing noise level during reverse sampling. Noise dominates the early

stage of reverse sampling, when excessive high-frequency enhancement would amplify it. By the later stage, the structural contours have largely recovered, and moderate high-frequency enhancement is needed to replenish details. Accordingly, this paper proposes the Time-Adaptive Frequency Modulation (TAFM) module. TAFM incorporates the diffusion timestep t and its corresponding cumulative signal-retention coefficient $\bar{\alpha}_t$ into the calculation of the frequency-domain modulation parameters. The timestep t controls the range of high-frequency enhancement, whereas $\bar{\alpha}_t$ regulates the amplitude gain. The range and strength of enhancement thus adapt to the progress of reverse denoising. The TAFM architecture is shown in Fig. 6.

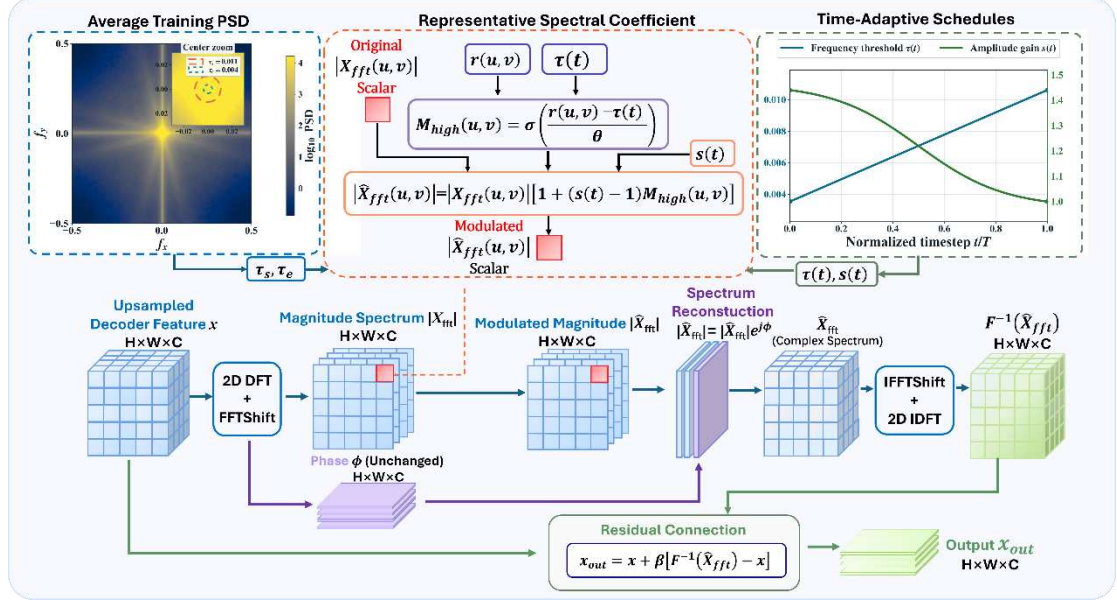


Fig. 6. Architecture of TAFM

As defined in the forward diffusion process, the cumulative signal-retention coefficient $\bar{\alpha}_t$ gradually decreases as the forward diffusion timestep t increases. Reverse sampling begins at $t=T$, when $\bar{\alpha}_t$ is small and the image contains strong noise. As sampling proceeds, t gradually decreases and $\bar{\alpha}_t$ correspondingly increases, so the image progressively enters a low-noise stage. TAFM therefore incorporates $\bar{\alpha}_t$ into the calculation of the amplitude-gain coefficient $s(t)$, so that frequency enhancement remains weak in the early stage of reverse sampling and high-frequency compensation gradually increases in the later stage. The amplitude-gain coefficient is defined as:

$$s(t) = 1 + a_{\max} \cdot \bar{\alpha}_t \quad (6)$$

where $a_{\max}$ is the learnable maximum gain-amplitude parameter. It is obtained by mapping a learnable variable through a sigmoid function, is constrained to the range from 0 to 0.5, and is updated together with the network parameters through backpropagation of the loss function.

In addition to the enhancement strength, TAFM must determine the high-frequency enhancement range from the frequency distribution of the training data. Therefore, the average power spectral density of the training images is computed as:

$$P_{\text{avg}}(u, v) = \frac{1}{KMN} \sum_{k=1}^K |F_k(u, v)|^2 \quad (7)$$

where K is the number of training images; M and N are the image height and width, respectively; and $F_k(u, v)$ is the two-dimensional Fourier transform of the k th image. The average power spectral density characterizes the frequency-energy distribution of the training set and provides a reference for initializing the high-noise-end parameter $\sigma(\tau_s)$ and low-noise-end parameter $\sigma(\tau_e)$. Both parameters are mapped to actual frequency thresholds through sigmoid functions and continue to be updated through backpropagation during training.

The range of high-frequency enhancement is controlled by the time-dependent threshold $\tau(t)$, which is obtained by linear interpolation between the two endpoint parameters:

$$\tau(t) = \sigma(\tau_s) \times t_{morm} + \sigma(\tau_e) \times (1 - t_{morm}) \quad (8)$$

where $t_{morm} = t / T$. Under the initial settings, the threshold is higher in the early sampling stage, so a narrower frequency range enters the enhancement region; the threshold gradually decreases in the later stage, allowing more high-frequency components to participate in detail recovery.

For a decoder output feature map x , TAFM first performs a two-dimensional Fourier transform to obtain the complex spectrum $F(u, v)$. Because the zero-frequency component is located at the beginning of the spectral array after the transform, a spectrum-centering operation (FFTShift) is then used to rearrange the frequency components and move the zero-frequency component to the center of the spectrum. The centered complex spectrum is denoted as:

$$X_{fft}(u, v) = FFTShift(F(u, v)) \quad (9)$$

FFTShift changes only the arrangement of the frequency components and does not alter their amplitudes or phases. Let the center of the centered spectrum be (u_c, v_c) . Using the maximum distance $r_{\max}$ from the spectral center to a corner for normalization, the frequency radius is defined as:

$$r(u, v) = \frac{\sqrt{(u - u_c)^2 + (v - v_c)^2}}{r_{\max}}, \quad r(u, v) \in [0, 1] \quad (10)$$

where a larger $r(u, v)$ indicates a higher corresponding frequency.

To avoid the nondifferentiability introduced by hard thresholding, a soft high-frequency mask is constructed from the current frequency threshold $\tau(t)$ and temperature coefficient θ :

$$M_{high}(u, v) = \sigma\left(\frac{r(u, v) - \tau(t)}{\theta}\right) \quad (11)$$

where θ is a learnable temperature coefficient that controls the transition width of the mask around the threshold. Its learnable variable is initialized to 0.05, and an absolute-value transform with a small positive offset ensures that its effective value remains positive. This parameter does not represent a physical temperature; it controls the smoothness of the sigmoid function. When the frequency radius is below the threshold, the mask weight approaches 0; above the threshold, it gradually approaches 1. The soft mask remains continuous at the frequency boundary, allowing both threshold endpoints and the temperature coefficient to receive gradient updates. The module then selectively modulates the spectral amplitude according to the mask:

$$|\hat{X}_{fft}| = |X_{fft}| \cdot (1 + (s(t) - 1) \cdot M_{high}) \quad (12)$$

where $|X_{fft}|$ and $|\hat{X}_{fft}|$ denote the amplitude spectra before and after modulation, respectively. This operation modifies only the spectral amplitude while preserving the phase. After the modulated amplitude is recombined with the original phase, the complex spectrum $\hat{X}_{fft} = |X_{fft}|e^{i\phi}$ is obtained. IFFTShift then restores the original ordering of the frequency components in the discrete Fourier transform, after which a two-dimensional inverse Fourier transform reconstructs the complex spectrum into a spatial-domain feature. Finally, the module output is obtained by fusing the reconstructed feature with the original feature through a residual connection:

$$x_{out} = x + \beta \cdot \left(F^{-1} \left(\hat{X}_{fft} \right) - x \right) \quad (13)$$

where β is the residual mixing coefficient. It limits the influence of the frequency-domain modulation branch on the original feature and reduces feature fluctuations caused by excessive modulation early in training.

(4) Loss function

The training objective of SFGD must simultaneously account for noise-prediction accuracy, structural fidelity in the image domain, and restoration quality of frequency-domain details. These three objectives correspond to the basic diffusion training paradigm and to the design goals of SCGM and TAFM. A joint loss comprising a noise-prediction loss, a structural-consistency loss, and a frequency-consistency loss is therefore designed. The three losses differ substantially in scale and optimization difficulty. The noise-prediction loss acts on pixel-level Gaussian residuals and converges stably, whereas the other two losses are based on intermediate estimates that are still inaccurate during training and thus fluctuate more strongly. Fixed weights would require repeated hyperparameter searches and could not adapt to dynamic changes in the relative difficulty of the three objectives. One term might dominate the gradients while the others are overwhelmed. This paper therefore introduces a task-uncertainty-based adaptive weighting strategy.

The noise-prediction loss L_{noise} is the fundamental training objective of the diffusion model and constrains the accuracy with which the network predicts the noise applied at each timestep:

$$L_{noise} = E_{x_0, t, \varepsilon} \left[\left\| \varepsilon - \varepsilon_\theta(x_t, c, t) \right\|^2 \right] \quad (14)$$

where $c = I_{blur}$ is the degraded condition image input to the network, ε is the Gaussian noise injected during the forward process, x_t is the noisy image obtained by diffusing the clean image x_0 to timestep t , and ε_θ is the noise predicted by the network. Because L_{noise} constrains only the noise space, it does not directly supervise final restoration quality in the image domain. A structural-consistency loss L_{struct} is therefore introduced. The predicted noise is used to derive an approximate clean image $\hat{x}_0$, which is compared directly with the ground truth x_0 :

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \varepsilon_\theta}{\sqrt{\bar{\alpha}_t}} \quad (15)$$

$$L_{struct} = 1 - SSIM(\hat{x}_0, x_0) \quad (16)$$

The SSIM-based loss constrains luminance, contrast, and structure simultaneously, providing direct supervision for restoration of the structural prior guided by SCGM^[32]. Computing $\hat{x}_0$ requires only one linear transformation and introduces no additional forward pass. However, the

estimation error of $\hat{x}_0$ is large when t is large. To prevent unreliable supervision from disrupting training, the calculation is restricted to timesteps satisfying $t < T_{\text{thresh}}$ ($T_{\text{thresh}} = T / 2 = 250$), at which L_{struct} is evaluated. The frequency-consistency loss L_{freq} directly supervises the spectral difference between the generated image and the ground truth:

$$L_{\text{freq}} = \left\| \left| F(\hat{x}_0) \right| - \left| F(x_0) \right| \right\|^2 \quad (17)$$

The loss constrains only the amplitude spectrum while leaving the phase unconstrained, consistent with TAFM's design principle of modulating the amplitude spectrum while preserving the phase spectrum. The three losses are combined using a task-uncertainty-based adaptive weighting strategy derived from the homoscedastic-uncertainty weighting method proposed by Kendall et al. for multitask learning^[33]. A learnable log-variance parameter $s_i = \log \sigma_i^2$ is introduced for each loss, and the joint loss is defined as:

$$L_{\text{total}} = \frac{1}{2} e^{-s_1} L_{\text{noise}} + \frac{1}{2} e^{-s_2} L_{\text{struct}} + \frac{1}{2} e^{-s_3} L_{\text{freq}} + \frac{1}{2} (s_1 + s_2 + s_3) \quad (18)$$

where s_1 , s_2 , and s_3 are learnable parameters optimized together with the model parameters through gradient descent. The exponential terms directly express the weighting rule that a larger variance corresponds to a smaller weight. When one loss remains large and difficult to reduce, gradient descent tends to increase its corresponding log-variance parameter and reduce its weight, allowing the network to prioritize objectives that are currently easier and more reliable. The regularization term in Eq. (18) prevents all weights from degenerating to zero. Compared with fixed weighting, this strategy eliminates manual hyperparameter searches, allows all three weights to be learned entirely from the training data, and has a clear Bayesian interpretation.

Experiments

(1) Experimental platform

γ -photon flow-field images are obtained through three stages. First, CFD simulation is used to obtain a numerical solution of the thermodynamic flow-field distribution. The result is then converted into a phantom and imported into the GATE (Geant4 Application for Tomographic Emission) platform, where γ -photon detection and acquisition are simulated using the Monte Carlo method^[34]. Finally, the coincidence-event data are reconstructed into two-dimensional flow-field images using ordered subsets expectation maximization (OSEM).

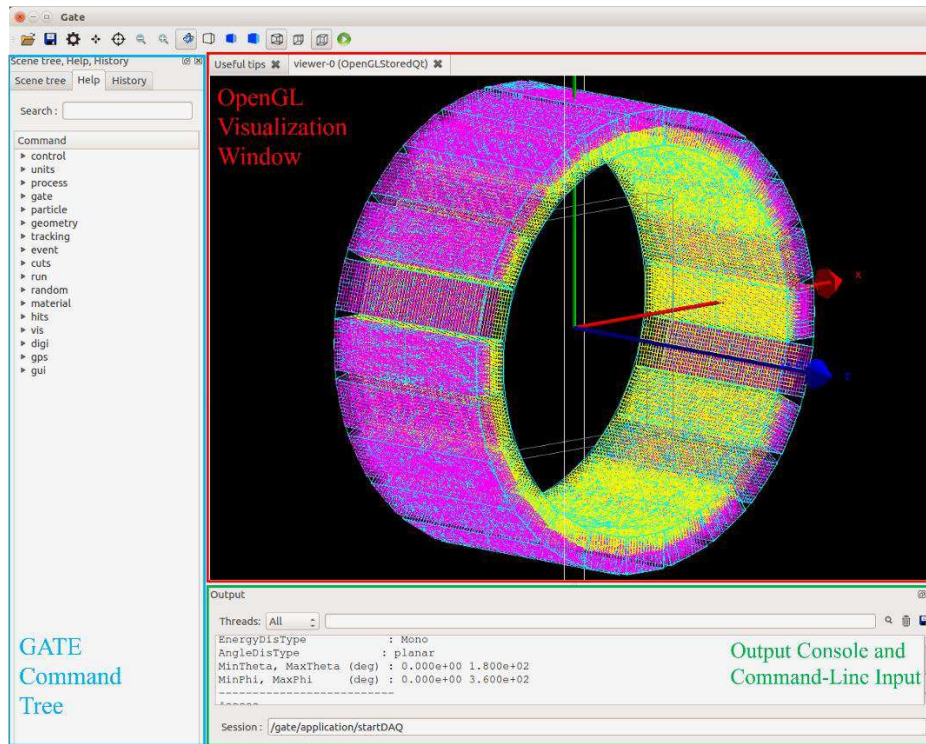


Fig. 7. GATE simulation platform

First, ANSYS CFD was used to perform whole-engine combustion simulations of an 8 kg-thrust turbojet engine under six operating conditions, yielding the corresponding flow-field temperature distributions. Because the compressor and combustion chamber contain many small, densely distributed flow structures, using the complete engine thermal map would introduce excessive fine-scale information. The exhaust-jet region was therefore selected as the data source. Flow-field sections were extracted at different locations, producing 97 temperature-distribution images with distinct flow-pattern characteristics across the six operating conditions. These images were converted to single-channel grayscale images, uniformly resized to 128×128 pixels, and used to construct the γ -photon source-density phantoms required by GATE. The corresponding grayscale images served as the reference images in the paired data.

The 97 phantoms were imported sequentially into GATE, in which an annular γ -photon detection system was constructed. GATE uses the Monte Carlo method to simulate the process from γ -photon generation in the phantom to recording by the detectors. The detector geometry and principal parameters are listed in Table 1. After simulation, the system records photon pairs satisfying the coincidence-selection criteria, ultimately yielding 97 sets of coincidence-event data.

After coincidence-event acquisition, the raw detection data must be converted into two-dimensional grayscale images using an image-reconstruction algorithm. OSEM was used for tomographic reconstruction, and all reconstructed γ -photon flow-field images were set to a resolution of 128×128 pixels. OSEM divides the complete projection data into multiple ordered subsets and uses them sequentially to update the image estimate. This strategy substantially accelerates convergence while maintaining reconstruction quality and is widely used for iterative reconstruction in γ -photon emission tomography.

Parameters	Value
Number of rings	52

Number of crystals per ring	312
Maximum ring difference	51
Projection Angle number	156
Projected radial distance number	156
Detection ring diameter	180 mm
Radial effective field of view	130 mm
Axial effective field of view	100 mm
Size of crystal	$1.9 \times 1.9 \times 10 \text{ mm}^3$
Crystal spacing	$0.2 \times 0.2 \text{ mm}^2$
Time resolution	1.5 ns
Spatial resolution	0.99 mm
Average energy resolution	15%
Energy window	350 – 625 keV
Acquisition time per dynamic frame	4 s

Table 1. Principal parameters of the γ -photon detection system

(2) Dataset construction

Because the detection system contains multiple axial crystal layers, OSEM reconstruction of each phantom produces several slices at different axial positions, each representing the radioactive distribution of the flow field at a different sectional depth. Images reconstructed from the same phantom are highly correlated. Using all of them as training samples would increase redundancy and could reduce the effective information content of the dataset. Therefore, three representative γ -photon flow-field images with relatively high image quality and complete flow structures were selected from each group. In total, 291 original image pairs were obtained and strictly paired one to one with their corresponding CFD grayscale flow-field images. The data were then divided at the phantom level into training, validation, and independent test datasets, with no slices from the same phantom shared across different datasets.

Training dataset

Slices reconstructed from the same phantom are strongly correlated. If individual images were randomly assigned, adjacent slices from the same phantom could appear in different datasets and cause data leakage. The 97 phantom groups were therefore split at the phantom level, ensuring that all slices from the same phantom belonged to the same dataset. The split was designed to cover different exhaust-jet morphologies across the training, validation, and independent test datasets while keeping the number of samples in each category relatively balanced. The training dataset comprised 67 phantom groups. Three slices with relatively high image quality and complete flow structures were selected from each phantom, producing 201 original training pairs. Each pair consisted of one γ -photon flow-field image and its corresponding CFD grayscale flow-field image.

A diffusion model must learn the relationship between noise and flow-field structure at different timesteps. Using only 201 original pairs would repeatedly expose the model to the same structural patterns and increase the risk of memorizing a limited sample set. To enrich local-structure samples, each pair was first subjected to four-quadrant, or cross-shaped, cropping: the complete image was divided into four nonoverlapping subregions, and every crop was resized to 128×128

pixels. A complete flow-field image contains extensive slowly varying regions, whereas vortex edges and local wake structures occupy a relatively small area. Cropping and resizing increase the relative scale of these local structures, allowing the model to learn local contours and their surrounding noise distributions more fully. The γ -photon flow-field image and CFD reference image were cropped and resized at exactly the same positions to maintain pixelwise correspondence. Flow structures may also appear in different orientations. If the training samples have relatively fixed orientations, the model may learn direction-specific feature responses. The original images were therefore rotated by 90° , 180° , and 270° , so that the same structures entered training in different orientations and improved the model's adaptability to changes in wake direction. Right-angle rotations preserve the original pixel grid and avoid interpolation errors caused by arbitrary-angle rotation; the input and reference images were transformed synchronously. Each original pair ultimately generated the original image, four cropped images, and three rotated images, yielding eight training pairs. The training set thus expanded from 201 to 1,608 pairs. Representative training samples are shown in Figs. 8 and 9.

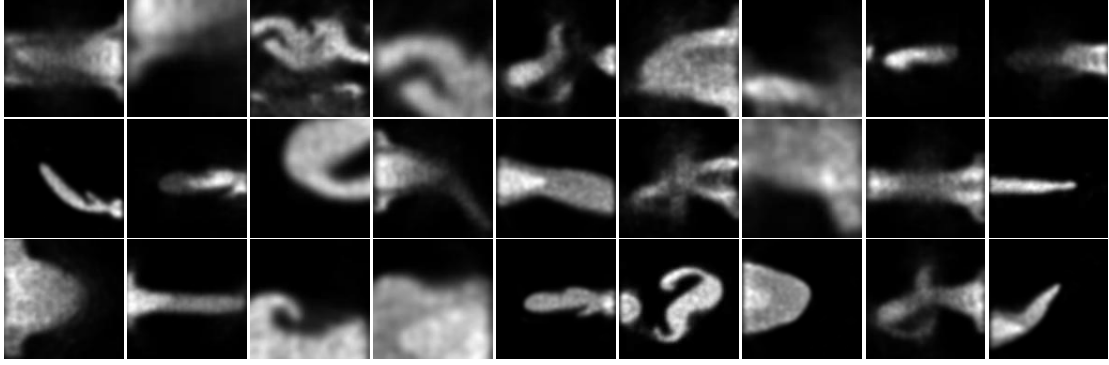


Fig. 8. Representative training samples (GATE)

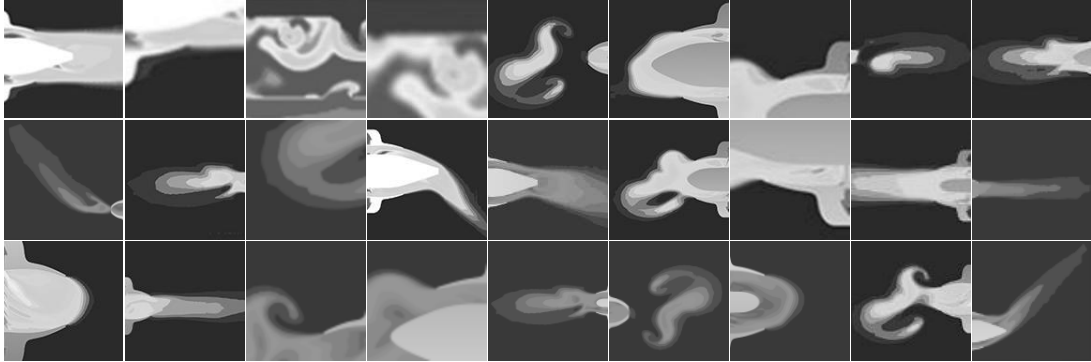


Fig. 9. Representative training samples (CFD)

Validation dataset

The validation dataset comprised 20 phantom groups obtained from the phantom-level split. Samples were selected according to the same flow-pattern coverage principle used for the training dataset. Different exhaust-jet morphologies were included, and the number of images in each category was kept as balanced as possible to prevent a single dominant flow pattern from biasing the validation results. Three reconstructed γ -photon flow-field images were selected from each phantom and paired with the corresponding CFD grayscale image, yielding 60 original validation pairs.

The validation set underwent the same paired transformations as the training set. Four-quadrant

cropping and right-angle rotation expanded it to 480 image pairs. Each γ -photon flow-field image and its corresponding CFD reference image were always transformed synchronously, ensuring exact correspondence in crop region and rotation direction. These transformed samples were not used for gradient updates. They were used only to evaluate denoising performance across different image orientations and local flow-field regions and to reduce the influence of any single image orientation on the evaluation. Representative validation samples are shown in Figs. 10 and 11.

At the end of each training epoch, the validation loss was computed on the fixed validation dataset. Whenever the current validation loss fell below the historical minimum, the program updated and saved the optimal model weights together with the corresponding training state. After training, the weights associated with the lowest validation loss were selected as the final model weights and subsequently used for the comparative and ablation experiments on the independent test dataset.

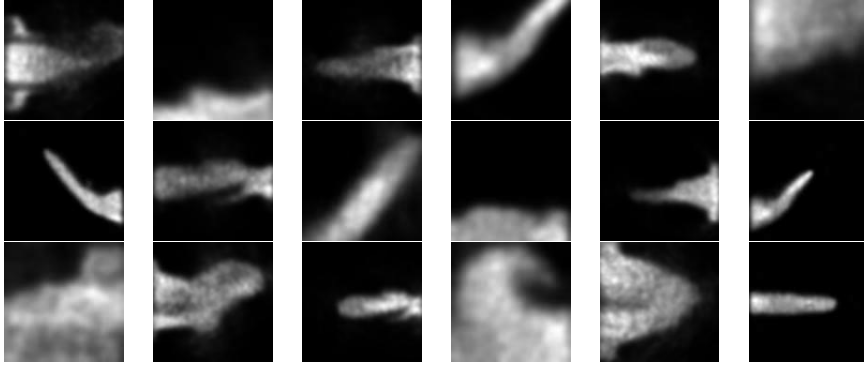


Fig. 10. Representative validation samples (GATE)

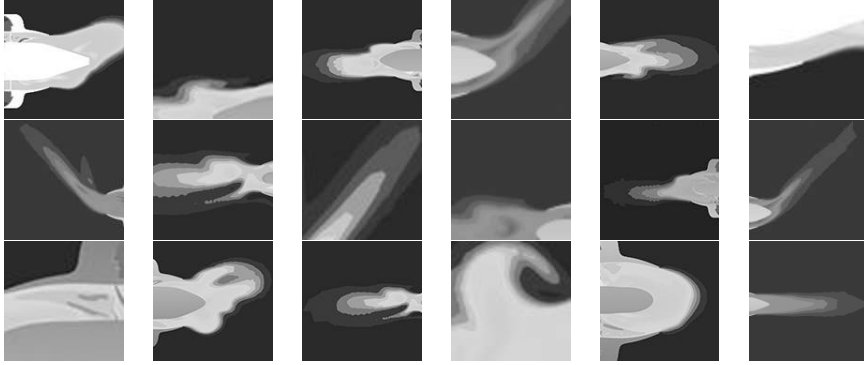


Fig. 11. Representative validation samples (CFD)

Independent test dataset

The independent test dataset comprised the remaining 10 phantom groups. Three reconstructed γ -photon flow-field images were selected from each phantom and paired with the corresponding CFD grayscale images, yielding 30 original test pairs. No cropping, rotation, or other data augmentation was applied to the independent test dataset. It was not used for model training, hyperparameter adjustment, or optimal-checkpoint selection and was used only for the final comparative and ablation evaluations.

(3) Experimental configuration

Model and training parameters

All experiments were conducted on a single NVIDIA RTX PRO 6000 GPU with 96 GB of memory.

The software environment comprised PyTorch 2.8.0, Python 3.12, and CUDA 12.8, and both model training and inference were GPU-accelerated. The model backbone and loss functions were implemented in PyTorch, and the frequency-domain transforms in TAFM were implemented using the torch.fft interface. Paired images were read with Pillow, converted to tensors using torchvision, and then normalized.

The Adam optimizer was used with an initial learning rate of 1×10^{-4} . Cosine annealing smoothly reduced the learning rate to 1×10^{-6} , preventing the loss oscillations that a fixed learning rate can cause late in training. The batch size was 16, and the model was trained for 500 epochs. In the joint loss, the structural-consistency term L_{struct} was computed only when the timestep satisfied $t < T_{thresh}$ ($T_{thresh} = T / 2 = 250$). This truncation was used because the approximate clean image inferred from the predicted noise has a large error at high-noise timesteps, where structural supervision would introduce unreliable gradients. Detailed training hyperparameters are listed in Table 2.

Parameters	Value
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
Final Learning Rate	1×10^{-6}
LR Scheduler	Cosine Annealing
Batch Size	16
Training Epochs	500
Timesteps T	500
Sampling Timesteps	125 (DDIM)
T_thresh	250

Table 2. Model training hyperparameters

Evaluation metrics

Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) were selected as quantitative measures of denoising performance. They independently evaluate pixel fidelity and perceptual structure, respectively, and their combined use reduces the limitations of relying on a single metric.

PSNR is defined from the mean squared error and measures the overall pixel-value deviation between a generated image and its reference image. It is calculated as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \quad (19)$$

where MSE is the mean squared error between the evaluated image and the ground-truth image, and MAX is the maximum image-pixel value (1.0 for the grayscale images used here). A higher PSNR indicates greater pixel-level reconstruction accuracy.

SSIM evaluates similarity in luminance, contrast, and structure and is therefore more consistent with subjective human perception of image quality. It is calculated as follows^[32]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (20)$$

where μ_x and μ_y are the local means of images x and y , respectively; σ_x and σ_y are their corresponding standard deviations; σ_{xy} is their covariance; and $C1$ and $C2$ are stability constants that prevent the denominator from approaching zero. SSIM ranges from -1 to 1 , with values closer to 1 indicating higher structural fidelity.

(4) Analysis of the model training process

The training and validation losses of SFGD were recorded throughout all 500 training epochs. PSNR and SSIM were also computed on a predefined subset of the validation dataset covering the principal flow patterns to analyze the training process and validation performance. The results are shown in Fig. 12.

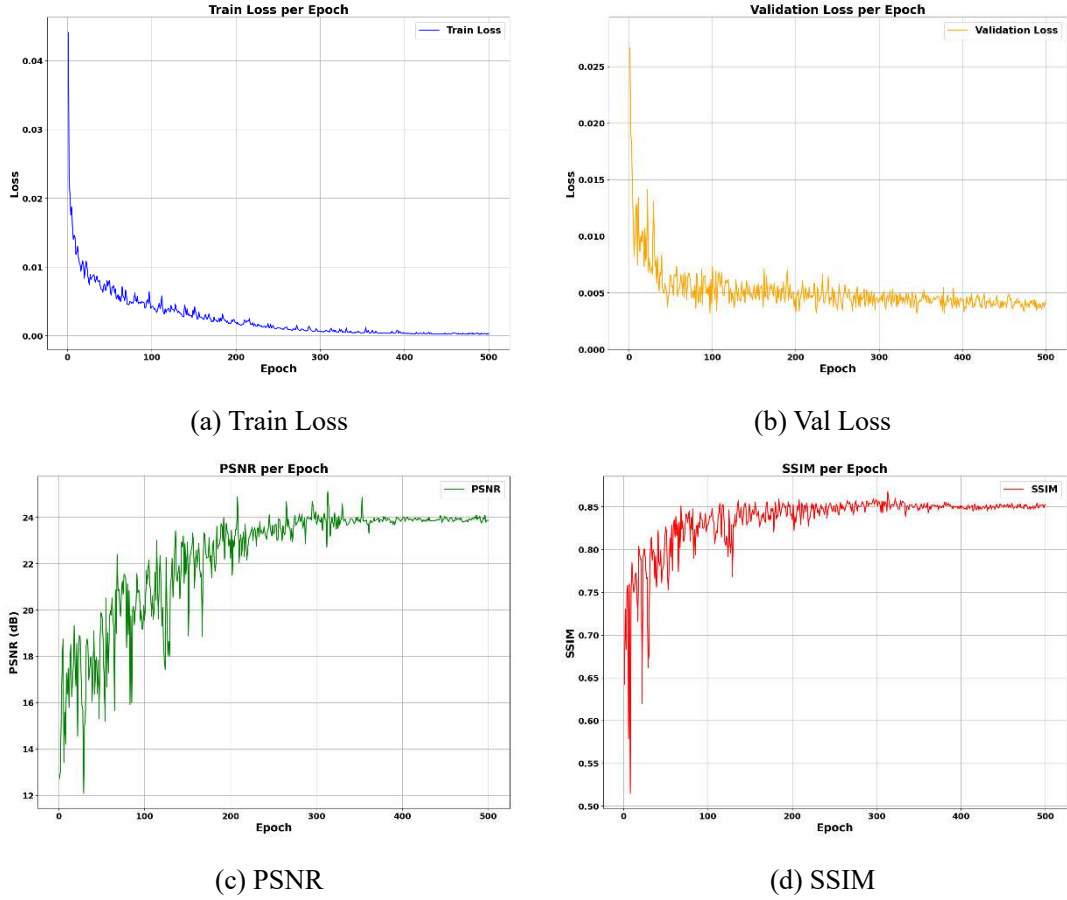


Fig. 12. Training and evaluation curves

At the beginning of training, the training and validation losses decreased rapidly and simultaneously, indicating that the model quickly learned the mapping between γ -photon flow-field images and their corresponding reference images. As the number of epochs increased, both losses declined more slowly and gradually stabilized. Their trends remained consistent, without a continued decrease in training loss accompanied by a sustained increase in validation loss, indicating stable training and no evident overfitting. The PSNR and SSIM curves generally followed the loss curves: both metrics improved rapidly early in training and exhibited small fluctuations later because of batch sampling and the stochasticity of the diffusion process, but overall performance progressively stabilized. The validation loss reached its minimum at epoch 416, by which point PSNR and SSIM on the fixed validation subset had converged to relatively high levels. This indicates that the model achieved a favorable balance between restoring the overall grayscale

distribution and preserving structural information. The weights saved at epoch 416 were therefore selected as the final model weights.

(5) Comparison with existing methods

To comprehensively evaluate denoising performance, five representative methods were selected: DnCNN and RIDNet as CNN-based denoisers, NAFNet as an efficient image-restoration network, FFTformer as a frequency-domain Transformer, and a conditional DDPM as the diffusion baseline most directly comparable with SFGD. For consistency, all models used the same training, validation, and independent test split, together with the same 128×128 single-channel input format, normalization procedure, and metric definitions. After training, the checkpoint with the lowest validation loss was selected for each model, and the model parameters were then frozen. Based on the orientation of the principal wake and the presence of local recirculation, the test samples were categorized into large-scale vortical wakes, oblique wakes, and horizontal wakes. One representative test sample from each category was selected for the qualitative and per-image quantitative comparisons shown in Figs. 13–15 and Table 3.

Qualitative analysis

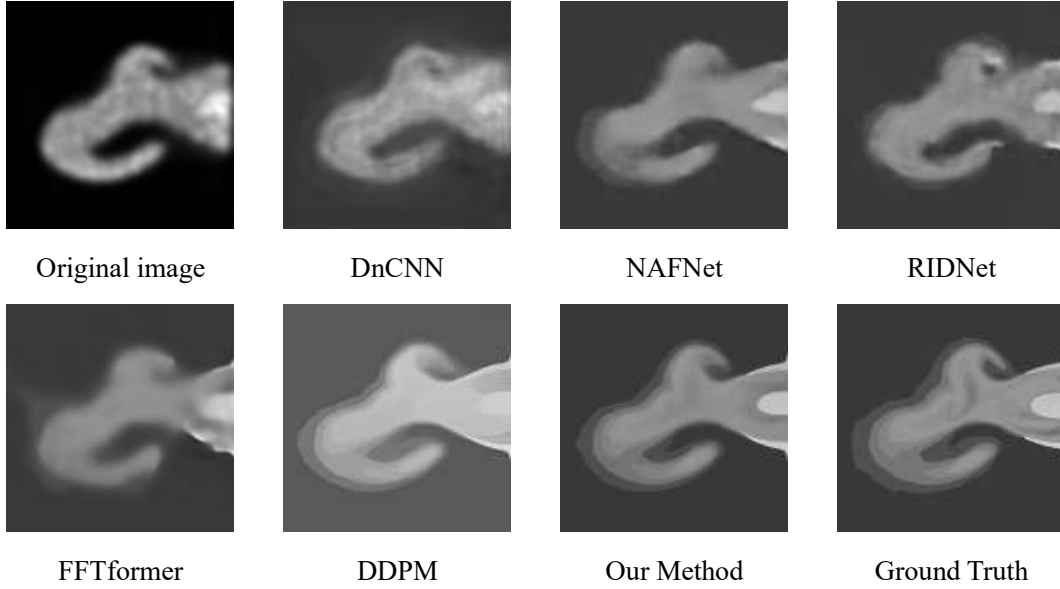
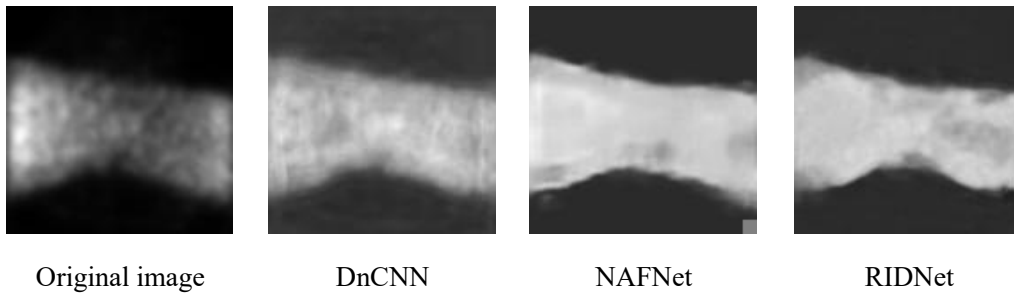


Fig. 13. Results for a large-scale vortical wake



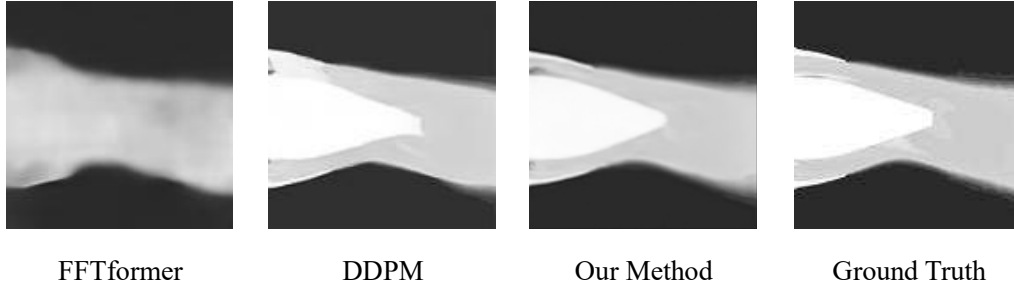


Fig. 14. Results for an oblique wake

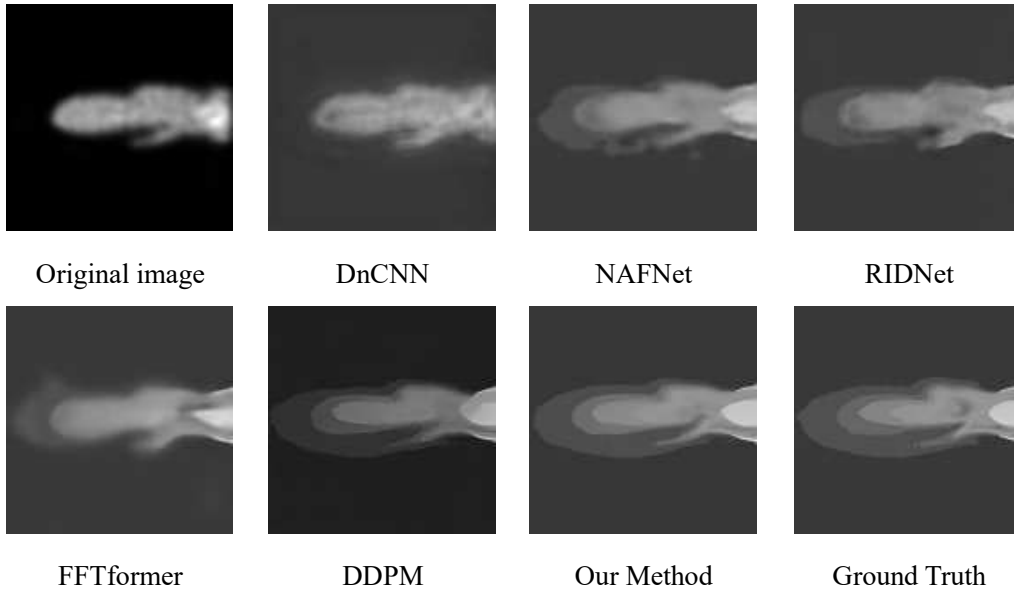


Fig. 15. Results for a horizontal wake

As shown in Figs. 13 – 15, the original γ -photon flow-field images have blurred overall contours, pronounced intensity spreading near their boundaries, and internal structures that are partly difficult to distinguish. DnCNN suppresses some noise interference, but its outputs remain strongly blurred, with broad flow-field boundaries and insufficient restoration of internal grayscale variations. NAFNet and RIDNet recover the principal contours more clearly and produce smoother outputs; however, some local regions are oversmoothed, several grayscale levels merge, and boundary continuity remains unstable. FFTformer produces noticeable structural smoothing in all three types of flow-field images. Local morphologies in the oblique and horizontal wakes are weakened, making it difficult to reproduce the detailed variations in the ground-truth images accurately. The conditional DDPM restores a relatively complete principal flow-field structure and achieves better boundary continuity than the CNN methods, but the contour thickness and grayscale distribution in some regions still differ from the ground truth. In comparison, SFGD provides stable restoration across the different flow patterns. In the large-scale vortical wake, the curved vortex cavity and its outer contour are preserved clearly. In the oblique wake, the upper and lower boundaries and the turning region more closely match the ground truth. For the horizontal wake, SFGD restores the internal grayscale hierarchy of the principal structure while maintaining good continuity in the tail. Across all three groups, SFGD suppresses noise in the γ -photon images while preserving flow-field contours and internal grayscale levels more consistently, yielding high visual

agreement with the ground truth.

Quantitative comparison

To quantify denoising performance further, PSNR and SSIM were calculated for the outputs of each model. The results are presented in Table 3.

Method	Fig. 13		Fig. 14		Fig. 15	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DnCNN	23.713	0.765	16.604	0.667	23.720	0.734
NAFNet	28.402	0.867	18.662	0.746	25.538	0.782
RIDNet	25.473	0.877	17.341	0.624	25.645	0.765
FFTformer	25.598	0.886	17.914	0.722	27.938	0.805
DDPM	27.459	0.913	20.686	0.804	31.247	0.845
Our Method	31.883	0.946	25.281	0.923	36.426	0.969

Table 3. Quantitative comparison on representative samples

Table 3 shows that SFGD achieves the highest PSNR and SSIM for all three flow-field images. For the large-scale vortical wake in Fig. 13, NAFNet performs well in pixel-level reconstruction accuracy, whereas DDPM has an advantage in structure preservation. This indicates that different models emphasize grayscale restoration and structural consistency to different degrees. The metrics for all methods in Fig. 14 are lower than those for the other two groups, possibly because the contour transitions in the input image are more diffuse and local grayscale levels are difficult to distinguish, leaving fewer structural cues for the models. Even with this weakly structured input, SFGD maintains comparatively strong restoration performance, indicating that combining structure guidance with frequency-domain modulation reduces the effect of insufficient structural information in the input. For the horizontal wake in Fig. 15, all methods perform better, but SFGD remains more stable in reconstructing the principal grayscale distribution and outer contour. Overall, the performance advantage of SFGD is not limited to one flow pattern. It maintains robust pixel-level reconstruction accuracy and structural consistency, in agreement with the qualitative observations above.

(6) Ablation study

SFGD introduces flow-field structural information through SCGM and regulates frequency-domain features at different diffusion timesteps through TAFM. Although the complete model reveals overall denoising performance, it does not isolate the contribution of each module. Ablation experiments were therefore conducted to analyze how structure guidance and frequency-domain modulation affect γ -photon flow-field image restoration. Both visual results and quantitative metrics were examined, with particular attention to the principal flow morphology, local edges, and grayscale reconstruction accuracy.

Qualitative ablation results

Four configurations were evaluated: the basic conditional DDPM, DDPM with TAFM only, DDPM with SCGM only, and the complete SFGD. The basic model retains the conditional diffusion denoising backbone while removing both SCGM and TAFM, providing a baseline without the proposed modules. The TAFM-only model assesses the independent contribution of time-adaptive frequency modulation to local high-frequency detail restoration in the absence of structure guidance.

The SCGM-only model examines how structure guidance constrains the principal flow-field contour and spatial relationships without frequency-domain modulation. The complete SFGD evaluates the complementary effect of using both modules together. All configurations used the same training and validation datasets, number of training epochs, and hyperparameter settings, with only the corresponding modules being changed. For each configuration, the checkpoint with the lowest validation loss was selected and then applied to the same independent test dataset. One representative sample from the independent test dataset was selected for the qualitative comparison shown in Fig. 16.

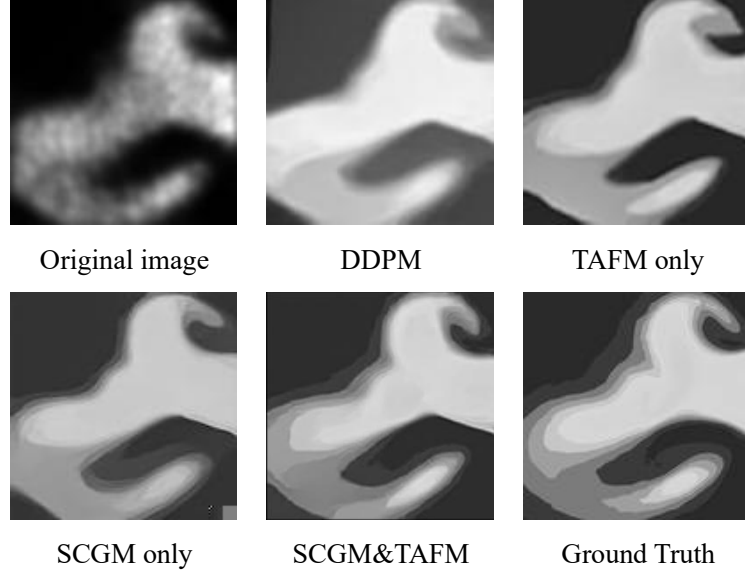


Fig. 16. Qualitative results of the ablation study

The basic diffusion model removes part of the noise from the original image, but its output remains noticeably smooth. The contour of the upper curved structure is relatively broad, and neither the boundary nor the internal grayscale levels of the lower closed region are fully recovered. Adding TAFM alone sharpens local edges and increases the grayscale contrast between the lower curved region and the background. This shows that frequency-domain modulation can replenish high-frequency information that tends to weaken during reverse sampling. However, without persistent constraints from structural features, the position and width of some contours still differ from the reference image. With SCGM alone, the principal flow-field morphology is restored more completely and the spatial relationship between the upper and lower structures more closely matches the reference image, but local grayscale transitions remain relatively smooth. The complete SFGD uses diffusion timestep t to adjust the range of high-frequency enhancement and the cumulative signal-retention coefficient $\bar{\alpha}_t$ to regulate amplitude gain. Consequently, the principal contour, local edges, and internal grayscale levels all more closely resemble the reference image. These results indicate that SCGM primarily improves structural localization, frequency-domain modulation replenishes local high-frequency information, and the time-adaptive strategy further coordinates modulation across different denoising stages.

Quantitative analysis

Visual comparison reveals morphological differences among model outputs but may be influenced by the observer. To quantify the contributions of SCGM and TAFM further, the mean PSNR and

SSIM of the four configurations were calculated over all 30 original image pairs in the independent test dataset. Table 4 also reports the changes in both metrics relative to the basic diffusion model, where ΔPSNR and ΔSSIM denote the corresponding differences from the baseline values.

Method	SCGM	TAFM	PSNR/dB	$\Delta\text{PSNR}/\text{dB}$	SSIM	ΔSSIM
DDPM			22.826	—	0.773	—
TAFM Only		✓	24.357	1.531	0.895	0.122
SCGM Only	✓		26.590	3.764	0.862	0.089
SCGM&TAFM	✓	✓	28.325	5.499	0.926	0.153

Table 4. Quantitative comparison of the proposed modules

Table 4 shows that both SCGM and TAFM improve the denoising performance of the basic diffusion model, although their effects on the evaluation metrics differ. With SCGM alone, the PSNR gain is larger than that of the TAFM-only model. This finding indicates that multi-scale structure guidance constrains the location and contour of the flow-field region and reduces grayscale error in the principal region, producing a greater improvement in overall pixel-level reconstruction accuracy. The TAFM-only model yields a smaller PSNR gain but a larger SSIM gain than the SCGM-only model. This result agrees with the design objective of TAFM: after high-frequency components are regulated according to the diffusion timestep, edge transitions and local morphologies more closely match the reference image, improving structural similarity. The complete SFGD achieves the highest PSNR and SSIM, demonstrating that the two modules address complementary aspects of γ -photon flow-field image denoising. SCGM provides stable constraints on the principal structure during reverse sampling, and TAFM supplements local edge information on that basis. The quantitative results correspond to the visual differences in Fig. 16 and confirm the rationale for combining structure guidance with frequency-domain modulation.

Conclusion

This study proposes a Structure-Frequency Guided Diffusion (SFGD) model to address structural distortion and detail oversmoothing in γ -photon flow-field image denoising. The Structure-Cross Guidance Module (SCGM) uses multi-scale structural extraction, window-based cross-attention, and redesigned skip connections to preserve the principal flow-field morphology and boundary continuity throughout reverse denoising, while the Time-Adaptive Frequency Modulation (TAFM) module dynamically regulates frequency-domain features according to the diffusion timestep to balance noise suppression and detail recovery. Comparative experiments on three representative flow-field images show that SFGD achieves mean PSNR and SSIM values of 31.197 dB and 0.946, respectively, outperforming the conditional DDPM by 4.733 dB and 0.092. On the independent test dataset, the full model reaches 28.325 dB in PSNR and 0.926 in SSIM, representing improvements of 5.499 dB and 0.153 over the basic conditional diffusion model. Ablation experiments further confirm that SCGM improves structural localization and continuity, whereas TAFM enhances edges and local grayscale variations, with their combination producing the best overall performance. These results demonstrate that SFGD more effectively balances noise removal, structural preservation, and detail restoration than the representative comparison methods, providing an effective denoising approach for γ -photon flow-field image analysis.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding

author upon reasonable request.

References

- [1] Xiao, H., Liu, Q., Xu, Y. B., Wang, M. & Liu, J. T. Research on a noise-suppression super-resolution enhancement module for positron flow field images based on convolution and SwinTransformer structures. *Sci. Rep.* **15**, 21443; 10.1038/s41598-025-07107-1 (2025).
- [2] Hudson, H. M. & Larkin, R. S. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans. Med. Imaging* **13**, 601–609; 10.1109/42.363108 (1994).
- [3] Zhao, M., Chen, H., Yao, M., Guo, R. & Xiao, H. Experimental study on combustion chamber flow field using positron emission tomography. *J. Chem. Eng. Jpn.* **53**, 729–738; 10.1252/jcej.19we150 (2020).
- [4] Cao, S. et al. Dynamic PET image reconstruction incorporating a median nonlocal means kernel method. *Comput. Biol. Med.* **139**, 104713; 10.1016/j.compbiomed.2021.104713 (2021).
- [5] Zhang, K., Zuo, W., Chen, Y., Meng, D. & Zhang, L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **26**, 3142–3155; 10.1109/TIP.2017.2662206 (2017).
- [6] Anwar, S. & Barnes, N. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 3155–3164; 10.1109/ICCV.2019.00325 (2019).
- [7] Saharia, C. et al. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4713–4726; 10.1109/TPAMI.2022.3204461 (2023).
- [8] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324; 10.1109/5.726791 (1998).
- [9] Tian, C. et al. Deep learning on image denoising: an overview. *Neural Netw.* **131**, 251–275; 10.1016/j.neunet.2020.07.025 (2020).
- [10] Tian, C., Xu, Y. & Zuo, W. Image denoising using deep CNN with batch renormalization. *Neural Netw.* **121**, 461–473; 10.1016/j.neunet.2019.08.022 (2020).
- [11] Tian, C. et al. Multi-stage image denoising with the wavelet transform. *Pattern Recognit.* **134**, 109050; 10.1016/j.patcog.2022.109050 (2023).
- [12] Chen, L., Chu, X., Zhang, X. & Sun, J. Simple baselines for image restoration. In *Computer Vision – ECCV 2022* 17–33; 10.1007/978-3-031-20071-7_2 (Springer, 2022).
- [13] Yang, L. et al. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput. Surv.* **56**, 105; 10.1145/3626235 (2024).
- [14] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning* **37**, 2256–2265 (PMLR, 2015).
- [15] Ho, J., Jain, A. N. & Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020).
- [16] Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations* <https://openreview.net/forum?id=StlgiaRCHLP> (2021).
- [17] Kawar, B., Elad, M., Ermon, S. & Song, J. Denoising diffusion restoration models. In

Advances in Neural Information Processing Systems **35**, 23593–23606; 10.52202/068431-1714 (2022).

- [18] Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241; 10.1007/978-3-319-24574-4_28 (Springer, 2015).
- [19] Wu, J., Wu, H. & Yuan, G. Detail-aware image denoising via structure preserved network and residual diffusion model. *Vis. Comput.* **41**, 639–658; 10.1007/s00371-024-03353-y (2025).
- [20] Liu, P., Zhang, H., Zhang, K., Lin, L. & Zuo, W. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 773–782; 10.1109/CVPRW.2018.00121 (2018).
- [21] Zou, W. et al. SDWNet: a straight dilated network with wavelet transformation for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* 1895–1904; 10.1109/ICCVW54120.2021.00216 (2021).
- [22] Qin, Z., Zhang, P., Wu, F. & Li, X. FcaNet: frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 783–792; 10.1109/ICCV48922.2021.00082 (2021).
- [23] Xie, W. et al. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 4308–4317; 10.1109/ICCV48922.2021.00427 (2021).
- [24] Huang, J.-J. & Dragotti, P. L. WINNet: wavelet-inspired invertible network for image denoising. *IEEE Trans. Image Process.* **31**, 4377–4392; 10.1109/TIP.2022.3184845 (2022).
- [25] Cui, Y. et al. Selective frequency network for image restoration. In *International Conference on Learning Representations* <https://openreview.net/forum?id=tyZ1ChGZIKO> (2023).
- [26] Kong, L., Dong, J., Ge, J., Li, M. & Pan, J. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5886–5895; 10.1109/CVPR52729.2023.00570 (2023).
- [27] Liu, Z. et al. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11976–11986; 10.1109/CVPR52688.2022.01167 (2022).
- [28] Zhang, B. & Sennrich, R. Root mean square layer normalization. In *Advances in Neural Information Processing Systems* **32**, 12381–12392 (2019).
- [29] Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* **30**, 5998–6008 (2017).
- [30] Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 10012–10022; 10.1109/ICCV48922.2021.00986 (2021).
- [31] Bachlechner, T., Majumder, B. P., Mao, H., Cottrell, G. & McAuley, J. ReZero is all you need: fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* **161**, 1352–1361 (PMLR, 2021).
- [32] Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612; 10.1109/TIP.2003.819861 (2004).

- [33] Kendall, A., Gal, Y. & Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7482–7491; 10.1109/CVPR.2018.00781 (2018).
- [34] Sarrut, D. et al. Advanced Monte Carlo simulations of emission tomography imaging systems with GATE. *Phys. Med. Biol.* **66**, 10TR03; 10.1088/1361-6560/abf276 (2021).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (General Program) (G. No. 62571239), Shenzhen Natural Science Foundation (General Program) (G. No. JCYJ20250604190734042), the Joint Fund of the National Natural Science Foundation of China (G. No. U2441285), and the Experimental Technology Research and Development Project of Nanjing University of Aeronautics and Astronautics (G. No.SYJS202604Z).

Author contributions

In this paper, Song YuLu is responsible for manuscript writing, Liu JianTang conceived the study, Xiao Hui conducted the GATE simulation experiment, Hou LiYing assisted in completing experimental data analysis. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.