WDFNET: ADAPTIVE FREQUENCY GUIDANCE FOR DIFFUSION-BASED ULTRASOUND IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing label diffusion classifiers rely primarily on deep semantic image conditioning, in which frequency information is learned only implicitly. Although deep networks can capture frequency-dependent patterns, prior studies have reported frequency biases in learned representations and shown that explicit frequency modeling can complement conventional spatial features in visual and medical image analysis. This issue is particularly relevant to ultrasound, where high-frequency responses may represent fine anatomical boundaries and tissue texture, but may also reflect speckle related patterns and acquisition variation. We propose WDFNet, an adaptive frequency conditioning framework for label diffusion classification that explicitly introduces wavelet-reconstructed high-frequency evidence into the image guidance pathway. Its Adaptive Frequency Guide (AFG) selects complementary frequency responses according to semantic context, while its Frequency Alignment Module (FAM) adjusts the fused condition according to the current noisy label, diffusion timestep, and auxiliary class prior. Experiments on six medical image datasets show consistent improvements over the evaluated discriminative and diffusion-based classifiers. Repeated runs, capacity-matched spatial controls, fixed-fusion controls, paired label-noise analyses, and controlled corruption experiments support adaptive frequency conditioning beyond fixed fusion and additional second-branch capacity within the same label-diffusion framework.

1 INTRODUCTION

Ultrasound classification requires representations that preserve structural details without allowing acquisition related variation to dominate prediction. Ultrasound is inexpensive, noninvasive, and real time, and supports applications ranging from cancer assessment to fetal monitoring (Noble & Boukerroui, 2006; Liu et al., 2019). Coherent B-mode formation produces tissue dependent speckle patterns together with structural details such as fine texture and rapidly varying anatomical boundaries (Burckhardt, 1978; Wagner et al., 1983; Thijssen, 2003). These components do not form a simple signal–noise pair: speckle can obscure anatomical structures, but its statistics may also characterize scattering properties and tissue texture, while a high-frequency response may correspond to a contour, local texture, a speckle pattern, or an acquisition artifact (Wagner et al., 1986; Lizzi et al., 2003; Damerjian et al., 2014). A useful ultrasound representation must therefore retain detailed evidence selectively rather than suppressing or indiscriminately amplifying high-frequency content.

Existing label diffusion classifiers such as CARD and DiffMIC rely mainly on semantic image conditioning, leaving frequency information to be learned implicitly within deep representations. CARD formulates prediction as diffusion over target variables under an input-dependent condition (Han et al., 2022), while DiffMIC and DiffMIC-v2 adapt this formulation to medical image classification by diffusing class label representations and conditioning the reverse process on semantic image features and an auxiliary class prior (Yang et al., 2023; 2025). Their forward processes operate in label space rather than on image pixels, and WDFNet adopts this established formulation instead of introducing a new label diffusion process. Deep encoders can implicitly capture frequency-dependent patterns, but prior studies have reported frequency biases in neural networks and shown that explicit frequency representations can provide complementary information to conventional spatial features (Rahaman et al., 2019; Qin et al., 2021; Jiao et al., 2024). These classifiers do not include an explicit frequency

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

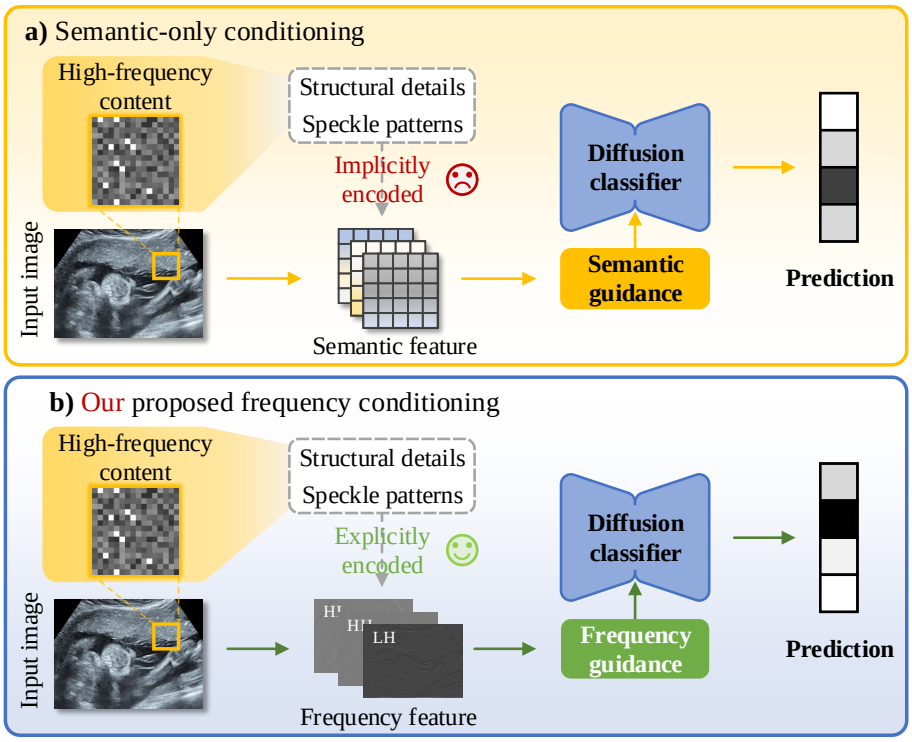

Figure 1: Conceptual comparison of image conditioning for label diffusion classification. High-frequency content includes structural details and speckle patterns. **a)** Semantic-only conditioning encodes this content implicitly within semantic features. **b)** WDFNet introduces explicit frequency features to complement semantic guidance. The prediction vectors are schematic.

representation for conditioning label-noise estimation. Fig. 1 contrasts semantic-only conditioning, which encodes structural details and speckle patterns implicitly, with WDFNet's explicit frequency guidance, which complements the semantic representation.

Explicit frequency modeling can complement semantic representations, but frequency evidence should not be treated as uniformly beneficial. Prior frequency-aware methods have demonstrated the value of spectral representations in image restoration, enhancement, segmentation, attention, and medical image representation learning (Huang et al., 2024; Jiang et al., 2023; Wu et al., 2024; Jiao et al., 2024; Qin et al., 2021). However, these objectives differ from conditional label denoising, and ultrasound high-frequency content is itself heterogeneous. A high-frequency response may correspond to an anatomical boundary or informative texture, but it may also reflect speckle related variation or acquisition artifacts (Thijssen, 2003; Damerjian et al., 2014). Directly adding an explicit frequency branch can therefore amplify irrelevant responses together with useful evidence. Effective label diffusion conditioning should not only expose frequency information explicitly, but also select it according to semantic context and adapt its influence to the evolving label state.

We propose the Wavelet Diffusion Frequency Guidance Network (WDFNet), an adaptive frequency-conditioning framework for label diffusion classification. Rather than altering the established label-space diffusion formulation, WDFNet redesigns the image guidance pathway. A semantic encoder extracts high-level image features and an auxiliary class prior, while a parallel wavelet branch reconstructs detail subbands into a spatially aligned high-frequency representation. WDFNet then decomposes adaptive conditioning into two complementary decisions: what frequency evidence should be retained and how it should influence the current reverse step. The Adaptive Frequency Guide (AFG) addresses the former through semantic-guided frequency selection, while the Frequency Alignment Module (FAM) addresses the latter by adapting the fused condition to the noisy label, diffusion timestep, and auxiliary class prior.

Adaptive Frequency Guide (AFG) determines what high-frequency evidence should modify the semantic representation. AFG reconstructs the wavelet detail components and processes them through a learnable frequency branch together with the semantic feature. Because the reconstructed response may contain anatomical boundaries, local texture, speckle related patterns, and acquisition artifacts, AFG estimates an elementwise gate from both semantic and frequency information rather than treating every high-frequency coefficient as equally useful. The selected frequency response is added through a semantic residual pathway, preserving the semantic representation while allowing complementary detail to contribute adaptively.

Frequency Alignment Module (FAM) determines how the fused image evidence should condition the evolving label state. FAM combines the noisy label, diffusion timestep, and auxiliary class prior to produce state-dependent modulation parameters for the fused image condition. The resulting scale and shift operations allow the same image evidence to contribute differently as the label representation changes throughout the reverse trajectory. AFG and FAM therefore address complementary aspects of adaptive frequency conditioning: AFG performs semantic-guided frequency selection, whereas FAM aligns the selected evidence with the current state of label recovery.

Our contributions are threefold. 1) We introduce explicit frequency conditioning for ultrasound label diffusion, using wavelet-reconstructed detail components to complement semantic image features without changing the label diffusion formulation. 2) We develop two complementary modules for adaptive conditioning: AFG selects high-frequency evidence according to semantic context, while FAM aligns the resulting condition with the noisy label, diffusion timestep, and auxiliary class prior. 3) We evaluate WDFNet on six medical image datasets and use wavelet ablations, progressive guidance comparisons, and label-noise prediction analyses to examine the contribution and empirical scope of the proposed design.

2 Related Work

2.1 Label Diffusion Classification

Diffusion models were introduced as generative models (Ho et al., 2020), but their denoising objectives have also been adapted for classification and conditional prediction (Li et al., 2023; Clark & Jaini, 2023). CARD formulates classification and regression as diffusion over target variables conditioned on the input (Han et al., 2022). DiffMIC and DiffMIC-v2 specialize this idea to medical image classification by diffusing class label representations while using image features and auxiliary predictions as conditions (Yang et al., 2023; 2025). WDFNet inherits this label space formulation; its contribution is an adaptive frequency image condition rather than a new forward diffusion process.

2.2 Frequency Based Medical Image Representation

Frequency transforms provide explicit access to spatial scales that may remain implicit in convolutional features. Adaptive wavelet pooling and frequency based networks show that learned models can benefit from separating or reweighting frequency components (Wolter & Garcke, 2021; Qin et al., 2021). In medical imaging, MedSegDiff uses a frequency parser to suppress distracting components during diffusion-based segmentation (Wu et al., 2024), whereas USFM introduces joint spatial and frequency masked modeling for ultrasound representation learning (Jiao et al., 2024). These methods motivate explicit frequency representations, but they differ from WDFNet in task and output space.

Frequency information serves different roles in these prediction pathways. FcaNet constructs channel attention descriptors from discrete cosine components (Qin et al., 2021); MedSegDiff filters dynamic conditioning features (Wu et al., 2024). WDFNet instead selects wavelet-reconstructed image detail to condition class-vector denoising. Its distinction is this conditioning design, not frequency modeling or adaptive conditioning in general.

2.3 Wavelet Diffusion and Ultrasound Speckle Modeling

WaveDM and DiffLL incorporate wavelet subbands into image restoration or enhancement diffusion models (Huang et al., 2024; Jiang et al., 2023). SDDPM instead defines a diffusion process for signal dependent multiplicative speckle denoising (Guha & Acton, 2023), and Speckle2Self learns

ultrasound despeckling without clean targets (Li et al., 2025). These approaches output restored images and are therefore not classification baselines matched to our task. WDFNet neither generates a despeckled image nor places its diffusion trajectory in wavelet space; it reconstructs high-frequency subbands only to condition label denoising.

3 METHODOLOGY

WDFNet is a conditional label diffusion classifier whose image guidance pathway combines semantic and high-frequency evidence. As shown in Fig. 2, an image encoder provides a semantic feature and an auxiliary class prior, AFG constructs an adaptive frequency condition, and FAM injects that condition into the label-noise estimator. Both forward and reverse diffusion act on class label vectors rather than on image pixels.

Training phase. The image I is processed once by the image encoder and AFG to obtain an image guidance vector V_g. An auxiliary classifier produces the prior $\hat{y}$, while Gaussian noise is added to the one-hot label y_0 to sample a noisy label y_t. The denoiser predicts the sampled label noise from $(y_t, t, \hat{y}, V_g)$ and is trained with a mean squared error objective. The auxiliary classifier is optimized separately with cross entropy loss.

Inference phase. The image-dependent quantities V_g and $\hat{y}$ are cached because they remain fixed across reverse steps. Sampling starts from a prior-centered noisy label vector, $y_T = \hat{y} + z$, with $z \sim \mathcal{N}(0, \mathbf{I})$. The learned reverse process refines the class representation using the configured sampling steps (Appendix A). Frequency guidance changes the condition used during label denoising; it does not denoise or modify the input ultrasound image.

3.1 CONDITIONAL LABEL DIFFUSION FORMULATION

The classifier follows the conditional label diffusion formulation of CARD and DiffMIC (Han et al., 2022; Yang et al., 2023). Let $y_0 \in \mathbb{R}^K$ denote a one-hot label for K classes and let $\hat{y} = p_\psi(I)$ denote the auxiliary prediction for image I. For a noise schedule $\{\alpha_t\}_{t=1}^T$ with $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, the forward process samples

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \left(1 - \sqrt{\bar{\alpha}_t}\right) \hat{y} + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

This construction gradually perturbs the target label around the image-dependent prior. Importantly, the input image is used only to compute $\hat{y}$ and the guidance vector V_g; the Gaussian variable ϵ is never added to the image.

The denoiser ϵ_θ estimates the sampled label noise from the label state and the image condition:

$$\hat{\epsilon} = \epsilon_\theta(y_t, t, \hat{y}, V_g), \qquad \mathcal{L}_{\text{diff}} = \mathbb{E}_{y_0, I, t, \epsilon}\left[\|\epsilon - \hat{\epsilon}\|_2^2\right]. \tag{2}$$

The auxiliary classifier is optimized with $\mathcal{L}_{\text{aux}} = \text{CE}(p_\psi(I), y_0)$ using its own optimizer. It provides the semantic feature and class prior, whereas the diffusion network learns conditional label-noise estimation. At inference, the estimated clean label and one reverse sample follow CARD and DiffMIC (Han et al., 2022; Yang et al., 2023):

$$\tilde{y}_0 = \frac{y_t - (1 - \sqrt{\bar{\alpha}_t})\hat{y} - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(y_t, t, \hat{y}, V_g)}{\sqrt{\bar{\alpha}_t}}, \tag{3}$$

$$y_{t-1} = \gamma_{0,t}\tilde{y}_0 + \gamma_{1,t} y_t + \gamma_{2,t}\hat{y} + \sqrt{\tilde{\beta}_t}\, z, \qquad z \sim \mathcal{N}(0, \mathbf{I}), \tag{4}$$

where $\gamma_{0,t} = \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}$, $\gamma_{1,t} = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t}$, $\gamma_{2,t} = 1 + \frac{(\sqrt{\bar{\alpha}_t}-1)(\sqrt{\alpha_t}+\sqrt{\bar{\alpha}_{t-1}})}{1-\bar{\alpha}_t}$, and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)$. The final step returns $\tilde{y}_0$ without additional sampling noise.

Discussion. Label diffusion separates the prediction state from the image domain. Gaussian noise is added only to the class representation, so ultrasound structures are never corrupted by the forward process. This formulation also isolates our contribution: WDFNet changes the image condition supplied to an established label denoising process rather than introducing a new diffusion trajectory.

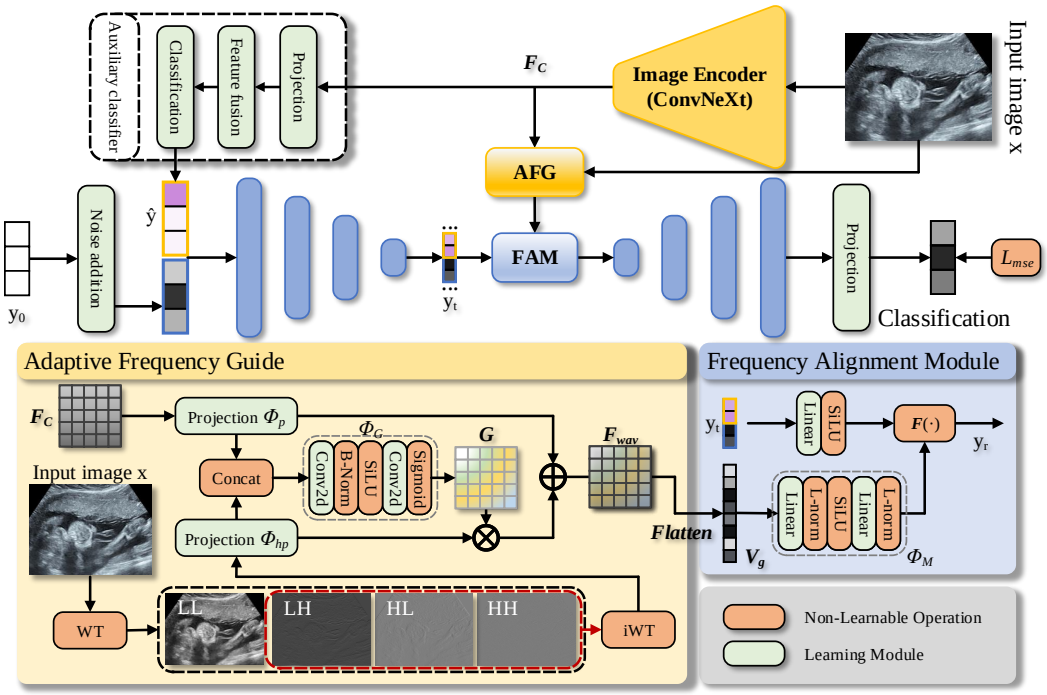

Figure 2: Overview of WDFNet. A ConvNeXt encoder produces semantic features and an auxiliary class prior. AFG combines semantic features with wavelet-reconstructed high-frequency features, and FAM modulates the resulting image condition according to the noisy label, timestep, and auxiliary prior to predict label noise.

3.2 Adaptive Frequency Guide

AFG constructs an image condition that exposes high-frequency cues while retaining a semantic residual path. Given an image $I \in \mathbb{R}^{C\times H\times W}$, a one level Haar transform produces the approximation subband I_{LL} and the horizontal, vertical, and diagonal detail subbands I_{LH}, I_{HL}, and I_{HH}. We set the approximation coefficients to zero and apply the inverse transform to reconstruct a spatially aligned detail image,

$$I_h = \mathcal{W}^{-1}(0, I_{LH}, I_{HL}, I_{HH}). \tag{5}$$

This operation makes the detail coefficients available to a conventional convolutional branch without treating them as an independently identifiable physical noise source.

The reconstructed image and the ConvNeXt semantic feature (Liu et al., 2022) are projected into a common feature space as $F_h = \Phi_h(I_h)$ and $F_s = \Phi_s(E(I))$. Because F_h can contain useful boundaries and texture as well as nuisance variation, AFG predicts an elementwise gate from both streams and preserves F_s as a residual:

$$G = \sigma(\Phi_g([F_s; F_h])), \tag{6}$$

$$F_{\text{fuse}} = F_s + G \odot F_h, \tag{7}$$

$$F_c = F_{\text{fuse}} + \text{GAP}(F_{\text{fuse}}). \tag{8}$$

Here $[\cdot;\cdot]$ denotes channel concatenation, $\odot$ denotes elementwise multiplication, and the pooled context is broadcast over spatial locations. The implementation learns where and how strongly frequency evidence should modify the semantic representation rather than assuming that every detail coefficient is beneficial.

Discussion. The inverse transform serves a representational rather than a denoising role. It converts the three detail subbands into a spatially aligned image that can be processed by a conventional branch. The learned gate then treats high-frequency responses as candidates rather than universally useful features. This distinction is important in ultrasound, where the same band can contain

anatomical boundaries, texture, and acquisition related variation. A qualitative visualization of the image-conditioned AFG gate is provided in Appendix E.

3.3 Frequency Alignment Module

FAM aligns the image condition with the evolving label state through time dependent feature modulation. The fused map is flattened and projected as $V_g = \Phi_M(\mathrm{Flatten}(F_c))$. The noisy label and auxiliary prior are concatenated, processed with a timestep-conditioned layer, layer normalization, and a SiLU activation to obtain the state vector z_t. An MLP maps z_t to scale and shift vectors that modulate the image guidance vector:

$$[s_t; b_t] = \Phi_{\mathrm{mod}}(z_t), \qquad h_t = V_g \odot (1 + s_t) + b_t. \tag{9}$$

Subsequent timestep-conditioned layers map h_t to $\hat{\epsilon}$. FAM consequently lets the same image condition contribute differently as the uncertainty of the label representation changes across reverse steps.

Discussion. AFG determines what image evidence is retained, whereas FAM determines how that evidence participates in the current reverse step. The scale and shift operation gives the denoiser a direct state-dependent adjustment without recomputing the image encoder. FAM therefore complements frequency selection rather than duplicating it.

4 Experiments

4.1 Experimental Protocol

Ultrasound datasets. The primary evaluation covers three ultrasound classification datasets. PMG2000 contains 2,098 placental ultrasound images in four classes. Following DiffMIC (Yang et al., 2023; 2025), we adopt the same training/test partition to ensure consistent evaluation. UCPM2200 comprises 2,171 placental ultrasound images in four maturity grades, with 1,738 training and 433 test images, collected under institutional review board approval. BUSI contains 780 breast ultrasound images from 600 women and defines normal, benign, and malignant classes (Al-Dhabyani et al., 2020). Dataset partitions are summarized in Appendix B.

Additional datasets. We also evaluate datasets for CT, dermoscopy, and MRI. Chest-CT contains 1,000 images of adenocarcinoma, large-cell carcinoma, squamous-cell carcinoma, and normal cases, with 613 training, 72 validation, and 315 test images (Hany, 2020). We use the predefined training and test partitions without using the validation set. Skin-Dermoscopic is a nine-class repackaged collection of ISIC images, with 2,239 training and 118 test images (rajivaiml, n.d.). This partition is distinct from the official ISIC challenge split. Brain-MRI contains glioma, meningioma, no-tumor, and pituitary-tumor classes, with 2,870 training and 394 test images (Bhuvaji et al., 2020).

Comparisons. We compare against convolutional and transformer classifiers—ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), Swin (Liu et al., 2021), PVT (Wang et al., 2021), GMIC (Shen et al., 2021), and ConvNeXt (Liu et al., 2022)—and the label diffusion classifiers DiffMIC and DiffMIC-v2 (Yang et al., 2023; 2025). Within each matched guidance comparison, variants share data augmentation, the training schedule, checkpoint-selection rule, and evaluation procedure. Appendix B specifies the checkpoint rule and distinguishes the full classification and compact diagnostic protocols.

Implementation. Images are center cropped and resized to 224×224, with random horizontal flips and rotations within 30° during training. We use PyTorch (Paszke et al., 2019), batch size 16, and Adam (Kingma & Ba, 2015). The full configuration uses pretrained ConvNeXt-Tiny, 600 main epochs, 1,000 training timesteps, and 100 sampling steps. Initial learning rates are 10^{-3} for the denoiser and 2×10^{-4} for the auxiliary classifier. Accuracy and macro F1 are the primary metrics. Repeated PMG2000 comparisons report mean $\pm$ standard deviation over three seeds; original benchmark entries remain distinct. Checkpoint selection for PMG2000 follows the protocol of DiffMIC (Yang et al., 2023), while all matched guidance variants within each dataset share the same checkpoint-selection rule and evaluation procedure.

4.2 Quantitative Analysis

WDFNet consistently improves the evaluated ultrasound benchmarks. Tables 1 and 2 report the benchmark results, using the three-run mean for WDFNet on PMG2000. WDFNet obtains the highest

Table 1: Accuracy and macro F1 on three ultrasound datasets. WDFNet's PMG2000 entry uses the three-run mean; baseline entries retain original benchmark values. The repeated DiffMIC-v2 comparison is reported separately in Table 4. Best reported values are highlighted.

Model	PMG2000		UCPM2200		BUSI	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
ResNet (He et al., 2016)	0.879	0.881	0.888	0.885	0.861	0.860
ViT (Dosovitskiy et al., 2021)	0.886	0.890	0.893	0.891	0.873	0.869
Swin (Liu et al., 2021)	0.893	0.892	0.891	0.889	0.852	0.849
PVT (Wang et al., 2021)	0.907	0.902	0.898	0.897	0.805	0.801
GMIC (Shen et al., 2021)	0.900	0.901	0.903	0.900	0.877	0.873
ConvNeXt (Liu et al., 2022)	0.916	0.916	0.906	0.907	0.881	0.880
DiffMIC (Yang et al., 2023)	0.931	0.926	0.919	0.911	0.895	0.871
DiffMIC-v2 (Yang et al., 2025)	0.935	0.933	0.924	0.920	0.901	0.889
Ours	**0.960**	**0.955**	**0.938**	**0.933**	**0.923**	**0.910**

Table 2: Accuracy and macro F1 score on three additional medical image datasets. The best result in each column is highlighted.

Model	Chest-CT		Skin-Dermoscopic		Brain-MRI	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
ResNet (He et al., 2016)	0.759	0.736	0.857	0.852	0.783	0.779
ViT (Dosovitskiy et al., 2021)	0.743	0.736	0.862	0.855	0.786	0.781
Swin (Liu et al., 2021)	0.721	0.714	0.859	0.849	0.789	0.782
PVT (Wang et al., 2021)	0.829	0.831	0.873	0.867	0.801	0.796
GMIC (Shen et al., 2021)	0.882	0.887	0.890	0.886	0.811	0.807
ConvNeXt (Liu et al., 2022)	0.917	0.917	0.915	0.913	0.814	0.813
DiffMIC (Yang et al., 2023)	0.899	0.899	0.912	0.909	0.819	0.815
DiffMIC-v2 (Yang et al., 2025)	0.902	0.905	0.918	0.911	0.823	0.819
Ours	**0.924**	**0.919**	**0.924**	**0.921**	**0.825**	**0.820**

reported accuracy and macro F1 in every listed dataset. The accuracy margin over DiffMIC-v2 is largest on PMG2000 and remains positive on UCPM2200 and BUSI. Because the diffusion baselines already operate in label space, these comparisons evaluate the proposed image conditioning pathway rather than the introduction of label diffusion itself.

The gains also appear on the three additional modalities, but they should be interpreted as evidence across multiple datasets rather than cross domain generalization. WDFNet improves the evaluated baselines on Chest-CT and Skin-Dermoscopic and yields a smaller gain on Brain-MRI. These experiments show that adaptive frequency guidance can be useful outside ultrasound under separate training for each dataset; they do not establish zero-shot or distribution shift generalization.

The improvement over DiffMIC-v2 remains stable across the evaluated seeds. Under the same PMG2000 training and checkpoint selection protocol, WDFNet obtains 0.960 ± 0.0015 accuracy and 0.955 ± 0.0013 macro F1 over three seeds, compared with 0.934 ± 0.0016 and 0.932 ± 0.0017 for DiffMIC-v2. WDFNet's means are used consistently in Tables 1 and 3 and Fig. 4(a).

4.3 Ablation Study and Model Analysis

Adaptive fusion and wavelet detail each improve the progressive guidance comparison. With original-image branch inputs, replacing fixed addition (C1) with adaptive gating (C2) increases accuracy/F1 from 0.924/0.918 to 0.939/0.931. Replacing the raw input with wavelet detail while retaining the gate yields WDFNet's 0.960/0.955. Label diffusion and FAM remain present throughout; these comparisons test fusion under raw inputs and the frequency representation under adaptive fusion, not removal of FAM. The ConvNeXt row is a separate reference (Appendix B).

Table 3: Progressive image guidance comparison on PMG2000. C1, C2, and WDFNet retain label diffusion and FAM; only the second-branch input and fusion differ. Their scores are run means. ConvNeXt is a separately evaluated reference, not a matched no-diffusion ablation.

Configuration	Accuracy	Macro F1
ConvNeXt (reference)	0.916	0.916
C1: Fixed fusion	0.924	0.918
C2: Raw spatial	0.939	0.931
WDFNet	**0.960**	**0.955**

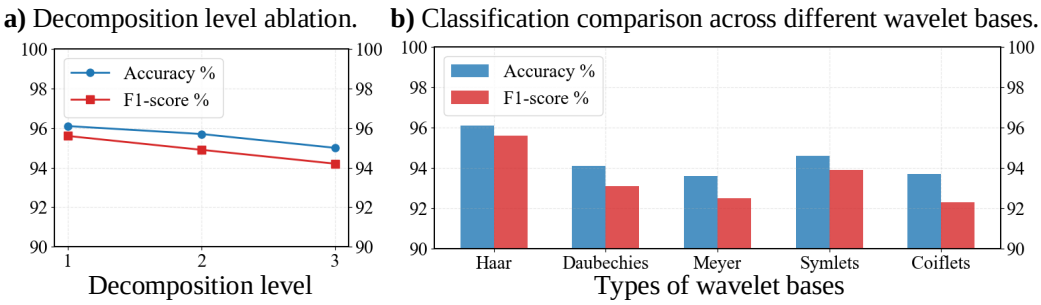

Figure 3: **a)** Decomposition level ablation on PMG2000. **b)** Accuracy and macro F1 score across the evaluated wavelet bases. These are the original single-run wavelet comparisons, not the repeated guidance comparison in Table 3.

The raw spatial comparison controls for a second learned image branch. C2 and WDFNet retain the same branch architecture, feature dimensions, adaptive gate, and label denoiser. Their difference is the branch input. The comparison therefore tests the frequency representation within the same learned architecture, rather than attributing a gain to additional branch capacity.

A shallow decomposition is preferred on PMG2000. Fig. 3(a) shows that increasing the Haar decomposition from one to three levels reduces accuracy from 0.960 to 0.950 and macro F1 score from 0.955 to 0.942. Deeper decomposition moves the selected detail bands to coarser spatial scales, which may reduce the fine detail available to the gate. We use one level in all main experiments.

Haar performs best among the evaluated wavelet bases, but the comparison does not establish universal optimality. Fig. 3(b) compares Haar, Daubechies, Symlets, Coiflets, and Meyer wavelets under the same PMG2000 setting. Haar obtains the highest observed accuracy and macro F1 score. This empirical result motivates our default choice; different datasets, resolutions, or acquisition devices may favor other bases.

Inverse reconstruction improves the usability of detail coefficients in the original single-run comparison. Reconstructing the detail subbands yields 0.960 accuracy and 0.955 F1, compared with 0.945 and 0.934 without the inverse transform. These original ablation values are distinct from the three-run means in Table 3. The comparison supports the image-aligned representation in Eq. (5).

4.4 Controlled Corruption and Noise Prediction

WDFNet retains higher mean classification accuracy under the evaluated corruptions. Fig. 4(a) extends the configurations in Table 3 to severe speckle and Gaussian corruption. WDFNet outperforms both controls in each condition. Raw spatial exceeds fixed fusion on original and severe-speckle inputs, but not on Gaussian inputs; adaptive gating alone therefore does not improve every corruption condition. WDFNet's comparison with raw spatial supports the wavelet representation beyond additional branch capacity.

The controlled corruption protocol changes the image condition while preserving the label diffusion task. Mild and severe multiplicative speckle use severity parameters of 0.10 and 0.20, respectively. The Gaussian sampling standard deviation is set per image to the standard deviation of the severe-

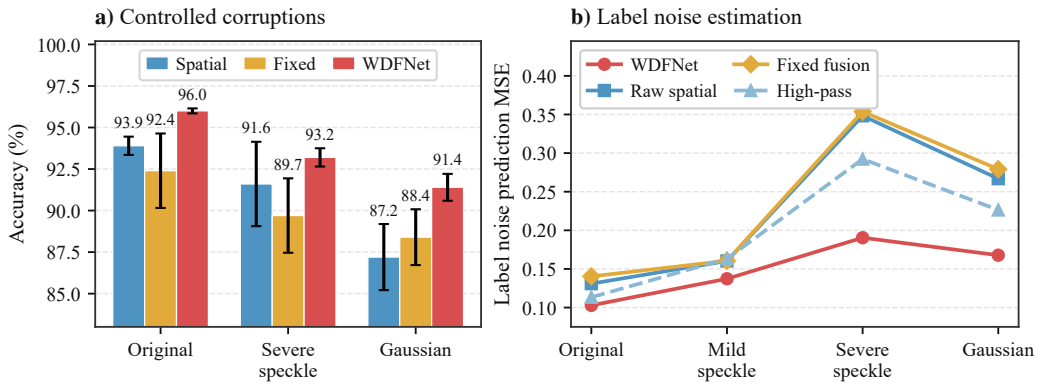

Figure 4: Controlled evidence on PMG2000. **a)** Classification accuracy on original images and under controlled corruptions; error bars show standard deviations over three seeds. Spatial and Fixed denote the raw spatial and fixed fusion configurations. **b)** Paired label-noise prediction MSE across image conditions; lower is better. Original denotes no added corruption; Gaussian denotes additive Gaussian corruption. See Appendix C for pairing and operator definitions.

speckle residual before clipping. Each classification condition in Fig. 4(a) is evaluated over the same three seeds and reports mean $\pm$ standard deviation. These corruptions are diagnostic stress tests rather than models of every clinical acquisition process.

The paired MSE diagnostic measures label-noise estimation under matched noise draws. All variants receive the same cases, timestep indices, and sampled label noise, while retaining their own auxiliary priors; their noisy-label states need not be identical. Squared prediction error is averaged over classes, five noise draws, five timesteps, and images. It measures neither image reconstruction error nor classification accuracy, and repeated noise draws do not constitute independent training runs.

WDFNet achieves the lowest observed label-noise prediction MSE under all four evaluated image conditions. WDFNet obtains MSE values of 0.103, 0.137, 0.190, and 0.168 on original images, mild speckle, severe speckle, and Gaussian corruption in Fig. 4(b). Under severe speckle, the corresponding values are 0.348 for raw spatial, 0.354 for fixed fusion, and 0.292 for High-pass. The latter substitutes a high-pass residual for wavelet detail to test an alternative high-frequency representation, not all possible high-pass filters. Appendix C specifies the operator. These point estimates do not establish statistical significance or universal superiority of Haar.

5 Limitations

WDFNet adapts the contribution of frequency features but uses a fixed, single-level Haar representation. Other spatial scales or bases may be preferable for different anatomy and acquisition settings, motivating adaptive multiscale extensions. Its learned gate serves classification rather than physical separation of anatomical detail and acquisition noise.

6 Conclusion

We introduced WDFNet, a label diffusion classifier that enriches semantic image conditioning with adaptively gated, wavelet-reconstructed high frequency cues. AFG determines how frequency evidence modifies the semantic representation, and FAM aligns the fused condition with the noisy label and diffusion timestep. Across six medical image datasets, WDFNet improves the evaluated discriminative and diffusion-based classifiers. Repeated runs, spatial controls, fixed-fusion controls, and corruption experiments support adaptive conditioning within the shared label-diffusion framework while defining its empirical scope. Future work will evaluate the design across centers and acquisition devices and extend adaptive frequency conditioning to dense prediction tasks.

AI USE STATEMENT

Generative AI tools were used primarily to polish manuscript language. They also assisted with experimental discussions, interpretation of exploratory results, and figure preparation. The authors are responsible for the final text, code, and reported findings.

ETHICS STATEMENT

WDFNet is intended as a research stage decision support method rather than an autonomous diagnostic system. Although improved classification could assist clinical image review, errors under distribution shift, unequal subgroup performance, or excessive reliance on model outputs could cause harm. Prospective use would therefore require privacy preserving data governance, external and subgroup validation, calibrated uncertainty, and clinician oversight.

REPRODUCIBILITY STATEMENT

Section 3 defines the label diffusion formulation and conditioning modules. Appendices A–D document implementation details, data protocols, corruption operators, and numerical results. They distinguish main classification from the compact diagnostic and identify unresolved provenance limitations. An anonymized executable reproduction package remains to be finalized.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (MRI). Kaggle dataset, 2020. URL https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.

Christoph B. Burckhardt. Speckle in ultrasound b-mode scans. *IEEE Transactions on Sonics and Ultrasonics*, 25(1):1–6, 1978. doi: 10.1109/T-SU.1978.30978.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36:58921–58937, 2023.

V. Damerjian, O. Tankyevych, N. Souag, and E. Petit. Speckle characterization methods in ultrasound images—a review. *IRBM*, 35(4):202–213, 2014. ISSN 1959-0318. doi: https://doi.org/10.1016/j.irbm.2014.05.003. URL https://www.sciencedirect.com/science/article/pii/S1959031814000797.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Soumee Guha and Scott T. Acton. Sddpm: Speckle denoising diffusion probabilistic models. *arXiv preprint arXiv:2311.10868*, 2023.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

Mohamed Hany. Chest ct-scan images dataset. Kaggle, 2020. URL https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yi Huang, Jiancheng Huang, Jianzhuang Liu, Mingfu Yan, Yu Dong, Jiaxi Lv, Chaoqi Chen, and Shifeng Chen. Wavedm: Wavelet-based diffusion models for image restoration. *IEEE Transactions on Multimedia*, 26:7058–7073, 2024.

Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics*, 42(6):1–14, 2023. doi: 10.1145/3618373.

Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96: 103202, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Xuesong Li, Nassir Navab, and Zhongliang Jiang. Speckle2self: Self-supervised ultrasound speckle reduction without clean data. *Medical Image Analysis*, 106:103755, 2025. doi: 10.1016/j.media.2025.103755. URL https://doi.org/10.1016/j.media.2025.103755.

Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022.

Frederic L Lizzi, Ernest J Feleppa, S Kaisar Alam, and Cheri X Deng. Ultrasonic spectrum analysis for tissue evaluation. *Pattern Recognition Letters*, 24(4):637–658, 2003. ISSN 0167-8655. doi: https://doi.org/10.1016/S0167-8655(02)00172-1. URL https://www.sciencedirect.com/science/article/pii/S0167865502001721.

J Alison Noble and Djamal Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on medical imaging*, 25(8):987–1010, 2006.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 783–792, October 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

rajivaiml. ISIC skin cancer dataset. Kaggle dataset, n.d. URL https://www.kaggle.com/datasets/rajivaiml/isic-skin-cancer-dataset.

Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Gene Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101908. URL https://www.sciencedirect.com/science/article/pii/S1361841520302723.

Johan M Thijssen. Ultrasonic speckle formation, analysis and processing applied to tissue characterization. *Pattern Recognition Letters*, 24(4-5):659–675, 2003.

Robert F Wagner, Stephen W Smith, John M Sandrik, and Hector Lopez. Statistics of speckle in ultrasound b-scans. *IEEE Transactions on sonics and ultrasonics*, 30(3):156–163, 1983. doi: 10.1109/T-SU.1983.31404.

Robert F. Wagner, Michael F. Insana, and David G. Brown. Unified approach to the detection and classification of speckle texture in diagnostic ultrasound. *Optical Engineering*, 25(6):738–742, 1986. doi: 10.1117/12.7973899.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 568–578, October 2021.

Moritz Wolter and Jochen Garcke. Adaptive wavelet pooling for convolutional neural networks. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1936–1944. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wolter21a.html.

Junde Wu, RAO FU, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant (eds.), *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 1623–1639. PMLR, 2024. URL https://proceedings.mlr.press/v227/wu24a.html. MIDL 2023, held 10–12 July 2023; proceedings published in 2024.

Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International conference on medical image computing and computer-assisted intervention*, pp. 95–105. Springer, 2023.

Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Zhaohu Xing, and Lei Zhu. Diffmic-v2: Medical image classification via improved diffusion network. *IEEE Transactions on Medical Imaging*, 44 (5):2244–2255, 2025.

A Implementation Details

Image preprocessing. Images are converted to RGB, cropped to a centered square, and resized to 224×224. Training uses random horizontal flipping and rotation within 30° across all six datasets, with the same augmentation for the compared methods. Channel normalization uses means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. Evaluation omits random augmentation. Corruptions are applied to the resized image in $[0, 1]$ before normalization.

Encoder and initialization. The encoder is ConvNeXt-Tiny with the timm identifier `convnext_tiny.in12k_ft_in1k`. The full configuration enables its ImageNet-12k pretrained, ImageNet-1k fine-tuned weights; the compact controlled diagnostic disables external pretraining. Both use 224×224 inputs rather than the weight identifier's reference resolution.

Frequency branch. The implementation uses one-level Haar decomposition with zero boundary extension. Approximation coefficients are set to zero before inverse reconstruction. Four 3×3 convolutions with stride 2 and padding 1 process the reconstructed image, with output channel counts 64, 128, 256, and 256. Each is followed by batch normalization and SiLU. Adaptive average pooling produces a 7×7 map, followed by batch normalization and SiLU. A 1×1 convolution, batch normalization, and SiLU project the final 768-channel ConvNeXt feature map to 256 channels.

Gate and vector projection. The gate processes the concatenated 512-channel semantic and frequency maps using a 3×3 convolution to 256 channels, batch normalization, SiLU, a 1×1 convolution, and sigmoid. Its output has shape $256 \times 7 \times 7$. After gated residual fusion and broadcast global average pooling, the flattened feature passes through a linear layer to 2,048 units, layer normalization, SiLU, a linear layer to the configured condition width, and layer normalization. The preserved full classification and compact diagnostic configurations use widths of 6,144 and 1,024, respectively.

State-dependent modulation. The noisy label and auxiliary probabilities are concatenated. A conditional linear layer multiplies its output by a learned timestep embedding, followed by layer normalization and SiLU. A linear map and SiLU produce FAM's scale and shift vectors. Two further conditional linear layers with layer normalization and SiLU precede the final class-dimensional prediction layer. AFG depends on the image; FAM additionally depends on the label state and timestep. AFG gate maps therefore are not timestep-varying attention maps.

Training updates. The denoiser minimizes squared noise prediction error averaged over batch and class dimensions. The implementation uses zero-based timestep indices, hence antithetic pairs are represented as t and $T - 1 - t$. It evaluates the auxiliary classifier to form the condition for a denoiser update, then updates the auxiliary classifier separately with cross entropy. The preserved configurations use Adam with $\beta_1 = 0.9$, $\epsilon = 10^{-8}$, zero weight decay, and gradient norm clipping at 1.0. Initial learning rates are 10^{-3} and 2×10^{-4} for the denoiser and auxiliary classifier. The auxiliary optimizer enables AMSGrad.

Sampling and caching. The auxiliary prediction and image condition can be reused across reverse steps. The sampler initializes the label state as $\hat{y} + z$, where $z \sim \mathcal{N}(0, \mathbf{I})$, and omits added noise in the last update. The locally retained sampler iterates over the requested number of contiguous schedule indices. A 100-step run should therefore not be assumed to use DDIM or uniformly respaced indices from a 1,000-step schedule. Exact sampler and checkpoint versions are necessary for executable reproduction.

B Data and Evaluation Protocols

Image partitions. Separate models are trained for each dataset. BUSI uses 586 training and 194 test images, with benign/malignant/normal counts of 328/158/100 and 109/52/33. UCPM2200 uses 1,738 training and 433 test images. Skin-Dermoscopic has 2,239 training and 118 test images; Brain-MRI has 2,870 training and 394 test images. Each matched comparison uses the same partition across guidance variants.

Progressive guidance configurations. Table 3 compares three label diffusion configurations that retain FAM. C1 (Fixed fusion) supplies the original image to the second branch and uses fixed addition. C2 (Raw spatial) retains that input and enables adaptive gating. WDFNet retains the gate and instead supplies wavelet-reconstructed detail. C1 versus C2 evaluates adaptive fusion for raw inputs; C2 versus WDFNet evaluates the frequency representation with adaptive fusion. Neither comparison removes FAM. The separately evaluated ConvNeXt row is a reference, not a controlled estimate of the effect of adding label diffusion.

Distinct training scales. The preserved full classification configuration uses 600 main epochs, width 6,144, 1,000 training timesteps, and 100 sampling steps, with 10 auxiliary initialization epochs and 40 denoiser warmup epochs. The compact noise diagnostic uses 160 main epochs, width 1,024, 100 training timesteps, 15 auxiliary initialization epochs, and 20 warmup epochs, without external pretrained initialization. Both use batch size 16. The compact diagnostic is not the main classification experiment repeated at the same training scale.

Metrics and uncertainty. Accuracy is the proportion of correct image predictions; macro F1 averages classwise F1 scores with equal weights. WDFNet's PMG2000 entries in Tables 1 and 3 use the three-run means 0.960 accuracy and 0.955 macro F1, consistent with Fig. 4(a). C1 and C2 report 0.924/0.918 and 0.939/0.931, respectively. Available standard deviations are given below; no missing F1 standard deviations are inferred. The original wavelet and inverse-reconstruction comparisons retain their single-run values, including 0.961/0.956 for the original full configuration. Repeated label-noise draws within one model measure evaluation variability, not variation across independent training runs. Updated MSE values are point estimates without confidence intervals.

C Controlled Corruption and Noise Prediction

Corruption operators. For $x \in [0,1]^{3\times224\times224}$, the preserved loader forms log-normal multiplicative speckle as $x_s = x \odot \exp(sz - s^2/2)$, where entries of z are standard normal, and clips the result to $[0,1]$. Mild and severe conditions use $s = 0.10$ and $s = 0.20$. For Gaussian corruption, it computes $\sigma_x = \mathrm{Std}(x_{0.20} - x)$ before clipping and applies $\mathrm{clip}(x + \sigma_x z', 0, 1)$ using another standard normal draw. Standard deviation is computed over all image entries using the population convention. This matches the Gaussian sampling scale to the realized speckle residual, not the exact realized variances after clipping. RGB entries are perturbed separately. Original denotes no additional corruption, not a noise-free acquired image.

Pairing and averaging. Image corruption uses seed 2,026 plus the image-list index. The preserved diagnostic evaluates five label-noise draws per image at schedule indices $\{10, 25, 50, 75, 90\}$ of a 100-step trained model. Label-noise generators use base seed 8,080 with deterministic case and replicate offsets, providing the same noise vectors to all variants. Each model retains its own auxiliary prior. Thus, fixed image, label, timestep, and noise do not imply identical noisy-label states across variants or corruption conditions, since the prior in the forward formula can change.

D Supplementary Numerical Results

Table 4: Repeated PMG2000 classification results. Mean ± standard deviation across three training runs, expressed in percentage points. These are retained reported summaries, not new evaluations.

Method	Accuracy (%)	Macro F1 (%)
DiffMIC-v2	93.4 ± 0.16	93.2 ± 0.17
WDFNet	96.0 ± 0.15	95.5 ± 0.13

Metric definition. The diagnostic averages $(\epsilon_{irk} - \hat{\epsilon}_{itrk})^2$ with equal weights over images i, noise draws r, classes k, and evaluated timesteps t. This is label-noise prediction MSE, not image reconstruction error, final label MAE, or classification accuracy. The updated values were supplied from a rerun using the same training and evaluation configuration. Confidence intervals from earlier results are not transferred to the updated point estimates.

Table 5: Classification accuracy (%) underlying Fig. 4(a). Mean ± standard deviation across three training runs. C1 and C2 correspond to the configurations in Table 3.

Configuration	Original	Severe speckle	Gaussian
WDFNet	96.0 ± 0.15	93.2 ± 0.55	91.4 ± 0.81
C2: Raw spatial	93.9 ± 0.55	91.6 ± 2.54	87.2 ± 1.99
C1: Fixed fusion	92.4 ± 2.24	89.7 ± 2.24	88.4 ± 1.68

Table 6: Label-noise prediction MSE underlying Fig. 4(b). Values are point estimates. High-pass uses the mean-filter residual defined in Appendix C.

Configuration	Original	Mild speckle	Severe speckle	Gaussian
WDFNet	0.102840	0.137298	0.190423	0.167786
Raw spatial	0.130964	0.160313	0.348283	0.266880
Fixed fusion	0.140385	0.160575	0.353801	0.279131
High-pass	0.113306	0.163260	0.292255	0.226384

Control definitions and scope. Fixed fusion, Raw spatial, and WDFNet form the progressive configurations defined in Table 3; all retain the label diffusion formulation and FAM. Fixed fusion versus WDFNet changes both input representation and fusion and is not a gate-only comparison. High-pass instead retains the learned branch and adaptive gate while replacing wavelet detail with $x - \mathrm{AvgPool}_{7\times7}(x)$, using stride 1 and padding 3. This mean-filter residual tests whether another high-frequency representation yields comparable noise prediction error. It does not exhaust alternative filters or establish a universally optimal frequency transform.

Evidence boundary. The summaries distinguish classification from the denoising objective. They do not identify separate physical speckle and Gaussian components, establish universal optimality of Haar, or measure transfer between modalities.

E Qualitative Analysis of Adaptive Frequency Guidance

Visualization protocol. Fig. 5 shows the AFG gate for two distinct images from the PMG2000 test list. Both examples are correctly classified by each illustrated model; they are selected for a readable qualitative comparison and do not constitute a representative accuracy sample. The 7×7 model has the standard AFG grid. The 14×14, 28×28, and 56×56 models are separately trained exploratory variants using the same image lists, one-level Haar transform, ConvNeXt-T auxiliary encoder, and 600-epoch schedule. All four have the same trainable parameter count. The 14×14 variant upsamples the semantic feature before fusion, whereas the 28×28 and 56×56 variants also change frequency-branch strides to obtain their native grids.

Each heatmap is the mean of the 256 gate channels G in Eq. (8), shown at its native spatial grid without an ultrasound overlay or interpolation. Within each model, the display interval is shared across the two images and its limits are printed above the column. Different models require different intervals: visual contrast across columns must not be interpreted as a common numerical gate scale.

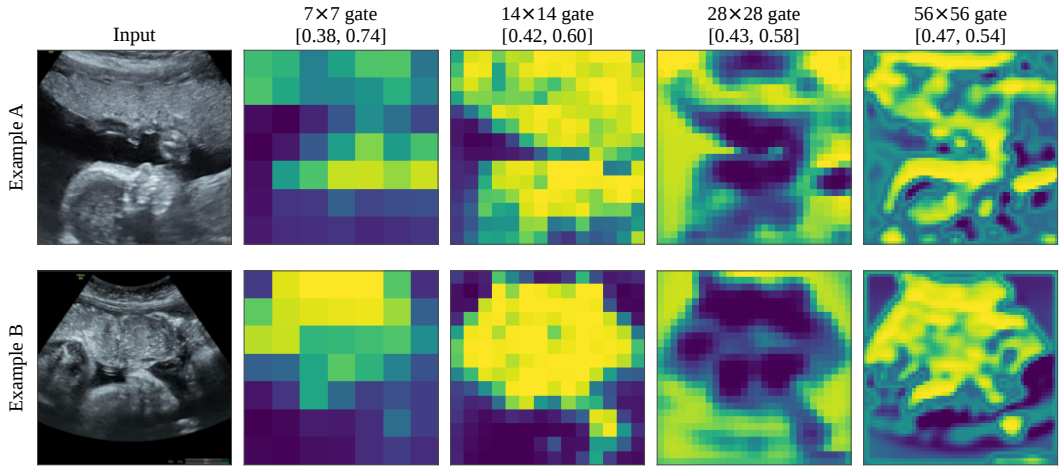

Figure 5: Image-conditioned AFG gates for two PMG2000 images. The spatial pattern differs between images within each trained model. The bracketed numbers give the display interval shared by each column's two gate maps; colors are not directly comparable across columns.