Wavelet-Based Mechanistic Interpretability of Vision Transformers via Frequency-Aware Ablations

Sophia J. Abraham[1], Jonathan D. Hauenstein[2], Walter J. Scheirer[1]
[1]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556
[2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556

Abstract

We explore a wavelet-based interpretability framework for Vision Transformers (ViT), aiming to analyze their reliance on frequency-specific representations. Through systematic ablations of wavelet subbands, we assess how different frequency components contribute to latent representations and attention mechanisms. Our empirical study on CIFAR-10 reveals that high-frequency details, particularly those captured by Haar wavelets, influence reconstruction fidelity and attention distributions. While preliminary findings suggest a frequency-dependent behavior in ViT representations, further investigation is needed to generalize across datasets and architectures. This study highlights the potential of frequency-based interpretability but also underscores the need for more rigorous evaluation in larger, more diverse settings. To encourage further exploration, all the experimentation and method code can be found on our GitHub repository[1].

1. Introduction

Vision Transformers (ViTs) leverage global self-attention mechanisms, making their internal representations challenging to interpret [4]. Unlike convolutional neural networks (CNNs), which process localized spatial features, ViTs operate on global token interactions, raising questions about their sensitivity to different frequency components in an image.

Existing interpretability methods focus largely on spatial analyses, such as attention visualization and patch-based ablations [1, 8], but often overlook frequency decomposition. This makes it unclear whether ViTs predominantly rely on high-frequency details (e.g., textures and edges) or low-frequency global structures for representation learning. Recent work has integrated wavelets into ViT architectures for efficiency improvements [13] and compositional analysis [10], but the direct impact of specific frequency components on ViT behavior remains underexplored.

To investigate this, we explore a *wavelet-based interpretability approach* that systematically ablates specific frequency subbands and examines their effect on ViT representations. By removing wavelet coefficients at different frequency levels and color channels, we aim to provide insights into how frequency-dependent features influence ViT attention patterns and semantic retention.

Our study contributes:

1. A **Wavelet-ViT Autoencoder** that provides a structured way to analyze how wavelet-decomposed image components are encoded in ViT latent representations.
2. A **systematic frequency ablation study**, measuring the effect of specific wavelet subbands on ViT reconstructions using CLIP similarity and mean squared error.
3. A **ViT attention analysis**, highlighting patterns in frequency-sensitive attention across different transformer layers.

Experiments on CIFAR-10 suggest that mid-to-high-frequency wavelet coefficients play a role in ViT feature representations and attention mechanisms. While these findings offer insights into ViT frequency sensitivity, further work is needed to confirm whether similar behaviors hold across datasets and architectures. Our approach serves as a first step toward frequency-aware interpretability in transformers, complementing traditional spatial-domain analysis.

2. Related Work

Wavelets in Deep Learning. Wavelet-based models have improved image generation [6, 9, 16] and restoration [15], leveraging frequency decompositions for more efficient sampling and reconstruction. However, these works focus on performance rather than interpretability. Our approach instead employs discrete wavelet transforms for *causal analysis*, systematically probing ViT reliance on specific frequency bands.

[1] https://github.com/sabraha2/Wavelet-Based-Mechanistic-Interpretability-of-Vision-Transformers-in-a-Latent-Diffusion-Setting

Frequency-Based Interpretability. Prior studies have examined neural networks' spectral sensitivity using Fourier analysis [2, 14], while wavelets have been explored for ViT compositionality [10]. However, these works provide mostly qualitative insights. We extend this by performing *quantitative ablations* of wavelet subbands to assess their direct impact on ViT representations.

Vision Transformer Interpretability. Existing ViT interpretability methods emphasize spatial analyses such as attention maps, token ablation, and feature attribution [1, 8, 12] but overlook frequency information. Our work addresses this gap by introducing *wavelet-based ablations*, explicitly isolating and evaluating frequency-dependent processing in ViTs.

CLIP as a Semantic Metric. CLIP-based similarity has been widely used for semantic evaluation in vision models [11]. Unlike prior works focusing on pixel-level fidelity, we use CLIP to *quantify semantic degradation* resulting from targeted frequency ablations, providing a novel perspective on how ViTs encode high- and low-frequency features. While wavelet-based techniques have been applied in generative tasks, our work is among the first to use them for mechanistic interpretability in ViTs, bridging frequency analysis with causal representation learning.

3. Method

Our approach integrates *wavelet decomposition* with a pretrained *Vision Transformer* (ViT) encoder and a *UNet-style decoder* for reconstruction. We systematically ablate specific frequency subbands to analyze their causal impact on pixel-wise accuracy (MSE), perceptual quality (VGG loss), and semantic alignment (CLIP similarity).

3.1. Wavelet Decomposition

We apply a two-level discrete wavelet transform (DWT, $\mathcal{W}$) to each RGB channel of an input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, yielding:

$$(\mathbf{L}, \{\mathbf{D}_1, \mathbf{D}_2\}) = \mathcal{W}(\mathbf{x}), \tag{1}$$

where $\mathbf{L}$ captures low-frequency structures, and $\mathbf{D}_i$ represent multi-scale high-frequency details (horizontal, vertical, and diagonal coefficients).

3.2. Hierarchical Latent Encoding

To extract frequency-informed representations, we resize $\mathbf{L}$ to 224 × 224 and process it through a pretrained ViT [4]:

$$\mathbf{z} = \mathrm{ViT}(\mathrm{Resize}(\mathbf{L})). \tag{2}$$

A learned projection maps $\mathbf{z} \in \mathbb{R}^{B\times 768}$ to a lower-dimensional latent space:

$$\mathbf{z}_{\text{latent}} = \mathbf{W}_p \mathbf{z}, \quad \mathbf{W}_p \in \mathbb{R}^{768\times 512}. \tag{3}$$

3.3. Wavelet-Based Reconstruction

A UNet-style decoder reconstructs the image $\hat{\mathbf{x}}$ from $\mathbf{z}_{\text{latent}}$, restoring spatial information across scales. The model is trained using a composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{perc}}\mathcal{L}_{\text{VGG}} + \lambda_{\text{edge}}\mathcal{L}_{\text{Sobel}}. \tag{4}$$

3.4. Wavelet Subband Ablation

To assess the role of specific frequency components, we systematically ablate detail coefficients at a chosen decomposition level i_{abl}:

$$\mathbf{D}_i^* = \begin{cases} \mathbf{D}_i, & \text{if } i \neq i_{\text{abl}} \\ \mathbf{0}, & \text{if } i = i_{\text{abl}} \end{cases} \implies \hat{\mathbf{x}}_{\text{abl}} = \mathcal{W}^{-1}(\mathbf{L}, \{\mathbf{D}_1^*, \mathbf{D}_2\}). \tag{5}$$

The reconstructed image $\hat{\mathbf{x}}_{\text{abl}}$ is evaluated for pixel-wise degradation (MSE) and semantic retention using CLIP similarity.

3.5. Semantic and Attention Analysis

To quantify semantic preservation, we compute the cosine similarity between the ablated reconstruction and a reference text prompt:

$$\mathrm{Sim}(\hat{\mathbf{x}}_{\text{abl}}, \text{"a photo of a scene"}). \tag{6}$$

Additionally, we inspect ViT attention maps before and after ablation, using attention rollout to reveal how token dependencies shift in response to frequency perturbations.

4. Experiments

We evaluate our wavelet-based interpretability framework on CIFAR-10, comparing *Haar* and *db2* wavelets in a two-level decomposition. Our experiments track training performance over 15 epochs, measuring mean squared error (MSE) and CLIP-based semantic similarity. We further conduct systematic frequency ablations to assess their impact on reconstruction quality and semantic alignment. Finally, we analyze ViT attention maps to investigate frequency-specific processing patterns.

4.1. Experimental Setup

We use CIFAR-10, resizing images from 32 × 32 to 64 × 64 before applying a two-level discrete wavelet transform (DWT) to each RGB channel. This decomposition produces a low-frequency subband ($\mathbf{L}$) and detail subbands ($\mathbf{D}_1, \mathbf{D}_2$). We compare *Haar* and *db2* wavelets to assess their interpretability impact. The low-frequency subbands (224×224) are processed through a pretrained ViT-base model [4], projecting embeddings to a 512-dimensional latent space. A UNet-style decoder then reconstructs the images, optimizing a composite loss that combines MSE, VGG-based perceptual loss, and Sobel edge loss.

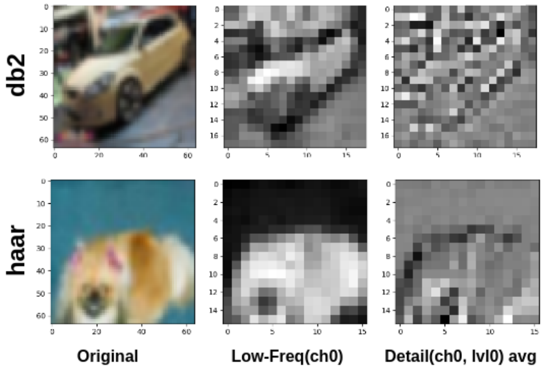

Figure 1. Wavelet-based decomposition and reconstruction. **Left**: Original images. **Middle**: Low-frequency subbands emphasizing global structure. **Right**: High-frequency details capturing fine textures.

We train separate models for the *Haar* and *db2* wavelet decompositions over 15 epochs using the Adam optimizer [7] with a learning rate of 5×10^{-4} and a batch size of 32. Throughout training, we track MSE, perceptual loss, edge loss, and CLIP similarity at each epoch to assess reconstruction performance and semantic fidelity. To evaluate how well semantic information is preserved, we compute the cosine similarity between reconstructed images and the text prompt `''a photo of a scene''` using CLIP [11]. This provides a high-level measure of how frequency-based modifications affect the retention of meaningful image features. Next, we investigate the contribution of specific frequency subbands by systematically removing individual wavelet detail subbands across RGB channels. We quantify their impact on reconstruction accuracy using MSE and assess semantic degradation through CLIP similarity. Additionally, we perform full-detail ablations to analyze the cumulative effect of removing all high-frequency components. Finally, to better understand how ViTs process frequency-based features, we examine attention maps at layers 0 and 11, focusing on heads 0 and 3. We selected layer 0 and layer 11 to compare early and late attention behavior, and chose heads 0 and 3 due to their differing attention spread observed in preliminary analysis. By applying attention rollout, we visualize whether certain heads specialize in high-frequency details (e.g., textures, edges) or low-frequency structures, revealing potential frequency-dependent mechanisms within the model.

5. Results

Table 1 compares Haar and DB2 wavelets in terms of CLIP similarity. Haar achieves a slightly higher original CLIP score (0.20735 vs. 0.20145), suggesting better initial semantic alignment. However, DB2 shows a greater increase in CLIP similarity post-reconstruction (0.01883 vs. 0.01211), indicating better semantic retention. This suggests that Haar may capture fine-grained details, while DB2 preserves broader semantic coherence.

Metric	Haar	DB2
CLIP Orig	0.20735	0.20145
CLIP Recon	0.21946	0.22028

Table 1. Comparison of Haar and DB2 wavelets based on CLIP similarity scores.

5.1. Ablation Experiments

To evaluate frequency-specific contributions, we ablate individual detail subbands at different wavelet decomposition levels and measure the impact on MSE and CLIP similarity. Table 2 summarizes results across RGB channels. Haar ablations lead to higher MSE, particularly at Level 1 (0.027), indicating that high-frequency details are crucial for pixel-wise accuracy. DB2 shows lower MSE (≤ 0.015), suggesting a more distributed feature representation. CLIP similarity drops more for Level 1 ablations, with Haar showing the largest degradation (up to 0.0171). This suggests that high-frequency details are critical for preserving semantic meaning. Removing all detail coefficients degrades Haar's reconstructions more severely (MSE: 0.0034, CLIP: 0.0084) compared to DB2 (MSE: 0.0014, CLIP: 0.0036), reinforcing that Haar's high-frequency components may play a larger role in maintaining both pixel-level accuracy and semantic integrity. These findings suggest that, within our framework, ViTs show increased sensitivity to high-frequency components, particularly those highlighted by Haar wavelets, which has implications for frequency-aware transformer design.

5.2. Attention Analysis

Figure 2 shows how ViT attention maps vary between Haar and DB2. The first four columns visualize raw attention weights from different layers (L0, L11) and heads (H0, H3). Haar wavelets exhibit stronger activations in structured, high-frequency regions, whereas DB2 produces more diffused attention distributions. The patch-wise rollout (fifth column) indicates that Haar preserves fine-grained spatial dependencies, while DB2 captures broader semantic areas. The final overlay highlights that Haar prioritizes object edges and textures, whereas DB2 spreads attention more evenly across the image. This suggests that ViTs process different frequency components with distinct attention patterns.

6. Conclusion and Future Work

This work presents a wavelet-based interpretability framework for analyzing Vision Transformers (ViTs) through

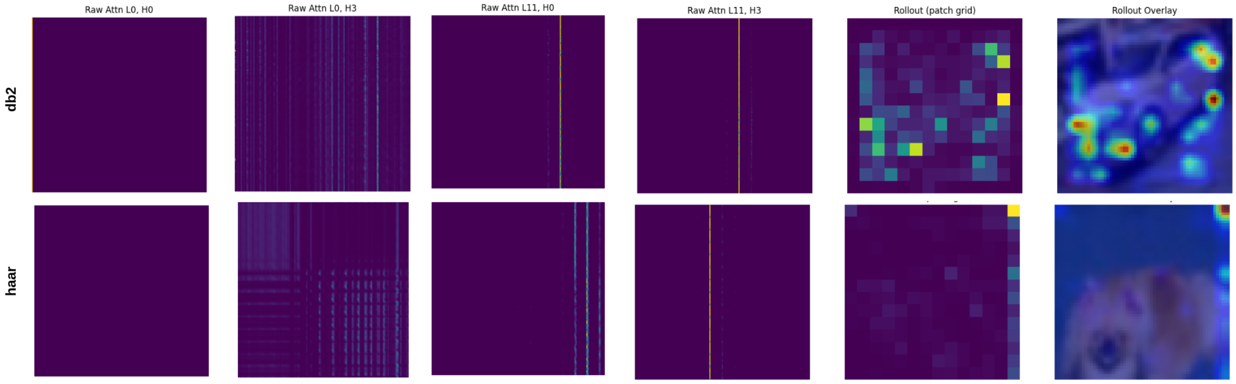

Figure 2. Comparison of ViT attention maps for Haar and DB2 wavelets. The first four columns represent raw attention from different layers and heads. The last two columns show attention rollout, with the final column overlaying attention on the original image. Haar wavelets exhibit higher sensitivity to high-frequency regions.

Wavelet	**Channel 0**		**Channel 1**		**Channel 2**	
	Level 0	**Level 1**	**Level 0**	**Level 1**	**Level 0**	**Level 1**
Haar	0.021 / -0.0005	0.027 / 0.0103	0.022 / 0.0104	0.027 / 0.0171	0.022 / 0.0076	0.027 / 0.0124
DB2	0.012 / 0.0032	0.015 / 0.0127	0.012 / 0.0103	0.015 / 0.0177	0.012 / 0.0109	0.015 / 0.0116
Combined (all)	Haar: 0.0034 / 0.0084		DB2: 0.0014 / 0.0036			

Table 2. Ablation results showing Avg MSE / Avg CLIP Δ for different wavelets and levels.

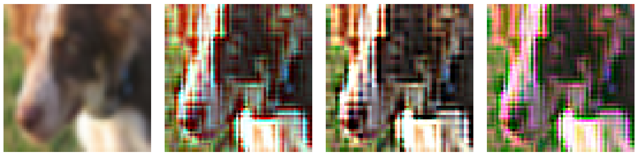

Figure 3. **Wavelet subband ablation effects. Left**: Original image. **Right**: Reconstructions after ablating specific wavelet subbands. (a) **Ablated (ch=0, lvl=0)**: Removal of horizontal details from channel 0 introduces blocky distortions. (b) **Ablated (ch=0, lvl=1)**: Removing higher-level details from channel 0 results in loss of fine texture. (c) **Ablated (ch=1, lvl=0)**: Vertical detail loss in channel 1 disrupts color consistency. (d) **Ablated (ch=1, lvl=1)**: Coarser structure loss in channel 1 affects semantic coherence.

frequency-aware ablations. Our findings suggest that high-frequency details, particularly those captured by Haar wavelets, may play a meaningful role in both pixel-level fidelity and semantic retention.

While our framework provides valuable insights, it has several limitations. First, our experiments are limited to the CIFAR-10 dataset, which restricts the generalizability of the results. Second, although we examine a subset of attention heads and layers, a more comprehensive analysis across the full ViT architecture would strengthen the conclusions. Third, our use of a UNet-style decoder may introduce reconstruction artifacts that are not fully disentangled from the effects of wavelet ablations; an ablation-free decoder baseline could help isolate this factor. Additionally, our semantic evaluation relies on a generic CLIP prompt (“a photo of a scene”), which may not accurately reflect the semantics of CIFAR-10 classes; using class-specific prompts could yield more precise alignment measures. Future research could extend this work in several directions. Scaling the analysis to larger datasets, such as ImageNet [3], would provide deeper insights into frequency-based interpretability across more complex visual domains. Exploring additional wavelet families, including Coiflet and other Daubechies variants [5], could reveal whether different wavelet properties influence ViT representations in distinct ways. Another promising avenue is the development of adaptive frequency-aware architectures, where frequency constraints are incorporated directly into ViT training to improve robustness and interpretability. Cross-model comparisons could further validate the framework by applying it to hybrid CNN-Transformer models or Diffusion Transformers. Additionally, expanding the methodology to task-specific domains, such as object detection, segmentation, or adversarial robustness, could enhance our understanding of how frequency-aware processing impacts different vision tasks.

ACKNOWLEDGMENTS

This work was funded by the DEVCOM Army Research Laboratory under the cooperative agreement W911NF-20-2-0218.[2]

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 1, 2

[2] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[5] Amara Graps. An introduction to wavelets. *IEEE computational science and engineering*, 2(2):50–61, 1995. 4

[6] 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. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[8] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 2

[9] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10199–10208, 2023. 1

[10] Akshad Shyam Purushottamdas, Pranav K Nayak, Yashmitha Gogineni, Sumohana S Channappayya, and Konda Reddy Mopuri. Exploring compositionality in vision transformers using wavelet representations. 1, 2

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3

[12] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 2

[13] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European conference on computer vision*, pages 328–345. Springer, 2022. 1

[14] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[15] Yong Yue, Mihai M Croitoru, Akhil Bidani, Joseph B Zwischenberger, and John W Clark. Nonlinear multiscale wavelet diffusion for speckle suppression and edge enhancement in ultrasound images. *IEEE transactions on medical imaging*, 25(3):297–311, 2006. 1

[16] Dimeng Zhang, JiaYao Li, Zilong Chen, and Yuntao Zou. Efficient image generation with contour wavelet diffusion. *Computers & Graphics*, 124:104087, 2024. 1

[2] We also acknowledge the assistance of ChatGPT, a language model trained by OpenAI, for providing helpful suggestions during the writing of this manuscript.