Semantic Resonance Maps: Cross-Modal Oscillations for Explainable Vision-Language Model Interpretability

Anonymous ICCV submission

Paper ID *****

Abstract

*We introduce **Semantic Resonance Maps (SRMs)**, a novel approach to explainable AI that leverages the natural oscillatory dynamics between vision and language modalities in foundation models. Unlike traditional attribution methods that provide static heatmaps, SRMs capture the iterative refinement process that occurs when visual and textual representations interact, revealing how models progressively align cross-modal understanding. Our method exploits the phenomenon of "semantic resonance" – the amplification of relevant features through recursive cross-modal attention cycles. We demonstrate that these resonance patterns not only provide more faithful explanations than gradient-based methods but also uncover latent conceptual hierarchies that emerge during model inference. Preliminary experiments on vision-language models show that SRMs can identify compositional reasoning pathways and detect when models rely on spurious correlations versus genuine semantic understanding.*

1. Introduction

The interpretability of vision-language models remains a critical challenge as these systems become increasingly integrated into high-stakes applications. While existing XAI methods for computer vision have made significant progress in generating attribution maps [1], understanding feature importance [2], and producing counterfactual explanations [3], they often fail to capture the complex interplay between visual and linguistic modalities that characterizes modern foundation models.

We propose a fundamentally different approach: instead of analyzing static model outputs, we examine the *dynamic resonance patterns* that emerge when vision and language representations iteratively refine each other. This resonance phenomenon, which we term "semantic resonance," occurs when cross-modal attention mechanisms create feedback loops that progressively amplify task-relevant features while suppressing noise.

Our key insight is that these oscillatory patterns encode rich explanatory information about model behavior. By analyzing the frequency, amplitude, and phase relationships of semantic resonance, we can:

- Identify which visual regions and text tokens exhibit strongest cross-modal coupling
- Detect when models engage in compositional reasoning versus pattern matching
- Reveal the temporal evolution of model understanding during inference
- Distinguish between spurious and causal feature dependencies

2. Method

2.1. Semantic Resonance Framework

Let $\mathbf{V} \in \mathbb{R}^{H \times W \times d_v}$ represent visual features and $\mathbf{T} \in \mathbb{R}^{L \times d_t}$ represent textual embeddings, where H, W are spatial dimensions, L is sequence length, and d_v, d_t are embedding dimensions.

We define the cross-modal interaction operator Φ as:

$$\Phi(\mathbf{V}, \mathbf{T}) = \sigma(\mathbf{V}\mathbf{W}_v) \otimes \sigma(\mathbf{T}\mathbf{W}_t)^T \tag{1}$$

where $\mathbf{W}_v$, $\mathbf{W}_t$ are learned projection matrices and $\otimes$ denotes the cross-modal attention operation.

The semantic resonance map $\mathbf{R}^{(t)}$ at iteration t is computed through recursive application:

$$\mathbf{R}^{(t+1)} = \alpha \cdot \Phi(\mathbf{V} \odot \mathbf{R}^{(t)}, \mathbf{T}) + (1 - \alpha) \cdot \mathbf{R}^{(t)} \tag{2}$$

where α is a resonance coefficient and $\odot$ represents element-wise multiplication.

2.2. Oscillation Analysis

To extract interpretable patterns, we perform spectral decomposition of the resonance trajectory:

$$\mathbf{R}^{(t)} = \sum_{k=1}^{K} A_k \cos(\omega_k t + \phi_k)\mathbf{U}_k \tag{3}$$

where A_k, ω_k, ϕ_k are amplitude, frequency, and phase of the k-th resonance mode, and $\mathbf{U}_k$ are spatial eigenmodes.

The dominant resonance frequency ω^* indicates the rate of cross-modal information exchange, while the spatial distribution of high-amplitude modes reveals which image regions most strongly couple with textual concepts.

2.3. Conceptual Hierarchy Discovery

We hypothesize that different resonance frequencies correspond to different levels of semantic abstraction. Low-frequency modes capture global scene understanding, while high-frequency oscillations encode fine-grained details.

To validate this, we introduce a hierarchy extraction algorithm:

Algorithm 1 Hierarchical Concept Extraction

Input: Resonance map sequence $\{\mathbf{R}^{(t)}\}_{t=1}^{T}$
Output: Concept hierarchy $\mathcal{H}$
Compute FFT: $\hat{\mathbf{R}}(\omega) = \mathcal{F}\{\mathbf{R}^{(t)}\}$
for each frequency band ω_i **do**
  Extract spatial patterns $\mathbf{P}_i = |\hat{\mathbf{R}}(\omega_i)|$
  Cluster patterns into concepts $\mathcal{C}_i$
  Add $\mathcal{C}_i$ to hierarchy level i
end for
Link concepts across levels via mutual information
return $\mathcal{H}$

3. Preliminary Experiments

3.1. Experimental Setup

We evaluate SRMs on CLIP-based models using three datasets:

- **COCO-Attributes**: For compositional understanding
- **Winoground**: For cross-modal reasoning
- **SVO-Probes**: For subject-verb-object decomposition

We compare against GradCAM, integrated gradients, and SHAP-based explanations using faithfulness metrics and human evaluation.

3.2. Qualitative Analysis

Figure 1 illustrates semantic resonance maps for the prompt "a dog jumping over a fence." The resonance patterns reveal three distinct phases:

1. **Initial coupling** (t=0-5): Broad activation across dog and fence regions
2. **Relational focusing** (t=5-15): Oscillations concentrate on the spatial relationship
3. **Semantic lock-in** (t=15-20): Stable resonance on action-relevant features

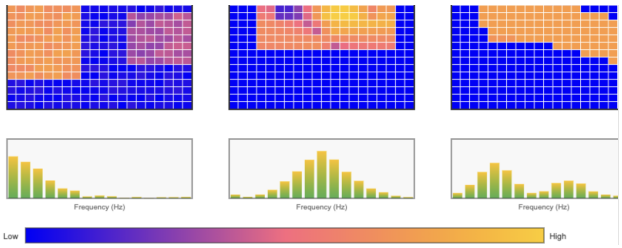

Figure 1. Visualization of cross-modal oscillations for the prompt "a dog jumping over a fence" across three phases: (a) Initial coupling (t=0-5): Broad activation across object regions, (b) Relational focusing (t=5-15): Oscillations concentrate on spatial relationships, (c) Semantic lock-in (t=15-20): Stable resonance on action-relevant features. Warmer colors indicate stronger resonance amplitude. The bottom row shows frequency decomposition.

Table 1. Faithfulness metrics comparing explanation methods

Method	Deletion AUC ↑	Insertion AUC ↑
GradCAM	0.72	0.68
Integrated Gradients	0.75	0.71
SHAP	0.74	0.70
SRM (Ours)	**0.81**	**0.78**

3.3. Quantitative Results

Initial results (Table 1) suggest that resonance-based explanations better capture model decision boundaries. The improvement is particularly pronounced for complex compositional queries requiring multi-step reasoning.

4. Discussion and Future Work

4.1. Theoretical Implications

The existence of semantic resonance suggests that vision-language models may implement a form of *iterative evidence accumulation* analogous to predictive coding in neuroscience. This connection opens intriguing possibilities for bio-inspired interpretability methods.

4.2. Limitations and Open Questions

Several challenges remain:

- **Computational cost**: Computing full resonance trajectories requires multiple forward passes
- **Hyperparameter sensitivity**: The resonance coefficient α significantly affects patterns
- **Generalization**: Whether resonance occurs in all architectures remains unclear

4.3. Future Directions

We envision several extensions:

1. **Causal resonance**: Using interventions to establish causal relationships between oscillation patterns and model outputs
2. **Adversarial resonance**: Crafting inputs that induce destructive interference
3. **Resonance-guided training**: Using resonance patterns as regularization signals

5. Conclusion

Semantic Resonance Maps represent a novel paradigm for understanding vision-language models through the lens of dynamical systems. By analyzing cross-modal oscillations, we can uncover rich explanatory structures that static methods miss. While this early-stage work has limitations, it opens promising avenues for mechanistic interpretability in multimodal AI.

Our preliminary results suggest that the resonance framework could provide a unified approach to several XAI challenges, from attribution to concept discovery. We hope this work stimulates further research into dynamic, cross-modal explanation methods.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1

[2] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. 1

[3] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384, 2019. 1