Towards Explainable Inverse Design for Photonics via Integrated Gradients

Junho Park[*1], **Taehan Kim**[*2], **Sangdae Nam**[2]

[1]University of Michigan, Ann Arbor
[2]University of California, Berkeley
junhop@umich.edu, terry.kim@berkeley.edu, nsd96@berkeley.edu

Abstract

Adjoint-based inverse design yields compact, high-performance nanophotonic devices, but the mapping from pixel-level layouts to optical figures of merit remains hard to interpret. We present a simple pipeline that (i) generates a large set of wavelength demultiplexers (WDMs) with SPINS-B, (ii) records each final 2D layout and its spectral metrics (e.g., transmitted power at 1310 nm and 1550 nm), and (iii) trains a lightweight convolutional surrogate to predict these metrics from layouts, enabling (iv) gradient-based attribution via Integrated Gradients (IG) to highlight specific regions most responsible for performance. On a corpus of sampled WDMs, IG saliency consistently localizes to physically meaningful features (e.g., tapers and splitter hubs), offering design intuition that complements adjoint optimization. Our contribution is an end-to-end, data-driven workflow—SPINS-B dataset, CNN surrogate, and IG analysis—that turns inverse-designed layouts into interpretable attributions without modifying the physics solver or objective, and that can be reused for other photonic components.

Introduction

Adjoint-based inverse design has produced compact nanophotonic components for filtering, coupling, and routing, but design interpretability lags behind performance. As search spaces grow (binary or continuous *pixel* patterns, topology and shape parameters), the resulting devices can appear unintuitive even to experts. We aim to reveal attributions, or which substructures of an inverse-designed layout most influence key metrics such as *transmitted power* at C-band (1550 nm).

We suggest a workflow for *interpretability-based* analysis: generate many demultiplexers with SPINS-B (Stanford Photonic INverse design Software), learn a predictive surrogate from layouts to metrics, then apply Integrated Gradients (IG) to attribute predictions back to pixels. Our contributions are:

- A scalable data pipeline that exports final SPINS-B layouts as 2D NumPy arrays, paired with simulated spectral responses, such as *power* at C-band (Su et al. 2019; spi 2022).
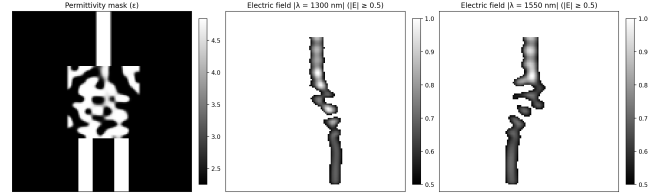
---

[*]These authors contributed equally.

Figure 1: Inverse Designed WDM: 1300nm wave is routed to right port and 1550nm wave is routed to left port.

- A CNN surrogate that reaches useful layout→metric accuracy, enabling pixel-wise IG attribution.
- A visual analysis mapping IG "hotspots" (critical design region) onto the inverse designed layout, offering physical intuition about which region influence power outputs.

Background: Inverse Design & Explainable AI

Wave Demultiplexer

A wavelength demultiplexer (WDM) is a photonic integrated circuit component, specifically an on-chip optical splitter: one input carries multiple wavelengths, and each wavelength exits a different port (e.g., 1310 nm → Port A, 1550 nm → Port B). Figure 1 illustrates inverse-designed WDM, successfully splitting two different wavelength inputs. Key performance terms include: *Power*—transmitted light power from input to the correct output (higher is better); *Insertion Loss (IL)*—loss from input to the correct output (lower dB is better).

Inverse Design of Photonics

We synthesize WDMs with an adjoint, density-based topology optimization design tool (SPINS-B). The design variable is a continuous field

$$u \in [0,1]^{H \times W}.$$

It is mapped to a permittivity distribution via filtering and a smoothed projection

$$\varepsilon(u) = \varepsilon_{\min} + \Pi_{\beta,\eta}\big(F_r * u\big)\big(\varepsilon_{\max} - \varepsilon_{\min}\big),$$

where F_r enforces minimum feature size and $\Pi_{\beta,\eta}$ (Heaviside with continuation) promotes binarization.
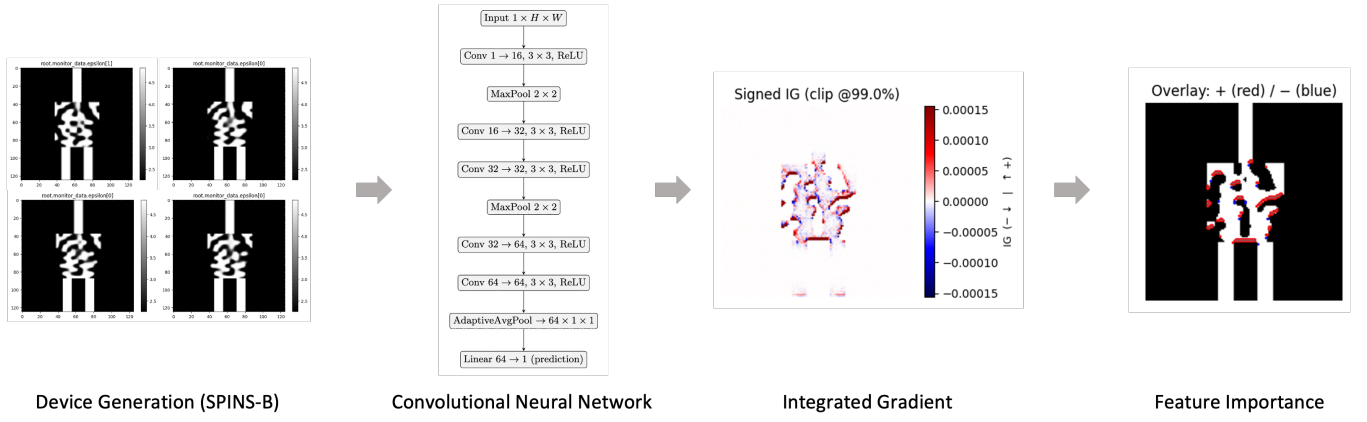
Figure 2: End-to-end pipeline: SPINS-B → dataset → CNN surrogate → IG → Feature Importance.

For each frequency ω_k (wavelength λ_k), we solve the frequency-domain Maxwell system

$$A\big(\varepsilon(u), \omega_k\big)\, \mathbf{E}_k = \mathbf{b}_k,$$

and evaluate power-overlap transmissions into target/output modes m,

$$T_{k,m}(u) \in [0, 1].$$

The objective aggregates multiple wavelengths and ports (throughput vs. crosstalk):

$$J(u) = \sum_k w_k\, \phi\big(T_{k,des}(u)\big) - \sum_k \sum_{m \neq des} \alpha_{k,m}\, \psi\big(T_{k,m}(u)\big).$$

Adjoint solves yield sensitivities w.r.t. permittivity, which are chained to the design field:

$$\nabla_u J = \frac{\partial J}{\partial \varepsilon}\, \frac{\partial \varepsilon}{\partial u}.$$

Related Work and Motivation

Explainable artificial intelligence (XAI) methods such as LIME (Ribeiro, Singh, and Guestrin 2016) and Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017) have been developed to provide post-hoc interpretability for neural networks by attributing model predictions to input features. Beyond these, SHAP provides a game-theoretic framework for attributions (Lundberg and Lee 2017), while Grad-CAM and SmoothGrad offer gradient-based visual explanations complementary to IG (Selvaraju et al. 2017; Smilkov et al. 2017).

The iterative and computationally intensive nature of inverse design frameworks such as SPINS-B makes their optimization process difficult to trace or interpret. In response, emerging studies have introduced interpretable machine learning methods to efficiency enhance transparency within photonic design pipelines. For instance, LIME-style interpretability for photonic (de)multiplexers underscores growing interest in explainable photonic design (Pira et al. 2025). However, such applications primarily focused on perturbation-based methods and did not fully leverage gradient-based attribution techniques for continuous, differentiable analysis. Superpixel perturbations modify regions

of the design at a time. This may blur or distort design features, producing inconsistent with fabrication or simulation constraints. In contrast, IG offers pixel-level, gradient-based attributions that remain consistent with the trained surrogate model F_θ and preserve the geometric integrity of the original structure.

Our approach extends this direction by employing IG to directly map learned sensitivities in the surrogate model onto physically meaningful regions, enabling consistent interpretation across wavelengths (1310 and 1550). Specifically, we overlay the IG maps onto the inverse designs, allowing for physical validation of the regions highlighted.

Problem Setup & Goals

We focus on interpretability for inverse-designed wavelength demultiplexers (WDMs). Given a finalized binary layout $\rho \in \{0, 1\}^{H \times W}$, our aim is to (i) predict scalar figures of merit (FoMs)—mainly *power* at 1310 and 1550—and (ii) attribute those predictions back to pixels to localize the geometry most responsible for performance.

Pipeline overview. Our approach consists of four stages (Fig. 2):

1. **Device generation (SPINS-B).** We run SPINS-B with randomized seeds/hyperparameters to produce a corpus of WDM layouts ρ and their simulated spectra/FoMs (Su et al. 2019; spi 2022).

2. **Surrogate learning (CNN).** We train a lightweight convolutional model $F_\theta(\rho)$ to regress *power* from layouts, enabling fast, differentiable predictions with 2D layout & FoMs.

3. **Attribution (Integrated Gradients).** We apply IG to F_θ to obtain pixel-wise saliency maps ("hotspots" - regions clipped at high thresholds) that attribute the predicted *power* to specific regions of ρ (Sundararajan, Taly, and Yan 2017).

4. **Feature Importance & Analysis.** We (a) visualize attribution hotspots overlaid on layouts, and (b) trace across interpolation steps $\alpha \in [0, 1]$ to characterize stable and emergent sensitivity regions.
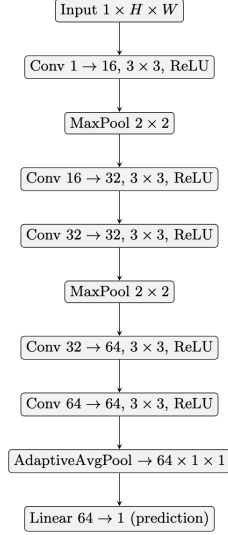
Figure 3: Compact CNN surrogate used to regress *power* from a binary layout and enable pixel-wise IG.

Setup and Goals We aim to have a surrogate F_θ that achieves useful accuracy on *power*, and guide IG attributions that align with physically meaningful substructures (e.g., tapers, splitter hubs) related to target metrics.

Methods
Data Generation with SPINS-B
We use SPINS-B to synthesize two-channel WDMs with fixed Input/Output port geometry and randomized seeds/hyperparameters to induce layout diversity. Each run outputs:

- final binary mask ρ as a 2D NumPy array,
- simulated spectral responses around 1310 and 1550,
- scalar FoMs (Figure of Merit): *power* at 1310 nm and 1550 nm.

Our corpus comprises $N = 500$ designs with various spectral response and *power*. We follow SPINS/SPINS-B best practices for architecture and logging (Su et al. 2019; spi 2022).

Surrogate Model
We train a surrogate model that we will perform integrated gradient on.

Architecture A lightweight CNN following the architecture (Fig. 3) with a single-output regression head that predicts the target metrics is used for *power*@1310 and *power*@1550.

Training We optimize mean squared error (MSE)

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \left(F_\theta(\rho^{(n)}) - y^{(n)} \right)^2, \tag{1}$$

where $F_\theta(\rho^{(n)})$ denotes the model prediction for the n-th input mask $\rho^{(n)}$, and $y^{(n)}$ is the corresponding target metric.

Here, N is the number of samples in the dataset, and θ represents all trainable parameters of the CNN. The network is trained using the Adam optimizer (learning rate 10^{-3}), with early stopping (patience $= 7$) and model selection based on validation MSE. The checkpoint with the lowest validation loss is restored for final evaluation.

Splits. Indices are shuffled with a fixed random seed (42), and the dataset is split into training, validation, and test sets in a 70/15/15 ratio.

Attribution via Integrated Gradients
Given baseline ρ_0 and input ρ, IG attributes

$$\text{IG}_i(\rho; \rho_0) = (\rho_i - \rho_{0,i}) \int_0^1 \frac{\partial F_\theta(\rho_0 + \alpha(\rho - \rho_0))}{\partial \rho_i} \, d\alpha. \tag{2}$$

This allows us to interpolate each feature of $\rho^{(n)}$ and obtain the contribution to the FoMs. We approximate the integral with $K=64$ steps. Attributions are visualized as magnitude maps, thresholded overlays, and binary "hotspot" masks.

Visualization. IG magnitudes are clipped at the 99th percentile; top-q hotspots (default $q=0.5\%$) are overlaid on ρ. For representative previews, test samples are stratified by *power* quantiles, with up to three per bin.
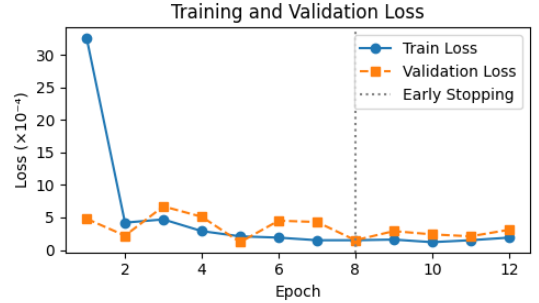
Results

Figure 4: Training and validation MSE loss curves (values $\times 10^{-4}$).

Model Convergence
The model exhibits convergence within ten epochs, as shown by the training and validation MSE loss curves (Fig. 4). Early stopping was triggered at epoch 8, indicating that the model reached an optimal fit without signs of overfitting or divergence. The reported MSE values (on the order of 10^{-4}) reflect the small numerical scale of the target values.

Tracing the α-Path
To interpret the model's learned response, we trace the interpolation path $\alpha \in [0, 1]$ between the baseline and the input using Integrated Gradients (IG). Figure 6 illustrates how the IG attributions accumulate as α increases. Regions that consistently retain high IG intensity correspond to stable, high-sensitivity features within the photonic structure, showing
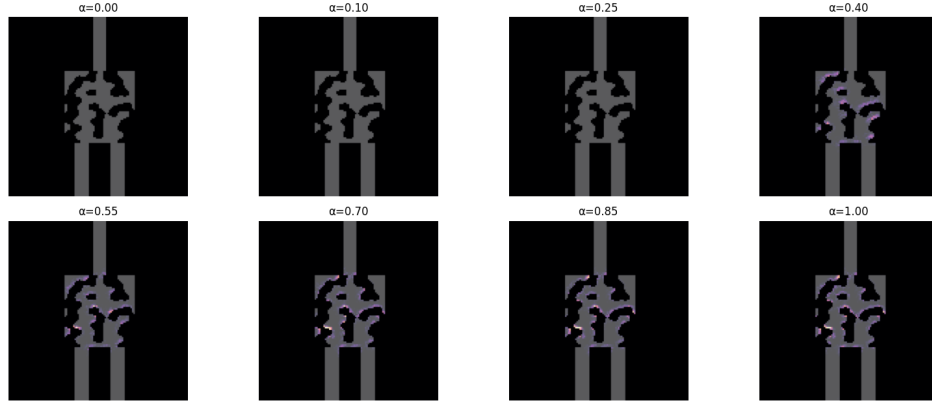
Figure 5: Tracing of α along the Integrated Gradients path. IG activations accumulate progressively, highlighting regions along the α.
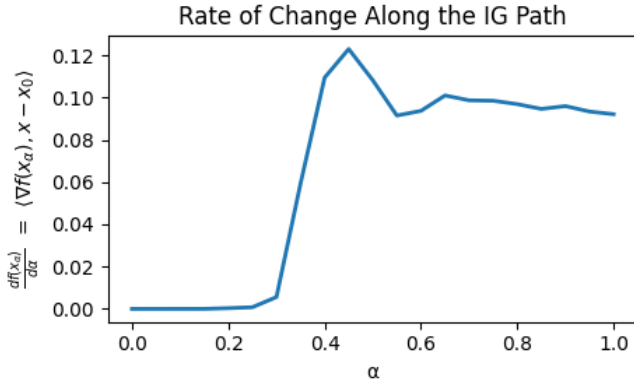

Figure 6: Visualization of the Integrated Gradients (IG) path tracing while interpolation parameter $\alpha \in [0, 1]$ transitions the layout.

how the model's prediction emerges along the interpolation path.

IG Hotspots and Physical Correspondence

Integrated Gradients (IG) maps concentrate on a small set of geometrically distinctive regions of the WDM mask. Across test devices (and along the α-path visualizations in Fig. 5 and the IG overlays), three patterns recur:

1. **High-curvature edges and sharp corners.** Inside corners, notches, and edges exhibit strong IG magnitude. These locations coincide with large index gradients $\nabla \varepsilon$ where mode mismatch and radiation are most sensitive to pixel-scale edits.

2. **Abrupt width transitions.** Interfaces between narrow and wide segments (micro-steps) are repeatedly highlighted. Small changes there alter local impedance and reflection, shifting the phase that determines how *power* divides at the splitter.

3. **Splitter/taper hub.** The central junction lights up early along the path ($\alpha \approx 0.25$–0.55) and remains salient, con-

sistent with a converter–distributor role: the junction sets the phase and mode shape that downstream geometry then routes.

These hotspots may be spatially corresponding to high field intensity and strong Poynting flow gradients in full-wave snapshots: regions with large IG magnitude tend to co-localize with (i) field compression at tapers, (ii) interference nodes/antinodes near the hub, and (iii) leakage bands along jagged outer edges. Taken together, the maps suggests that a small set of curvature- and width-controlled features dominantly regulates transmitted *power*.

Potential Application

The IG analysis pipeline can be adapted for practical edits and design-loop integrations aimed at increasing transmitted *power*:

Design loop integration.

- *Hotspot-aware constraints:* enforce larger filter radii or minimum-width constraints guided by IG mask during adjoint updates.

- *Counterfactual checks:* in simulation, compare identical edit budgets applied to (a) top-$q\%$ IG highlighted regions vs. (b) IG de-prioritized regions matched in size/location; expect larger *power* gains for (a).

- *Fabrication validation (future work):* Experimentally A/B test whether IG hotspots can meaningfully guide and improve measured throughput.

Limitations & Future Work

In summary, we suggest a potential framework of using IG maps to provide pixel-level guidance on photonics device design. The motivation of this work is to present a methodology that integrates explainable methods into inverse photonic design to improve upon the traditionally time-consuming process by guiding specific features. However, several limitations should be noted.

(i) Surrogate vs. physics. IG explanations describe sensitivities of the learned surrogate model F_θ and should not

be interpreted as direct physical ground truth. As this is an on-going work, we plan to validate these insights experimentally, assessing whether design modifications guided by IG correspond to measurable improvements in fabricated devices.

(ii) Data scale and coverage. The current dataset ($N=500$ layouts) is relatively small and structurally narrow, which may lead to overfitting and limit attribution stability. Future work will expand the dataset to 10^3–10^4 layouts with broader geometric diversity, allowing for more robust and generalizable feature interpretations.

Acknowledgement

References

2022. SPINS-B: Open-source inverse design framework (documentation). https://spins-b.readthedocs.io/en/latest/. Accessed 2025-10-19.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pira, L.; Antony, A.; Prathap, N.; Peace, D.; and Romero, J. 2025. Enhanced Photonic Chip Design via Interpretable Machine Learning Techniques. LIME-based interpretability for photonic (de)multiplexers, arXiv:2505.09266.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, 1135–1144.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *IEEE International Conference on Computer Vision (ICCV)*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: Removing Noise by Adding Noise. *arXiv preprint arXiv:1706.03825*.

Su, L.; Vercruysse, D.; Skarda, J.; Sapra, N. V.; Petykiewicz, J. A.; and Vučković, J. 2019. Nanophotonic Inverse Design with SPINS: Software Architecture and Practical Considerations. arXiv:1910.04829.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.