PRISM: Position-encoded Regressive Inverse Spectral Model for Multilayer Thin-Film Design

Runtian Wang [*1] **Renhao Xue** [*1 2] **Baige Chen** [1] **Hao Wu** [1]

Abstract

The inverse problem of multilayer thin-film optical coatings design represents a complex combinatorial-continuous optimization challenge. We present **PRISM** (Position-encoded Regressive Inverse Spectral Model), a unified decoder-only autoregressive transformer that streamlines this process by jointly predicting discrete material selection and continuous thickness regression within a single backbone. PRISM introduces two primary architectural innovations: (1) *spectrum prefix conditioning*, which utilizes standard prefix tokens for in-context target injection, and (2) *cumulative-depth Rotary Position Embeddings*, which encode continuous thickness directly into the positional representation to preserve the physical spatial relationships of the stack. Our benchmarks demonstrate that a PRISM-13M model reduces MAE by over 50% compared to other transformer baselines while utilizing only one-fifth of the parameters. Furthermore, a 44M-parameter variant achieves state-of-the-art performance (MAE = 0.010) on our in-distribution validation benchmark and operates significantly faster than simulated annealing, offering a highly efficient alternative to classical optimization methods.

1. Introduction

Multilayer thin-film optical coatings are ubiquitous in modern photonics. Anti-reflection coatings, bandpass filters, dichroic mirrors, and neutral-density filters all rely on precisely engineered stacks of dielectric and metallic layers whose interference effects shape the spectral response (Macleod, 2010). The *forward* problem of computing the reflectance and transmittance spectrum of a given stack is solved analytically by the Transfer Matrix Method (TMM) (Born & Wolf, 2019). The *inverse* problem on the other hand, is far harder. It is a mixed discrete-continuous optimization over material choices (combinatorial) and layer thicknesses (continuous), with a highly non-convex landscape and many degenerate solutions.

Traditional approaches to inverse thin-film design rely on iterative numerical optimization. Simulated annealing (Kirkpatrick et al., 1983), genetic algorithms (Martin et al., 1995), and needle optimization (Tikhonravov et al., 1996) search the design space directly, evaluating thousands of candidate structures via TMM at test time. While these methods can achieve high spectral fidelity, they are computationally expensive, requiring minutes to hours per design, making them impractical for real-time or high-throughput applications.

Recent neural approaches have sought to amortize the cost of inverse design with varying success. Tandem networks (Liu et al., 2018) utilize joint MLP-surrogate training but are limited by fixed-length representations, while conditional generative models (So & Rho, 2019) often struggle with mode coverage.

Transformer-based sequence modeling has emerged as a promising alternative, yet current frameworks, OptoGPT (Ma et al., 2024) and OptoFormer (Wu et al., 2026), face important efficiency and precision bottlenecks. OptoGPT is a decoder-only autoregressive transformer that serializes each layer into a joint material-thickness token, which requires discretizing thickness and creates a large joint vocabulary. OptoFormer instead uses an encoder with a dual-decoder architecture to separately predict material and thickness sequences, reducing the joint-vocabulary issue but adding architectural complexity. PRISM addresses these limitations by combining factored material prediction and continuous thickness regression within a single decoder-only backbone. We explore these limitations in detail in Section 2.

We propose **PRISM** (Position-encoded Regressive Inverse Spectral Model), an autoregressive transformer that addresses these limitations. PRISM generates thin-film designs layer by layer using causal self-attention, with two key innovations:

[*]Equal contribution [1]Independent Researcher [2]Work does not relate to position at Amazon. Correspondence to: Runtian Wang <w.henry4@live.com>, Renhao Xue <lukasxue@amazon.com>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

- **Spectrum prefix conditioning** (Section 3.3): The target spectrum is projected through a linear layer into a single token prepended to the decoder sequence. The entire model uses causal self-attention only, simplifying the architecture while keeping the conditioning always visible in the attention window.
- **Cumulative-depth RoPE** (Section 3.4): Rotary Position Embeddings (Su et al., 2024) use the cumulative physical depth of the film stack (in nm) rather than sequential token index. This gives the attention mechanism a physically meaningful distance metric that directly reflects the geometric depth in the stack, providing an elegant solution to the problem of continuous thickness input for the transformer architecture.

These innovations are realized in a unified decoder backbone with a per-material thickness regression head, enabling joint beam search over (material, thickness) pairs at inference.

We evaluate PRISM across a broad design space of 17 dielectric and metallic materials. Our results demonstrate that PRISM achieves state-of-the-art accuracy while maintaining high parameter efficiency. Our models substantially outperform both neural baselines and simulated annealing on in distribution benchmarks while staying competitive against traditional methods on practical targets with much faster inference.

2. Related Work

Classical inverse thin-film design. The inverse design of multilayer optical coatings has a long history in optical engineering. Needle optimization (Tikhonravov et al., 1996) iteratively inserts thin layers at positions that maximally reduce the merit function, then locally optimizes thicknesses. Genetic algorithms (Martin et al., 1995) and simulated annealing (Kirkpatrick et al., 1983) perform global stochastic search over the joint discrete-continuous design space. Gradient-based methods using differentiable TMM implementations (Luce et al., 2022) enable L-BFGS optimization of thicknesses for fixed material sequences. All these methods optimize directly against each target spectrum at test time, achieving high fidelity at the cost of minutes to hours per design.

Neural inverse design in nanophotonics. Neural networks have been applied to inverse design across nanophotonics, including metasurfaces (Jiang et al., 2019), plasmonic nanostructures (Malkiel et al., 2018), and thin films. Tandem networks (Liu et al., 2018) address the one-to-many nature of inverse problems by jointly training an inverse network with a forward surrogate that enforces spectrum consistency. However, they use fixed-length representations (e.g., padded to a maximum layer count) and cannot naturally handle variable-length designs. Conditional generative models, including conditional GANs (So & Rho, 2019) and CVAEs (Sohn et al., 2015), learn distributions over designs conditioned on target spectra but often struggle with mode coverage and require multiple samples for good results.

Autoregressive models for inverse design. OptoGPT (Ma et al., 2024) pioneered the use of autoregressive transformers (Vaswani et al., 2017) for thin-film inverse design, treating the problem as sequence generation. It uses a decoder-only autoregressive architecture with spectrum conditioning and a joint material-thickness vocabulary, where each token encodes both a material and a discretized thickness. While effective, this approach has two drawbacks: (i) the vocabulary scales as $|\text{materials}| \times |\text{thickness bins}|$, reaching 904 tokens for 18 materials and 50 bins; and (ii) thickness precision is limited to the bin width (10 nm).

Concurrent work. OptoFormer (Wu et al., 2026), developed concurrently and independently, shares the motivation of moving beyond joint material-thickness vocabularies. It addresses the vocabulary explosion by factoring generation into separate material and thickness streams via a dual-decoder architecture: a spectrum encoder produces a latent representation consumed by two independent decoders. This separation avoids the combinatorial vocabulary but introduces substantial architectural complexity through an encoder and two decoders. PRISM arrives at a different solution, achieving both factored prediction and continuous thickness regression within a single decoder backbone. As neither code nor pretrained models for OptoFormer are publicly available, we were unable to include it as a baseline; we compare against it qualitatively on architectural design choices.

Rotary Position Embeddings. RoPE (Su et al., 2024) encodes position information by rotating query and key vectors in the complex plane, with rotation angles proportional to position. Originally designed for sequential token positions, RoPE has been extended to longer contexts via Position Interpolation (Chen et al., 2023), NTK-aware scaling (bloc97, 2023), and YaRN (Peng et al., 2024). Our work repurposes RoPE for a non-sequential domain: we use cumulative physical depth (in nm) as the position, giving the attention mechanism a distance metric grounded in the physics of thin-film interference.

3. Method

3.1. Problem Formulation

A multilayer thin-film stack consists of L layers, each specified by a discrete material m_ℓ and positive continuous thickness d_ℓ (in nm), deposited on a glass substrate. The forward

model f computes the optical spectrum consisting of reflectance and transmittance via the Transfer Matrix Method (TMM) (Born & Wolf, 2019):

$$\mathbf{s} = f\big((m_1, d_1), \ldots, (m_L, d_L)\big) \in [0,1]^{142}, \tag{1}$$

where $\mathbf{s}$ is the list of concatenated reflectance and transmittance values sampled at set intervals across the EM spectrum. The inverse problem seeks a design $(m_1, d_1), \ldots, (m_L, d_L)$ whose spectrum $f(\cdot)$ matches a given target $\mathbf{s}^*$.

We frame this as auto-regressive sequence generation. Given $\mathbf{s}^*$, the model generates layers left to right:

$$p\big((m_\ell, d_\ell) \mid \mathbf{s}^*, (m_1, d_1), \ldots, (m_{\ell-1}, d_{\ell-1})\big), \tag{2}$$

terminating when it emits an `EOS` token. Material selection is a categorical distribution over 17 materials; thickness is a continuous regression target.

3.2. Architecture Overview

PRISM is a decoder-only transformer with a single shared backbone feeding two output heads (Figure 1a). The input sequence is:

$$[\underbrace{\texttt{SPEC}}_{\text{prefix}}, \underbrace{(m_1, d_1),\ \ldots,\ (m_L, d_L),\ \texttt{EOS}}_{\text{layer tokens}}], \tag{3}$$

where `SPEC` is a learned projection of the target spectrum and each layer token is embedded from its material ID only (thickness is encoded via positional embeddings).

3.3. Spectrum Prefix Conditioning

The target spectrum $\mathbf{s}^*$ is projected into a single d_{model}-dimensional token via a linear layer:

$$\mathbf{h}_{\text{spec}} = W_{\text{spec}}\, \mathbf{s}^* + \mathbf{b}_{\text{spec}}. \tag{4}$$

This token is prepended to the decoder sequence. Under causal masking, the spectrum prefix attends only to itself, while all subsequent tokens attend to the prefix and to all preceding tokens. This replaces the encoder and cross-attention of standard encoder-decoder designs with a simpler architecture as the conditioning is always within the causal attention window.

3.4. Cumulative-Depth RoPE

Standard RoPE (Su et al., 2024) assigns integer positions $0, 1, 2, \ldots$ to tokens. In thin-film design, the physically relevant quantity is not the layer index but the cumulative geometric depth. Two layers separated by a thick intervening layer interact differently (via interference) than two layers in close proximity, even if they are adjacent in the sequence.

We define the position of each token as the cumulative thickness up to that layer:

$$p_0 = 0, \quad p_\ell = \sum_{j=1}^{\ell} d_j, \quad \ell = 1, \ldots, L. \tag{5}$$

The spectrum prefix sits at position $p_0 = 0$. These continuous positions are used directly in place of integer indices in the standard RoPE formulation (Su et al., 2024), applied to both queries and keys in every attention layer. Figure 1b contrasts this scheme with standard integer-indexed RoPE.

This encoding has two advantages: (1) the attention dot product between two tokens depends on their *physical separation* in nm, not their sequential distance, aligning the inductive bias with thin-film interference physics; and (2) the model can naturally handle variable total depths without retraining, since positions are continuous rather than bounded integers.

3.5. Dual Output Heads

The shared transformer backbone produces hidden states $\mathbf{h}_\ell \in \mathbb{R}^{d_{\text{model}}}$ at each position. The vocabulary V comprises the 17 materials plus the special tokens `PAD` and `EOS`. Two heads operate on the hidden states:

Material head. A linear projection produces logits over the vocabulary:

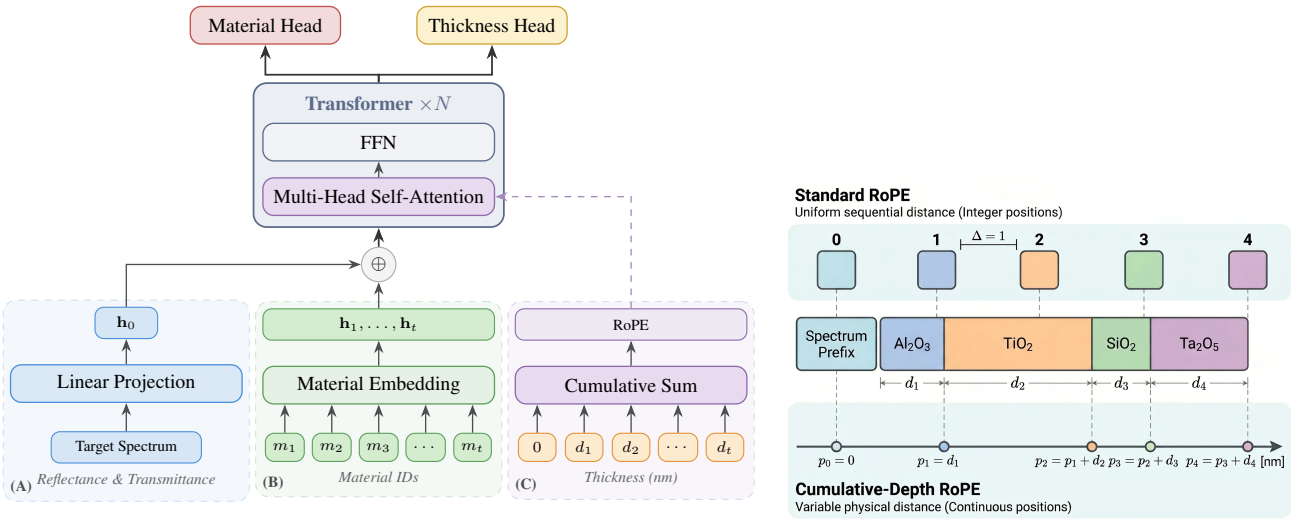
$$\mathbf{z}_\ell^{\text{mat}} = W_{\text{mat}}\, \mathbf{h}_\ell \in \mathbb{R}^{|V|}. \tag{6}$$

Per-material thickness head. A multi-layer MLP predicts one thickness for every entry of the vocabulary simultaneously:

$$\hat{\mathbf{d}}_\ell = \text{THK_MIN} \cdot \exp\big(\text{softplus}\big(\text{MLP}(\mathbf{h}_\ell)\big)\big) \in \mathbb{R}_{>0}^{|V|} \tag{7}$$

The MLP consists of two hidden layers with GELU activations and dropout. The softplus activation ensures positivity in log-space; the exponential maps back to nm. The output dimension matches the material head for indexing convenience; only entries corresponding to physical materials are used at decode time. When material m is selected at position ℓ, the prediction $\hat{d}_\ell^{(m)}$ is used.

This design is the key enabler for **beam search**. The augmented regression head allows the model to predict thickness values conditioned on the material token chosen for the layer. This avoids the two-stage decode required when thickness depends on the chosen material.



(a) PRISM architecture. (A) The target spectrum is projected into a single prefix token via a linear layer. (B) Material tokens are embedded via learned embeddings and the spectrum embedding is appended as the first token. (C) Cumulative-depth RoPE encodes positions as the running sum of layer thicknesses in nm. The shared transformer backbone feeds two output heads: a linear material head producing next-material logits, and a thickness regression head predicting continuous thicknesses.

(b) Cumulative-depth RoPE. A multilayer stack with per-layer thicknesses $d_1, \ldots, d_L$ and cumulative depths $p_0, \ldots, p_L$. Standard RoPE assigns integer positions $0, 1, \ldots, L$ to the embedding sequence. Cumulative-depth RoPE assigns each token a continuous position equal to the running sum of thicknesses, aligning the attention geometry with physical separation in nm.

Figure 1. PRISM architecture (left) and cumulative-depth RoPE positional encoding (right).

3.6. Training

Material loss. We use label-smoothed KL divergence (Szegedy et al., 2016) with smoothing $\epsilon = 0.1$:

$$\mathcal{L}_{\text{mat}} = \text{KL}\big(\tilde{p} \,\|\, \hat{p}\big), \quad \tilde{p}_i = \begin{cases} 1-\epsilon & i = y \\ \epsilon/(V-2) & i \neq y, i \neq \texttt{PAD} \end{cases}, \tag{8}$$

where $\hat{p}$ is the softmax of material logits and y is the ground-truth material.

Thickness loss. We compute MSE in log-space, masked to non-padding positions and gathered at the ground-truth material index:

$$\mathcal{L}_{\text{thk}} = \sum_{\ell : y_\ell \neq \texttt{PAD}} \big(\tilde{d}_\ell^{(y_\ell)} - \log(d_\ell/\text{THK_MIN})\big)^2, \tag{9}$$

where $\tilde{d}_\ell^{(y_\ell)}$ is the log-space prediction at the ground-truth material index. Log-space training prevents large thicknesses from dominating the loss.

Total loss. Both $\mathcal{L}_{\text{mat}}$ and $\mathcal{L}_{\text{thk}}$ are sums over non-padding positions; the combined loss is normalised once by the number of non-padding tokens:

$$\mathcal{L} = \frac{\mathcal{L}_{\text{mat}} + \alpha \cdot \mathcal{L}_{\text{thk}}}{N}, \tag{10}$$

where α is the thickness loss weight and N is the number of non-padding tokens.

3.7. Decoding and Re-ranking

At inference, PRISM supports greedy decoding and beam search, as enabled by the per-material thickness head described in Section 3.5.

Crucially, during beam search, decoded designs are **re-simulated via TMM** to obtain physically accurate spectra. This re-simulation enables a powerful **TMM-reranking** strategy: rather than relying solely on model log-probability to select among beam candidates, we rank the K candidates by their true spectral error against the target. Since TMM evaluation is cheap, re-ranking adds negligible cost while selecting for physical fidelity rather than model confidence. This allows the model to benefit from diversity in its beam since a high-quality design need only appear among the K candidates to be selected, even if the model assigns it lower probability than a spectrally inferior alternative.

4. Experimental Setup

4.1. Design Space

Our design space comprises 17 materials (dielectrics: Al_2O_3, AlN, HfO_2, MgF_2, MgO, Si_3N_4, SiO_2, Ta_2O_5, TiO_2, ZnO, ZnS, ZnSe; semiconductors: Ge, Si; metals: Al, ITO, TiN) deposited on a glass substrate. Layer thicknesses range from 10 to 500 nm in 10 nm steps, with 1–20 layers per stack. Spectra are computed at 71 wavelengths (400–1100 nm, 10 nm spacing) for both reflectance and transmittance (142

values total), using TMM with incoherent substrate treatment (500 μm glass, s-polarization, normal incidence). Refractive index data ($n + ik$) for each material is loaded from tabulated measurements and interpolated to the wavelength grid via cubic splines.

4.2. Data Generation

Training data is generated by uniformly sampling material sequences and thicknesses, then simulating spectra via TMM. Layer counts are sampled with probability $P(L) \propto L$ to oversample longer sequences. We generate up to 30M training samples, 100K development samples, and 10K validation samples.

4.3. Model Configuration

We train two models to study the effect of scale:

PRISM-13M. $d_{\text{model}} = 256$, $d_{\text{ff}} = 1024$, 4 attention heads, 4 transformer layers, dropout 0.1, 2-hidden-layer thickness MLP head. Trained for 30 epochs on 10M samples to enable direct comparison with OptoGPT under matched data conditions.

PRISM-44M. $d_{\text{model}} = 768$, $d_{\text{ff}} = 3072$, 8 attention heads, 6 transformer layers, dropout 0.1, 2-hidden-layer thickness MLP head. Trained for 60 epochs on 30M samples for maximum performance.

Both models use batch size 1024, thickness loss weight $\alpha = 1.0$, and AdamW optimizer (Loshchilov & Hutter, 2019) with cosine annealing learning rate schedule (Loshchilov & Hutter, 2017).

4.4. Baselines

We compare against five baselines spanning optimization and neural approaches, all evaluated on the same validation set and practical targets with identical TMM re-simulation:

Simulated Annealing (SA) (Kirkpatrick et al., 1983). Stochastic global optimization with 8 restarts $\times$ 5,000 steps each. Moves include thickness perturbation ($\sigma = 30$ nm), material swap, layer insertion, and layer removal. Exponential temperature schedule from $T = 0.1$ to 10^{-4}.

Differentiable TMM (Diff-TMM) (Luce et al., 2022). Gradient-based optimization via a PyTorch-differentiable TMM implementation. L-BFGS with 32 random restarts across layer counts $\{3, 5, 7, 10, 14, 18\}$, 300 iterations per restart. Thicknesses parameterized in log-space.

OptoGPT (Ma et al., 2024). Pretrained 63.9M-parameter autoregressive transformer with cross-attention. Joint material-thickness vocabulary with approximately 900 structure tokens. We use the published checkpoint (epoch 146) with greedy decoding.

Tandem Network (Liu et al., 2018). Joint inverse-forward MLP (1.58M parameters). The inverse network predicts fixed-length (20-layer) material logits, thicknesses, and layer count; the forward network reconstructs the spectrum for consistency loss. Trained on 10M samples for 30 epochs.

CVAE. Conditional VAE (1.41M parameters, 64-dim latent), following the generative design approach of So & Rho (2019). Encoder maps (spectrum, structure) to latent distribution; decoder generates fixed-length designs from spectrum + sampled z. KL weight annealed from 0 to 0.1.

4.5. Evaluation Protocol

Metrics. All metrics are computed over the full 142-dimensional spectrum vector after TMM re-simulation of predicted designs. We report Mean Absolute Error (MAE) and coefficient of determination (R^2) on both benchmarks. On the practical targets benchmark we additionally report a spectral earth-mover's distance (EMD), a shape-sensitive metric that measures the minimum cost to transport mass between the predicted and target spectra along the wavelength axis (computed independently on the reflectance and transmittance components and summed). EMD complements pointwise MAE by tolerating small wavelength shifts of correctly-shaped spectral features. Importantly, we never compare predicted structures to ground-truth structures directly.

Decoding configurations. For PRISM, we report two decoding variants: (1) greedy, (2) TMM-reranked-best (beam seach with spectrum error re ranking)

Benchmarks. We evaluate on two benchmark categories:

- **Generated targets (in-distribution)**: 10,000 randomly generated structures with 1–20 layers, matching the training distribution.
- **Practical targets**: 84 practical optical filter spectra spanning a broad range of categories (narrowband, broadband, edge, notch, bandstop, dichroic, multi-bandpass, hot/cold mirrors, broadband HR mirrors, beam splitters, linear variable filters, color filters, and other specialty filters), which do not appear in the training distribution.

5. Results

5.1. In-Distribution Performance

Table 1 compares PRISM at both scales against all baselines on both the validation set (in-distribution) and the 84 practical targets (out-of-distribution). The practical set comprises practical filter spectra whose shapes are distributionally distinct from the training data, testing whether each method can generalize to practical optical designs.

Table 1. Comparison on in-distribution validation and out-of-distribution practical targets. All spectra are TMM re-simulated. †Optimization methods evaluated on $n = 1{,}000$ on validation due to computational cost. Practical target evaluation uses all 84 targets. PRISM uses beam search with width 5 for TMM-reranked.

		Validation (in-dist, $n = 10{,}000$)		Practical (OOD, $n = 84$)		
Method	Type	MAE ↓	R^2 ↑	MAE ↓	R^2 ↑	EMD ↓
SA†	Optim.	0.016	0.978	**0.146**	**0.711**	75.41
Diff-TMM†	Optim.	0.031	0.954	0.204	0.539	94.86
OptoGPT	Neural	0.059	0.715	0.356	−0.20	97.25
Tandem	Neural	0.068	0.771	0.466	−0.54	125.49
CVAE	Neural	0.161	0.066	0.390	−0.39	134.67
PRISM-13M (greedy)	Neural	0.027	0.945	0.303	0.247	96.17
PRISM-13M (TMM-reranked)	Neural	0.024	0.957	0.272	0.362	90.83
PRISM-44M (greedy)	Neural	0.012	0.989	0.244	0.450	77.63
PRISM-44M (TMM-reranked)	Neural	**0.010**	**0.992**	0.228	0.533	**72.20**

On the validation set, PRISM-44M greedy outperforms all methods including SA, making it the first neural method to surpass iterative optimization on this benchmark. Even the smaller PRISM-13M outperforms all neural baselines by a wide margin, despite having $4.9\times$ fewer parameters than OptoGPT (13M vs. 63.9M).

5.2. Practical Targets Performance

On the practical targets, PRISM substantially outperforms all prior neural baselines on every metric. SA achieves the lowest pointwise error (MAE 0.146, R^2 0.711), benefiting from direct per-target optimization, but PRISM-44M TMM-reranked achieves the lowest *shape-sensitive* error: its EMD of 72.2 is below SA's 75.4 even though SA enjoys an MAE advantage. As we discuss in Section 6, pointwise MAE systematically under-credits neural methods on this benchmark. Figure 3 shows spectral comparisons across methods on a representative subset.

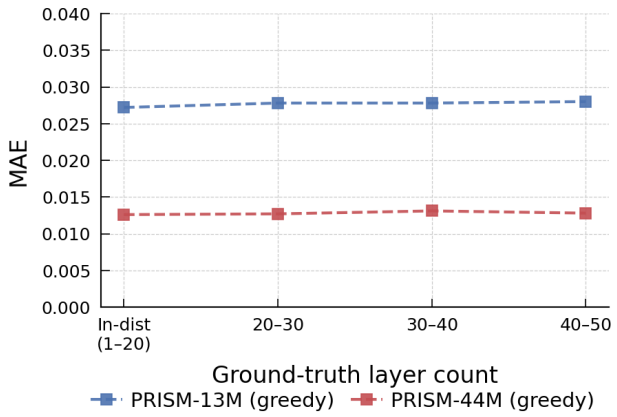


Figure 2. Out-of-distribution generalization to longer sequences ($n = 10{,}000$ per condition, 10 nm thickness steps). Both PRISM-13M and PRISM-44M exhibit minimal degradation when extrapolated to generated sequence up to $2.5\times$ longer than training.

5.3. Out-of-Distribution Generalization

Figure 2 summarizes PRISM's performance on out-of-distribution sequence lengths for both model sizes. PRISM exhibits robust out-of-distribution generalization, maintaining greedy MAE on sequences up to 2.5× longer than training. Scaling from 13M to 44M parameters halves greedy MAE across all conditions, with the same qualitative behaviors at both scales.

6. Analysis

On the practical benchmark, SA achieves the lowest pointwise MAE and R^2 (Table 1). However, qualitative inspection of the decoded spectra (Figure 3) tells a different story: the curves produced by the neural methods, and PRISM in particular, track the target shape more faithfully than the optimization baselines, capturing sharper band edges, peak positions, and passband widths that pointwise MAE under-credits.

The mismatch arises because MAE rewards average agreement, not structural alignment. SA's per-target optimization converges to spectra that minimize average deviation by staying close to a smoothed/mean shape, even when key structures are lost. Under EMD the ranking flips at the top. We view EMD as a critical addition to the evaluation regime for real world target cases where specific parts of the spectrum are regions of interest while the other parts are generally flat.

7. Conclusion

We presented PRISM, an autoregressive transformer for inverse thin-film optical design that introduces two architectural innovations: spectrum prefix conditioning and

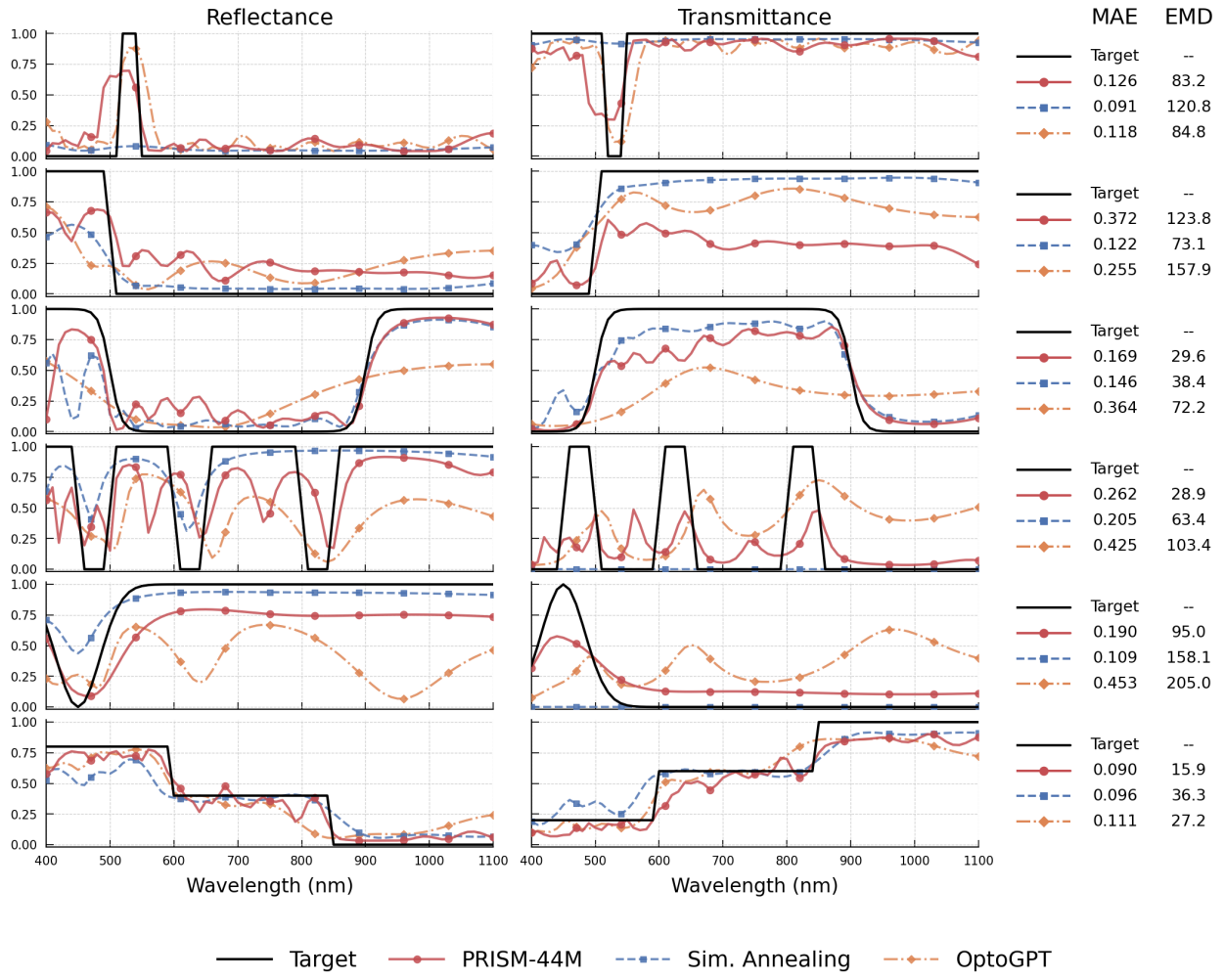


Figure 3. Qualitative comparison on practical target spectra. Each panel shows a target (black) and the TMM-re-simulated spectrum of the design produced by each method. While SA achieves lower MAE on the practical benchmark overall, we can see that the solutions generated by the transformer models and PRISM follow the target curves more closely.

cumulative-depth RoPE. A 44M-parameter model achieves greedy MAE = 0.012 on in-distribution data, surpassing all baselines including simulated annealing while running significantly faster. On 84 out-of-distribution practical targets, PRISM-44M TMM-reranked also achieves the lowest spectral earth-mover's distance, edging out simulated annealing on this shape-sensitive metric despite a higher pointwise MAE.

PRISM exhibits robust out-of-distribution generalization, maintaining greedy MAE on sequences up to $2.5\times$ longer than training. Scaling from 13M to 44M parameters halves greedy MAE across all conditions, with the same qualitative behaviours at both scales.

Limitations. PRISM is conditioned on a full target spectrum across all 71 wavelengths at normal incidence. Practical filter design often involves partial specifications, such as constraining only a specific wavelength subrange while leaving the remainder free, which our current conditioning scheme does not support. Similarly, the model does not accept angle of incidence as an input, precluding designs that require angular selectivity or must perform across a range of incident angles. Both extensions would require richer conditioning interfaces and are left for future work. Beyond these, promising directions include RLVR post-training with TMM-based rewards to directly optimize spectral fidelity on practical targets, which cannot naturally occur in synthetic data.

Reproducibility. All model architectures, training procedures, hyperparameters, and evaluation protocols are fully described in this paper. Code is available at `https://github.com/wang-henry4/prism` and pre-trained checkpoints at `https://huggingface.`

co/flying-iwik/Prism-44M-Mixed.

References

bloc97. NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and target length training data. Reddit post, r/LocalLLaMA, 2023.

Born, M. and Wolf, E. *Principles of Optics*. Cambridge University Press, 7th edition, 2019.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Jiang, J., Sell, D., Hoyer, S., Hickey, J., Yang, J., and Fan, J. A. Free-form diffractive metagrating design based on generative adversarial networks. *ACS nano*, 13(8): 8872–8878, 2019.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science*, 220(4598): 671–680, 1983.

Liu, D., Tan, Y., Khoram, E., and Yu, Z. Training deep neural networks for the inverse design of nanophotonic structures. *ACS Photonics*, 5(4):1365–1369, 2018.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Luce, A., Mahdavi, A., Marquardt, F., and Wankerl, H. Tmm-fast, a transfer matrix computation package for multilayer thin-film optimization: tutorial. *Journal of the Optical Society of America A*, 39(6):1007–1013, 2022. doi: 10.1364/JOSAA.450928.

Ma, T., Wang, H., and Guo, L. J. Optogpt: A foundation model for inverse design in optical multilayer thin film structures. *Opto-Electronic Advances*, 7:240062, 2024. doi: 10.29026/oea.2024.240062.

Macleod, H. A. *Thin-Film Optical Filters*. CRC Press, 4th edition, 2010.

Malkiel, I., Mrejen, M., Nagler, A., Arieli, U., Wolf, L., and Suchowski, H. Plasmonic nanostructure design and characterization via deep learning. *Light: Science & Applications*, 7(1):60, 2018.

Martin, S., Rivory, J., and Schoenauer, M. Synthesis of optical multilayer systems using genetic algorithms. *Applied Optics*, 34(13):2247–2254, 1995.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

So, S. and Rho, J. Designing nanophotonic structures using conditional deep convolutional generative adversarial networks. *Nanophotonics*, 8(7):1255–1261, 2019.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *CVPR*, pp. 2818–2826, 2016.

Tikhonravov, A. V., Trubetskov, M. K., and DeBell, G. W. Application of the needle optimization technique to the design of optical coatings. *Applied optics*, 35(28):5493–5508, 1996.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wu, H., Yuan, S., Li, B., Peng, J., Zhong, H., and Zhang, Y. Dual-decoder transformer for inverse design of optical multilayer films. *Intelligent Opto-Electronics*, 2:250018, 2026. doi: 10.29026/ioe.2026.250018.