CHROMA: Conversational Human-Readable Optical Multilayer Assembly for Natural Language-Driven Inverse Design of Structural Coloration

Mingxuan Li
University of Pittsburgh
Pittsburgh, PA 15261, USA
mil152@pitt.edu

Jungtaek Kim
University of Wisconsin–Madison
Madison, WI 53706, USA
jungtaek.kim@wisc.edu

Oliver Hinder
University of Pittsburgh
Pittsburgh, PA 15261, USA
ohinder@pitt.edu

Paul W. Leu
University of Pittsburgh
Pittsburgh, PA 15261, USA
pleu@pitt.edu

Abstract

We present *CHROMA*, a human-in-the-loop system that turns natural language-based requests into manufacturable multilayer thin-film designs for structural coloration. CHROMA combines a frozen large language model encoder with a compact, trainable Transformer decoder over a discrete material–thickness vocabulary, and couples decoding with a differentiable transfer-matrix-method verifier. Users write prompts like "teal reflection at 60°, six layers, include ZnO, avoid Al next to Mn." Then, CHROMA proposes a stack, simulates the optical response, and enforces hard constraints. We describe dataset construction from physics-generated spectra paired with templated natural-language paraphrases, report scaling trends over encoder size and decoder depth, and demonstrate angle-resolved analyses of selected stacks. CHROMA's conversational interface provides exceptional flexibility for defining design targets.

1 Introduction

Structural coloration, which creates color from interference, diffraction, and scattering rather than absorption, has reemerged as a compelling alternative to pigments and dyes [6, 11, 10, 3], inspired by natural photonic systems such as butterfly wings, peacock feathers, and iridescent beetles [2, 12, 4]. Unlike organic colorants that can fade or impose environmental burdens, structural colors are encoded in geometry and materials, often spanning a gamut beyond standard RGB spaces [1]. While metasurfaces push the state of the art in gamut, multilayer thin films offer clear advantages in manufacturability and robustness, making them ideal for many practical optical components. The core challenge lies in inverse design: determining the precise sequence of materials and thicknesses to achieve a target optical response.

Conventional multilayer design relies on iterative black-box optimization (e.g., genetic algorithms and Bayesian optimization) [5, 7] or supervised inverse models [9]. These methods, however, face significant challenges: optimization can be computationally expensive and struggle with complex, non-convex landscapes, while supervised models often require vast, well-structured datasets. We argue that a fundamental limitation is the interface itself. We pursue a different paradigm: *let people*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Design.

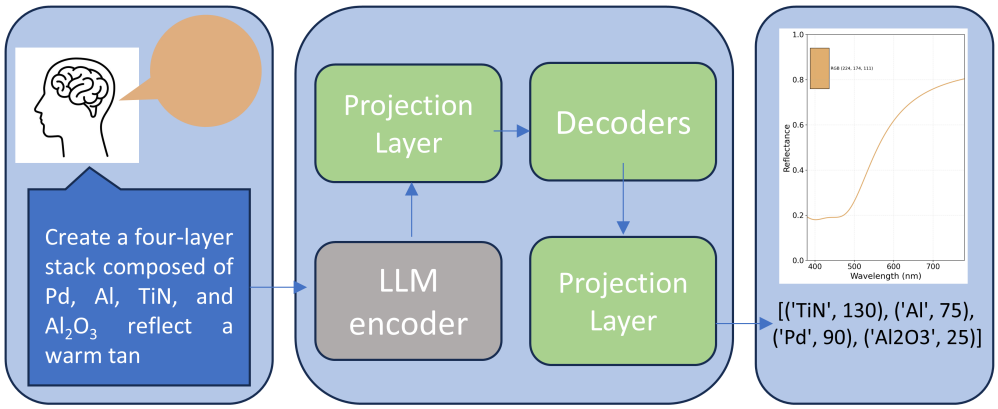

Figure 1: **CHROMA schematic.** Natural-language prompt is translated by a frozen LLM encoder and a compact decoder into a candidate design from a discrete vocabulary.

speak color. Our work, CHROMA, introduces a conversational system that translates free-form text, rich with constraints and intent, into manufacturable stacks; see Figure 1.

2 Methodology

System Overview. CHROMA is designed as a multi-stage pipeline that bridges the gap between high-level human language and low-level physical design. A user specifies a target color and viewing geometry along with optional constraints (i.e., layer count, material inclusion/exclusion, forbidden adjacencies, thickness bounds). This natural-language prompt is first embedded by a frozen Llama encoder. Its latent representation is then passed to a compact, trainable Transformer decoder, which autoregressively proposes a multilayer stack from a discrete material–thickness vocabulary. This generated stack is immediately verified by a transfer-matrix method (TMM) module that simulates its spectral response ($R(\lambda)$ or $T(\lambda)$), converts the spectrum to standard RGB, and computes the color error (ΔE) against the target.

Model and Training. We formalize a design as a sequence $\mathcal{S} = [(m_1, t_1), \ldots, (m_L, t_L)]$, where materials m_ℓ are drawn from a library $\mathcal{M}$ of 25 popular optical materials and thicknesses t_ℓ are chosen from a quantized set $\mathcal{T} = \{5, 10, \ldots, 200\}$ nm (i.e., 40 values). This discretization is key to ensuring manufacturability. Our model's vocabulary consists of all 1,000 material–thickness pairs, plus special tokens for padding and sequence boundaries (i.e., <PAD>, <BOS>, <EOS>), totaling 1,003 tokens.

The architecture balances linguistic understanding with specialized, efficient learning. The encoder (i.e., 1B, 3B, or 8B Llama variant) is kept frozen. This strategic choice leverages its powerful, pre-trained understanding of language for stability and sample efficiency, avoiding catastrophic forgetting. Its outputs are projected to a fixed dimension ($d = 512$) and consumed by a compact Transformer decoder ($d = 512$, 8 heads, $N \in \{2, 4, 8\}$ blocks), which is trained from scratch. Training employs a standard teacher-forcing scheme with the cross-entropy. Unless otherwise stated, we use batch size 560, learning rate 10^{-4}, and train for 20 epochs.

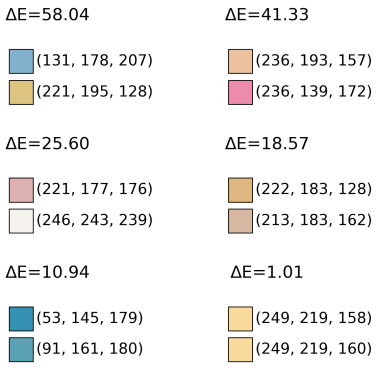

Figure 2: **Visualizing Color Difference (ΔE).** Each subplot compares a ground truth color (top) to a simulated color (bottom). The ΔE value quantifies their perceptual difference, which becomes nearly indistinguishable as ΔE approaches zero.

Color Metric (CIE76). Color evaluation follows the standard CIE pipeline: (i) convert reflectance $R(\lambda)$ to tristimulus values (X, Y, Z) using 1931 color-matching functions and a D65 illuminant; (ii) transform (X, Y, Z) to the perceptually more uniform CIE $L^*a^*b^*$ space; and (iii) calculate color difference using the CIE76 formula:

$$\Delta E = \sqrt{(L_2^* - L_1^*)^2 + (a_2^* - a_1^*)^2 + (b_2^* - b_1^*)^2}. \tag{1}$$

This metric represents the Euclidean distance in CIELAB space, where a common just-noticeable-difference (JND) threshold is $\Delta E \approx 2$. This is visualized in Figure 2.

Dataset and Evaluation Splits. Our dataset is synthesized through a physics-first workflow: (i) generate random multilayer stacks; (ii) simulate their optical spectra with JaxLayerLumos; (iii) generate corresponding natural-language descriptions using a set of templates and paraphrasing techniques. This ensures every text prompt is paired with a physically valid structure. In addition, we simulate spectra using JaxLayerLumos [8] on a 128-point grid from 380–780 nm.

We evaluate the model on two distinct splits to test different capabilities:

- Training-paraphrase (TP): Rephrasings of prompts for structures seen during training. This split probes the model's ability to generalize to varied linguistic expressions for known designs.
- Validation (held-out structures): A set of 900 base stacks not used in training, each with 8 paraphrases. This split tests the model's ability to perform inverse design for new optical targets.

The training set contains 72,000 prompts (i.e., 9,000 base structures $\times$ 8 templates), and the validation set has 7,200 prompts (i.e., 900×8).

3 Results and Discussion

Our experiments investigate how model performance scales with the size of the frozen encoder and the depth of the trainable decoder. Figure 3 plots the mean CIE76 ΔE against the number of trainable parameters, which is primarily a function of decoder depth; see Tables 1 and 2.

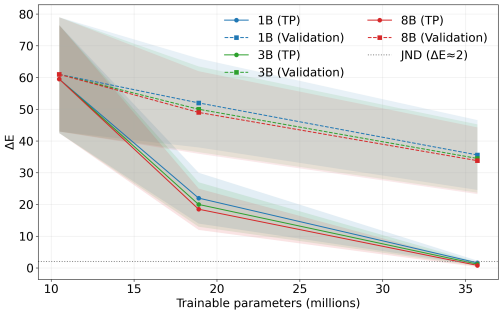

Figure 3: **Scaling Analysis.** Mean ΔE versus trainable parameters for decoder depths (2, 4, and 8 layers) and encoder sizes (1B, 3B, and 8B parameters). Solid lines represent TP performance, while dashed lines show validation on held-out structures. The dotted line marks the JND threshold.

Two clear patterns emerge. First, decoder depth is the dominant factor in performance. A shallow decoder ($N = 2$) struggles, saturating at a high $\Delta E \approx 60$ on both splits, indicating it lacks the capacity to map complex linguistic and physical constraints to the correct token sequence. Increasing depth to $N = 4$ causes a dramatic performance jump, with TP error falling to $\Delta E \approx 20$ and validation error dropping to ~ 50. A deep decoder ($N = 8$) achieves sub-JND accuracy on the TP split ($\Delta E < 2$) and further improves validation error to $\Delta E \approx 34$. This indicates the bottleneck lies in mapping a compact semantic representation to a constrained, combinatorial token sequence; deeper decoders provide longer effective receptive fields and greater capacity for constraint-aware sequence planning, which the shallow model lacks.

Second, increasing the frozen encoder's size provides only marginal benefits. Moving from a 1B to an 8B parameters encoder yields small but consistent reductions in ΔE and slightly tighter variance bands. This suggests larger encoders help normalize linguistic variation, but the fixed ($d = 512$) projection creates an information bottleneck that limits gains from scaling the frozen encoder; performance is thus governed primarily by decoder capacity under constraints.

The persistent gap between TP and validation performance ($\Delta E \approx 1$ vs. ≈ 34 for the best model) reveals the core challenge of generalization in this domain. Several factors contribute: (i) quantization

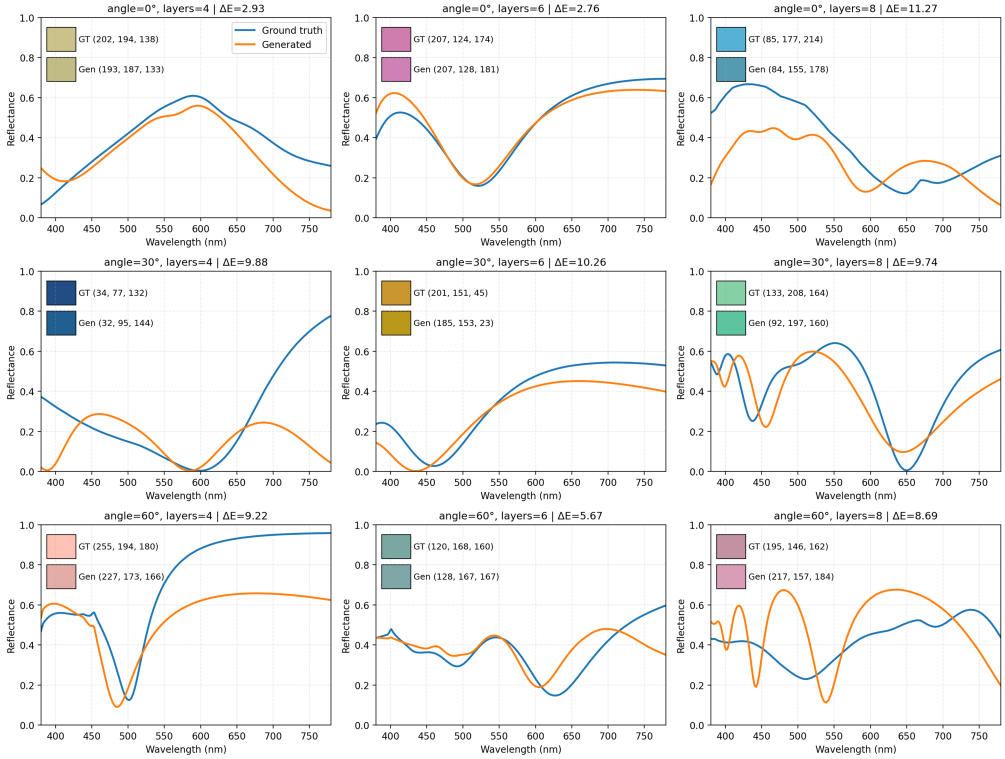

Figure 4: **Successful validation examples.** Examples of successful predictions on the held-out validation set, demonstrating CHROMA's ability to generate novel, high-fidelity designs with low ΔE across different angles and layer counts.

error from the discrete thickness vocabulary; (ii) a mismatch between the training objective (token-level cross-entropy) and the final evaluation metric (spectrally-derived ΔE); (iii) the combinatorial complexity introduced by hard constraints in the prompts; and (iv) the high sensitivity of optical spectra to small changes in the layer stack, especially at non-normal incidence angles. These insights suggest clear avenues for future work, such as incorporating sequence-level, physics-informed losses and developing more continuous parameterizations for layer thicknesses.

4 Conclusion

We have introduced CHROMA, a conversational framework that successfully bridges the gap between natural language and the inverse design of optical structures. By integrating a frozen LLM encoder with a compact, trainable decoder, CHROMA translates complex, language-based requirements into physically-valid and constraint-aware designs. This approach moves beyond traditional optimization by offering an intuitive and flexible interface, empowering users to embed practical manufacturing considerations directly into prompts. To support reproducibility, we have detailed our dataset construction, model parameters, and evaluation protocols.

Reproducibility Notes. Our implementation employs JaxLayerLumos [8] for spectral modeling, incorporates natural-language templates for labeling, and adopts the training configuration shown in the appendices.

Limitations. CHROMA currently operates in a discrete design space, which introduces quantization error. Its TMM assumes ideal, planar layers, omitting real-world effects like surface roughness and process variability. We report CIE76 ΔE, which is a simplified perceptual metric. Also, the template-based prompts and frozen encoder limit the scope of language understanding, and a generalization gap remains between seen and unseen structures. Future work will explore continuous parameterizations, richer physical models, and advanced training strategies to address these limitations.

Acknowledgments and Disclosure of Funding

We received support from the MDS-Rely Center, funded by NSF IUCRC awards EEC-2052662, EEC-2052776, and EEC-2435402. In addition, the University of Pittsburgh Center for Research Computing and Data (RRID:SCR_022735) provided computational resources for this project, supported by NIH award S10OD028483 and NSF award OAC-2117681.

References

[1] Z. Dong, J. Ho, Y. F. Yu, Y. H. Fu, R. Paniagua-Dominguez, S. Wang, A. I. Kuznetsov, and J. K. W. Yang. Printing Beyond sRGB Color Gamut by Mimicking Silicon Nanostructures in Free-Space. *Nano Letters*, 17(12):7620–7628, 2017.

[2] A. Ingram and A. Parker. A review of the diversity and evolution of photonic structures in butterflies, incorporating the work of John Huxley (The Natural History Museum, London from 1961 to 1990). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363 (1502):2465–2480, 2008.

[3] K. M. Jablonka, S. M. Moosavi, M. Asgari, C. Ireland, L. Patiny, and B. Smit. A data-driven perspective on the colours of metal–organic frameworks. *Chemical science*, 12(10):3587–3598, 2021.

[4] Z. Jia, M. C. Fernandes, Z. Deng, T. Yang, Q. Zhang, A. Lethbridge, J. Yin, J.-H. Lee, L. Han, J. C. Weaver, K. Bertoldi, J. Aizenberg, M. Kolle, P. Vukusic, and L. Li. Microstructural design for mechanical–optical multifunctionality in the exoskeleton of the flower beetle Torynorrhina flammea. *Proceedings of the National Academy of Sciences*, 118(25):e2101017118, 2021.

[5] J. Kim, M. Li, O. Hinder, and P. W. Leu. Datasets and benchmarks for nanophotonic structure and parametric design simulations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 4685–4715, 2023. Datasets and Benchmarks Track.

[6] S. Kinoshita, S. Yoshioka, and J. Miyazaki. Physics of structural colors. *Reports on Progress in Physics*, 71(7):076401, 2008.

[7] M. Li, J. Kim, and P. W. Leu. Discovering multi-layer films for electromagnetic interference shielding and passive cooling with multi-objective active learning. In *Neural Information Processing Systems Workshop on AI for Accelerated Materials Design (AI4Mat)*, 2024.

[8] M. Li, J. Kim, and P. W. Leu. JaxLayerLumos: A JAX-based differentiable optical and radio frequency simulator for multilayer structures. *Journal of Open Source Software*, 10(114):8572, 2025.

[9] T. Ma, H. Wang, L. J. Guo, et al. OptoGPT: A foundation model for inverse design in optical multilayer thin film structures. *Opto-Electronic Advances*, 7(7):240062–240062, 2024.

[10] J. Sun, B. Bhushan, and J. Tong. Structural coloration in nature. *RSC Advances*, 3(35): 14862–14889, 2013.

[11] Y. Zhao, Z. Xie, H. Gu, C. Zhu, and Z. Gu. Bio-inspired variable structural color materials. *Chemical Society Reviews*, 41(8):3297–3317, 2012.

[12] J. Zi, X. Yu, Y. Li, X. Hu, C. Xu, X. Wang, X. Liu, and R. Fu. Coloration strategies in peacock feathers. *Proceedings of the National Academy of Sciences*, 100(22):12576–12578, 2003.

A Model and Training Details

This section provides a detailed breakdown of the model's architecture and the hyperparameters used during training, intended to ensure full reproducibility of our results.

Table 1: **Trainable parameter components**.

Component	#Parameters
Custom token embedding	513,536
Classifier head	514,539
Encoder projection	1,049,088
One decoder block	$\approx$ 4,204,032

Architectural Components. Table 1 itemizes the sources of trainable parameters in our model. The majority of parameters reside within the Transformer decoder blocks, with a fixed overhead for the token embedding, the projection layer (which maps the frozen encoder's output to the decoder's dimension), and the final classifier head.

Table 2: **Parameter scaling vs. decoder depth** (frozen encoder).

#Decoder Layers	Decoder Parameters	Total Trainable Parameters
2	8.4M	**10.5M**
4	16.8M	**18.9M**
8	33.6M	**35.7M**

Scaling of Trainable Parameters. Table 2 shows the total number of trainable parameters for each of the decoder depths (N = 2, 4, and 8) explored in our scaling analysis; see Figure 3. Since the LLM encoder is frozen, the model's trainable capacity is determined almost entirely by the number of decoder layers.

Table 3: **Default training configuration.**

Parameter	Value	Parameter	Value
batch_size	560	learning_rate	1×10^{-4}
max_text_len	256	max_custom_seq_len	20
custom_embed_dim	512	decoder_hidden_dim	512
decoder_ff_dim	2048	num_decoder_heads	8
dropout_rate	0.1	num_decoder_layers	8
freeze_llama_base	true	num_epochs	20

Hyperparameters. Table 3 lists the default hyperparameters used for training our models, unless otherwise specified in the main text. These settings were chosen to ensure stable training and effective learning within our computational budget.

B Dataset Generation and Material Properties

The foundation of CHROMA is a dataset that reliably links natural language to physical structures. This section details how prompts are generated and the optical properties of the materials used.

> Rephrase the following description of a multi-layer thin-film structure used for structural coloration. The content must remain unchanged, but the wording should be revised. Additionally, replace `<COLOR_NAME>` with a natural-language color name corresponding to the RGB value [211, 219, 181].
>
> Create an optical structure that consists of a four-layer stack composed of GaAs, InP, Pd, TiN. At 0 degrees, the reflected color observed from this multilayer configuration corresponds to `<COLOR_NAME>`.

Figure 5: Example template used for generating paraphrased natural language prompts. A base description is programmatically varied to create a diverse set of prompts for each physically simulated structure.

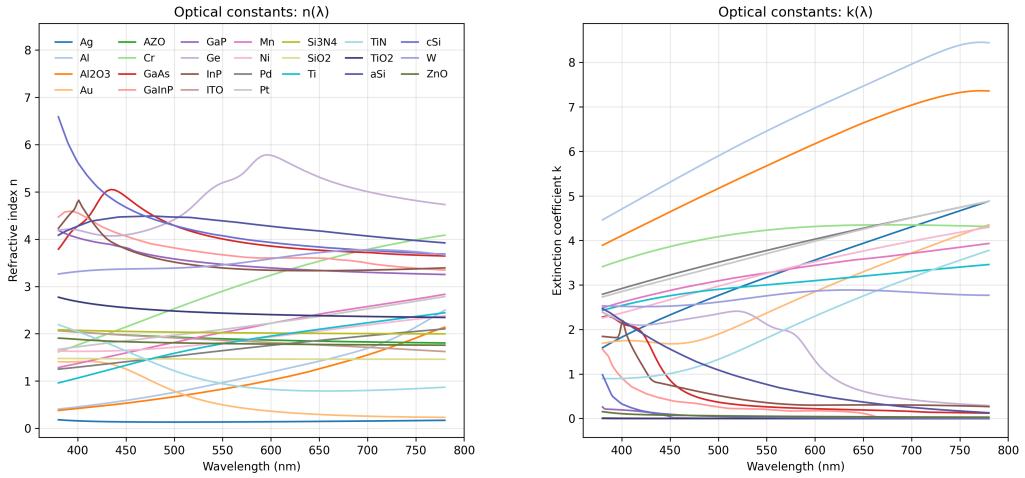

Figure 6: **Optical-constant library.** Wavelength-dependent refractive index $n(\lambda)$ and extinction coefficient $k(\lambda)$ for the 25 materials in our library, spanning the visible spectrum (380–780 nm). These data are the basis for all physical simulations.

Prompt Generation and Optical Constants. Figure 5 shows an example of the templates used to generate varied natural language prompts for a given structure, ensuring linguistic diversity in our dataset. Figure 6 plots the complex refractive index (n and k) data for the 25 materials available to the model. Accurate optical constants are critical for the physics-in-the-loop verification step.

C Qualitative Performance Analysis

To provide a more concrete understanding of the model's behavior, we present representative examples of its successes and failures on the held-out validation set.

Table 4: **Representative validation cases with low ΔE.** Each row reports the viewing angle, layer count, and both the ground-truth and generated stacks. Visual RGB comparisons and ΔE are shown in Figure 4. Bracket notation: [material, thickness nm; ...].

Angle	Layers	Ground-truth Stack	Generated Stack
0°	4	[Mn,10; SiO_2,200; Ge,125; InP,160]	[SiO_2,200; Ge,25; Mn,190; InP,195]
0°	6	[ZnO,85; SiO_2,155; Cr,165; AZO,5; TiN,165; ITO,70]	[ZnO,40; AZO,155; Cr,175; ZnO,35; AZO,10; Cr,195]
0°	8	[GaInP,35; Al_2O_3,15; Cr,150; GaAs,155; Ge,70; AZO,45; W,40; Si_3N_4,175]	[Si_3N_4,115; GaAs,80; GaInP,35; GaAs,70; W,70; Al_2O_3,30; Cr,195]
30°	4	[SiO_2,95; aSi,65; ZnO,85; TiN,180]	[ZnO,160; TiO_2,65; aSi,15; aSi,85]
30°	6	[ZnO,180; Mn,175; GaAs,40; Ag,185; W,200; Ge,190]	[ZnO,175; W,85; Ag,120; W,45; Ag,25; Al,105]
30°	8	[cSi,95; Al_2O_3,5; SiO_2,120; TiN,145; InP,85; W,40; Au,20; ZnO,60]	[cSi,110; W,95; InP,40; W,50; ZnO,20; Al_2O_3,150; W,155]
60°	4	[GaP,85; Au,200; GaAs,135; Ti,25]	[GaP,85; Ti,150; Au,70; GaP,30]
60°	6	[InP,185; AZO,105; Ag,110; GaInP,165; GaP,135; TiN,195]	[GaInP,35; InP,160; TiO_2,30; GaP,85; Ag,25; GaP,160]
60°	8	[GaInP,90; W,10; Ge,190; cSi,150; Cr,150; Al,60; GaP,160; aSi,165]	[cSi,150; Mn,100; GaAs,50; W,100; Ge,200; GaInP,15; W,35; Al,45]

Table 5: **Prompts for the rows in Table 4.** Each prompt corresponds to the same (Angle, Layers) entry; see Figure 4 for visual comparisons.

Angle	Layers	Prompt
0°	4	Design an optical multilayer structure made up of four layers arranged in the sequence SiO2, InP, Mn, and Ge. When viewed at normal incidence (0 degrees), the reflected hue from this stack appears as a muted gold. The thickness of the Mn layer must be less than that of the InP layer, while the SiO2 layer should have a greater thickness than the Mn layer.
0°	6	Design an optical multilayer structure made up of six layers: SiO2, AZO, Cr, ZnO, TiN, and ITO. When viewed at a 0-degree angle, the reflected color from this stack appears as a soft pink. The thickness of each individual layer should be kept at or below 165 nm.
0°	8	Design an optical multilayer structure featuring an eight-layer stack made up of GaInP, GaAs, Ge, W, Al2O3, Si3N4, AZO, and Cr. When viewed at a 0-degree angle, the color reflected by this multilayer arrangement appears as a sky blue hue.
30°	4	Design an optical multilayer structure made up of four layers arranged as SiO2, ZnO, aSi, and TiN. When viewed at a 30-degree angle, the reflected color produced by this multilayer assembly matches the shade of deep blue. In the layer stack, the thickness of aSi must be less than that of TiN, while the SiO2 layer should be thicker than the ZnO layer.
30°	6	Design an optical structure made up of a six-layer stack including Mn, GaAs, W, Ge, Ag, and ZnO. When viewed at a 30-degree angle, the color reflected from this multilayer arrangement appears as goldenrod. The Mn layer must be thinner than the Ge layer, and the Ge layer should be thicker than the ZnO layer.
30°	8	Design an optical multilayer structure made up of eight layers arranged in the following sequence: cSi, ZnO, InP, Al2O3, W, TiN, Au, and SiO2. When viewed at an angle of 30 degrees, the reflected color produced by this multilayer assembly appears as a soft green.
60°	4	Design an optical structure featuring a four-layer stack made up of GaP, Ti, Au, and GaAs. When viewed at a 60-degree angle, the reflected color produced by this multilayer arrangement appears as a soft peach. It is essential to ensure that Au is not positioned directly next to Ti.
60°	6	Design an optical multilayer structure featuring a six-layer assembly of GaP, InP, AZO, Ag, TiN, and GaInP. When viewed at an angle of 60 degrees, the reflected color produced by this stack appears as teal. It is essential to ensure that Ag is not positioned directly next to InP.
60°	8	Design an optical multilayer structure made up of eight sequential layers: GaP, cSi, aSi, W, Cr, Al, Ge, and GaInP. When viewed at a 60-degree angle, the reflection from this stack exhibits a muted mauve hue. The layer of W must be positioned directly next to the GaInP layer.

Successful Cases. Tables 4 and 5 detail several successful predictions on the validation set, corresponding to the visuals in Figure 4. These examples show that even when the generated stack differs significantly from the ground-truth stack in materials and thicknesses, CHROMA can often find alternative solutions that produce a very similar optical response, demonstrating a capacity for creative, non-obvious design.

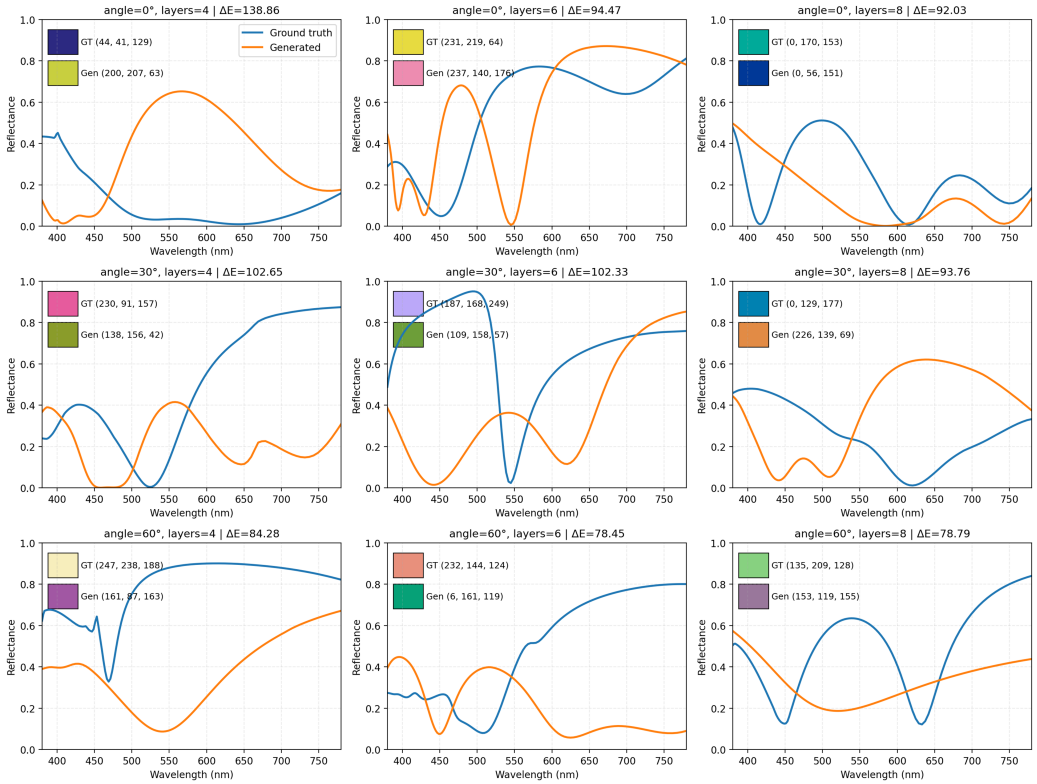

Figure 7: **Worst-case validation examples.** Cases from the held-out validation set where predicted spectra deviate significantly from ground truth, resulting in large ΔE. These highlight challenging regions of the design space and provide insight into model failure modes.

Table 6: **Representative validation cases with high ΔE.** Each row reports the viewing angle, layer count, and both the ground-truth and generated stacks. Visual RGB comparisons and ΔE are shown in Figure 7. Bracket notation: [material, thickness nm; . . .].

Angle	Layers	Ground-truth Stack	Generated Stack
0°	4	[ITO,80; InP,105; W,35; Al_2O_3,85]	[ITO,145; InP,70; Al_2O_3,115; W,155; InP,185]
0°	6	[ZnO,65; TiO_2,5; AZO,130; Si_3N_4,65; Au,90; cSi,75]	[TiO_2,120; cSi,75; Au,30; AZO,35; cSi,75; TiO_2,70]
0°	8	[TiO_2,190; Ge,90; Ti,195; Mn,130; ZnO,185; Pd,30; InP,105; GaAs,5]	[TiO_2,55; Ge,160; Pd,135; Ge,25; Pd,130; GaAs,100; GaP,50; ZnO,115; GaP,125]
30°	4	[SiO_2,80; GaInP,35; ITO,125; Pt,130]	[SiO_2,125; ITO,200; GaInP,135; Pt,110; Ag,50]
30°	6	[Al,5; SiO_2,30; Si_3N_4,200; Ag,105; Ti,15; aSi,85]	[Si_3N_4,160; aSi,90; Ag,40; Al,185; Ag,200; Si_3N_4,25]
30°	8	[AZO,100; aSi,160; GaP,105; TiN,105; Ti,50; Si_3N_4,30; Mn,140; Pd,75]	[AZO,170; GaAs,70; Mn,80; Mn,85; aSi,85; TiO_2,30; AZO,20]
60°	4	[GaP,85; Al_2O_3,165; Ti,135; SiO_2,150]	[GaAs,10; TiN,85; Al_2O_3,5; TiO_2,70]
60°	6	[Si_3N_4,45; ITO,140; TiN,85; cSi,165; SiO_2,75; aSi,25]	[ITO,140; Si_3N_4,35; ITO,140; aSi,25; Ti,135; SiO_2,160; TiO_2,60]
60°	8	[GaP,65; TiO_2,95; TiN,185; Pt,170; Ge,55; AZO,110; Al,195; Ti,30]	[AZO,55; Ti,155; Ge,140; AZO,55; Ti,40; TiO_2,105; GaAs,10; W,50]

Table 7: **Prompts for the rows in Table 6.** Each prompt corresponds to the same (Angle, Layers) entry; see Figure 7 for visual comparisons.

Angle	Layers	Prompt
0°	4	Design an optical assembly featuring a four-layer stack made up of Al_2O_3, ITO, W, and InP. When viewed at 0 degrees, the reflected color exhibited by this multilayer setup is a deep indigo. The tungsten (W) layer must be positioned next to the Al_2O_3 layer.
0°	6	Design an optical structure featuring a six-layer assembly made up of Si_3N_4, TiO_2, Au, ZnO, AZO, and cSi. When viewed at 0 degrees, the reflected color produced by this multilayer arrangement appears as golden yellow. The Au layer must be positioned next to the Si_3N_4 layer.
0°	8	Design an optical multilayer structure made up of eight layers arranged in the following sequence: TiO_2, InP, GaAs, Pd, Ge, Ti, ZnO, and Mn. When viewed at normal incidence (0 degrees), the reflected color produced by this stack appears as teal. The outermost (final) layer should be GaAs.
30°	4	Design an optical multilayer structure made up of four layers arranged in the following order: SiO_2, ITO, Pt, and GaInP. When viewed at an angle of 30 degrees, the reflected color from this stack appears as a vivid pink.
30°	6	Design an optical multilayer structure comprising six layers arranged as Si_3N_4, Ti, aSi, Al, Ag, and SiO_2. When viewed at an angle of 30 degrees, the reflected color produced by this stack appears as a light lavender. The Si_3N_4 layer must be positioned directly next to the Ag layer.
30°	8	Design an optical stack made up of eight layers arranged in the following order: Ti, Pd, aSi, GaP, Si_3N_4, AZO, Mn, and TiN. When viewed at an angle of 30 degrees, the reflected color produced by this multilayer structure appears as deep sky blue.
60°	4	Design an optical configuration featuring a four-layer stack made up of GaP, SiO_2, Al_2O_3, and Ti. When viewed at a 60-degree angle, the reflected color produced by this multilayer assembly appears as pale gold. The combined optical thickness must be limited to no more than 540 nm.
60°	6	Design an optical structure featuring a six-layer stack made up of Si_3N_4, ITO, cSi, TiN, SiO_2, and aSi. When viewed at a 60-degree angle, the color reflected by this multilayer arrangement appears as a soft coral.
60°	8	Design an optical system featuring an eight-layer stack arranged from the first layer as GaP, followed by TiN, Ge, Al, TiO_2, AZO, Ti, and Pt. When viewed at a 60-degree angle, the color reflected from this multilayer assembly appears as a soft green.

Failure Cases. In contrast, Tables 6 and 7 document failure modes, visually represented in Figure 7. These high-ΔE cases often occur when prompts involve complex constraints (e.g., specific adjacencies or layer ordering) that are difficult to satisfy simultaneously with the color target. In some instances, the model generates sequences of incorrect length or repeats materials in unphysical ways. Analyzing these failures is crucial for identifying weaknesses and guiding future improvements to the model architecture and training process.