MetaDiT: Enabling Fine-grained Constraints in High-degree-of Freedom Metasurface Design

Hao Li[1], **Andrey Bogdnaov** [1,2]

[1]Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao 266000, Shandong, China
[2]School of Physics and Engineering, ITMO University, St. Petersburg 197101, Russia
haolicq.ai.research@gmail.com, a.bogdanov@hrbeu.edu.cn

Abstract

Metasurfaces are ultrathin, engineered materials composed of nanostructures that manipulate light in ways unattainable by natural materials. Recent advances have leveraged computational optimization, machine learning, and deep learning to automate their design. However, existing approaches exhibit two fundamental limitations: (1) they often restrict the model to generating only a subset of design parameters, and (2) they rely on heavily downsampled spectral targets, which compromises both the novelty and accuracy of the resulting structures. The core challenge lies in developing a generative model capable of exploring a large, unconstrained design space while precisely capturing the intricate physical relationships between material parameters and their high-resolution spectral responses. In this paper, we introduce MetaDiT, a novel framework for high-fidelity metasurface design that addresses these limitations. Our approach leverages a robust spectrum encoder pretrained with contrastive learning, providing strong conditional guidance to a Diffusion Transformer-based backbone. Experiments demonstrate that MetaDiT outperforms existing baselines in spectral accuracy, we further validate our method through extensive ablation studies.

Code — https://github.com/JessePrince/metadit
arXiv — https://arxiv.org/abs/2508.05076

1 Introduction

Metasurfaces are ultrathin, engineered materials composed of nanostructures that manipulate light in ways natural materials cannot (Jeong, Kim, and Lee 2024; Koshelev et al. 2023, 2018). Unlike bulky traditional optics (e.g., lenses), metasurfaces achieve precise wave control at subwavelength scales (Kildishev, Boltasseva, and Shalaev 2013; Khorasaninejad and Capasso 2017), enabling applications like ultracompact cameras (Kim et al. 2024; Park et al. 2024), AR/VR displays (Aththanayake et al. 2025; Tian et al. 2025), communications (Xu et al. 2025; Fu et al. 2025) and optical computing (Zhou et al. 2024; Hu et al. 2024). However, designing these materials is challenging due to their high-dimensional parameter space. Traditional approaches rely heavily on human intuition and iterative trial-and-error, which are often inefficient and suboptimal. To overcome these limitations, researchers have turned to inverse design methods: leveraging computational optimization (Wang, Zhao, and Zhang 2023; Li, Lin, and Hsu 2023) and machine learning (Al-Zawqari, Vandersteen, and Ferranti 2023; Tian et al. 2024; Chen et al. 2025) to discover metasurface structures that achieve target electromagnetic (EM) responses. These techniques not only accelerate the design process but also enable the discovery of novel configurations. Inverse design thus serves as a critical pathway toward scalable and optimized metasurface engineering.

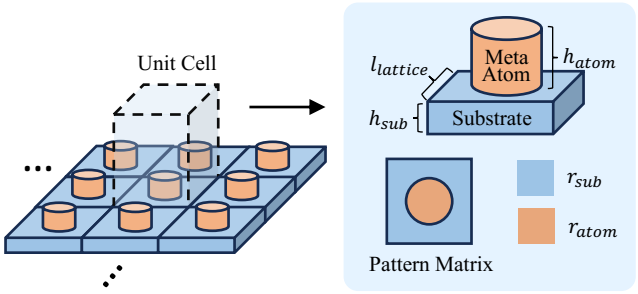

Figure 1: Illustration of metasurface material and the unit cell. The color represents different refractive index.

Recent advances in artificial intelligence have spurred significant interest in applying deep learning models to inverse design problems (Tanriover et al. 2022; Yang et al. 2025; Dong et al. 2025; Saifullah et al. 2025). Over the past decade, we have witnessed remarkable progress in this area, particularly with the emergence of generative models such as generative adversarial networks (GANs) (Goodfellow et al. 2014; Liu et al. 2018; So and Rho 2019; Yeung et al. 2021), variational autoencoders (VAEs) (Kingma, Welling et al. 2013; Tran, Nanthakumar, and Zhuang 2025; Kojima et al. 2023), and diffusion models (Ho, Jain, and Abbeel 2020; Niu, Phaneuf, and Mojabi 2023; Zhang et al. 2023, 2024). These approaches have enabled the generation of multiple solutions for novel material designs that meet desired optical behaviors, proving to be a powerful engine for creating novel metasurface materials.

Designing metasurfaces are designing their constituent materials, geometric configurations, and structural parame-

ters. While recent advances have leveraged powerful generative models for metasurface design, we identify two critical limitations in the current paradigm. First, prior works (Zhang et al. 2023, 2024; Seo et al. 2025; Niu, Phaneuf, and Mojabi 2023) typically formulate the design task as a conditional generation problem, wherein key attributes, such as meta-atom thickness and lattice constants are manually fixed. This effectively reduces the model's role to mapping a target EM response to a geometry within a constrained, pre-selected subspace, thereby limiting the diversity and novelty of generated designs. Second, these approaches often rely on heavily downsampled target spectra (e.g. 12x downsampled in (Zhang et al. 2023)), which simplifies the optimization objective but compromises the physical fidelity of the resulting structures. Although low reconstruction error may be achieved on the coarse target, the final designs can exhibit undesirable frequency-dependent behaviors. We argue that the core challenge lies in developing a high-capacity generative model that can operate over a large, unconstrained design space while faithfully capturing the **underlying physical relationships** between material parameters and their high-resolution spectral responses, as governed by Maxwell's equations. This leads us to the central question of our work:

How can we develop a generative framework that simultaneously optimizes all design parameters while precisely satisfying high-resolution spectral constraints?

In this paper, we propose **MetaDiT**, a novel generative framework to address these limitations. We first develop a spectrum encoder equipped with sequence attention and channel attention mechanisms to extract rich semantic features from input spectra. To align spectral and structural representations, the encoder is trained using a contrastive learning objective alongside a Vision Transformer (ViT) (Dosovitskiy et al. 2020) that processes material geometries. Subsequently, we encode the metasurface material design as an image with three channels, and the pretrained spectrum encoder is leveraged to guide a Diffusion Transformer (DiT) based (Peebles and Xie 2023) diffusion model for material generation. To enable fine-grained conditioning, we introduce a coarse-to-fine conditioning scheme: the coarse spectral embedding is injected via adaLN (Perez et al. 2018), while the fine-grained embedding is concatenated with image tokens, facilitating in-context learning through self-attention. Furthermore, we employ the Accumulated Absolute Error (AAE) to capture localized failures that may be overlooked by conventional averaging metrics. We also introduce the AAE&K metric, defined as the maximum AAE over K independently generated designs for a given target spectrum, to assess the model's consistency in producing diverse yet accurate solutions.

Experimental results show that MetaDiT significantly outperforms existing methods in designing novel metasurface materials with all variable parameters in dataset and fine-grained spectral targets. Specifically, MetaDiT reduces MAE and AAE by **52.2%** compared to a vanilla DiT baseline, and surpasses MetaDiff (Zhang et al. 2023), a model specifically designed for metasurface generation, by **39.1%**. We conduct comprehensive ablation studies to assess the impact of each component in MetaDiT, confirming the effectiveness of our architectural design and training strategy. Moreover, we explore the scalability of MetaDiT: Can we obtain better performance by increasing model capacity?

Our contributions can be summarized as follows:

- We propose MetaDiT, a novel method that generates design parameters under high-resolution spectral constraints, with the flexibility to optimize all parameters when available.
- We propose novel metrics and perform extensive experiments to evaluate MetaDiT's performance and systematically analyze the impact of each design component.
- We open-source our code and all model weights in the hope of paving the way for the community to develop more powerful models.

2 Related Works

2.1 Diffusion Generative Models

Diffusion models (Ho, Jain, and Abbeel 2020) have emerged as highly capable generative frameworks, driving significant advances in image (Dhariwal and Nichol 2021; Nichol et al. 2021; Rombach et al. 2022; Tumanyan et al. 2023) and video generation (Blattmann et al. 2023; Guo et al. 2023; Tu et al. 2024; Wang et al. 2024). While early diffusion models primarily relied on U-Net (Ronneberger, Fischer, and Brox 2015) backbones, the Diffusion Transformer (DiT) (Peebles and Xie 2023) has demonstrated superior training stability and scalability, becoming the dominant architecture in many recent works (Esser et al. 2024; Brooks et al. 2024; Kong et al. 2024). Owing to their strong generative capacity, diffusion models have also been adopted in scientific domains, including molecular (Wang et al. 2025; Liu et al. 2025) and material generation (Xie et al. 2021; Zeni et al. 2023). Recently, researchers have begun applying diffusion models to metasurface design (Zhang et al. 2023, 2024; Seo et al. 2025; Niu, Phaneuf, and Mojabi 2023), demonstrating their potential to generate novel, high-performance structures. However, existing approaches typically generate only a subset of design parameters and rely on downsampled, coarse spectral constraints, simplifying the problem but ultimately limiting the novelty and physical fidelity of the generated materials.

2.2 Inverse Design of Metasurfaces

Designing metasurfaces entails selecting constituent materials, configuring geometric layouts, and tuning structural parameters to achieve desired EM behavior across a frequency range. Early approaches employed computational optimization techniques (Wang, Zhao, and Zhang 2023; Li, Lin, and Hsu 2023), which, while provably convergent, require costly forward–adjoint field evaluations and struggle to scale in high-dimensional design spaces. To address these limitations, researchers have explored machine learning-based methods, framing inverse design as a conditional generative task. GANs(Liu et al. 2018; So and Rho 2019; Yeung et al. 2021) and VAEs(Kingma, Welling et al. 2013; Tran, Nanthakumar, and Zhuang 2025; Kojima et al. 2023) can generate diverse candidate structures in a single forward pass.

However, GANs suffer from training instability, while VAEs often produce blurry reconstructions that compromise spectral fidelity. More recently, diffusion models have emerged as state-of-the-art in metasurface material design. Works such as (Zhang et al. 2023, 2024; Seo et al. 2025) have demonstrated that diffusion-based frameworks can outperform GAN and VAE baselines. Nevertheless, as discussed in Section 2.1, existing models simplify the design task by restricting parameter coverage and using coarse spectral constraints. To address these limitations, we propose MetaDiT, which integrates a contrastively pretrained spectrum encoder with a Diffusion Transformer based backbone. MetaDiT enables fine-grained control and exploration of a more complete metasurface design space.

3 Preliminaries

3.1 Metasurfaces and Scattering Spectrum

Metasurfaces are planar arrays of subwavelength dielectric structures, where each periodic unit cell U is defined by a set of geometric and material parameters. As shown in Figure 1, a unit cell consists of two key components:

1. A **substrate** with refractive index $r_{\text{sub}} \in \mathbb{R}$ and thickness $h_{\text{sub}} \in \mathbb{R}$.
2. A **meta-atom** with refractive index $r_{\text{atom}} \in \mathbb{R}$, thickness $h_{\text{atom}} \in \mathbb{R}$, and a binary geometric pattern matrix $P \in \mathbb{F}_2^{n \times n}$ encoding its structure (where n is the spatial resolution of the pattern).

The lattice constant $l_{\text{lattice}} \in \mathbb{R}$ governs the periodicity of the array. Together, these parameters fully describe the metasurface's optical properties.

The optical responses of a metasurface is characterized by its **scattering spectrum**, a complex-valued function $S(f) \in \mathbb{C}$ that describes the amplitude and the phase of scattered light at frequency f. This spectrum is goverened by the aforementioned material parameters, which collectively determine resonant scattering behavior.

3.2 Problem Definition

The goal of metasurface inverse design is to automatically generate a unit cell U that achieves a desired scattering spectrum $S(f)$. Traditional approaches rely on human expertise and iterative trial-and-error, which are often computationally expensive. Recent advances leverage leraned models M to directly predict unit cell geometries $\hat{U} = M(S)$ from target spectra. The generated designs are then validated via electromagnetic simulations, through which we can calculate the spectrum of a metasurface material. For convenience, we also refer U as the material design or structure.

4 Method

4.1 Dataset and Encoding

Following (Zhang et al. 2023), we adopt the dataset introduced in (An et al. 2020), which contains 170k+ metasurface designs with a high degree of geometric variability generated via a randomized algorithm. Each metasurface unit cell is represented by a binary pattern matrix of size 64×64, along with three continuous design parameters: the atom refractive index r_{atom}, atom thickness h_{atom}, and lattice constant l_{lattice}. The substrate refractive index $r_{\text{sub}} = 1.4$ and height $h_{\text{sub}} = 2\mu$m are held constant throughout the dataset.

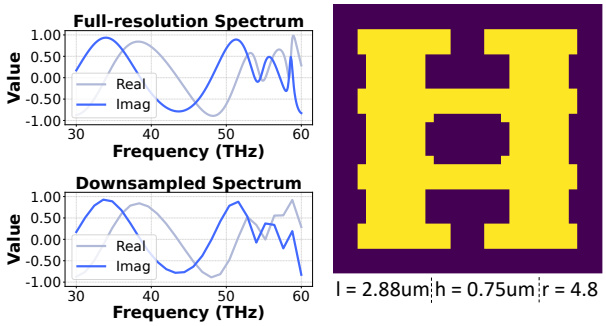

Figure 2: **An example of a metasurface material sample.** (Top left) Full-resolution scattering spectrum. (Bottom left) Downsampled spectrum, where the peak near ∼58 THz is attenuated due to loss of resolution. (Right) Corresponding pattern matrix; yellow regions (ones) indicate the meta-atoms, while dark purple (zeros) denotes the substrate. The l, h, r represents l_{lattice}, h_{atom} and r_{atom} respsectively.

We encode each unit cell as a three-channel image $U \in \mathbb{R}^{3\times64\times64}$, where the three channels respectively represent the design parameters r_{atom}, h_{atom}, and l_{lattice}. Within each channel, the original binary pattern matrix is preserved in structure: positions with value one are replaced by the corresponding design parameter, while zeros remain unchanged. This encoding scheme enables the model to generate all structural parameters jointly, in contrast to prior works that treat r_{atom}, h_{atom}, and l_{lattice} as fixed conditioning inputs.

The target in the dataset is the transmission scattering spectrum. It is provided at 301 discrete frequency points and is encoded as a two-channel sequence $S \in \mathbb{R}^{301\times2}$, where the channels represent real and imaginary values. Unlike prior works (Zhang et al. 2023, 2024), which downsample the spectrum to simplify the target, we preserve the full resolution to enable the model to capture and satisfy fine-grained spectral constraints. An example of a material structure and its corresponding spectrum is shown in Figure 2. Notably, the downsampled spectrum exhibits attenuation of the peak near ∼58 THz, highlighting the loss of fine-grained spectral information due to resolution reduction.

4.2 Architecture of MetaDiT

The core objective of MetaDiT is to learn the intrinsic relationship between a material's structure and its corresponding spectral response. To this end, we design the key components of MetaDiT by addressing two central questions that guide our architectural and training decisions.

1. How can we effectively encode fine-grained spectral conditions? Prior works (Zhang et al. 2023, 2024) encode the spectrum as a single global feature vector for conditioning (i.e. D-dimensional vector and is projected to the model hidden size), which we argue overlooks the semantic struc-

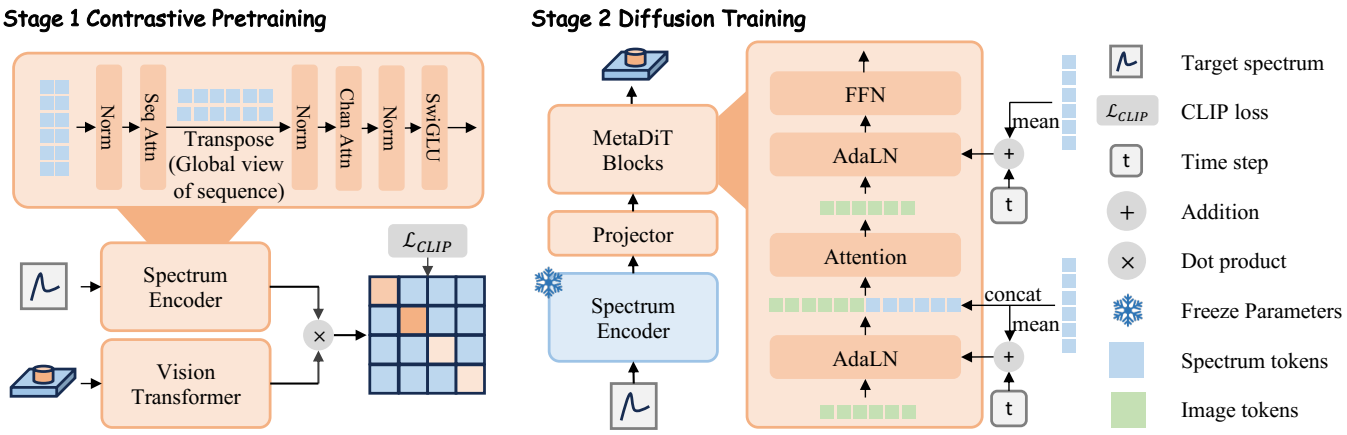

Figure 3: **Overview of MetaDiT**. We first train a spectrum encoder using contrastive learning, then the spectrum feature is pre-computed and fed into MetaDiT backbone for diffusion training.

ture inherent in spectral data. Preserving the sequence format allows the model to attend to localized spectral patterns that are critical for accurate material generation.

We propose to use a Transformer-based (Vaswani et al. 2017) encoder to encode the spectrum in its sequence format. To enable effective representation learning within the Transformer encoder, we first project the input spectrum from its raw form $S \in \mathbb{R}^{301\times 2}$ into a higher-dimensional feature $X_S \in \mathbb{R}^{301\times D_S}$ using a learnable linear embedding layer, where D_S is the hidden size of the spectrum encoder. We then apply position embedding and stack L layers of Transformer-based encoder to extract the feature.

However, we observe that the scattering spectrum is inherently a dual-variate sequence, where the embedded representation encodes both amplitude and phase components that are intrinsically coupled by physical laws. This coupling introduces rich inter-channel dependencies that are crucial for accurate modeling. While in typical Transformer encoder, the self-attention applied along the frequency (sequence) axis captures long-range dependencies and dispersion patterns, it alone is insufficient to fully represent the complex-valued nature of the spectrum. To address this limitation, we incorporate an additional attention mechanism along the channel axis, enabling the model to globally attend over the entire spectrum and explicitly model the inter-channel relationships. This dual-attention design facilitates more effective feature re-weighting and enhances the model's capacity to capture the underlying physical structure of the data. Our encoder block is formulated as follows:

$$\begin{aligned} X_l^{(1)} &= X_l + \text{Attn}_{\text{seq}}(\text{Norm}(X_l)) \\ X_l^{(2)} &= X_l^{(1)} + \left[\text{Attn}_{\text{chan}}(\text{Norm}((X_l^{(1)})^\mathsf{T}))\right]^\mathsf{T} \\ X_{l+1} &= X_l^{(2)} + \text{FFN}(\text{Norm}(X_l^{(2)})) \end{aligned} \tag{1}$$

Here, X_l denotes the input feature at layer l, Attn_{seq} and $\text{Attn}_{\text{chan}}$ represent self-attention along the sequence and channel dimensions respectively, Norm is a normalization layer, FFN is the feed-forward network, and T denotes the transpose operation.

The resulting feature forms a sequence with the same shape as X_S, where each token is a D_S-dimensional vector. We refer to these as spectrum tokens.

2. How can we enable finer condition injection in DiT? DiT uses adaLN for time and condition injection. While this strategy is effective for class conditioning, it forbids direct interaction between image tokens and spectrum tokens. In MetaDiT, our goal is to enhance the model's ability to capture the relationship between the spectrum and the corresponding material design. Therefore, we seek token-level interaction between tokens from these two modalities.

We first project the spectrum tokens to the hidden size of DiT using a lightweight projector, then the spectrum tokens and image tokens are concatenated along the sequence dimension, establishing an in-context conditioning paradigm. Let $\mathbf{X_S} \in \mathbb{R}^{N_s\times D_T}$ denote the embedded spectrum tokens, and $\mathbf{X_I} \in \mathbb{R}^{N_i\times D_T}$ denote the embedded image tokens, where N_s and N_i are the number of tokens for the spectrum and image, respectively, and D_T is the embedding dimension of DiT. The concatenated self-attention is

$$\mathbf{X}_{\text{attn}} = \text{Attn}\left([\mathbf{X_I}; \mathbf{X_S}] \in \mathbb{R}^{(N_i+N_s)\times D_T}\right) \tag{2}$$

we then discard the spectrum portion and retain only the updated image tokens:

$$\mathbf{X_I}' = \mathbf{X}_{\text{attn}}[:N_i] \in \mathbb{R}^{N_i\times D_T} \tag{3}$$

This design enables the model to jointly attend to both modalities and dynamically integrate contextual information into each token representation through self-attention. Notably, it introduces no additional parameters compared to explicitly adding a cross-attention module. Furthermore, we empirically show that this in-context self-attention mechanism outperforms the cross-attention baseline in Section 5.3.

We further inject coarse condition control into the model. Let $\mathbf{t} \in \mathbb{R}^{D_T}$ denote the timestep embedding. We first compute a pooled representation of the spectrum:

$$\mathbf{s}_{\text{pool}} = \frac{1}{L_s}\sum_{i=1}^{L_s} \mathbf{X_S}[i,:] \tag{4}$$

We then combine this with the timestep embedding:

$$\mathbf{z} = \mathbf{t} + \mathbf{s}_{\text{pool}} \in \mathbb{R}^{D_T} \tag{5}$$

the resulting condition signal $\mathbf{z}$ is then used for adaLN modulation. This coarse-to-fine conditioning allows the model to learn at two different levels of granularity.

4.3 Training Strategy

We conduct a two-stage training strategy for MetaDiT.

Stage 1: Contrastive Pretraining. In the first stage, we aim to train the spectrum encoder to learn semantically rich representations by aligning spectral and material features. We adopt a CLIP-style (Radford et al. 2021) contrastive training paradigm, jointly optimizing the spectrum encoder and a Vision Transformer that encodes the corresponding material design U. We extract the representation of U using the `[CLS]` token from ViT. Let $\mathcal{E}_S$ and $\mathcal{E}_U$ denote the spectrum encoder and the ViT encoder, respectively. The training objective is defined as:

$$\begin{aligned} \mathcal{L}_{\text{CLIP}} &= \frac{1}{2}\left[\text{CE}(e^{\tau} \cdot \mathbf{U}^{\mathsf{T}}\mathbf{S}) + \text{CE}(e^{\tau} \cdot \mathbf{S}^{\mathsf{T}}\mathbf{U})\right] \\ \mathbf{U} &= \frac{\mathcal{E}_U(U)}{\|\mathcal{E}_U(U)\|_2} \quad \mathbf{S} = \frac{\mathcal{E}_S(S)}{\|\mathcal{E}_S(S)\|_2} \end{aligned} \tag{6}$$

where τ is a leanbale temperature parameter and CE represents Cross Entropy loss.

Stage 2: Diffusion Training. In the second stage, we leverage the pre-computed spectral features from $\mathcal{E}_S$ to train the MetaDiT model via a denoising diffusion objective. Given a material structure U, we apply the forward diffusion process to corrupt it at timestep t: $U_t = \sqrt{\bar{\alpha}_t}\, U + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, α_t is the noise schedule, and $\epsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise.

The model is trained to predict the added noise, conditioned on the spectrum S, using the following objective:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{U,t,\epsilon,S}\, \|\epsilon - \epsilon_\theta(U_t, t, S)\|_2^2 \tag{7}$$

where ϵ_θ denotes the noise prediction model.

To further enhance conditional generation quality, we adopt classifier-free guidance (Ho and Salimans 2022) during diffusion training and sampling. Specifically, we randomly drop the spectral condition S during training, replacing it with a null embedding. At inference time, model predicts the noise using the following equation$\hat{\epsilon}_\theta = \epsilon_\theta(U_t, t, \varnothing) + w(\epsilon_\theta(U_t, t, S) - \epsilon_\theta(U_t, t, \varnothing))$, where $\epsilon_\theta(U_t, t, S)$ is the noise prediction conditioned on the spectrum S, $\epsilon_\theta(U_t, t, \varnothing)$ is the unconditional prediction, and w is the guidance scale hyperparameter that controls the strength of conditioning. The overview of the proposed architecture and the training strategy are shown in Figure 3.

4.4 Evaluation

In this paper, we use Accumulated Absolute Error (AAE), defined as $\text{AAE} = \sum_f |S_{\text{gt}}(f) - \hat{S}(f)|$, where $S_{\text{gt}}(f)$ is the ground truth spectrum. Our objective is to assess model accuracy across the entire frequency spectrum. While a model may perform well at the majority of frequency points, it can still fail at specific frequencies that are critical for the desired functionality. Conventional approaches that average performance across the frequency range can obscure such localized failures, masking important discrepancies that may significantly impact practical applications.

Furthermore, we propose an average AAE metric across K independently generated designs for the same target spectrum, defined as $\text{AAE\&K} = \max_i\{\text{AAE}_i\}_{i=1}^{K}$. This metric is designed to evaluate the model's ability to consistently generate multiple distinct yet accurate solutions.

Following (Zhang et al. 2024, 2023), we adopt a surrogate model M_{sur} to predict the spectral response of a given unit cell U, using its output $S_{\text{gt}} = M_{\text{sur}}(U)$ as a substitute for computationally expensive electromagnetic simulations. Specifically, we employ StarNet (Ma et al. 2024) equipped with an MLP head to perform the spectrum prediction. As demonstrated in Table 1, the surrogate model achieves remarkably low prediction error, validating its effectiveness as a fast and reliable approximation.

Model	#Param	MAE
PNN-1 (An et al. 2020)	-	0.0539
PNN-2 (Zhang et al. 2023)	-	0.0426
StarNet-MLP (Ours)	1.90M	**0.0084**

Table 1: **Prediction error of the surrogate model.** Our model achieves significantly lower spectrum estimation error compared to prior works that employ a dedicated Prediction Neural Network (PNN) for surrogate modeling. Results adopted from the original paper.

5 Experiment

In this section, we aim to answer the following questions: (1) Can MetaDiT outperform all baselines? (2) Are all components of MetaDiT essential? (3) Can we simply scale up MetaDiT for better performance? We first elaborate our experimental setups in Section 5.1 and answer the above questions in Section 5.2, Section 5.3 and Section 5.4.

5.1 Experimental Setups

Dataset. We adopt the dataset from (An et al. 2020) and encode it following the procedure detailed in Section 4.1. The dataset is randomly split into training, validation, and test sets with a ratio of 8:1:1.

Baselines. We establish two key baselines for comparison: (1) a standard DiT baseline and (2) a carefully reproduced version of MetaDiff (Zhang et al. 2023), trained under our data encoding framework while maintaining fidelity to the original methodology. We also implement an average predictor that predicts the average of the spectrum values as a basic baseline.

Implementation. For the spectrum encoder, we stack four layers of spectrum encoder blocks and set the hidden dimension to $D_S = 256$. The FFN the encoder is enhanced with SwiGLU (Shazeer 2020). For MetaDiT, we vary the model

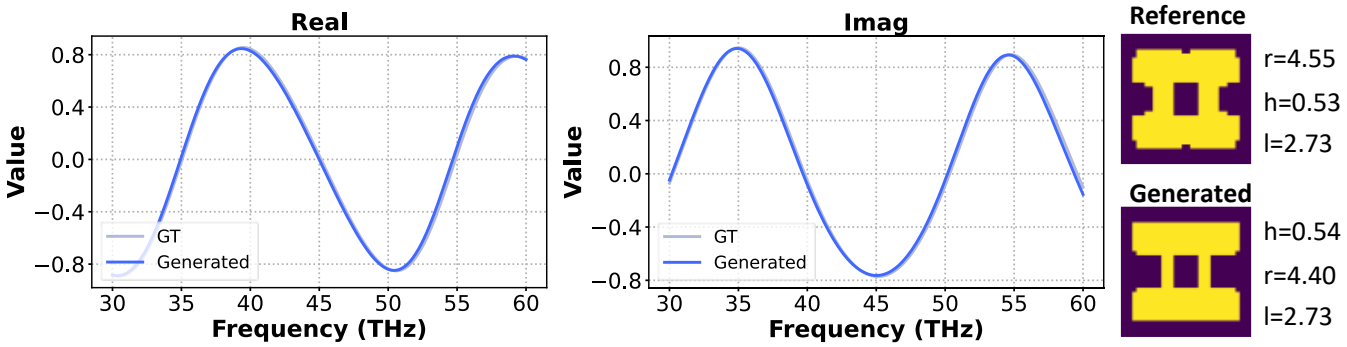

Figure 4: **Visualization of the generated results.** (Left) Comparsion of the ground truth transmission spectrum and the generated one. (Right) Comparsion bettwen reference material and the generated material. r, h, l represents r_{atom}, h_{atom}, l_{lattice} respectively. The unit of h_{atom} and l_{lattice} is μm.

Model	#Param	MAE	AAE	AAE&2	AAE&4
AVG[1]	-	0.5860	352.7424	352.7424	352.7424
MetaDiff	32.56M	0.1861	112.0591	170.6253	258.9915
MetaDiff-HR[2]	33.41M	0.1315	79.1365	100.2521	125.4889
DiT	32.80M	0.1677	100.9437	138.0702	187.7744
MetaDiT	32.57M	**0.0801**	**48.2495**	**58.8007**	**68.7275**

Table 2: **Main results of the reproduced baselines and MetaDiT**. [1] means averge baseline, we calculate the MAE and AAE when model designs the average of the spectrum value. [2] means high resolution spectrums are used.

Method	MAE	AAE
MetaDiT	0.0801	48.2495
w/o Pretrained Encoder	0.1370	82.4838
w/o Coarse condition	0.0996	59.9662
w/ Cross-attention	0.0927	55.8119

Table 3: **Ablation results of MetaDiT.** Encoded spectrum features significantly improves the performance, In-context condition and coarse-to-fine conditioning are also essential for MetaDiT.

capacity by adjusting the hidden size, the projector is a simple linear layer. Both models employ sinusoidal positional embeddings (Vaswani et al. 2017). All experiments are optimized using AdamW (Loshchilov and Hutter 2017), with a learning rate of 10^{-4} and trained for 500 epochs. We apply cosine annealing as the learning rate schedule. For simplicity and fair comparison, we omit Exponential Moving Average (EMA) and weight decay, although these techniques can further enhance performance. More implementation details can be found in Appendix.

Environment. We implement the model using PyTorch (Paszke 2019), with DeepSpeed ZeRO 2 (Rajbhandari et al. 2020) for better training efficiency. We use standard acceleration techniques like bfloat16 (Kalamkar et al. 2019) and gradient checkpointing (Herrmann et al. 2019), diffusion training is conducted using full precision. All experiments are conducted using 4×Nvidia A100 80GB. For consistency, we fix the random seed to 0.

5.2 Main Results

In this section, we demonstrate that *MetaDiT can outperform all baselines*. In our setting, the model is required to design all parameters while adhering to high-resolution spectral constraints. As shown in Table 2, directly using high-resolution spectra as input improves the performance of MetaDiff by approximately 29.3%, underscoring the importance of fine-grained conditioning. However, MetaDiff still struggles to fully capture the relationship between spectrums and material structures. MetaDiT further improves the performance by an additional **39.1%** over MetaDiff-HR.

Following MetaDiff, we also implement a vanilla DiT baseline conditioned on high-resolution spectra represented as a single feature vector. As shown in Table 2, this model performs even worse than MetaDiff-HR, highlighting the necessity of more expressive condition injection mechanisms. By incorporating the techniques described in Section 4, MetaDiT achieves a substantial improvement of **52.2%** over the vanilla DiT baseline.

To evaluate AAE&K, we sample multiple designs by varying the random seed across $\{0, 7, 42, 3407\}$. As shown in Table 2, MetaDiT demonstrates greater robustness in generating diverse metasurface designs. Even under the worst-case scenario, which is comparing the maximum AAE across four different samples, MetaDiT achieves substantially lower error than all baselines. We visualize several generated results in Figure 4, see Appendix for more results.

5.3 Ablation Studies

In this section, we ablate key components of MetaDiT and provide insights into the architectural choices and training strategies that contribute to its performance.

How important is proper spectrum encoding? We investigate the impact of spectrum encoding by replacing the pretrained encoder and projector with a simple MLP projector and retraining the model. As shown in Table 3, this leads to a substantial performance drop of 41.5%, highlight-

Model	#Param	Width	#Layer	#Head
MetaDiT	32.57M	384	12	6
MetaDiT-B	57.78M	512	12	8
MetaDiT-L	129.73M	768	12	12

Table 4: **Model Specifications of MetaDiT**, we vary the width and the number of attention heads to implement different sizes of MetaDiT.

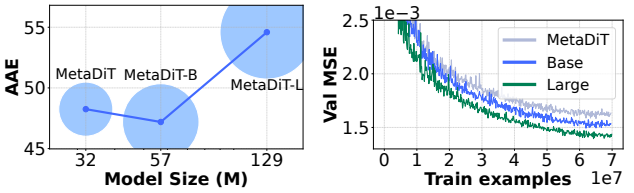

Figure 5: **Scaling the MetaDiT size.** (Left) The final AAE of MetaDiT in different sizes. (Right) The MSE loss of model predicted noise on validation set.

ing the critical role of the pretrained encoder and sequential spectrum representation. Notably, the performance remains 18.3% higher than the vanilla DiT baseline, further underscoring the benefits of sequential formatting and the interaction between spectrum and image tokens.

Is in-context conditioning effective? To assess the effectiveness of in-context conditioning, we remove it and instead introduce an additional cross-attention layer after the self-attention on image tokens. In this setup, spectrum tokens serve as keys and values, while image tokens act as queries. This modification increases the parameter count by approximately 21.7%. However, as shown in Table 3, this approach underperforms compared to in-context conditioning, validating the effectiveness of our design for enabling direct interaction between image and spectrum tokens.

Is coarse condition injection necessary? While the in-context conditioning mechanism enables fine-grained, token-level interaction between image and spectrum tokens, we further introduce a coarse condition by pooling the spectrum and injecting it via the adaLN modulation. To assess its impact, we ablate the coarse condition by removing it from the adaLN inputs, leaving only the timestep embedding. As shown in Table 3, this results in a performance drop of 10.6% compared to MetaDiT, validating our coarse-to-fine design. This demonstrates that learning across two levels of granularity enhances model performance.

5.4 Is simple scaling effective?

In this section, we aim to verify that whether better performance can be obtained by simply scaling up the size of the diffusion backbone. Specifically, we adjust the model width (hidden size), number of layers, and number of attention heads. As summarized in Table 4, we construct two additional variants of MetaDiT. All models are trained under identical settings and evaluated consistently to assess the impact of scale on performance.

The results are presented in Figure 5 (Left). Scaling up to MetaDiT-B, which introduces a 77.4% increase in parameters over MetaDiT, yields a modest performance gain of 2.3%. However, further scaling to MetaDiT-L, with an additional 124.5% increase in parameters relative to MetaDiT-B, results in a performance drop of 15.7%. These findings suggest that simply enlarging the diffusion backbone offers diminishing returns and may even degrade design accuracy.

As illustrated in Figure 5 (Right), scaling up the model size consistently lowers the diffusion MSE on the validation set, suggesting improved reconstruction quality. However, faithfully matching the target spectrum **requires more than reconstruction fidelity**; it demands precise alignment with the input condition. This is analogous to instruction-following in the visual generation community (Ghosh, Hajishirzi, and Schmidt 2023), where success is measured not solely by image quality, but by how well the output adheres to the input prompt. This insight reinforces our central claim: the fundamental challenge lies in designing a model that effectively captures the intricate relationship between material structure and its corresponding scattering spectrum.

6 Conclusion

This paper introduces MetaDiT, a novel framework capable of designing the complete set of available metasurface material parameters while accurately satisfying high-resolution scattering spectra. By integrating contrastive pretraining, a dual-attention architecture, and a coarse-to-fine condition injection strategy, MetaDiT achieves state-of-the-art performance across all evaluated baselines. Through ablation studies, we highlight the contribution of each design choice. Moreover, scaling experiments reveal that superior design accuracy does not solely stem from improved reconstruction fidelity, but rather from the model's enhanced ability to learn the underlying physical relationship between material structure and its corresponding scattering spectrum.

Acknowledgements

This research was supported by Priority 2030 Federal Academic Leadership Program.

References

Al-Zawqari, A.; Vandersteen, G.; and Ferranti, F. 2023. Gaussian process regression for the modeling of metalenses. In *PIERS*, 49–53. IEEE.

An, S.; Zheng, B.; Shalaginov, M. Y.; Tang, H.; Li, H.; Zhou, L.; Ding, J.; Agarwal, A. M.; Rivero-Baleine, C.; Kang, M.; et al. 2020. Deep learning modeling approach for metasurfaces with high degrees of freedom. *Optics Express*, 28(21): 31932–31942.

Aththanayake, A.; Lininger, A.; Strangi, C.; Griswold, M. A.; and Strangi, G. 2025. Tunable holographic metasurfaces for augmented and virtual reality. *Nanophotonics*, (0).

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.;

Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; et al. 2024. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators. OpenAI technical report.

Chen, W.; Sun, R.; Lee, D.; Portela, C. M.; and Chen, W. 2025. Generative inverse design of metamaterials with functional responses by interpretable learning. *Advanced Intelligent Systems*, 7(6): 2400611.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Dong, Y.; An, S.; Jiang, H.; Zheng, B.; Tang, H.; Huang, Y.; Zhao, H.; and Zhang, H. 2025. Advanced deep learning approaches in metasurface modeling and design: A review. *Progress in Quantum Electronics*, 100554.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Fu, X.; Wang, P.; Liu, Y.; Fu, Y.; Cai, Q.; Wang, Y.; Yang, S.; and Cui, T. J. 2025. Fundamentals and applications of millimeter-wave and terahertz programmable metasurfaces. *Journal of Materiomics*, 11(1): 100904.

Ghosh, D.; Hajishirzi, H.; and Schmidt, L. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 36: 52132–52152.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*, 27.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Herrmann, J.; Beaumont, O.; Eyraud-Dubois, L.; Hermann, J.; Joly, A.; and Shilova, A. 2019. Optimal checkpointing for heterogeneous chains: how to train deep neural networks with limited memory. *arXiv preprint arXiv:1911.13214*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, J.; Mengu, D.; Tzarouchis, D. C.; Edwards, B.; Engheta, N.; and Ozcan, A. 2024. Diffractive optical computing in free space. *Nature Communications*, 15(1): 1525.

Jeong, H.-D.; Kim, H.; and Lee, S.-Y. 2024. Review of metasurfaces with extraordinary flat optic functionalities. *Current Optics and Photonics*, 8(1): 16–29.

Kalamkar, D.; Mudigere, D.; Mellempudi, N.; Das, D.; Banerjee, K.; Avancha, S.; Vooturi, D. T.; Jammalamadaka, N.; Huang, J.; Yuen, H.; et al. 2019. A study of BFLOAT16 for deep learning training. *arXiv preprint arXiv:1905.12322*.

Khorasaninejad, M.; and Capasso, F. 2017. Metalenses: Versatile multifunctional photonic components. *Science*, 358(6367): eaam8100.

Kildishev, A. V.; Boltasseva, A.; and Shalaev, V. M. 2013. Planar photonics with metasurfaces. *Science*, 339(6125): 1232009.

Kim, Y.; Choi, T.; Lee, G.-Y.; Kim, C.; Bang, J.; Jang, J.; Jeong, Y.; and Lee, B. 2024. Metasurface folded lens system for ultrathin cameras. *Science Advances*, 10(44): eadr2319.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kojima, K.; Koike-Akino, T.; Wang, Y.; Jung, M.; and Brand, M. 2023. Inverse design of two-dimensional freeform metagrating using an adversarial conditional variational autoencoder. In *Photonic and Phononic Properties of Engineered Nanostructures XIII*, volume 12431, 48–56. SPIE.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Koshelev, K.; Lepeshov, S.; Liu, M.; Bogdanov, A.; and Kivshar, Y. 2018. Asymmetric metasurfaces with high-Q resonances governed by bound states in the continuum. *Physical review letters*, 121(19): 193903.

Koshelev, K. L.; Sadrieva, Z. F.; Shcherbakov, A. A.; Kivshar, Y. S.; and Bogdanov, A. A. 2023. Bound states in the continuum in photonic structures. *Phys.-Usp*, 93: 528–553.

Li, S.; Lin, H.-C.; and Hsu, C. W. 2023. Inverse Design of Nonlocal Metasurfaces Using Augmented Partial Factorization. In *ACES*, 1–2. IEEE.

Liu, Z.; Luo, Y.; Huang, H.; Zhang, E.; Li, S.; Fang, J.; Shi, Y.; Wang, X.; Kawaguchi, K.; and Chua, T.-S. 2025. NEXT-MOL: 3d diffusion meets 1d language modeling for 3d molecule generation. *arXiv preprint arXiv:2502.12638*.

Liu, Z.; Zhu, D.; Rodrigues, S. P.; Lee, K.-T.; and Cai, W. 2018. Generative model for the inverse design of metasurfaces. *Nano letters*, 18(10): 6570–6576.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; and Fu, Y. 2024. Rewrite the stars. In *CVPR*, 5694–5703.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Niu, C.; Phaneuf, M.; and Mojabi, P. 2023. A diffusion model for multi-layered metasurface unit cell synthesis. *IEEE Open Journal of Antennas and Propagation*, 4: 654–666.

Park, J.-S.; Lim, S. W. D.; Amirzhan, A.; Kang, H.; Karrfalt, K.; Kim, D.; Leger, J.; Urbas, A.; Ossiander, M.; Li, Z.; et al. 2024. All-glass 100 mm diameter visible metalens for imaging the cosmos. *ACS nano*, 18(4): 3187–3198.

Paszke, A. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PmLR.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Saifullah, Y.; Wu, N.; Wang, H.; Zheng, B.; Qian, C.; and Chen, H. 2025. Deep learning in metasurfaces: from automated design to adaptive metadevices. *Advanced Photonics*, 7(3): 034005–034005.

Seo, D.; Um, S.; Lee, S.; Ye, J. C.; and Chung, H. 2025. Physics-guided and fabrication-aware inverse design of photonic devices using diffusion models. *arXiv preprint arXiv:2504.17077*.

Shazeer, N. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

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

Tanriover, I.; Lee, D.; Chen, W.; and Aydin, K. 2022. Deep generative modeling and inverse design of manufacturable free-form dielectric metasurfaces. *ACS Photonics*, 10(4): 875–883.

Tian, Z.; Yang, Y.; Zhou, S.; Zhou, T.; Deng, K.; Ji, C.; He, Y.; and Liu, J. S. 2024. High-dimensional Bayesian optimization for metamaterial design. *Materials Genome Engineering Advances*, 2(4): e79.

Tian, Z.; Zhu, X.; Surman, P. A.; Chen, Z.; and Sun, X. W. 2025. An achromatic metasurface waveguide for augmented reality displays. *Light: Science & Applications*, 14(1): 94.

Tran, T. V.; Nanthakumar, S.; and Zhuang, X. 2025. Deep learning-based framework for the on-demand inverse design of metamaterials with arbitrary target band gap. *npj Artificial Intelligence*, 1(1): 2.

Tu, S.; Dai, Q.; Cheng, Z.-Q.; Hu, H.; Han, X.; Wu, Z.; and Jiang, Y.-G. 2024. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 7882–7891.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 1921–1930.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, C.; Zhao, Z.; and Zhang, X. S. 2023. Inverse design of magneto-active metasurfaces and robots: Theory, computation, and experimental validation. *Computer Methods in Applied Mechanics and Engineering*, 413: 116065.

Wang, L.; Rong, Y.; Xu, T.; Zhong, Z.; Liu, Z.; Wang, P.; Zhao, D.; Liu, Q.; and Wu, S. 2025. DiffSpectra: Molecular Structure Elucidation from Spectra using Diffusion Models. *arXiv preprint arXiv:2507.06853*.

Wang, W.; Liu, J.; Lin, Z.; Yan, J.; Chen, S.; Low, C.; Hoang, T.; Wu, J.; Liew, J. H.; Yan, H.; et al. 2024. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*.

Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; and Jaakkola, T. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.

Xu, Y.; Zhao, T.; Chen, G.; Liu, A.; Huang, Z.; Wu, J.; Zhang, C.; Jin, B.; Chen, J.; Wu, P.; et al. 2025. A dual-band programmable metasurface for terahertz beam steering. *Applied Physics Letters*, 126(19).

Yang, G.; Xiao, Q.; Zhang, Z.; Yu, Z.; Wang, X.; and Lu, Q. 2025. Exploring AI in metasurface structures with forward and inverse design. *iScience*, 28(3).

Yeung, C.; Tsai, R.; Pham, B.; King, B.; Kawagoe, Y.; Ho, D.; Liang, J.; Knight, M. W.; and Raman, A. P. 2021. Global inverse design across multiple photonic structure classes using generative deep learning. *Advanced Optical Materials*, 9(20): 2100548.

Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Shysheya, S.; Crabbé, J.; Sun, L.; Smith, J.; et al. 2023. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*.

Zhang, Z.; Yang, C.; Qin, Y.; Feng, H.; Feng, J.; and Li, H. 2023. Diffusion probabilistic model based accurate and high-degree-of-freedom metasurface inverse design. *Nanophotonics*, 12(20): 3871–3881.

Zhang, Z.; Yang, C.; Qin, Y.; Zheng, Z.; Feng, J.; and Li, H. 2024. Addressing high-performance data sparsity in metasurface inverse design using multi-objective optimization and diffusion probabilistic models. *Optics Express*, 32(23): 40869–40885.

Zhou, H.; Zhao, C.; He, C.; Huang, L.; Man, T.; and Wan, Y. 2024. Optical computing metasurfaces: applications and advances. *Nanophotonics*, 13(4): 419–441.