Supplementary Information for Artificial Intelligence-Generated Photonics: Mapping Optical Properties to Subwavelength Structures Directly via a Diffusion Model

Shijie Rao, Kaiyu Cui*, Jiawei Yang, Yali Li, and Shengjin Wang, Xue Feng, Fang Liu, Wei Zhang, Yidong Huang

* Correspondent author: kaiyucui@tsinghua.edu.cn
Affiliation: Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China

Supplementary Note 1. Comparison with existing optical inverse design neural network models

Early works on DNN-based inverse design usually adopt fully connected networks (FCNs) [20,21,23,52]. The FCNs are easier to train, but they do not have generative abilities and can only achieve a one-to-one mapping. Therefore, they are usually performed in a predefined and limited design space where the inverse problem is a one-to-one mapping.

Generative models have become an ideal DNN-based approach to solve the problem of non-uniqueness and enlarge the design space. Generative models like generative adversarial networks (GANs) can be designed to generate outputs from the given inputs and a set of random noise. The introduced randomness makes it possible to map a single input to multiple outputs. GANs based on convolutional neural networks (CNNs) are especially popular because usually photonic structures can be described graphically. Several GANs have been proposed in recent years to achieve a direct-mapping inverse design on certain photonic devices [27-29,31,33,34]. Besides GANs, CNNs, and FCNs, Transformer models have also been adopted to perform optical inverse design [32]. It is particularly efficient in designing 1D structures. The Transformer architecture also has the potential to be scaled to large size models.

These methods have shown incredible abilities but they still face some common challenges from the direct-mapping inverse design itself and the limitations of DNNs:

1. The solution of one specific inverse design problem may be non-unique, or none at all. In most cases, we cannot find the photonic structure that exactly matches the given optical property. Instead, we should try to find compatible solutions, which results in a fuzzy search problem. However, one DNN is usually trained to fit an accurate mapping function. It may not perform well on fuzzy search.
2. The design space of the DNNs relies on the training dataset. The generated photonic structures are usually restricted to a specific distribution that depends on the training dataset. The ability of GANs to create new unseen structures is not strong enough.
3. The acceptable input optical properties are limited. During the training process, the optical properties are usually acquired directly by forward prediction and fed to the network as input. In this way, the trained DNN may only accept realistic optical properties as inputs, which is not practical enough. From the perspective of the algorithm users, for example, if we want to design a long-wave pass filter, we do not know the complete actual response of the filter, but only some design requirements such as the cut-off frequency and the pass band. However, such abstract requirements may not be acceptable to DNNs because they are usually unseen during training.

We proposed several techniques to solve these problems. Different from DNNs, CNNs, or Transformer networks, diffusion network generates images from a random noise. It is a generative network rather than a discriminative network. Therefore, it has the potential to achieve a one-to-many mapping. Compared with existing works, our method has many unique features has several advantages:

1. **This is the first work achieving optical inverse design by diffusion network**, to the best of our knowledge. The diffusion network is much easier to train and more powerful than GANs. As the latest and strongest generative network, it can break the limitations of GANs, such as the difficulties in training and scaling up. Diffusion networks and Transformer-based models have been adopted and proved their superiority in modern large-size image and video generation models. Therefore, the proposed network also has the potential to be scaled to large-size models and provide extreme generative capabilities. Moreover, the proposed method can also be easily generalized to design other subwavelength structures, such as quasicrystal, aperiodic, and supercell structures.

2. **The proposed network ensures fabrication limits and provides a much larger design space.** It can generate freeform shapes with various topologies rather than regular or predefined shapes. More importantly, conforming to the fabrication limits are very important to inverse design methods. Most of the existing GAN-based inverse design methods, such as Refs. 26, 28, 29, and 30, cannot ensure the fabrication limits. However, the generated shapes are fabricable based on our method, since the fabrication limits are learned by the network during our training process. The fabrication of inverse-designed meta-atoms is also verified by experiments shown in Fig. 5.

3. **The proposed method solves the domain gap between complete optical properties and abstract design requirements by the specially designed prompt encoder network and self-supervised learning.** It is user-friendly and can accept abstract design requirements for realistic problems. The domain gap between training data and practical usage is a huge challenge for the practical usage of the deep learning-based inverse design method. Among previous works, Ref. 29 has attempted to solve this problem by proposing the ‘contrast vector’. However, the ‘contrast vector’ is also a predefined property that has great limitations. Moreover, two networks are trained to accept normal transmission response and ‘contrast vector’ separately in Ref. 29. Instead, the proposed prompt encoder network in this work solves the gap between abstract design demands and realistic optical properties for the first time by self-supervised learning. We have achieved a single network that can accept both complete optical properties and abstract design requirements, as shown in Figs. 4c-f. Moreover, the neural network model can still function normally when encountering unseen inputs due to the fuzzy search ability. These features make our method truly practical.

4. **Our method is not just aimed at a specific application** (such as structural color or polarizer), but can be used as a general approach across a range of applications

with the proposed mask mechanism, as demonstrated in Figs. 4c-h. It provides very high flexibility and thus empowers AI for photonics.

Supplementary Note 2. Fine-tuning the model via transfer learning.

In this work, our results are mainly based on the inverse design of 220-nanometer-thick silicon meta-atoms. Here, we take the inverse design of 600-nanometer-thick TiO_2 meta-atoms as an example to demonstrate that our method can be easily converted to design meta-atoms based on other materials and other thicknesses via transfer learning. The substrate of the TiO_2 meta-atoms is also SiO_2.

Our inverse design method involves four DNN models: an image encoder-decoder network, a forward prediction network, a prompt encoder network, and a latent diffusion network. Note that the image encoder-decoder network and prompt encoder network are relatively universal. We do not need to retrain them if the material or the thickness of the meta-atom changes. Only the training process of the forward prediction model requires the data acquired from a numerical simulation. During the training procedure of the latent diffusion model, the optical properties are quickly predicted by the forward prediction model, and only a dataset of 2D geometries is needed. Therefore, to convert our method to design TiO_2 meta-atoms, the most important step is to fine-tune the forward prediction model.

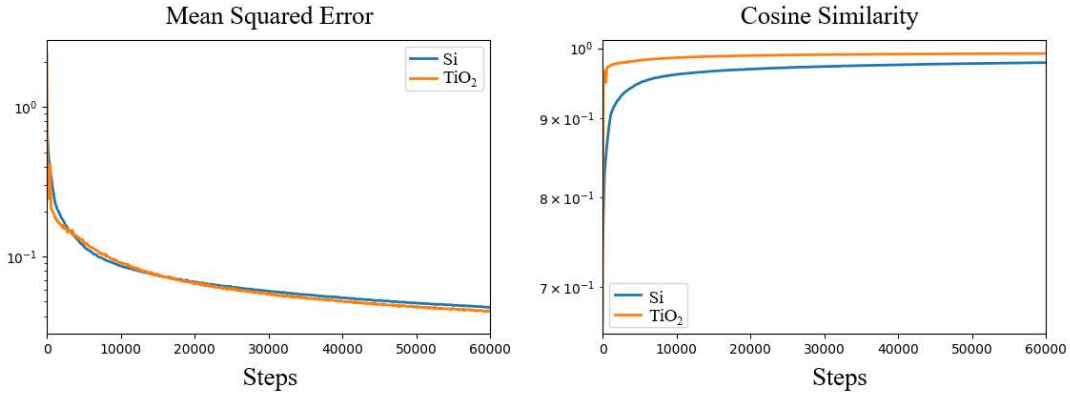

Fig. S1 The mean squared error and cosine similarity produced during testing.

During the training process of the forward prediction network for silicon meta-atoms, we build a training set containing 10^5 samples and train the network for approximately 90000 steps with a batch size of 512. To fine-tune the forward prediction network for TiO_2 meta-atoms, 20000 samples are simulated for constructing the training set. Then, we fine-tune the network for 60000 steps with the same batch size. The mean squared error and cosine similarity produced on the testing set during training are shown in Fig. S1. Through transfer learning, the forward prediction network can reach the same performance as that attained before in fewer training steps while using less training data. After the training process, the mean squared error and cosine similarity reached 0.043 and 99.22% during testing, respectively.

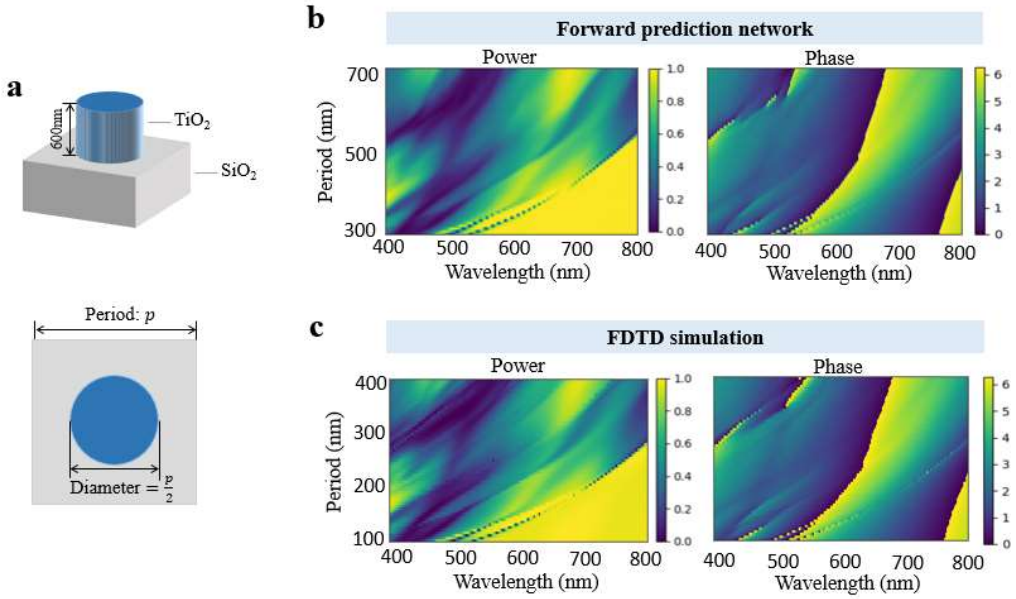

Fig. S2 Parameter sweeping results obtained for TiO_2 meta-atoms by the fine-tuned network. a. The specifications of the meta-atom used for parameter sweeping. **b.** Parameter sweeping results obtained by our proposed forward prediction network. **c.** Parameter sweeping results were obtained via FDTD simulation.

To test the performance achieved by the fine-tuned forward prediction network for TiO_2 meta-atoms, we also conduct a parameter sweeping task. We sweep the period of the 600-nanometer-thick TiO_2 cylinder meta-atom from 300 nm to 700 nm, and the diameter of the TiO_2 cylinder is set to half of the period, as shown in Fig. S2a. The sweeping results obtained from our forward prediction network (Fig. S2b) also fit well with the FDTD simulation outputs (Fig. S2c). These results indicate the excellent performance of our forward prediction network and the transfer learning strategy.

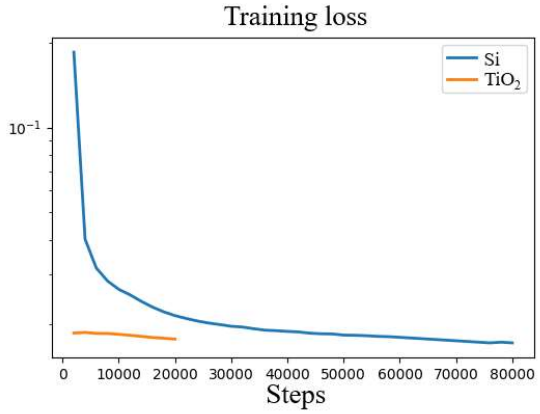

Fig. S3 Training losses induced by the latent diffusion network when inversely designing silicon and TiO_2 meta-atoms.

The next step is to fine-tune the latent diffusion network using the fine-tuned forward prediction network. We use the same 2D geometry dataset that we employed to train the latent diffusion network for silicon meta-atoms. The training loss is plotted in Fig. S3. We train the original latent diffusion network for silicon meta-atoms for approximately 80,000 steps with a batch size of 128 until the loss converges. During the fine-tuning process for TiO_2 meta-atoms, we find that the loss converges very quickly; therefore, we stop the training procedure after 20000 steps. Therefore, these experiments indicate that transfer learning can greatly reduce the incurred training cost.

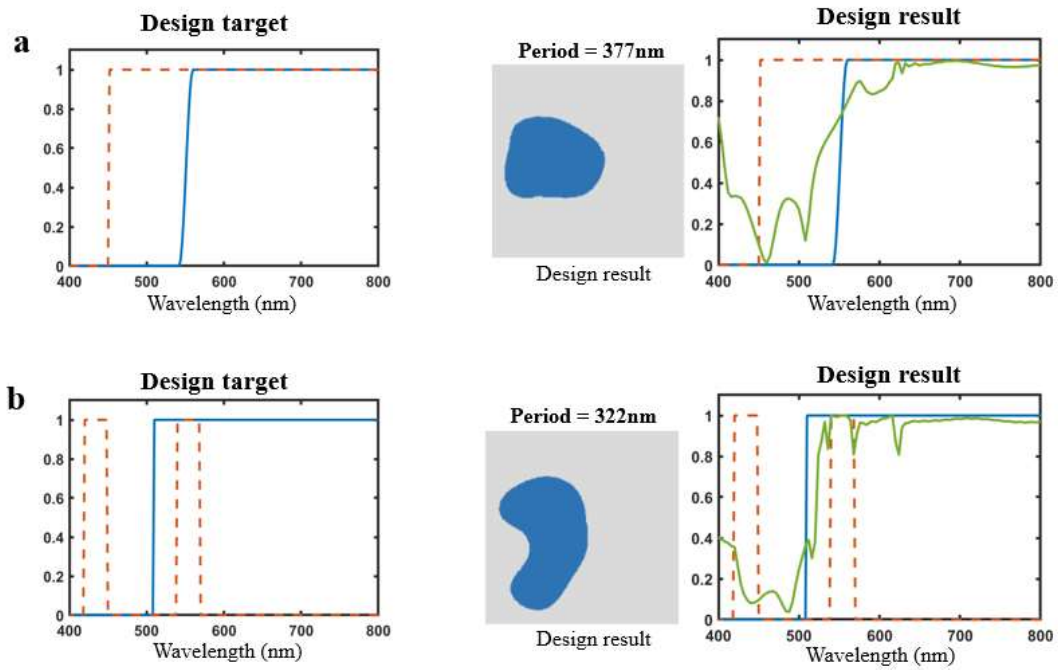

Fig. S4 Inverse design examples involving TiO_2 meta-atoms. a. The inverse design results of a longwave pass filter. **b.** The inverse design results obtained with the requirements that the power transmission rate should be high at approximately 550 nm and low at approximately 430 nm.

To illustrate the inverse design capability of the fine-tuned network, several inverse design examples involving TiO_2 meta-atoms are shown in Fig. S4. These meta-atoms are generated without the C4 symmetry constraint subject to the given transmission power responses. The FDTD simulation results (green curves) of the inversely designed meta-atoms agree well with the given design targets (blue curves). The direct mapping ability of the inverse design network trained via transfer learning can be confirmed.

Supplementary Note 3. Design of the forward prediction network.

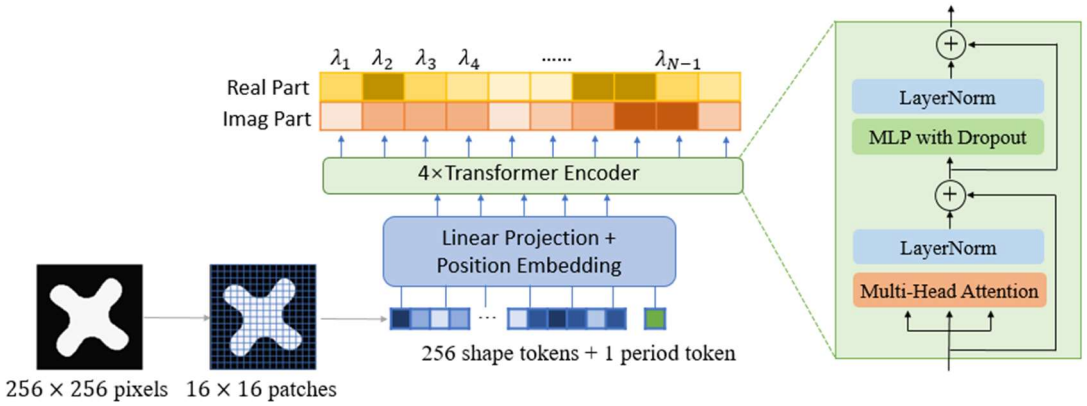

Fig. S5. The architecture of the transformer-based forward prediction network.

Our forward prediction network takes the 2D geometry and period of the target meta-atom as inputs. As shown in Fig. S5, the 2D geometry is described by a binary image with a size of 256×256. We slice the binary image into 16×16 patches. Each patch contains 16×16 pixels and is projected to a token with a feature dimensionality of 256. Each patch works as an input token of the transformer. The period of the meta-atom is randomly chosen from 300 nm to 700 nm and attached to the end of the token sequence. The period value is also projected to a token with the same dimensionality. Therefore, 256 shape tokens and 1 period token form an input feature sequence $\boldsymbol{x} \in R^{257\times256}$. Sinusoidal positional embeddings are then added to the feature sequence. Furthermore, the feature sequence is processed by 4 transformer encoder[1] blocks. Each block has 8 heads, and the head size is set to 256. In this way, the output sequence $\boldsymbol{y} \in R^{257\times256}$ has 257 tokens with a dimensionality of 256. Finally, each output token is mapped to a vector with 2 dimensions by another linear projection, and we can obtain the final sequence $\in R^{257\times2}$. The first dimension of $\boldsymbol{z}$ indicates 257 sample points with wavelengths from 400 nm to 800 nm, and the second dimension contains the real and imaginary parts of the transmission response. The network is trained by the Adam[2] optimizer with a batch size of 512 for 320 epochs. The learning rate is set to 0.0002, and the weight decay rate is set to 0.0001.

It is trained to be a polarization-sensitive structure. To predict the transmission response generated under vertically polarized incident light, we can simply rotate the geometry of the meta-atom by 90 degrees and conduct prediction again using the same forward prediction network.

Supplementary Note 4. Performance of the forward prediction network.

According to our quantitative analysis, the mean squared error and cosine similarity of our forward prediction network reach 0.040 and 98.2%, respectively. Our forward prediction network requires approximately 42 ms on an Intel Core i7-11700 CPU @ 2.5 GHz and approximately 1.4 ms on an NVIDIA RTX 2080Ti GPU, while the FDTD simulation requires approximately 70 s on average on the same CPU. Our network, which is powered by a GPU, can operate 50000 times faster than the FDTD simulation.

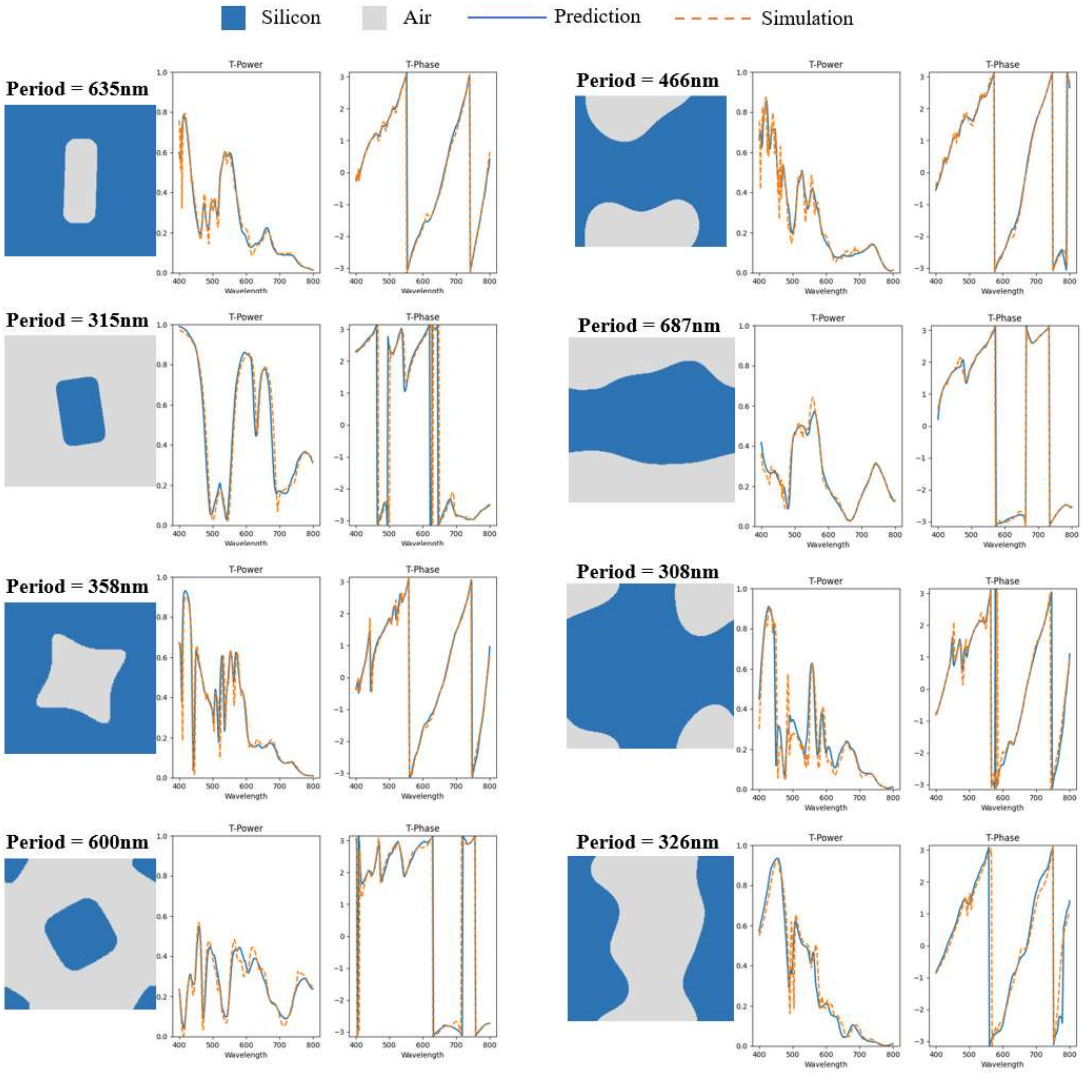

Fig. S6 Visualization results produced by the forward prediction network.

A more intuitive visualization of the results is shown in Fig. S6. We randomly select some samples from the testing set. These meta-atoms include rectangular pillars, rectangular holes, C4 symmetric pillars, C4 symmetric holes, and some randomly shaped structures. The incident light is set to horizontal polarization. The predicted transmission responses match well with the FDTD simulation outputs. These results illustrate that our forward prediction network can greatly accelerate the forward

prediction process while maintaining acceptable prediction error. This acceleration is vital for effectively training the latent diffusion network.

Supplementary Note 5. Analysis of the attention maps produced by the forward prediction network.

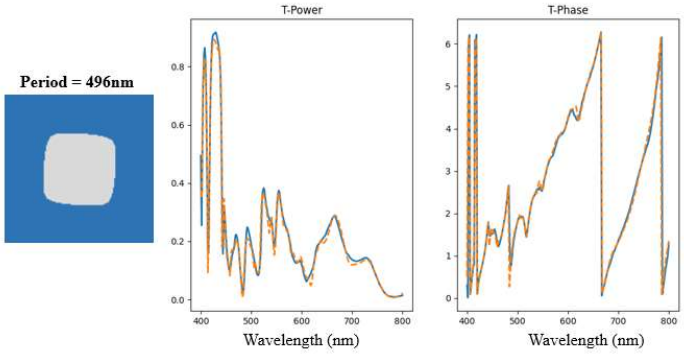

Fig. S7 The meta-atom used for analyzing the attention maps of the forward prediction network.

We choose a random meta-atom as an example to analyze the attention maps yielded by the forward prediction network and try to understand what the network has learned to accurately predict the transmission response. The meta-atom is shown in Fig. S7. The prediction results (blue curves) obtained for the transmission power and phase response highly match the FDTD simulation results (orange dashed curves).

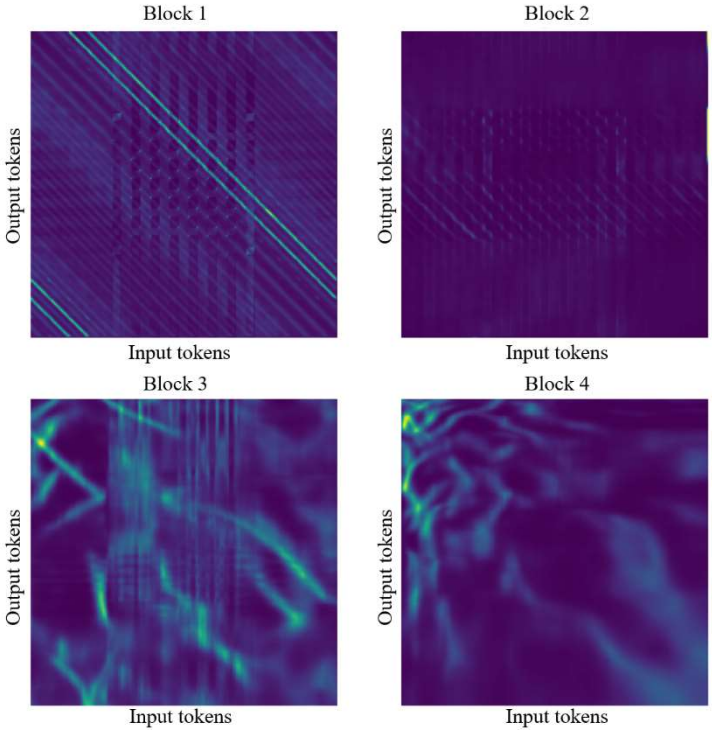

Fig. S8 The attention maps of the four transformer blocks.

Four transformer encoder blocks are contained in our forward prediction network, and each block generates an attention map that maps the input tokens to the output tokens via a self-attention mechanism. The generated attention maps are shown in Fig. S8. During the test, we find that the first two attention maps are relatively stable under different inputs, while the last two attention maps are highly related to the input data.

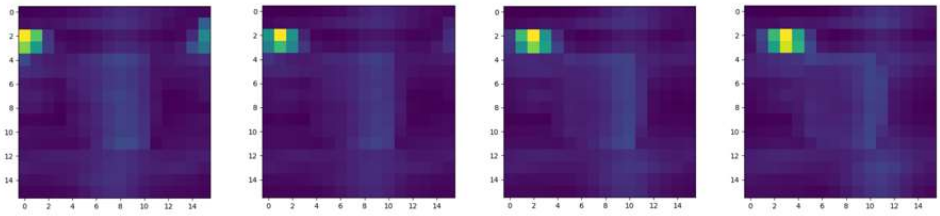

Fig. S9 The reshaped attention weights of the first four rows from the attention map of block 1.

The attention map of block 1 can be effectively interpreted because the input tokens have clear physical meanings. The first 256 tokens correspond to the 256 image patches, and the last token represents the period. We can reshape the first 256 attention weights of each row of the attention map to a 16×16 matrix. Some of the results are displayed in Fig. S9 and indicate that the first transformer encoder block mainly extracts features from the spatially adjacent patches. This is understandable because light fields are usually localized in these subwavelength dielectric structures, and the light field of each patch highly depends on its neighboring patches because these patches can provide boundary conditions. This is similar to the FDTD simulation output, where at every time step, the electrical field of each cell is calculated from the magnetic fields of the neighboring cells.

The attention map of block 2 can also be interpreted. As the output tokens of block 1, which are also the input tokens of block 2, come from the adjacent patches, we can still assume that each input token of block 2 correlates to one of the 16×16 spatial locations. Therefore, for the attention map of block 2, we can also reshape the first 256 attention weights in each row to a 16×16 matrix (Fig. S10a). Different from the results of block 1, these results show that each output token of block 2 no longer only depends on the adjacent patches but is calculated from a relatively global feature. Each output token represents a different global feature. This can be attributed to block 2 mainly analyzing the nonlocal light field. We find that the visualized attention weight matrix (Fig. S10a) exhibits some similarities to the electrical field distribution produced at different wavelengths for the meta-atom simulated by FDTD (Fig. S10b). Therefore, we can assume that block 2 tries to predict the global light field distribution at each wavelength, which is important to the transmission response.

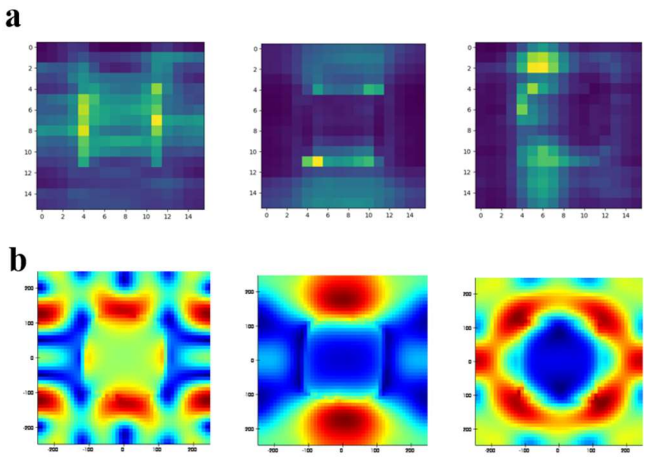

Fig. S10. a. The reshaped attention weights of some rows from the attention map of block 2. **b.** The distributions of the electrical fields inside the meta-atoms simulated by FDTD at different wavelengths.

The attention maps of blocks 3 and 4 are less interpretable because we cannot find a clear physical explanation for each token. The output tokens of block 2, which are also the input tokens of block 3, no longer have a one-to-one correspondence with one of the 16×16 spatial locations. Instead, each token represents a high-level abstract feature, and finally, the tokens are transformed from the spatial domain to the spectral domain because each output token of block 4 is bound to a certain wavelength. This domain transfer task is mainly accomplished by block 4. The role of block 3 may be to map the local and nonlocal light fields to the desired optical properties.

Supplementary Note 6. Implementation details of the prompt encoder network.

Our prompt encoder network is implemented by the same transformer encoder block shown in Fig. S5. The encoder network has 4 transformer encoder blocks with 4 heads, and the head size is set to 128. An additional decoder network is also designed to train the encoder network. The decoder network has 2 transformer encoder blocks with 4 heads, and the head size is 64. The inputs of the prompt encoder network are the property vector (the transmission power spectrum, transmission phase spectrum, or complex-valued transmission spectrum) and the property mask. Every element of the property vector is regarded as an input token of the transformer encoder. All of the tokens in the masked bands (where $mask = 0$) are replaced by the same trainable masked token. Sinusoidal positional embeddings are also added to the input tokens.

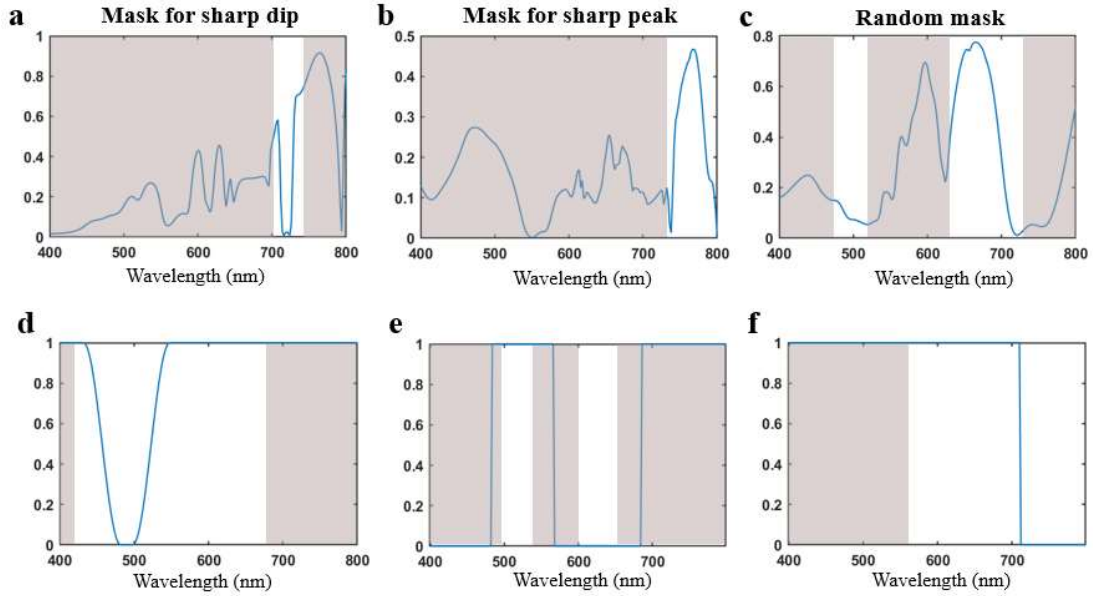

Fig. S11. Samples of training data for prompt encoder-decoder network. a, b, c. Samples of realistic data. **d, e, f.** Samples of human-crafted data.

The training dataset mainly contains two types of data: realistic data and human-crafted data. The realistic data are from FDTD simulation. As for masking strategy, we apply both human-labeled masks and random masks. For human-labeled masks, we use the mask to indicate the important features in the optical properties, such as sharp dips and peaks in the transmission power responses shown in Fig. S11a and Fig. S11b. For random masks, we randomly generate masks at 1~3 random bands with random width, such as the data sample shown in Fig. S11c. The human-crafted data are generated from abstract design demands. For example, the working band and cutoff frequency of a filter. We translate these abstract design demands to optical properties and apply masks that fit the demands (Fig. S11d-f).

During training, masks are randomly generated and dropped with a dropout rate of 0.3. If a mask is dropped, it is set to all ones, and none of the input tokens are replaced by the masked token. The training data are acquired from the simulated meta-atom dataset used to train the forward prediction network. We train the networks with a batch size of 512 for 25000 steps. The Adam optimizer is also adopted to train the network. After training, the mean absolute error and root mean square error in the unmasked bands are about 0.0061 and 0.0075. Then the prompt decoder network is discarded, and the prompt encoder network is utilized to train the latent diffusion network.

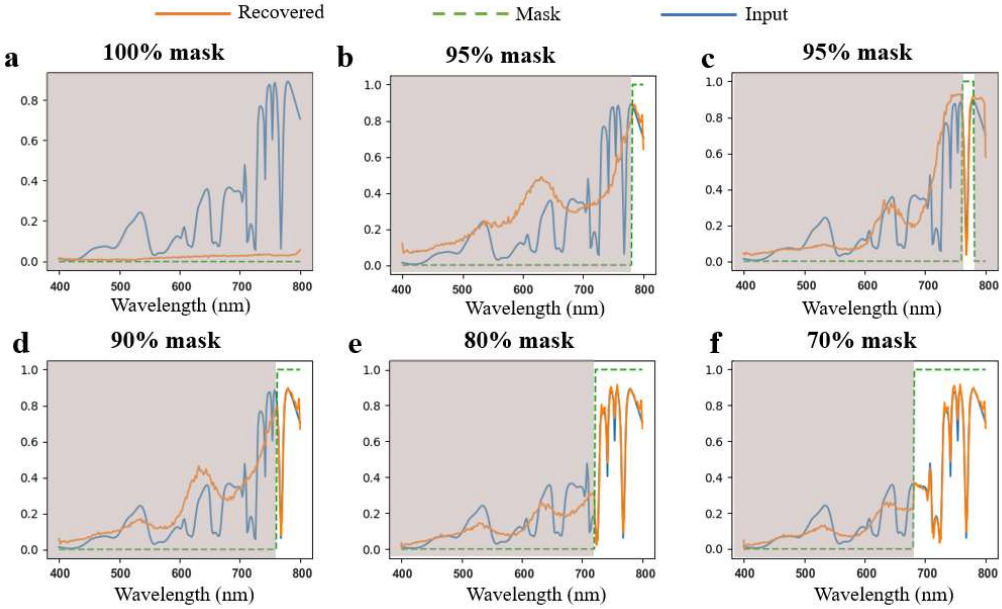

Fig. S12 Results of the prompt encoder-decoder network.

While the embeddings are hard to interpret, we can visualize the decoded results to understand the behavior of the prompt encoder network. If we mask the entire bands, the outputs of the decoder are close to zeros in the entire band (Fig. S12a). However, the values at longer wavelengths are slightly higher than those at shorter wavelengths. This is likely because, overall, the transmittance of Si-based meta-atoms is lower at shorter wavelengths and higher at longer wavelengths, and the prompt encoder-decoder network have learned this bias. If we mask 95% of the bands, the decoder can precisely recover the unmasked bands, and roughly predict the masked bands (Fig. S12b, c). If some important features (such as digs or peaks) are unmasked, the prediction can be more accurate in the masked bands. And, if we decrease the proportion of the masked bands, the prediction can also be more accurate in the masked bands (Fig. S12d-f).

Therefore, the proportion and position of the mask determine the possible solution space. Provided more unmasked features or more distinctive features, the solution can be found more accurately. The prediction in the masked bands may represent the average of all possible solutions. Given different mask values $\{m_1, m_2, \dots, m_n\}$, every optical response y can be encoded to different embeddings

$\{g(y, m_1), g(y, m_2), \ldots, g(y, m_n)\}$ by the prompt encoder network $g(\cdot)$. If two responses y_1 and y_2 have some same features, then in the embedding space, they can have two close embeddings: $g(y_1, m_1) \approx g(y_2, m_2)$. In this way, our inverse design method can achieve fuzzy search and accept flexible inputs.

Supplementary Note 7. Additional inverse design results provided by the diffusion network.

First, to demonstrate the generative capabilities of our latent diffusion network, some results that are randomly generated without any conditional inputs are shown in Fig. S13. The darker region represents air, the lighter region represents dielectric, and the degree of darkness represents the period of the meta-atom. Our latent diffusion network can generate various shapes with and without C4 symmetry.

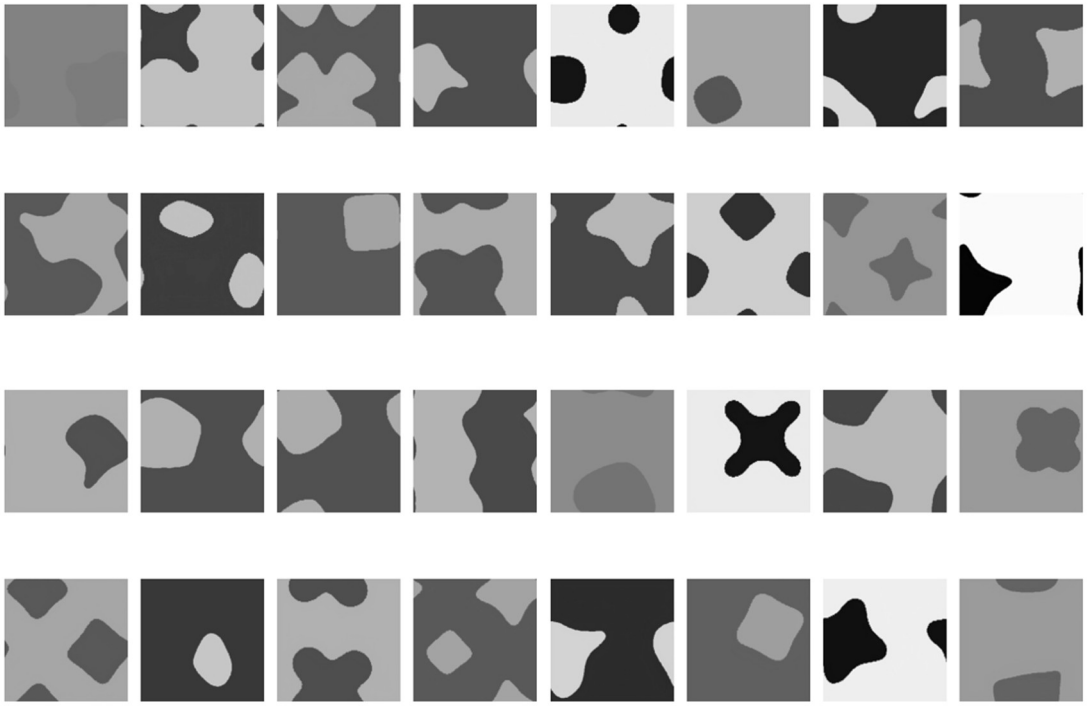

Fig. S13 The shapes randomly generated by the latent diffusion network.

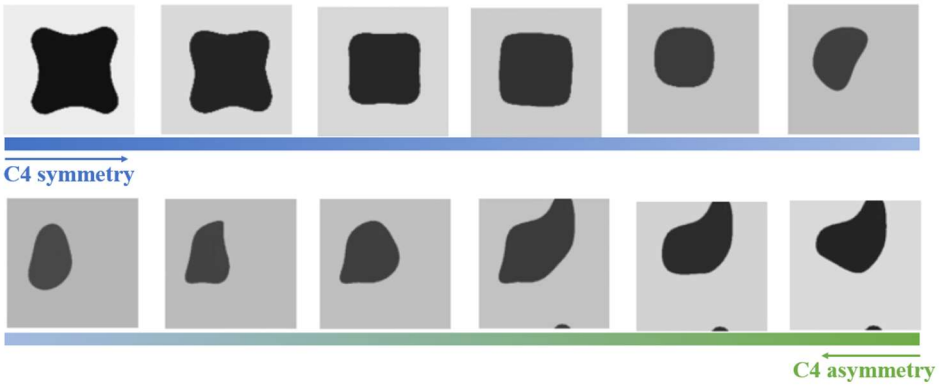

Fig. S14 The interpolated shapes between a C4 symmetric shape and a C4 asymmetric shape generated by a latent diffusion network.

By proceeding through the latent space, we can see how the diffusion network utilizes interpolation to generate new shapes. We choose an initial noise value n_1 that can generate a C4 symmetric shape and another initial noise value n_2 that can generate a C4 asymmetric shape and employ the latent diffusion network to generate shapes using linear interpolations between n_1 and n_2. The interpolated shapes are shown in

Fig. S14. Diffusion models have powerful abilities to model images from different domains. The results show that our latent diffusion network can establish a method for interpolating between C4 symmetric and C4 asymmetric shapes, thus generating new shapes that are not included in the training set. This ability to generate new shapes can greatly enlarge the search space.

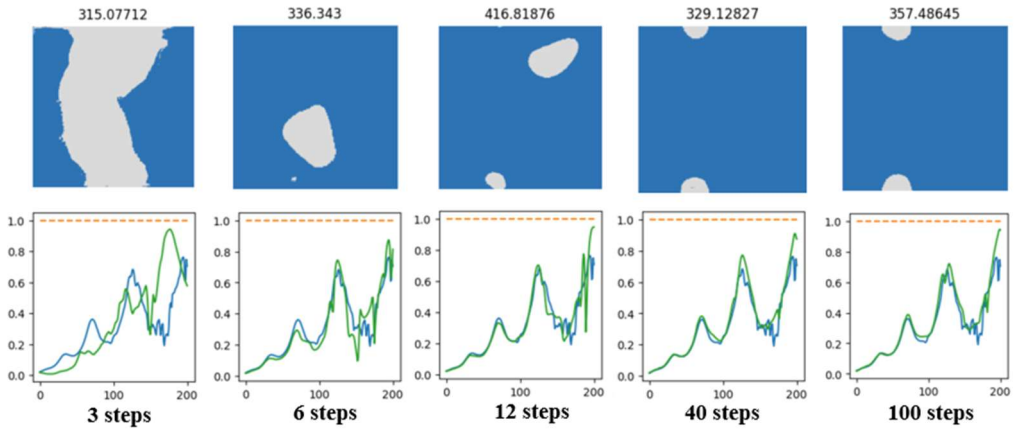

Fig. S15 Inverse design results obtained using different numbers of diffusion steps.

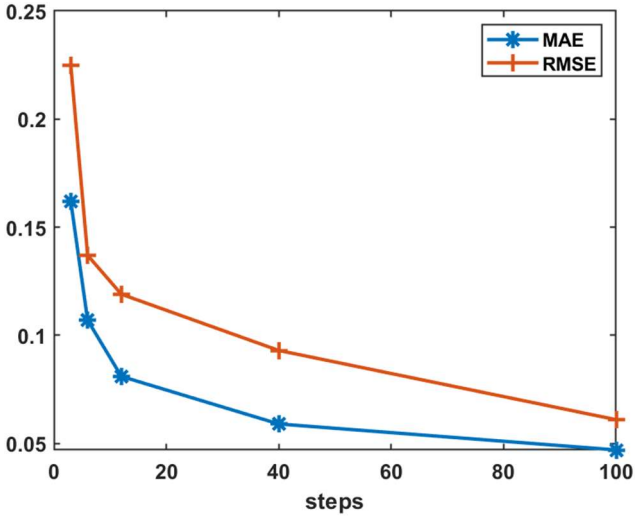

Fig. S16. Evaluation of inverse design performance using different diffusion steps. We evaluate the results shown in Fig. S15 using two metrics: MAE and RMSE. The curves indicate that the inverse design result starts to converge when $N \geq 40$. At $N = 40$, the MAE and RMSE are 0.059 and 0.093.

The number of diffusion steps is an important hyperparameter in diffusion models. As we adopt a continuous diffusion time to train the network, the number of diffusion steps N can be dynamically changed during inference. A larger N indicates a more refined denoising process. We conduct the same inverse design task as that shown in Fig. 4c but change N to different values (the results shown in Fig. 4c are generated under $N = 40$). The inverse design results obtained under different N values are shown in Fig. S15. The first row shows the generated meta-atoms with their periods, and the second row shows the corresponding design targets (blue curves), input masks

(orange curves), and design results (green curves). The results show that when $N = 6$, the inverse design result starts to satisfy the imposed requirement. When $N = 12$, we can obtain a relatively reliable inverse design result. When $N \geq 40$, the inverse design result starts to converge. Our latent diffusion network can effectively obtain inverse design results with few denoising steps.

We evaluate the results shown in Fig. S15 using two metrics: mean absolute error (MAE) and root mean square error (RMSE). The results are shown in Fig. S16. The curves also indicate that the inverse design result starts to converge when $N \geq 40$. At $N = 40$, the MAE and RMSE are 0.059 and 0.093. The inverse design performance of several existing methods is reported in Ref. 34. Due to the differences in the evaluation data, working bands, sampling interval, etc., the values of MAE and RMSE are not comparable directly for different methods. However, the absolute values of these metrics are similar to existing methods reported in Ref. 34, indicating that our method can achieve comparable performance while providing much more complex functions such as fuzzy search and partial inputs.

Then, we set $N = 40$, and we can see how the diffusion network denoises the initial random noise to form the desired output. This denoising process is shown in Fig. S17.

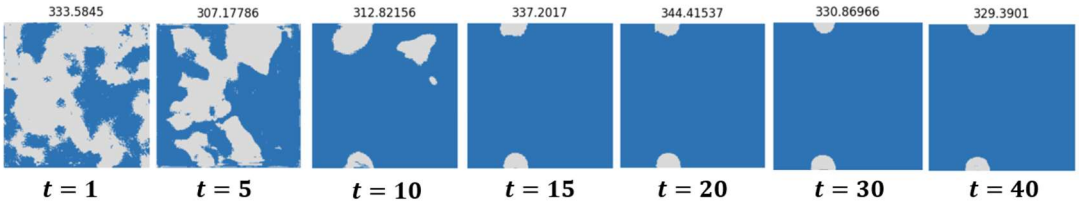

Fig. S17 Denoising process of the diffusion network.

Supplementary Note 8. Inverse design results of bandstop and bandpass filters

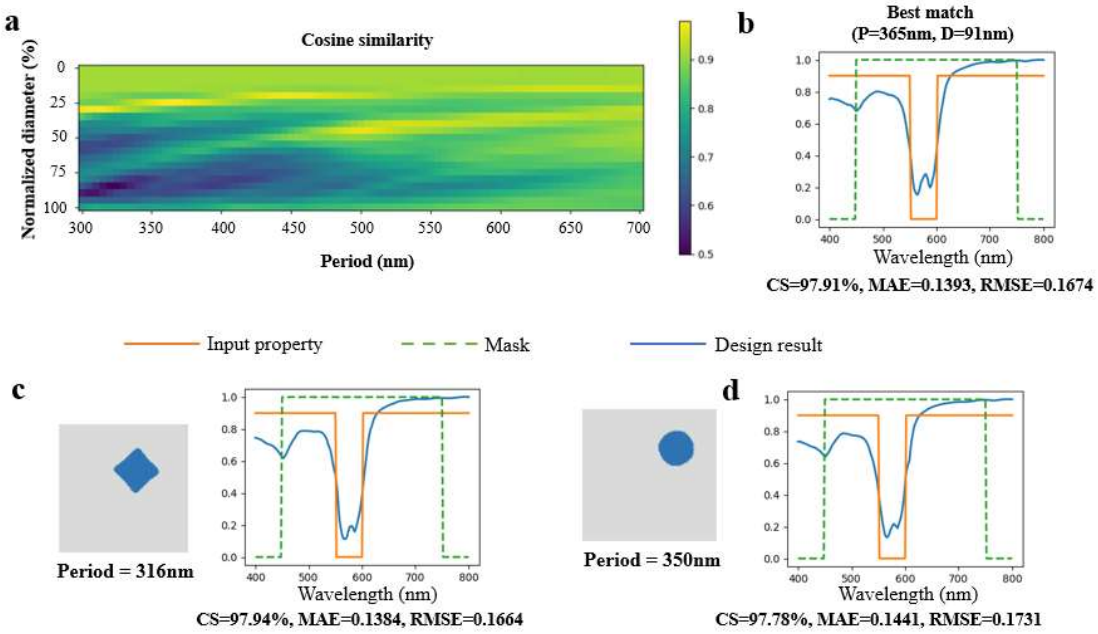

Fig. S18. Design results of bandstop filters by parameter sweep and diffusion network. a, Parameter sweep results of cylindrical pillars. We sweep period P from 300 nm to 700 nm at 5 nm intervals. Diameter D is normalized using P, and we sweep the normalized diameter from 0% to 100% at 5% intervals. **b,** The best match found by parameter sweep evaluated by CS. **c, d.** The inverse design results generated by diffusion network.

If we want to design a bandstop filter (shown as orange curves in Fig. S18) based on 220 nm SOI, a human designer may attempt to achieve the meta-atom by regular shapes such as cylindrical or rectangular pillars and find the best design parameters by parameter sweep. Taking cylindrical pillars as example, we need to determine two parameters: period (P) and diameter (D). We sweep P from 300 nm to 700 nm at 5 nm intervals. D is normalized using P, and we sweep the normalized diameter from 0% to 100% at 5% intervals. To complete the sweep, we need to run FDTD simulation 1600 times and each simulation requires several minutes. After simulation, we evaluate the difference between the design target and design results by three metrics: cosine similarity (CS), MAE, and RMSE. The sweep result of the CS metric is shown in Fig. S18a. Then, we can find the best match by the highest CS. We find that the best match also has the lowest MAE, and lowest the RMSE, which is shown in Fig. S18b.

The proposed direct-mapping inverse design method can make the design process much easier. We only need to set the design target to be similar to an ideal filter, and the diffusion network can output the design results in a few seconds. Although such an ideal filter cannot be physically implemented, the diffusion network can output several candidates that meet the requirements as much as possible. It can find a solution similar

to the best match found by parameter sweep (Fig. S18d). Moreover, the network can generate freeform shapes instead of only circles. Therefore, it can further achieve better results (Fig. S18c). Note that many of the existing neural networks may fail to function properly when the input is an ideal filter. Because the response of the ideal filter does not satisfy the condition of independent and identically distributed. However, our prompt encoder network makes it possible to accept flexible inputs.

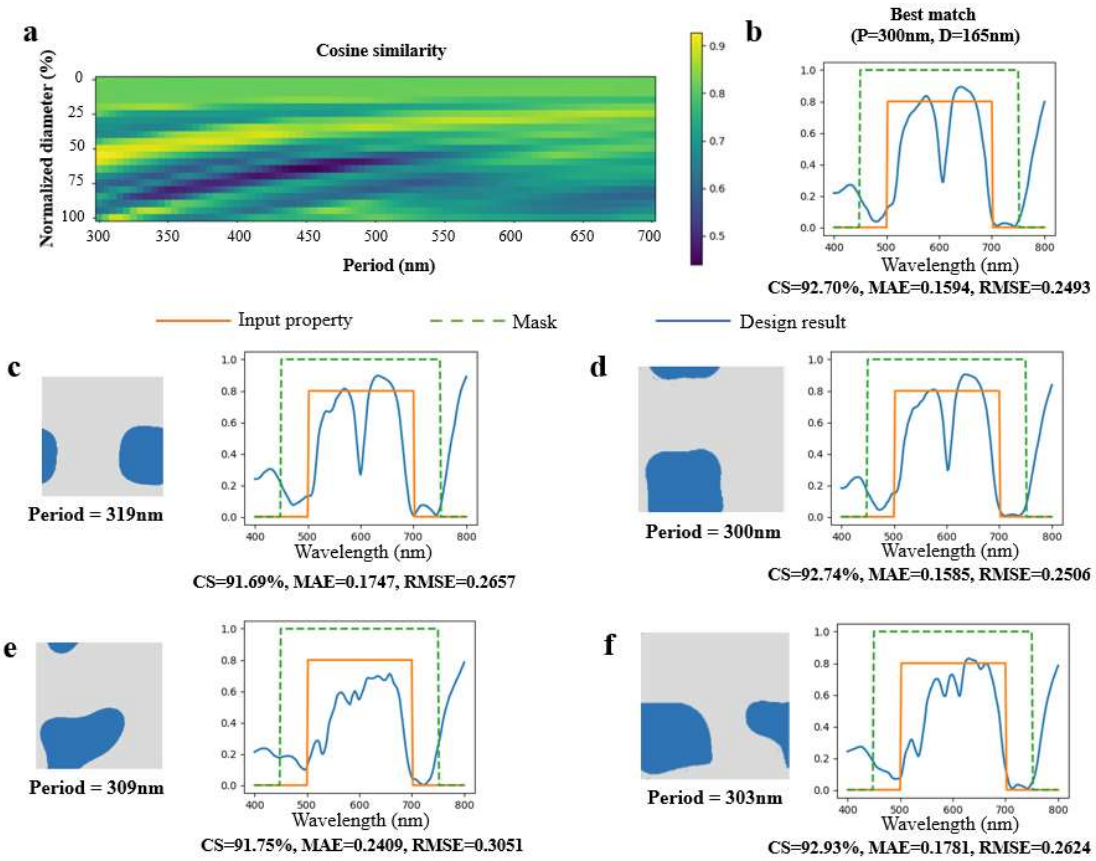

Fig. S19. Design results of bandpass filters by parameter sweep and diffusion network. a, Parameter sweep results of cylindrical pillars. We sweep period P from 300 nm to 700 nm at 5 nm intervals. Diameter D is normalized using P, and we sweep the normalized diameter from 0% to 100% at 5% intervals. **b,** The best match evaluated by CS found by parameter sweep. **c-f.** The inverse design results generated by diffusion network.

To further demonstrate the inverse design capability of the diffusion network, we also employ the network to design bandpass filters (shown as orange curves in Fig. S19). Fig. S19b shows the result designed by the same parameter sweep method. Our diffusion network can also find a cylinder meta-atom similar to the best match found by parameter sweep (Fig. S19c). However, these cylinder meta-atoms have a low transmittance at around 600 nm. If we remove the C4 symmetry constraint to give the diffusion network more design freedom, it can generate freeform meta-atoms directly that better meet the design target (Fig. S19e Fig. S19f). These results indicate that the

proposed inverse design method can reliably generate the required meta-atoms. It can greatly accelerate and simplify the design process. The large design space introduced by freeform shapes leads to powerful inverse design capabilities.

Supplementary Note 9. Inverse design results of matched filters

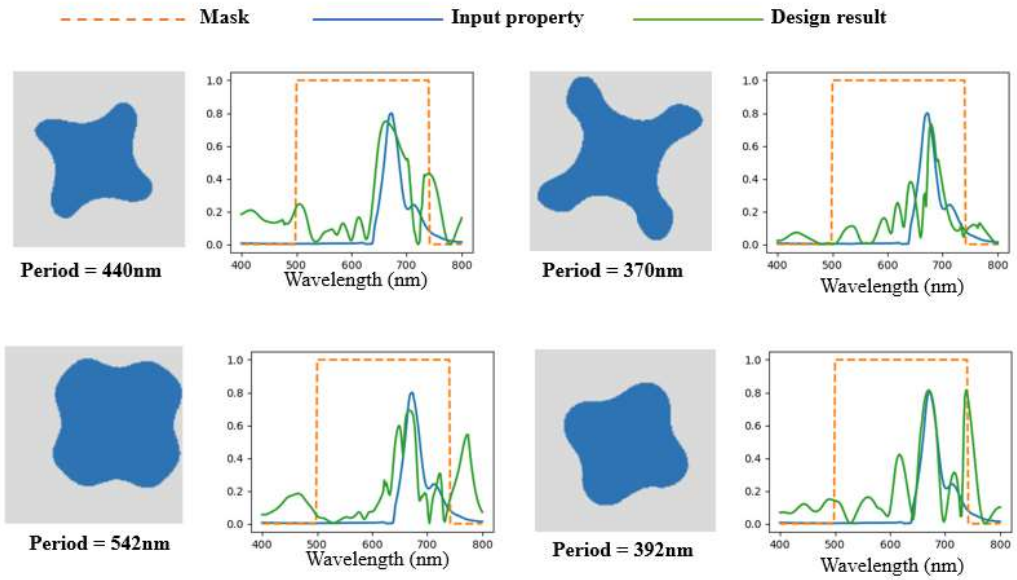

Fig. S20. Inverse design results of filters that can detect red fluorescence protein in bioluminescent area.

Besides designing typical filters and polarizers, the proposed inverse design method also has advantages in other practical applications. For example, if we want to detect a red fluorescence protein in bioluminescent area (shown as blue curves in Fig. S20), the best approach is to design a polarization-independent matched filter. While a human designer may find it difficult to design a meta-atom whose transmission response matches the fluorescence spectrum, our diffusion network can output several candidates in a few seconds (shown in Fig. S20). It can greatly accelerate the research related to subwavelength structures.

Supplementary Note 10. Details of fabrication and testing of the inverse-designed structures

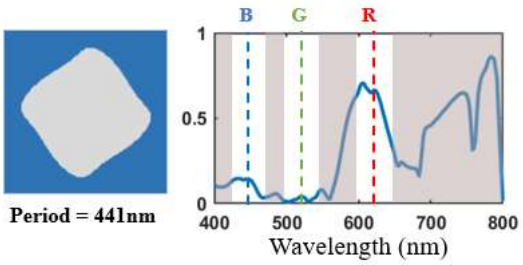

Fig. S21 Inverse design results of the red structural color.

The 64 structural colors are achieved by manipulating the transmission response at three different wavelengths (623 nm for red, 528 nm for green, and 461 nm for blue). We set 4 different transmission rates at each wavelength. The rates are *(0.7, 0.46, 0.23, 0.0), (0.35, 0.23, 0.12, 0.0)*, and *(0.35, 0.23, 0.12, 0.0)* for red, green, and blue color. For example, to generate the color with a 24-bit RGB value of (255, 0, 32), we first convert the 24-bit RGB value to the 6-bit RGB value of (3, 0, 1). Then, the 6-bit RGB value is mapped to the transmission rates of *0.7, 0.0,* and *0.12* at 623 nm, 528 nm, and 461 nm. Finally, the transmission rates work as the input optical properties and guide the diffusion network to generate the desired meta-atom shown in Fig. S21.

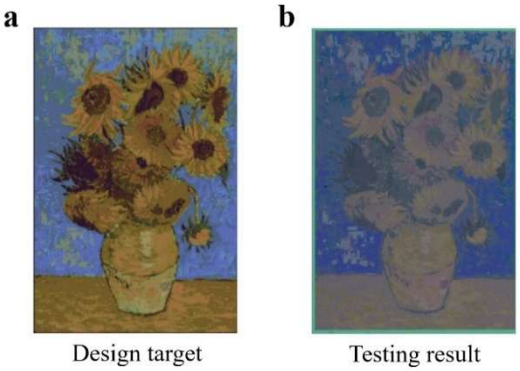

Fig. S22 Design target and testing result of the painting of sunflowers.

We utilize the designed 64 meta-atoms with different structural colors to construct the painting of sunflowers shown in Fig. S22a. The painting has 423×282 pixels. Each pixel is $6.9 \times 6.9\ \mu m^2$ and is achieved by a specific meta-atom. After fabrication, the chip is put on an LED screen that displays the color with an RGB value of (127, 255, 255) and observed through a microscope. The microscope image is shown in Fig. S22b. The sapphire layer results in a little chromatism between the testing result and the design target.

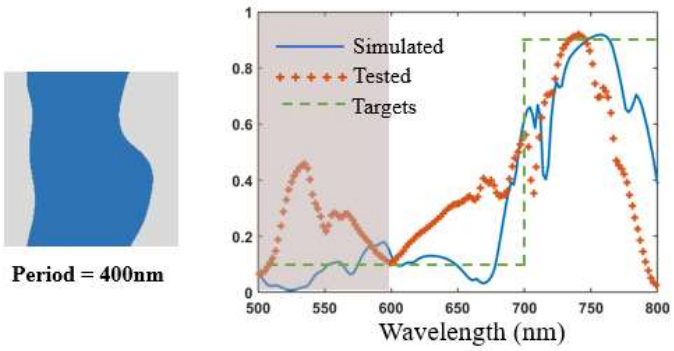

Fig. S23. Inverse design results of a longwave pass filter.

The design of the longwave pass filter results in a grating-like meta-atom, which is shown in Fig. S23. The green dashed line is the optical property condition given to the diffusion model. The model maps the optical property to a meta-atom directly, and the simulated transmission response of the meta-atom is displayed as the blue curve. We fabricated the meta-atom and measured its transmission spectrum, which is shown as the red dotted curve. The meta-atom is designed to have relatively high transmission in 700~800 nm bands and low transmission in 600~700 nm bands. The measured transmission spectrum roughly meets the design targets. However, the transmission response at 600~700 nm bands is higher than the simulation, which might be caused by fabrication error.

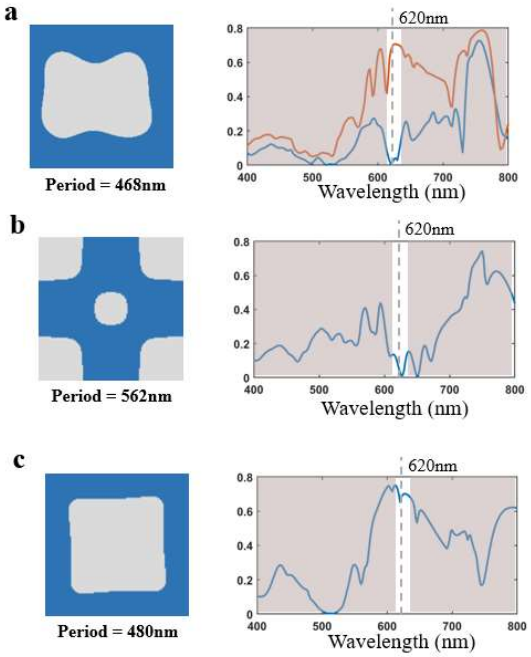

Fig. S24. Inverse design results of the polarization manipulating meta-atoms

Fig. S24 shows the inverse design results of three meta-atoms for polarization manipulation. Fig. S24a is the designed polarization-sensitive meta-atom that has a high and low transmission for horizontally and vertically polarized light at 620 nm. Fig. S24b

and Fig. S24c are the designed C4 symmetry meta-atoms that have high and low transmissions at 620 nm, respectively. Utilizing these three meta-atoms, we encode two different patterns into two different polarization directions at the same wavelength of 620 nm. The fabricated chip can present different patterns under horizontally and vertically polarized light.

Supplementary Note 11. The influence of the surrogate simulator.

We propose a forward prediction network to replace FDTD simulation. In this way, we can train the diffusion model at high-efficiency. The performance of the forward prediction network is demonstrated in Supplementary Note 5. Although it can achieve high-performance, its predictions inevitably contain some deviations. To study the influence of forward prediction network precision on the final inverse design performance, we have trained two auxiliary diffusion networks. The first one (referred to as D-fdtd) is trained using the same dataset that is adopted to train the forward prediction network. The dataset contains about 1×10^5 meta-atoms and their corresponding transmission responses simulated by FDTD. The second one (referred to as D-dnn) is also trained by the same dataset. However, we replace the simulated responses by the predictions from the forward prediction network. It is worth noting that the original inverse design network (referred to as D-0) is trained using a dataset that only contains about 2×10^5 shapes. The periods are randomly selected and transmission responses are predicted by the forward prediction network. The three networks: D-fdtd, D-dnn, and D-0, are trained using the completely same strategy.

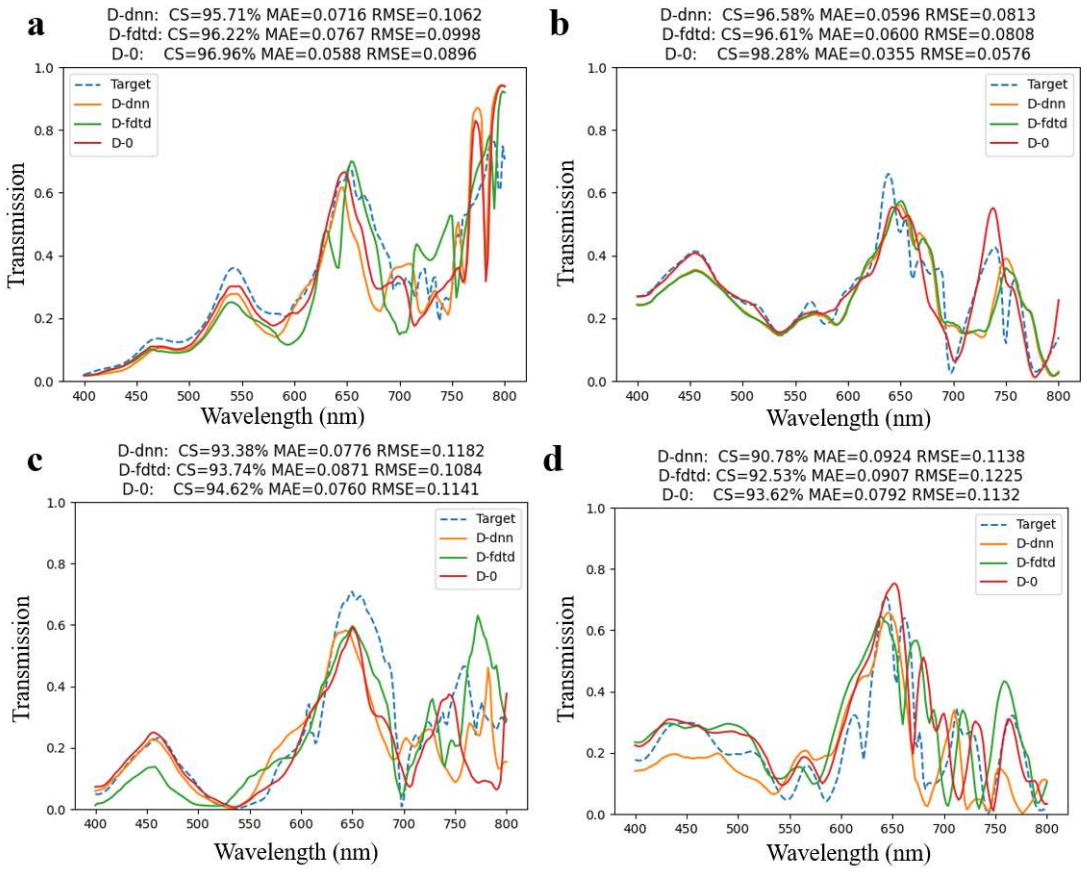

Fig. S25. Performance of D-dnn, D-fdtd, and D-0.

Then, we test the inverse design performance of D-fdtd, D-dnn, and D-0. Fig. S25 shows four inverse design results. The inverse design performance is also evaluated by CS, MAE, and RMSE metrics. The results indicate that the performance of D-fdtd and

D-dnn is similar, while D-0 achieves the best results. Further, the generated shapes are encoded to latents by the image encoder network, and we evaluate the diversity of the generated shapes by the standard deviation (STD) of the latents. The results are shown in Fig. S26a. D-0 reaches the highest STD score, indicating that its generated shapes have the highest diversity. D-0 also achieves better results on MAE and RMSE metrics. We then visualize the distribution of the generated shapes corresponding to the inverse design task shown in Fig. S25a by PCA in Fig. S26b, which also indicates that D-fdtd and D-dnn achieve similar diversity while D-0 has the highest diversity.

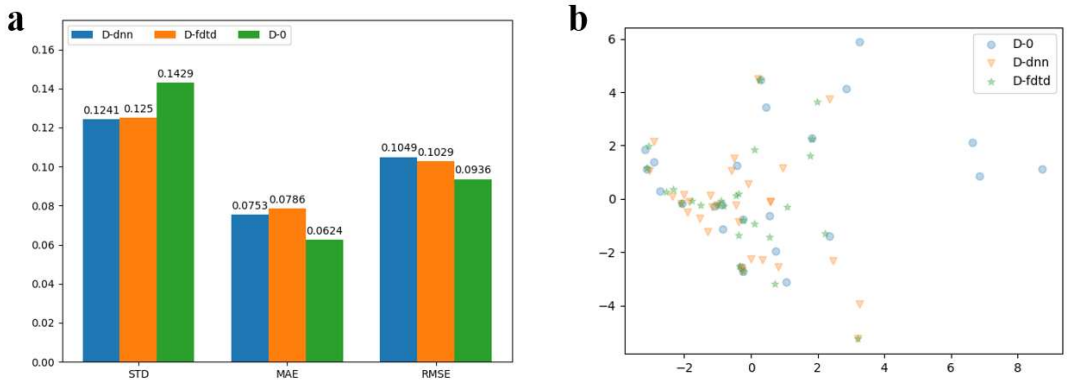

Fig. S26. a, The evaluation results of D-fdtd, D-dnn, and D-0. **b,** The distribution of the generated shapes corresponding to the inverse design task shown in Fig. S25a are visualized by PCA.

These results show that the surrogate simulator (i.e. the forward prediction network) has little impact on the final inverse design performance. Furthermore, by adopting the forward prediction network, we can utilize the whole freeform shapes dataset to train the inverse design network, thus achieving much more design freedom. And more design freedom leads to better inverse design performance and diversity as the results shown in Fig. S26.

Supplementary Note 12. Ablation study on prompt encoder network.

The prompt encoder is indeed a key innovation in our work, designed to enhance the practicality of the inverse design framework. To further illustrate the role of prompt encoder network, we removed the prompt encoder and retrained the model using the same hyperparameters. We systematically compared the inverse design performance of the models with and without the prompt encoder.

When provided with a physically realistic, full-bandwidth target transmission response, both models achieved satisfactory inverse design results, as shown in Fig. S27. The model with the prompt encoder achieved a mean MAE of approximately 0.056, while the model without it achieved a comparable MAE of about 0.054. In this example, the retrained model demonstrated similar convergence speed and final performance on training and validation loss with and without the prompt encoder.

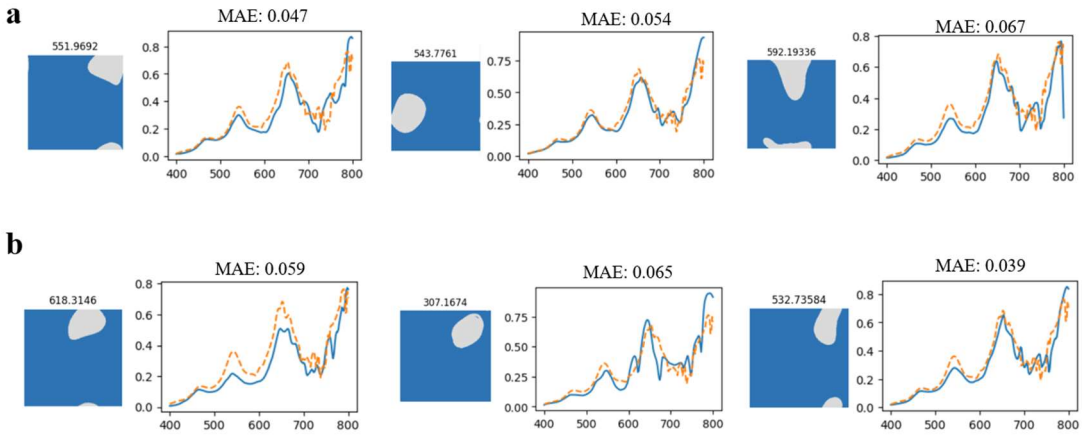

Fig. S27. Inverse design results of a physically realistic, full-bandwidth transmission response. a, Design results of the model with prompt encoder. **b,** design results of the model without prompt encoder.

Moreover, a critical test lies in a more practical scenario where the user only has abstract design parameters rather than a full target spectrum. Using the bandpass filter design from Supplementary Note 8 as an example, the model with the prompt encoder successfully generated valid structures, as shown in Fig. S28(a). In contrast, the model without the prompt encoder completely failed in this task, as shown in Fig. S28(b), because it lacks the ability to interpret and perform a fuzzy search based on abstract concepts. These comparative results indicate the importance of the prompt encoder for handling real-world design requirements.

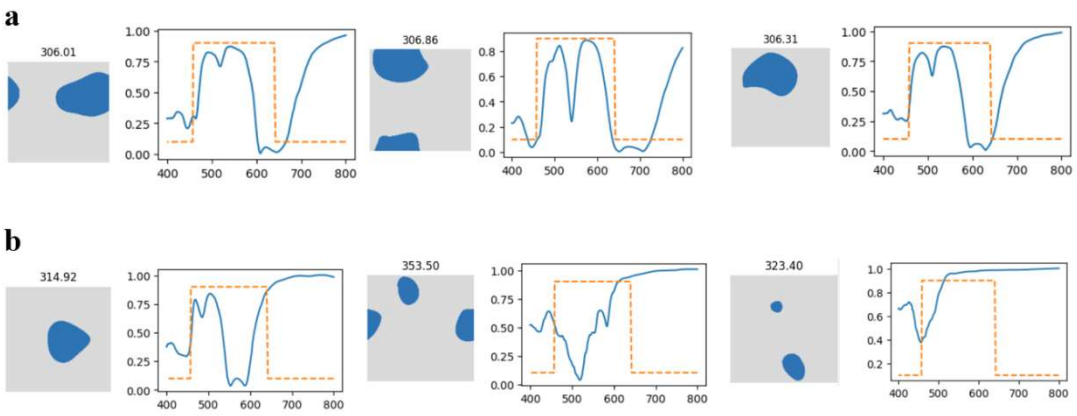

Fig. S28. Inverse design results of a bandpass filter. a, Design results of the model with prompt encoder. **b,** design results of the model without prompt encoder.

Supplementary Note 13. Explore of physical mechanism of AIGP.

Based on Maxwell's equations, a well-defined structure necessarily yields a deterministic optical response. While analytical solutions remain elusive for complex photonic systems, numerical methods such as finite-difference time-domain (FDTD) simulations provide reliable access to their optical characteristics. This establishes a one-to-one mapping between structural parameters and optical responses that neural networks can effectively learn.

Specifically, in designing subwavelength photonic structures, the absence of analytical solutions necessitates reliance on Maxwell-equation-based numerical simulations. FDTD simulation essentially acts as a mapping function—from structure to optical response—under Maxwell's constraints. Our forward network approximates this function, while the inverse network learns its inverse mapping. When trained on FDTD data, the AI model inherently internalizes the optical behavior of subwavelength structures as governed by Maxwell's equations.

Consequently, we position artificial intelligence as a powerful auxiliary tool in photonic inverse design — accelerating design exploration, revealing non-intuitive structural configurations in expanded parameter spaces, and ultimately enhancing device performance.

It is also important to note that while regular meta-atoms (e.g., cylinders, squares) allow some analytical understanding of light propagation, freeform subwavelength structures—which offer greater design freedom and superior optical performance—inherently lack straightforward analytical solutions. This interpretability challenge is intrinsic to freeform structures themselves, not solely to the design method. Even traditional optimization algorithms, like topology optimization, produce final structures whose specific functional roles are often difficult to interpret piecewise.

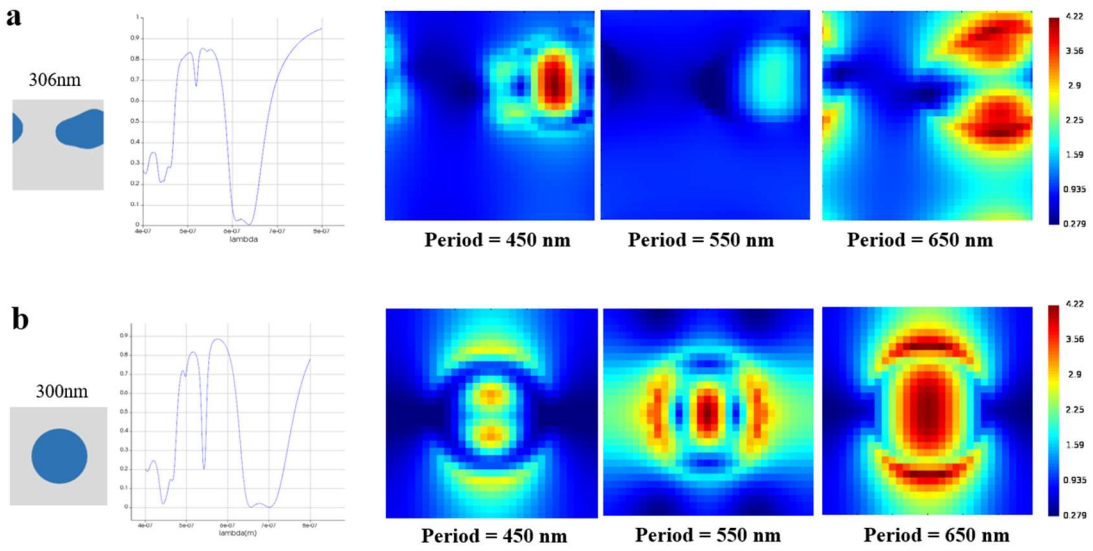

Fig. S29. Electrical field distribution of the designed meta-atom. a, Meta-atom designed by inverse design model. **b,** Meta-atom designed by parameter sweep.

Nevertheless, we have performed simulation-based analysis to understand the inverse-designed structures. Taking the bandpass filter from Supplementary Note 8 as an example, the cylindrically symmetric meta-atom designed by parameter sweep exhibits an undesirable dip in the passband (Fig. S29a). In contrast, the freeform structure designed by our model mitigates this dip (Fig. S29b). To understand this improvement, we simulated the electric field distributions at different wavelengths (Fig. S29). Both structures exhibit resonances and low transmittance at 450 nm and 650 nm, where the field is strongly localized. However, the cylindrical structure also excites an additional resonance mode around 550 nm, creating the dip. The freeform structure, by breaking symmetry, suppresses this specific resonance, thereby flattening the passband. This demonstrates how our inverse design framework can discover non-intuitive structures that outperform conventional designs.

Supplementary Note 14. Computational cost comparison of AIGP.

a rigorous comparison is crucial to justify the efficiency claims of our AIGP framework. To quantitatively compare the computational costs across different methods and the various stages of our AIGP framework, we define one unit of computational cost as the amount required for a single FDTD simulation. Taking the bandpass/bandstop filter design in Supplementary Note 8 as an example: the traditional parameter sweep method required 1,600 FDTD simulations, corresponding to 1,600 units. For our AIGP approach, the data generation stage involved 10^5 meta-atoms, equating to 10^5 units. In the inference stage, the average time for a complete inverse design plus forward validation is about 20+0.15=20.15 seconds, which is equivalent to only 0.0672 units when benchmarked against the 300 seconds (5 minutes) required for a single FDTD simulation on a same CPU. The training stage, conducted with a batch size of 512 for 90,000 steps, involved approximately 4.6×10^7 forward and backward passes. Considering a backward pass is about twice as computationally intensive as a forward pass, the total cost for training is estimated at 10^7 units. In summary, the quantified costs are: parameter sweep, 1,600 units; AIGP data generation, 10^5 units; AIGP training, 10^7 units; and AIGP inference per design, 0.0672 units. These results are summarized in Supplementary Table 1.

Supplementary Table 1. Comparison of computational costs

Method	Operation	Cost (unit)
Traditional	FDTD simulation	1
	Parameter sweep	1600
AIGP	Data generation	10^5
	Network training	10^7
	AIGP inference per design	0.0672

The computational burden is shifted to a one-time, upfront cost during preparation, which dramatically reduces the cost per design during the deployment and inference phase. While the total cost for training the AIGP model exceeds that of a single parameter sweep, the inference stage requires no iterative calculations and consumes only about 1/20,000 of the computational cost of one parameter sweep per design. Crucially, the training process is a one-time effort. Once the model is trained, it can generate vast numbers of different structures in a very short time (for example, about 3 samples per minute on CPU and 128 samples per minute on GPU), which holds great potential for significantly accelerating the development speed of subwavelength photonic structures. In summary, although the training and data generation require substantial resources initially, this one-time investment enables a dramatic and perpetual speed-up for countless future designs.

Furthermore, evaluating the cost of the AIGP method solely based on abstract "computational units" is not entirely comprehensive, as the real-world cost of these units varies significantly depending on the computing platform. AI models are

inherently suited for parallel computing and can be easily deployed and accelerated on high-performance hardware like GPUs. In contrast, traditional iterative optimization algorithms, due to their sequential nature, are often unable to leverage massive GPU parallelization effectively. This means that the actual wall-clock time required to train the AIGP model on a GPU can be comparable to, or even shorter than, the time needed for a single run of a traditional optimization algorithm on a CPU. This represents another significant advantage of the AIGP framework.

Supplementary Note 15. Simplified description of the AIGP framework.

The complexity of the original Fig. 1b compromises its clarity. To significantly improve readability, we provide a simplified version below that outlines the main workflow of the overall framework.

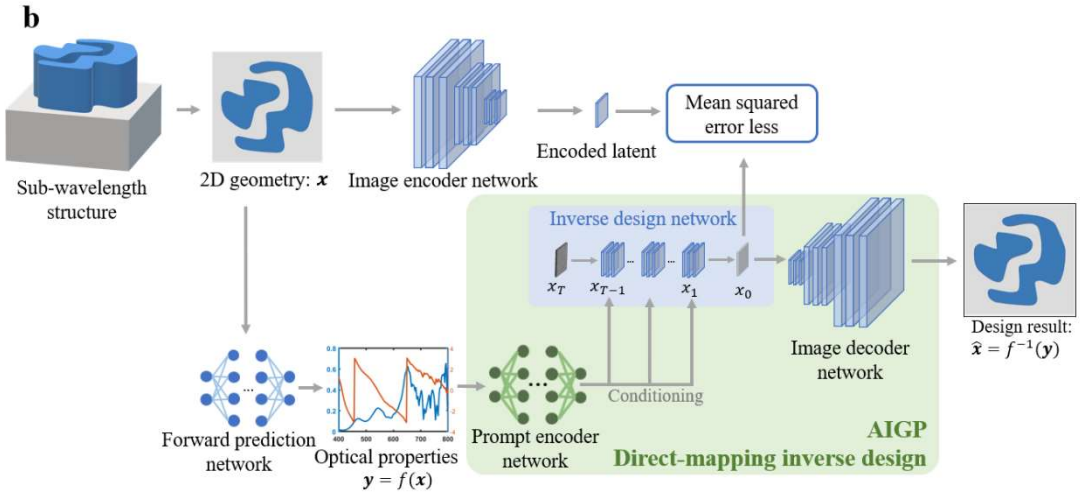

Fig. S30. Simplified schematic framework of the AIGP framework.

Supplementary Note 16. Chip fabrication details and fabrication error analysis.

The designed subwavelength structures are fabricated in Tianjin H-Chip Technology Group Corporation. The structures are formed on a silicon-on-sapphire (SOS) chip. The intrinsic silicon layer is 230 nm thick and the sapphire layer is 475 μm thick with flatness < 15 μm. The fabrication process was as follows: Electron-beam (EB) lithography was performed at a beam current of 2 nA using a 250 nm thick ZEP 520A resist layer. Subsequently, the pattern was transferred via ICP etching with a gas mixture of CHF_3 and SF_6. Finally, the chromium (Cr) mask was removed using a dry etching process with a combination of CHF_3 and O_2. The linewidth precision of the fabricated samples is specified as ±5% for features above 200 nm and ±10 nm for features at or below 200 nm. And the minimum linewidth is 90nm. The Manufacturing Process Parameters Table is added as Supplementary Table 2 and is also provided below:

Supplementary Table 2. Manufacturing Process Parameters Table

Process	Parameters
EB Lithography	2 nA beam current, ZEP 520A resist (250 nm thickness)
ICP Etching	Gas chemistry: CHF_3 + SF_6
Cr Removal	Dry etching with CHF_3 + O_2

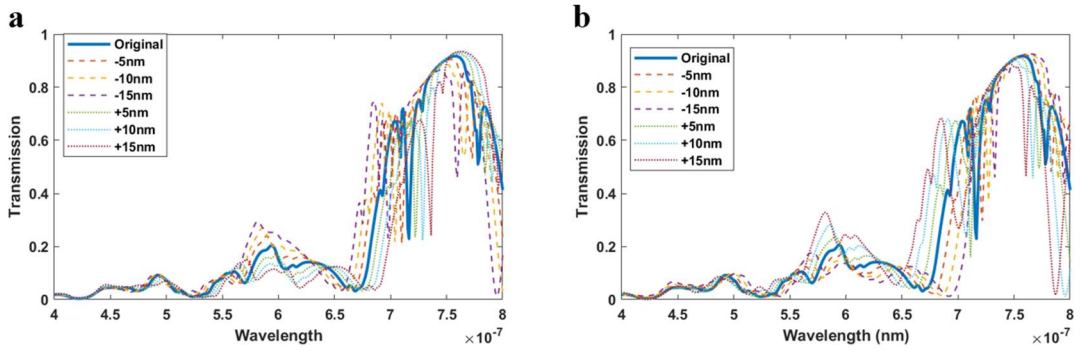

Fig. S31. The impact of fabrication variations on key optical metrics. a, Variation of transmission responses with different period errors. **b,** Variation of transmission responses with different linewidth errors.

Our inverse design method is engineered to ensure that all generated structures adhere to practical fabrication constraints. Specifically, for our fabrication experiments, the minimum feature size achievable with our equipment is 90 nm. Consequently, the model was explicitly trained to generate optical structures whose minimum linewidth and curvature radius comply with this manufacturability requirement. However, inherent fabrication imperfections can still introduce deviations between the measured transmission response and the simulated ideal. To further investigate the impact of these fabrication errors on device performance, we performed a systematic FDTD simulation

study using the long-pass filter from Fig. 5b as a test case. We quantified the specific effects of period and linewidth variations (±5 nm, ±10 nm, and ±15 nm) on the transmission spectrum, with the results detailed in Fig. S31.

Supplementary Table 3. Impact of fabrication variations on key optical metrics

Variation	MAE	Similarity	Peak shift (nm)
Period -15 nm	0.1378	0.9229	10.868
Period -10 nm	0.1031	0.9562	5.9055
Period -5 nm	0.0638	0.9783	2.3733
Period +5 nm	0.0586	0.9629	2.3883
Period +10 nm	0.0884	0.9452	7.2105
Period +15 nm	0.1109	0.9272	8.4257
Linewidth -15 nm	0.1159	0.9320	8.4257
Linewidth -10 nm	0.0967	0.9548	5.9993
Linewidth -5 nm	0.0656	0.9766	4.7918
Linewidth +5 nm	0.0723	0.9732	1.1885
Linewidth +10 nm	0.1295	0.9306	4.7318
Linewidth +15 nm	0.1602	0.8970	10.564

Furthermore, we quantified the practical impact of fabrication errors using specific metrics, as summarized in Supplementary Table 3. Both period and linewidth variations were found to influence the transmission spectrum of the structure. The designed minimum linewidth of the structure is approximately 200 nm. Under actual fabrication, the period error was about 1 nm and the linewidth error was around 15 nm. The resulting measured peak wavelength shift was approximately 10 nm. Furthermore, the measured transmission spectrum in the 600–700 nm range remained high, which aligns well with the simulation results corresponding to a +15 nm linewidth error.

Supplementary Note 17. Workflow of the AIGP design process.

Our inverse design model is intentionally designed to be user-friendly by accepting a wide range of input data without upfront feasibility checks. Instead of validating the input, the system directly evaluates whether the output satisfies the design target.

Taking the inverse design of optical structures from a target transmission spectrum as an example (Fig. S32), the workflow is as follows: The user can provide either a specific target spectrum or abstract design parameters (which are first converted into an ideal filter response by the prompt encoder network). The inverse design network then generates a batch of k candidate structures. These structures are rapidly evaluated by the forward prediction network to obtain their predicted transmission spectra. The results are compared against the input target to calculate a performance metric. If this metric meets a pre-defined threshold, the design is accepted. Otherwise, the process repeats with different initialization parameters to generate a new batch of structures. If after n batches no satisfactory design is found, the input target is deemed infeasible under the current constraints (e.g., fabrication limits, material properties). This approach ensures both flexibility in input and reliability in output.

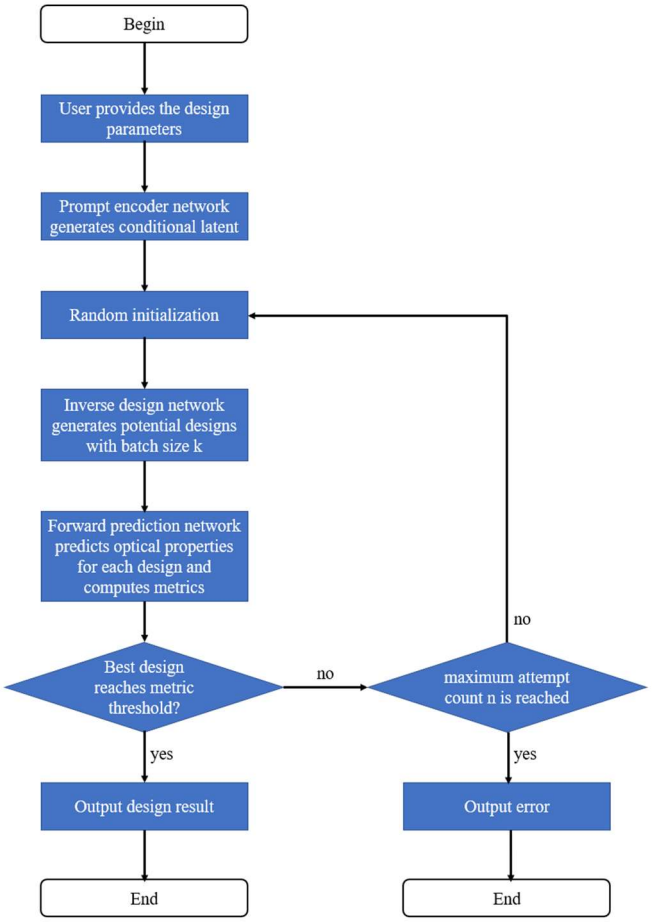

Fig. S32. The workflow of the inverse design process in practice.

In the practical design workflow, we determine whether an inversely designed structure meets the specifications by comparing the forward prediction network's output with the original design target. Since the forward prediction network itself achieves high performance metrics (see Supplementary Note 4), the metric comparing the forward prediction and the design target can effectively serve as a reliable confidence measure for the final design.

In our experiments, we use *cosine similarity* to evaluate the validity of a design. A threshold of 0.95 is applied when the user provides a full target spectrum. When the input consists of abstract design parameters (which are compared against an ideal filter shape), a slightly lower threshold of 0.9 is used.

Regarding the batch size k and the maximum number of attempts n, we recommend that the product $n \cdot k$ be no less than 32. Users can adjust the values of k and n based on their available computational resources.

Even if the inverse design model does not yield a result that meets the predefined threshold, it will still output all generated candidates along with their corresponding evaluation metrics for user assessment. Based on these outputs, the user can determine an appropriate adjustment strategy. For instance, if all generated structures deviate significantly from the target, this suggests that the design objective may be infeasible under the given constraints, and the user may need to relax the target specifications or modify structural parameters such as material or thickness. If certain candidates are close to the target but fail to meet the metric threshold, the user could consider switching to a more suitable evaluation metric—for example, using peak wavelength shift instead of cosine similarity for a bandpass filter—or appropriately lowering the acceptance threshold to ultimately identify a viable design solution.

Supplementary Note 18. Discussion on how to perform inverse design that simultaneously constrains both amplitude and phase.

Multi-response inverse design capability is crucial for practical applications such as multifunctional metasurfaces. Our model is theoretically capable of performing inverse design under multi-dimensional input constraints. However, due to the time required for dataset generation, model training, and performance evaluation, we focus here on demonstrating the feasibility and general workflow of our framework for inverse design that simultaneously constrains both amplitude and phase. This serves to illustrate the versatility of our proposed approach.

As outlined in the main text, our model framework consists of four key components: the forward prediction network, the prompt encoder-decoder network, the image encoder-decoder network, and the inverse design network. To achieve inverse design that simultaneously constrains both amplitude and phase, the architecture and implementation of the forward prediction network remain largely unchanged, though retraining may be necessary. For optical devices such as metalenses that require control over both phase and amplitude, materials with low optical loss—such as TiO_2 or Si_3N_4—are typically preferred, while silicon is less commonly used. Therefore, to adapt to other materials, a dataset must first be generated via FDTD simulations. The forward prediction network can then be fine-tuned through transfer learning, following the methodology described in Supplementary Note 2. The image encoder-decoder network, being independent of the material properties and dependent only on the geometric structure, can be reused directly without any retraining.

Modifications are required, however, in the prompt encoder network. Since amplitude and phase are fundamentally different physical dimensions—with amplitude being an absolute value and phase a relative one—their encoding strategies should differ accordingly. While our current implementation supports separate encoding of amplitude and phase, their joint encoding requires further investigation. A straightforward approach would be to concatenate or add their respective encoded vectors. Alternatively, a dedicated joint encoding scheme could be developed—for instance, by combining amplitude and phase into a complex number and encoding it directly, or transforming them into trigonometric form via Euler's formula. While these methods are theoretically feasible, empirical studies are needed to determine the optimal strategy. Once the prompt encoder network is finalized, the inverse design network can also be retained without structural changes, requiring only fine-tuning via transfer learning.

In summary, while fully achieving inverse design under simultaneous amplitude and phase constraints requires further dataset construction and model refinement, we have provided a comprehensive theoretical and architectural roadmap within our framework. The current work establishes a versatile foundation, and we have outlined clear pathways—including complex number encoding—for future implementation, demonstrating the extensibility of our method.