Research Article Vol. 33, No. 16 / 11 Aug 2025 / Optics Express 34568

Optics EXPRESS

Optimizing the uniformity and efficiency of AR waveguide displays with an open-source beam tracing method

YEFU ZHANG,[1,†] YUQIANG DING,[1,†] ZHENYI LUO,[1] SEOK-LYUL LEE,[2] AND SHIN-TSON WU[1,*]

[1] *College of Optics and Photonics, University of Central Florida, Orlando, FL 32816, USA*
[2] *AUO Corporation, Hsinchu Science Park, Hsinchu 300, Taiwan*
[†] These authors contributed equally
[*] *swu@creol.ucf.edu*

Abstract: Low optical efficiency and poor brightness uniformity are presently two major challenges in waveguide-based augmented reality (AR) displays, often presenting a trade-off that limits the system's performance. Addressing these issues requires a simulation framework that is both physically accurate and computationally efficient. In this paper, we present a framework for optimizing the efficiency and uniformity of waveguide-based AR systems, built on a newly developed open-source beam tracing model. The algorithm accounts for grating multi-interaction effects and traces each beam only once for reuse during optimization, while parallel CPU (central processing unit) processing and data compression are employed to accelerate the simulation speed. The proposed method achieves an average evaluation time of 0.417 seconds, which is more than 100 times faster than non-sequential ray tracing tools, making it highly desirable for iterative waveguide design and optimization.

1. Introduction

Augmented reality (AR) display technology has advanced rapidly over the past two decades [1–6], evolving from bulky headsets to compact and lightweight glasses for applications such as spatial computing, digital twins, medical imaging, and navigation. Key performance metrics include eyebox size [7], brightness uniformity, and optical efficiency [8], which influence image stability, brightness consistency, and power consumption [9]. To overlay virtual content with the real world, optical combiners, particularly waveguide combiners [10,11], are widely used due to their compactness, transparency, and ability to support a large eyebox via beam steering and exit pupil expansion (EPE) [12,13], making them ideal for wearable AR glasses.

Waveguide-based AR displays use an in-coupler to couple the light from microdisplay and an out-coupler to extract the light, and a folding-coupler to enable 2D EPE [14,15]. Both geometric and diffractive couplers have been developed. The geometric coupler consists of several cascaded partial reflectors with different reflectivity to achieve eyebox uniformity. It offers relatively high optical efficiency (~10%), but it often suffers from stray light and ghost images due to unintended reflections. In contrast, diffractive couplers, such as surface relief gratings (SRGs) [16–18] and polarization volume gratings (PVGs) [19–21], use micro- or nano-scale periodic structures, offering greater control with reduced stray light.

Although waveguide-based AR displays have advanced significantly in recent years [22–24], they continue to face persistent challenges related to uniformity and optical efficiency. More importantly, the fundamental performance limits of these systems remain unclear. Key questions include: What is the maximum efficiency achievable for a given waveguide design? What is the trade-off between uniformity and efficiency? How many coupler zones are required for acceptable performance, and does the slicing direction of these zones matter? Additionally, if the design

#570463 https://doi.org/10.1364/OE.570463
Journal © 2025 Received 11 Jun 2025; revised 18 Jul 2025; accepted 28 Jul 2025; published 5 Aug 2025

of the waveguide must be preserved, requiring limitations on grating efficiency, how does that impact overall system performance?

Addressing these questions requires a deeper understanding of the ideal angular response of each coupler in the system. A critical factor in this analysis is the multi-interaction behavior of gratings [25], especially at the in-coupler and folding-couplers. After the initial diffraction, residual light may continue to propagate and interact with the same grating multiple times. Therefore, incorporating grating's multi-interactions into the simulation is essential for accurate evaluation and reliable optimization of AR waveguide systems.

Recently, a beam tracing method was proposed for waveguide AR evaluation [26]. This approach, implemented in MATLAB, introduces a recursion algorithm for beam propagation combined with targeted optimization strategies. It demonstrates efficient computation and promising design insights. However, it does not consider multi-interaction effects, and beams must be retraced for each evaluation, limiting both accuracy and computational efficiency. Moreover, commonly used non-sequential ray tracing tools also fail to model multi-interaction accurately, as their ideal grating assumption neglects repeated interactions within the waveguide.

To address these limitations, in this paper, we propose a field-of-view (FoV)-based beam tracing method (BTM) specifically tailored for diffractive waveguide displays. Our approach uses polygon-based beam modeling, incorporates grating multi-interaction analysis, and leverages parallel CPU processing for high-speed simulation. Notably, here multi-interactions refer to repeated diffraction events that change the beam's propagation direction. Each beam is traced only once and stored for reuse, dramatically reducing the computation time. This method enables fast and accurate evaluation of both efficiency and uniformity, laying the groundwork for coupler design, system optimization, and comprehensive exploration of the efficiency-uniformity trade-off in AR waveguides. The algorithm is implemented as an open-source framework, making it a valuable tool for both academic research and industrial applications.

2. Conventional optimization approach for AR waveguide design

Before introducing the BTM, we first describe the conventional optimization approach. This method relies on ray tracing and is typically implemented using commercial non-sequential ray tracing software such as Zemax or LightTools. It is an important technique because it takes all FoV angles into account during optimization. Additionally, it considers the actual performance of different grating structures through rigorous coupled-wave analysis (RCWA) simulation [27,28].

The complete process is illustrated in the flowchart shown in Fig. 1. The first step is to define the design targets, which typically include FoV, eye-box size, and operation wavelength of the system. Choosing the FoV is especially important because the refractive index limits the total internal reflection (TIR) region in the k-diagram.

To demonstrate how this method works, we present a full design example. In this case, the design targets are FoV = 20°×15° and eye-box size = 12 mm × 8 mm. The waveguide glass used in the system has a thickness of 0.7 mm and a refractive index of 1.52. The in-coupler has a radius of 2 mm. As shown in Fig. 2(a), the K-diagram confirms that the FoV lies entirely within the TIR region. The grating periods for both in-coupler and out-coupler are set at 0.437 µm. The corresponding k-vectors are oriented at −38° and −142° relative to the x-axis, respectively. The next step involves designing the in-coupler, folding-coupler, and out-coupler. The eye relief which is defined as the distance from the user's eye to the out-coupler is set to be 20 mm. Using the FoV and eye relief, we calculate the size of the out-coupler by back-tracing the eye-box. Once the position and size of the out-coupler are determined, we continue the back-tracing process to find the appropriate shape and position of the folding-coupler. Based on the required in-coupler size and the system's FoV, we can also design the projection lens and display panel. After completing these steps, the full system layout is established, as Fig. 2(b-c) depicts. For clarity, the couplers are labeled in Fig. 2(b). With the couplers' sizes and positions now defined, we can

Optics EXPRESS

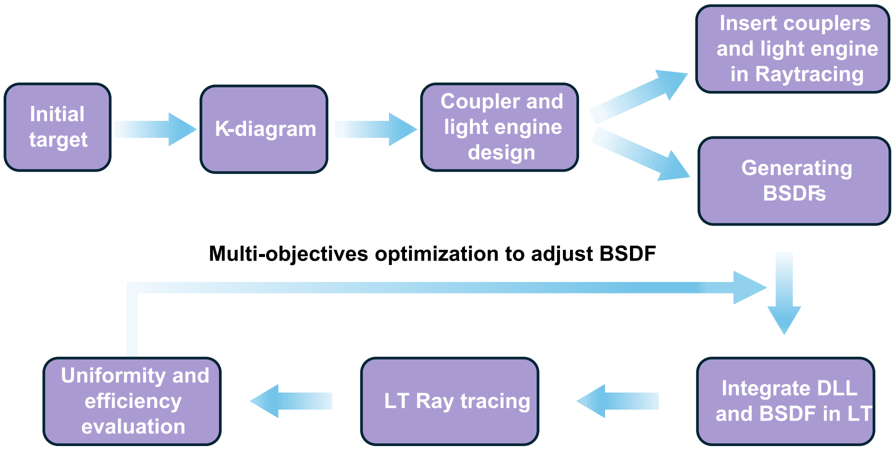
Initial target
K-diagram
Coupler and light engine design
Insert couplers and light engine in Raytracing
Generating BSDFs
Multi-objectives optimization to adjust BSDF
Integrate DLL and BSDF in LT
LT Ray tracing
Uniformity and efficiency evaluation

Fig. 1. Flowchart of conventional optimization process for waveguide systems. BSDF: Bidirectional scattering distribution function, LT: LightTools, and DLL: dynamic-link library.

generate a waveguide model in non-sequential ray tracing software. To streamline this process, we develop a MATLAB-based Application Programming Interface (API) for LightTools. After inputting these basic design parameters, the API automatically creates a 3D waveguide model in LightTools, ready for further optimization, as illustrated in Fig. 2(d).

After the desired 3D model is generated, the next step is the optimization process. As shown in Fig. 2(b), the folding-coupler is divided into 7 zones and the out-coupler into 6 zones. Depending on the type of grating used, either transmissive or reflective, the API assigns optical properties to the front or back surfaces of each coupler zone. These properties are set as 'user-defined', and a custom dynamic-link library (DLL) for the gratings is uploaded to define their behavior. During optimization, the bidirectional scattering distribution function (BSDF) is generated based on the specified grating period. The optimization variables depend on the grating type. For polarization volume gratings (PVGs), the key variables are the slant angle and thickness. For slanted binary surface-relief gratings (SRGs), the variables include the duty ratio, wall depth, and slant angle. More complex trapezoidal SRGs involve additional variables, providing greater flexibility in optimization.

This method takes all FoV angles into account and incorporates the actual optical response of the gratings, making it a comprehensive and accurate approach. However, its main limitation is the significant computational cost. Each optimization step requires regenerating the BSDF files, and the ray tracing for such a complex system in LightTools is time consuming. Without access to a high-performance computing system, the process can become extremely slow. For this reason, we propose a new algorithm BTM to significantly improve the efficiency of the optimization process.

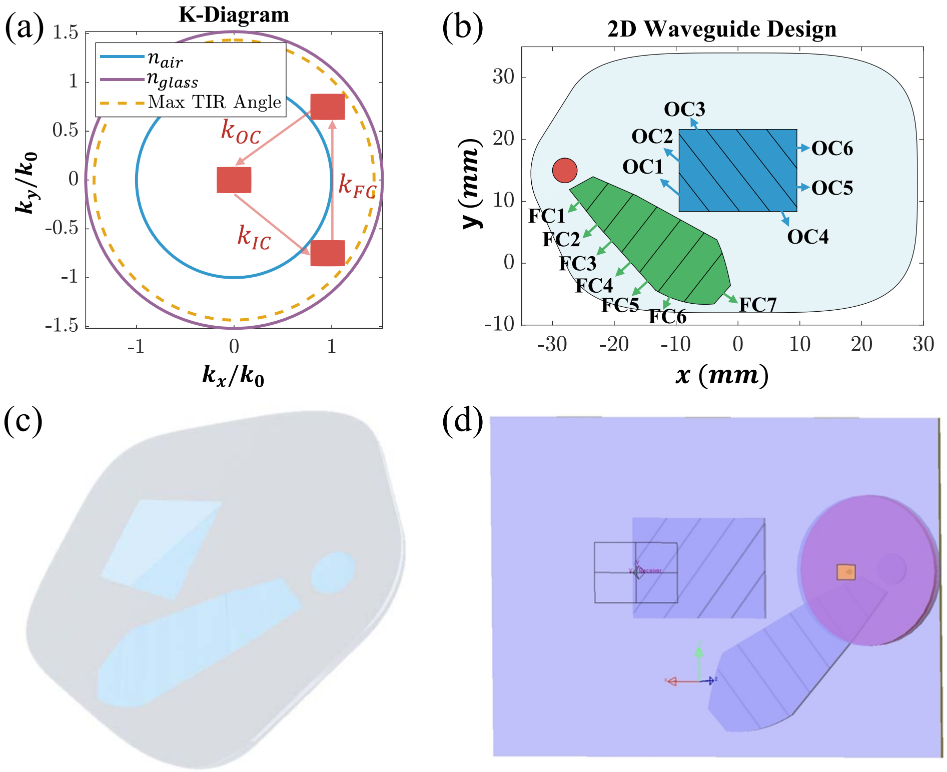

Fig. 2. Design example of an AR waveguide system. (a) K-diagram showing that the target FoV of 20°×15° lies entirely within the TIR region. (b) 2D schematic of the waveguide layout, with labeled folding-coupler, and out-coupler positions. (c) 3D waveguide model visualized using the Python Open3D package. (d) 3D waveguide model generated using a MATLAB-based API for LightTools, enabling further optical simulation and optimization.

3. FoV-based optimization using beam tracing method

3.1. *Overview of FoV-based optimization*

Before explaining how the FoV-based optimization method works, we must first establish a complete system design. For consistency and clarity, the design described in Section 2 will be used throughout the following explanation and optimization process. It should be noted that the current beam tracing approach is designed for systems using spatially separated 1D folding and 1D output couplers. Adapting the method to configurations such as single 2D gratings or crossed 1D gratings would require modifying the calculation logic. The core idea of FoV-based optimization is to optimize the optical performance for each field-of-view (FoV) angle individually. As illustrated in Fig. 3(a), the light from microdisplay panel first passes through an ideal projection lens and becomes a collimated beam. This collimated light then encounters the in-coupler. Each position on the display panel corresponds to a specific propagation direction, which determines the angle at which the light enters the waveguide system and interacts with the in-coupler grating. Once coupled into the waveguide, the light changes direction, and its new propagation path is defined by a different k-vector. Figure 3(b) shows the range of polar and azimuthal angles at which light reaches the folding-coupler. This is defined as the effective angular response region of the folding-coupler. Only the light within this angular region contributes to image formation, so grating efficiency outside this region has no impact on the system performance and is excluded from consideration. A similar effective angular region is defined for the out-coupler, as shown in Fig. 3(c).

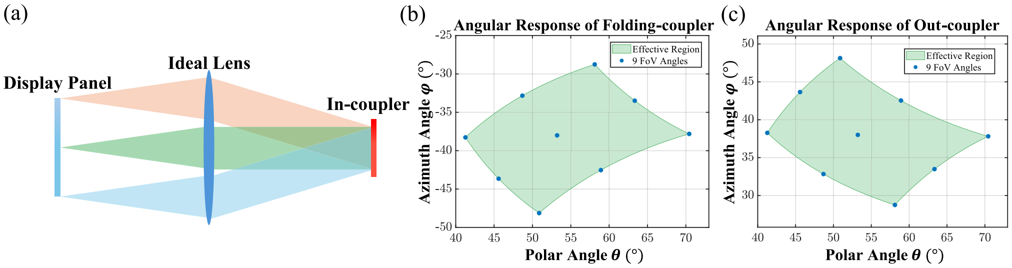

Fig. 3. (a) Illustration of light propagation from the display panel through an ideal lens to the in-coupler. (b) Effective angular response region for the folding-coupler. (c) Effective angular response region for the out-coupler.

For each specific FoV direction, there is a corresponding point in the effective region of every coupler. Since the incident angles (both polar and azimuthal) are fixed for each direction, the grating response as an ideal grating operating at a known angle. This simplifies the optimization problem into a set of discrete angular targets. The goal of the optimization is to determine the ideal angular response for each coupler that results in uniform brightness across the entire eyebox while maintaining high optical efficiency. Here, we illustrate our method using nine representative FoV angles, which are marked as blue dots in Fig. 3(b) and 3(c). These nine FoV angles correspond to nine points on the display panel: the four corners, the center, and the midpoints of each edge. The display is aligned along the x–y plane and collimated by a projection lens before the light enters the waveguide. As a result, the nine selected rays enter the waveguide with the following angular directions (horizontal, vertical): (10°, 7.5°), (0°, 7.5°), (−10°, 7.5°), (10°, 0°), (0°, 0°), (−10°, 0°), (10°, −7.5°), (0°, −7.5°), and (−10°, −7.5°). Although nine angles are sufficient to illustrate the concept, a denser sampling, such as 25 or 36 angles, would improve the accuracy in the final design. Once the target angular responses are identified for all couplers, the final step is to design or fine-tune the gratings so that their actual responses match the targets as closely as possible. This ensures optimized light propagation and high-quality image delivery across the entire FoV.

After discussing the angular response of the couplers, the next key aspect is to identify the effective folding-coupler and out-coupler regions for different FoV angles [26]. As described in Sec. 2, both folding-coupler and out-coupler are determined by back-tracing the eyebox through the waveguide. For each FoV angle, the eyebox is traced back to a unique position on the waveguide, and the collection of these back-traced areas defines the regions where light must be extracted. As a result, each FoV angle corresponds to a specific effective region on the folding-coupler and out-coupler. Only the light that interacts with these effective regions can be successfully guided out and reach the viewer's eye. Any light coupled outside these regions is wasted, as it does not contribute to the eyebox output.

As Fig. 4 shows, the effective regions for the nine representative FoV angles are highlighted in red. The couplers that overlap with these regions are highlighted to indicate their relevance to the optimization. During the optimization process, the algorithm selectively considers only the beams passing through these effective regions. To handle couplers located near, but not within these regions, we define folding-couplers with red or yellow edges, which are also marked in Fig. 4. Red edges represent folding-couplers just before the effective region, typically on the side closer to the in-coupler. Since light reaching these couplers is unlikely to enter the eyebox, their first-order diffraction efficiency is set to be zero for that specific FoV angle. In contrast, yellow edges mark folding-couplers beyond the effective region, after the light has already passed through the intended exit path. These yellow-edge folding-couplers do not affect the system's efficiency or uniformity, so their diffraction efficiency can be set to any value. This classification helps ensure that only useful beams are considered and prevents wasted light from affecting the optimization time.

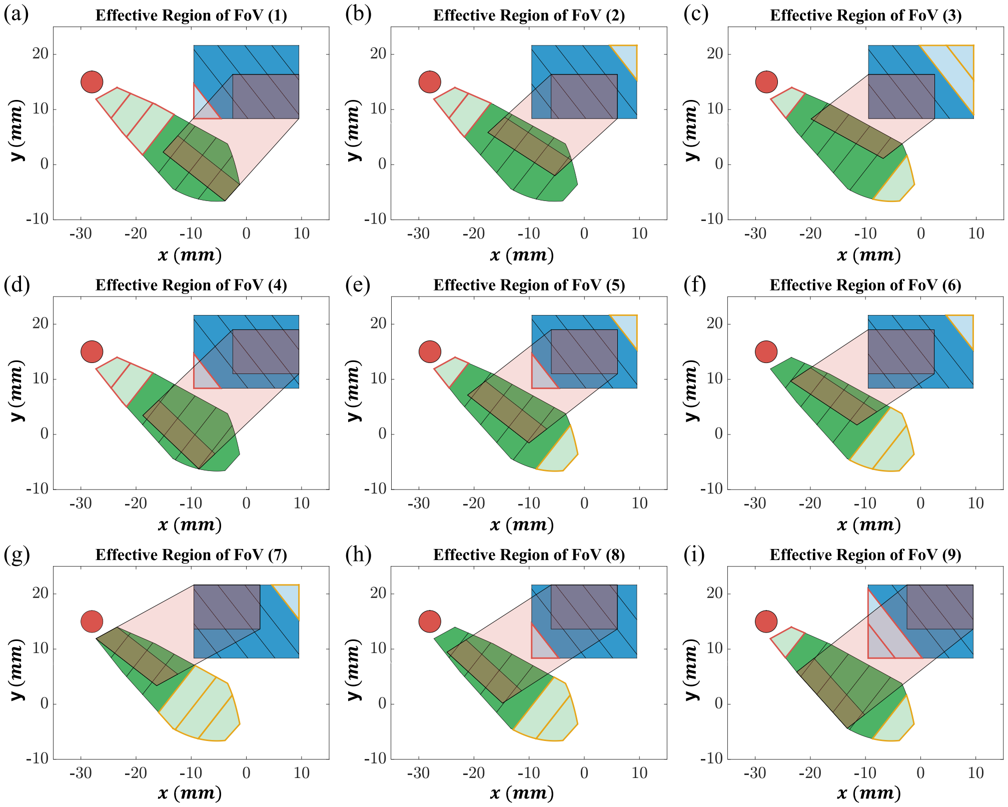

Fig. 4. Effective coupling regions for nine FoV angles, shown in subplots (a)-(i) corresponding to FoV (1) to FoV (9). Red regions indicate where light must be extracted to reach the eyebox. Couplers overlapping these regions are highlighted. Red-edge folding-couplers are excluded due to low contribution, while yellow-edge folding-couplers have no impact on performance.

In our simulation, we assume zero first-order diffraction efficiency in the red-edged folding-couplers that do not contribute light to the eyebox. This assumption represents an idealized scenario where the grating is designed to minimize the efficiency at non-useful angles, allowing us to estimate the upper limit of system performance. Users may modify these effective regions as needed when applying the open-source code.

3.2. Beam tracing in AR waveguides

The BTM algorithm is developed in a Python environment. One of the key aspects of this method is the multi-interaction behavior of the grating. In a practical waveguide system, light may interact with a grating multiple times due to internal reflections. According to previous studies, the sum of the first-order diffraction efficiency from the initial grating interaction and the zero-order diffraction efficiency from subsequent interactions cannot exceed 1, due to the principle of energy conservation [25]. This relationship can be expressed mathematically as:

$$\eta_{+1}(\theta) + \eta_0(\theta) \leq 1, \tag{1}$$

where θ is the TIR angle of the first-order beam after its initial interaction with the grating. In this paper, we make two simplified assumptions to better illustrate how the BTM algorithm works. First, we assume the in-coupler has 100% first-order diffraction efficiency. This means that all the energy is coupled into the waveguide on the first interaction, and any subsequent interaction results in complete energy loss. As a result, the beam coupled into the waveguide forms a crescent-shaped distribution, and only the remaining energy is traced, as shown in Fig. 5(a). The percentage values in Fig. 5(a) indicate the fraction of energy that remains after interacting with the in-coupler, which varies depending on the FoV angle due to the differences in TIR angles. The second assumption is that $\eta_{+1}(\theta) + \eta_0(\theta) = 1$. These assumptions simplify the algorithm and clarify the behavior of multi-interactions. However, for a more general and realistic optimization, the in-coupler efficiency should be based on simulated data, and the efficiency values for multi-interactions should vary during the optimization process.

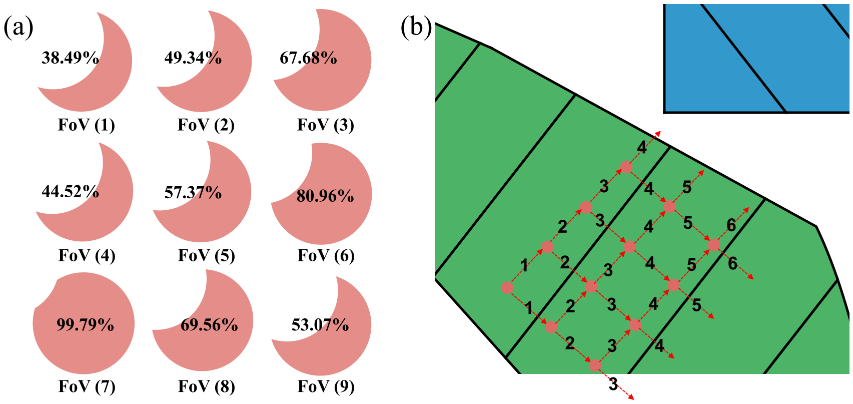

Fig. 5. (a) Crescent-shaped beam after in-coupling for nine FoV angles. (b) Illustration of multi-interaction behavior in the folding-coupler region.

Within the folding-coupler region, multi-interactions also occur, as illustrated in Fig. 5(b). This figure shows how a single light beam can interact with the grating multiple times. The number next to each beam indicates how many times it has interacted with the grating. The beam's propagation path is determined by the TIR angle. Each time the beam interacts with the grating, it splits into two new beams, one for each diffraction order, which then continue propagating. Because of this, the number of beams grows exponentially, especially for beams with smaller TIR angles. In contrast, there is no multi-interaction in the out-coupling region. Once light is coupled out of the waveguide, it exits the system and does not undergo further internal reflections. Importantly, this intrinsic multi-interaction behavior cannot be accurately modeled in commercial non-sequential ray tracing software like LightTools. In these tools, gratings are treated as ideal elements that only interact with light once. After the initial diffraction, light continues to propagate via TIR without re-interacting with the grating, which limits the ability to simulate realistic behavior in waveguide systems.

In our method, each light beam is characterized by four key properties: shape, position, energy, and propagation direction (represented by the k-vector). The shape and position of a beam are described using its edge coordinates, effectively treating the beam as a polygon. To handle geometric operations such as propagation, slicing, and overlap detection, we use the Python package Shapely, which provides robust tools for polygon-based computations. This allows us to model beam interactions and transformations efficiently and accurately. The treatment of energy in this framework will be discussed in Sec. 3.3.

Due to the multi-interaction behavior of gratings, the computational workload in beam tracing becomes extremely high, especially for beams with small TIR angles. These beams travel a shorter distance between interactions and therefore encounter gratings more frequently. Additionally, for certain FoV angles where the effective region in the folding-coupler is located far from the out-coupler, such as FoV (9) shown in Fig. 4(i), the number of interactions increases significantly compared to other FoV directions. As a result, the total number of beams grows rapidly. Beyond multi-interactions, beam complexity also increases when a single beam overlaps with two or more couplers. In such cases, the beam must be sliced into multiple segments, each propagating separately and carrying a different portion of the original energy. This further contributes to the exponential growth in the number of beams during simulation. To handle this computational demand, we implement parallel processing using Python's multiprocessing package. This allows the algorithm to run in parallel CPU mode, significantly improving the simulation speed and making the process more scalable for a complex waveguide system.

The beam tracing process within the folding-coupler region is outlined in the flowchart shown in Fig. 6. The process begins when the in-coupled beam propagates toward the folding-coupler. All operations enclosed in the red region of Fig. 6 are collectively referred to as one layer of the tracing algorithm. These operations are described in detail as follows. Each propagation step is governed by the TIR angle and the waveguide thickness, which jointly determine the beam's travel distance. After each step, the algorithm checks for overlap between the propagating beam and all effective folding-coupler regions. If an overlap is detected, the $+1^{\text{st}}$ order beam is generated and saved. Its propagation direction (k-vector) is then updated to point toward the out-coupler. Meanwhile, the 0^{th} order beam continues to propagate and is checked for overlaps at each subsequent step. This loop continues until the 0^{th} order beam completely exits the effective region. At that point, the operations for the first layer are complete.

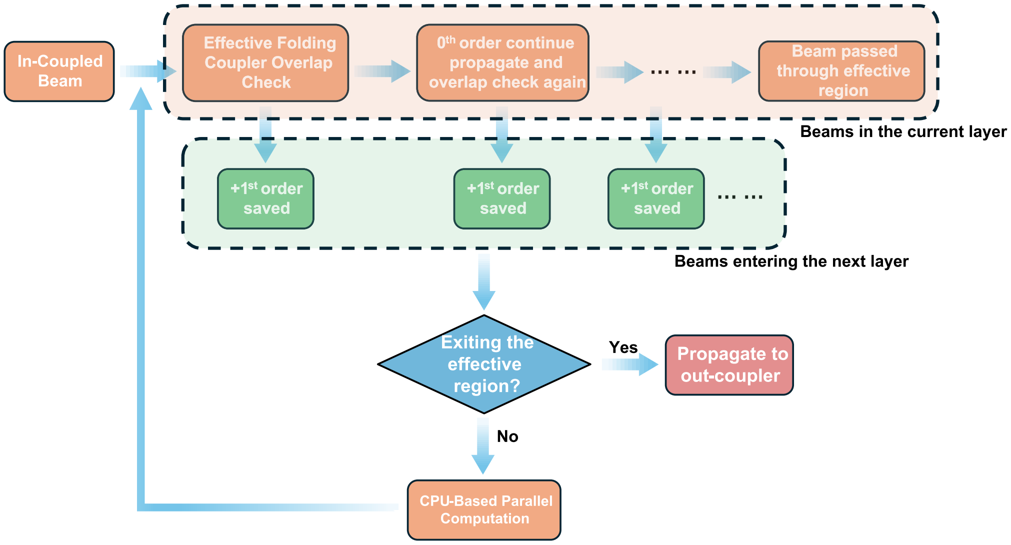
In-Coupled Beam
Effective Folding Coupler Overlap Check
0th order continue propagate and overlap check again
… …
Beam passed through effective region
Beams in the current layer
+1st order saved
+1st order saved
+1st order saved
… …
Beams entering the next layer
Exiting the effective region?
Yes
Propagate to out-coupler
No
CPU-Based Parallel Computation

Fig. 6. Flowchart of the beam tracing process in the folding-coupler region.

Next, the saved $+1^{\text{st}}$ order beams, indicated in the green region of Fig. 6, propagate one TIR distance toward the out-coupler direction. After this propagation, another overlap check is performed. If a beam no longer overlaps with any effective folding-coupler region, it is assumed to have exited the folding-coupler entirely and proceeds directly toward the out-coupler.

However, if an overlap still exists, the beam enters the second layer, and the same set of operations from the first layer (as outlined in the red region) is repeated. Beginning from the second layer onward, the algorithm transitions to parallel beam tracing, where multiple beams are processed simultaneously using CPU-based multiprocessing. This layer-by-layer approach significantly improves computational efficiency, as each layer processes a group of beams propagating in the same direction. As illustrated in Fig. 7(a), beams within each layer move uniformly along a fixed direction. After completing the operations for one layer, the newly generated beams are passed to the next layer and the process repeats. This iterative propagation continues until all beams have exited the effective folding-coupler region.

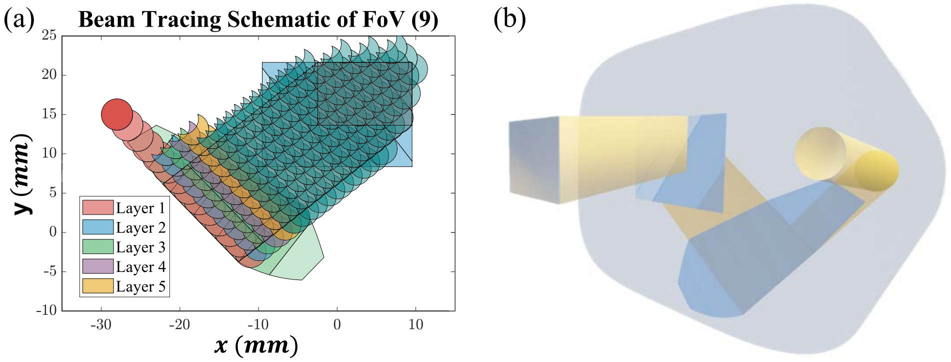

Fig. 7. (a) Layer-by-layer beam propagation in the waveguide system. Beams in each layer propagate in the same direction and are processed in parallel to improve computational efficiency. (b) 3D visualization of light propagation and out-coupling for FoV (9), rendered using Python's Open3D package.

After passing through the folding-coupler, the light continues to propagate toward the out-coupler. Since there is no multi-interaction in the out-coupler region, light will exit the waveguide after a single interaction, the beam tracing process becomes significantly more straightforward. All beams propagate in the same direction, simplifying the calculations, as illustrated in Fig. 7(a). As in the folding-coupler stage, CPU-based parallel computation is used to accelerate the beam tracing in this region. This ensures that even with many beams, the process remains efficient. Figure 7(b) presents a 3D model that visualizes the effective light propagation within the waveguide for FoV (9). It clearly shows how the beams exit through the out-coupler and contribute to the final image formation in the eyebox.

3.3. Beam storage and compression strategy

As mentioned earlier, our method requires beams to be traced only once. All traced beams are stored and later utilized for evaluating the optical efficiency and uniformity of the system. Since many beams propagate within the waveguide, only those contributing to the final output, specifically, beams entering the effective region of the out-coupler, are retained.

The size of the effective region in the out-coupler is identical to that of the eyebox, as it is derived directly from back-tracing the eyebox, as shown in Fig. 7(b). For different FoV angles, these effective regions are located at different positions across the out-coupler. Once all beams exit the folding-coupler, the tracing toward the out-coupler begins. When a beam overlaps with the effective region of the out-coupler, it will be coupled out, corresponding to the $+1^{\text{st}}$ diffraction order of the grating. The remaining 0^{th} order continues to propagate. If the out-coupling point lies within the effective region, the corresponding beam data will be recorded.

To enable efficient evaluation of system performance, each valid beam is mapped onto a discretized matrix representing the effective out-coupling region. This matrix uses binary values to indicate spatial occupancy: subregions intersected by the beam are assigned a value of 1, while all other subregions are set to be 0. In addition to this binary representation, the algorithm also records the geometric area of the overlap and the corresponding optical energy. For this waveguide design, the eyebox is defined as 12 mm × 8 mm, and the effective region is discretized with a sampling resolution of 0.1 mm. As a result, the effective region is represented by a matrix of size 121×81. As beams enter this region, the algorithm identifies the overlapping subregions, marks the corresponding matrix elements, and calculates the associated area. An example of this process is illustrated in Fig. 8, where the green region indicates the portion of the beam intersecting the effective out-coupling area.

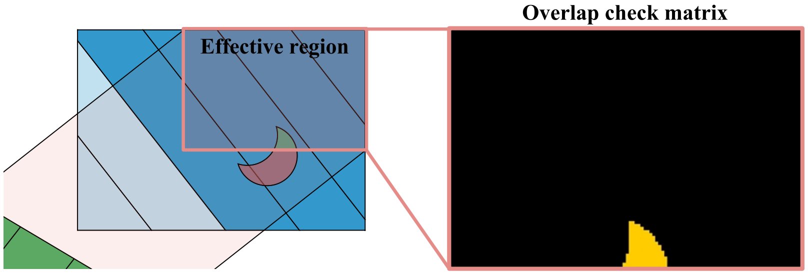

Fig. 8. Example of overlap detection between a beam and the effective out-coupling region. The effective region is discretized into a 121×81 binary matrix, with intersected subregions assigned a value of 1.

Another critical parameter that must be recorded for each beam is its energy. During the optimization process, the diffraction efficiency of each coupler varies dynamically. As a result, it is impractical to assign fixed energy values during beam tracing; otherwise, the entire beam tracing process would need to be repeated each time the coupler efficiencies are updated. To address this, our method records the number of interactions each beam has with the relevant couplers, rather than the energy itself. As discussed above, we assume $\eta_{+1}(\theta) + \eta_0(\theta) = 1$ for simplicity. For example, in the case of FoV (9), there are five effective folding-couplers and four effective out-couplers, as illustrated in Fig. 4(i). Considering both $+1^{\text{st}}$ and 0^{th} diffraction orders for each coupler, a total of 18 interaction terms must be recorded to determine the energy contribution of a given beam. Each beam's interaction history is stored as a 1×18 vector, $[n_1, n_2, \ldots, n_{18}]$, where each element n_i represents the number of times the beam interacts with a specific diffraction order of a coupler. The final energy of the beam is computed based on this energy vector using the following equation:

$$\begin{aligned} E = {} & [\eta_{fc2}^{n_1}(1-\eta_{fc2})^{n_2} \cdot \ldots \cdot \eta_{fc6}^{n_9}(1-\eta_{fc6})^{n_{10}}] \cdot \\ & [\eta_{oc3}^{n_{11}}(1-\eta_{oc3})^{n_{12}} \cdot \ldots \cdot \eta_{oc6}^{n_{17}}(1-\eta_{oc6})^{n_{18}}], \end{aligned} \tag{2}$$

where $\eta_{fc2}, \ldots, \eta_{fc6}$ represent the $+1^{\text{st}}$ order diffraction efficiencies of the effective folding-couplers, and $\eta_{oc3}, \ldots, \eta_{oc6}$ denote the $+1^{\text{st}}$ order efficiencies of the effective out-couplers. During the optimization process, these efficiencies are treated as variables. For the current FoV angle, a total of 13 optimization variables are considered, corresponding to the relevant coupler interactions. Based on Eq. (2), the energy of each beam can be computed rapidly using the stored energy vector and the current values of the efficiency variables. As a result, the beam tracing procedure only needs to be performed once and stored. Subsequent evaluations of optical efficiency and uniformity can be carried out without retracing, significantly reducing the overall computational time during optimization.

Given the complexity of the beam tracing process, it is common for multiple beams to have identical energy. That is, the same energy vector $[n_1, n_2, \ldots, n_{18}]$ after being stored. To reduce data redundancy and further accelerate subsequent computations, a compression step is introduced. In this step, all beams sharing the same energy vector are merged by summing their corresponding overlap check matrices and their respective overlap areas. The energy vector itself remains unchanged. As a result, these multiple beams are combined into a single representative beam that encapsulates the total contribution of the group. This compression significantly reduces the number of beams stored and processed, thereby improving the computational efficiency of the subsequent efficiency and uniformity evaluations during the optimization process.

3.4. Evaluation of efficiency and uniformity

As above mentioned, each saved beam in the system is associated with three key data components: the overlap check matrix, the overlap area, and the energy vector. During the efficiency calculation, the overlap area is used to determine the optical contribution of each beam. The system efficiency is then computed using the following equation:

$$\eta = \frac{E_{out}}{E_{in}} = \frac{E_1A_1 + E_2A_2 + E_3A_3 + \ldots}{1 \times A_{in}}, \tag{3}$$

where $E_1, E_2, \ldots$ represent the energies of the individual beams, calculated using Eq. (2), $A_1, A_2, \ldots$ denote their corresponding overlap areas within the effective region, and A_{in} is the area of the input beam, which is equivalent to the area of the in-coupler. The overall system efficiency is therefore determined by summing the weighted energy contributions of all output beams relative to the energy input.

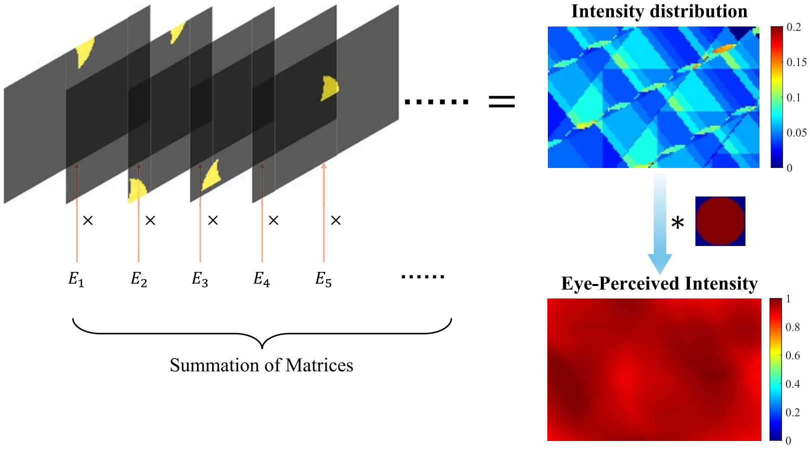

Fig. 9. Procedure for eyebox uniformity evaluation. Each beam's overlap matrix is weighted by its energy and summed to form the intensity distribution. A 4 mm diameter circular aperture is then applied via convolution to simulate the human eye's perception, yielding the final normalized intensity profile used for uniformity calculation.

To evaluate the eyebox uniformity, the stored overlap check matrices are utilized. As illustrated in Fig. 9, each matrix is multiplied by the corresponding beam energy to generate the beam-specific intensity distribution. These weighted matrices are then summed up to obtain the overall intensity distribution within the eyebox. However, to approximate the perceived intensity from the perspective of a human observer, a convolution with a circular aperture is required. This step accounts for the finite size of the human pupil. In this study, the aperture is modeled as a circular mask with a diameter of 4 mm, and its spatial resolution is set to be 0.1 mm, matching the sampling resolution of the effective region. Following the convolution, the resulting distribution represents the perceived intensity profile across the eyebox. The uniformity of this profile is then calculated using the following equation:

$$U = \frac{I_{min}}{I_{max}}, \tag{4}$$

where I_{min} and I_{max} are the minimum and maximum values of the human eye-observed intensity distribution, respectively. This metric provides a quantitative measure of brightness consistency across the eyebox, with values closer to 1 indicating better uniformity.

4. Results and discussion

After presenting the details of the algorithm, we discuss the results of the beam tracing and optimization process. Table 1 summarizes the number of effective beams for each FoV angle, along with the corresponding calculation time required for a single evaluation of efficiency and uniformity. These calculations were performed using the CPU-based parallel processing mode. The average computation time across the nine FoV angles is 0.417 seconds, demonstrating a significant improvement in speed compared to traditional non-sequential ray tracing software. All simulations were conducted on a computer equipped with a 12th Gen Intel Core i7-12700 processor.

Table 1. Effective beam count and evaluation time per FoV angle using CPU-based parallel processing

FoV	1	2	3	4	5	6	7	8	9
Number of beams	26,043	4,025	452	422,652	19,824	1,613	1,582	46,334	1,002,258
Calculation Time (s)	0.065	0.016	0.008	1.02	0.061	0.01	0.01	0.113	2.45

Notably, due to multi-interactions within the waveguide, the number of beams increases exponentially as the FoV range expands or as more folding and out-coupler slices are introduced. This exponential growth significantly increases the memory requirement, making large Random Access Memory (RAM) capacity essential for supporting the complex configurations during the optimization process. The optimization process is carried out using the NSGA-II algorithm implemented via the Python-based PyMOO package. NSGA-II (Non-dominated Sorting Genetic Algorithm II) is a widely used multi-objective evolutionary algorithm that identifies a diverse set of Pareto-optimal solutions through non-dominated sorting and crowding distance mechanisms. In our optimization framework, two objectives are defined: optical efficiency and eyebox uniformity. The decision variables correspond to the $+1^{st}$ order diffraction efficiencies of the effective couplers, with each variable constrained within the range between 0 and 1. These variables are adjusted during the optimization process to identify solutions that achieve the best trade-off between efficiency and uniformity.

The optimization begins with a rough optimization phase applied to all FoV angles individually. In this phase, a limited number of evaluations are performed. The goal is to identify the FoV angle with the lowest achievable efficiency, under the condition that all FoVs maintain a comparable level of eyebox uniformity. From this process, FoV (4) is identified as the angle with the minimum optical efficiency among all candidates. The resulting Pareto front for FoV (4) is shown in Fig. 10(a), illustrating the trade-off between efficiency and uniformity. In this study, a minimum uniformity threshold of 80% is imposed to ensure acceptable visual quality. Under this constraint, the maximum efficiency achieved is approximately 18.37%. It is important to note that, beyond eyebox uniformity, FoV uniformity (or image uniformity across different viewing directions) is also critical for consistent display performance. Therefore, the optimization strategy requires that all FoV angles achieve the same efficiency. This ensures a uniform image brightness across the field of view. Based on the rough optimization results, the minimum efficiency value (18.37%) is selected as a constraint. All other FoV angles are then optimized with this fixed efficiency target, with the objective of maximizing uniformity. The final results are presented in Fig. 10(b), where all FoV angles achieve 18.37% efficiency while maintaining high uniformity.

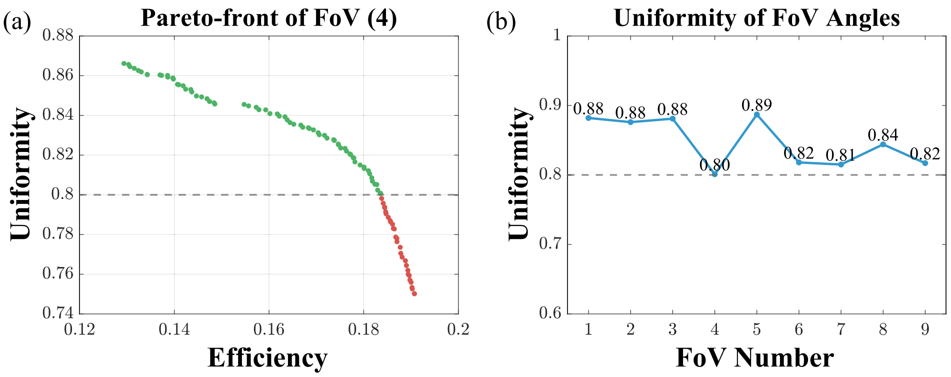

Fig. 10. (a) Pareto front for FoV (4) showing the trade-off between optical efficiency and eye-box uniformity. (b) Final optimization results for all nine FoV angles, each achieving 18.37% efficiency while maintaining high uniformity.

The final optimization results lead to a system with 18.37% efficiency and uniformity exceeding 80% for all FoV angles. While these results are significant, the primary objective of this study is to determine the angular response of each coupler that supports this performance. The derived efficiency values for the folding-couplers are illustrated in Fig. 11. Since nine FoV angles are used in the optimization, each folding-coupler has nine corresponding efficiency values, one for each direction. These are represented as histograms, where the height indicates the +1$^{\text{st}}$ order diffraction efficiency at a given angle. Notably, in some plots, certain FoV directions are shown as dashed lines instead of filled bars. These correspond to the yellow-edged folding-couplers defined in Fig. 4, which indicate areas where the beam does not contribute to the eyebox. As such, the efficiency at these angles does not influence waveguide performance and can be assigned arbitrary values. A similar analysis is performed for the out-couplers, and the resulting efficiency distributions are plotted in Fig. 12. These results collectively define the target angular response for each coupler, forming the basis for practical grating design and fabrication.

Thus far, the performance results for the designed waveguide system have been presented. Next, we will investigate how slicing the folding-coupler and out-coupler impacts the overall system performance. Initially, the system starts with a single folding coupler and a single out-coupler (i.e., no slicing). The number of slices is then increased to 10 folding-couplers and 9 out-couplers. Slicing is performed in the direction perpendicular to the propagation path of normally incident light entering the in-coupler. Each coupler is divided into equal segments from the starting edge to the endpoint. Throughout this process, the uniformity is maintained above 80%, while the corresponding optical efficiency is evaluated. Results are plotted in Fig. 13(a). As the number of couplers increases, efficiency increases rapidly and then gradually saturates, reaching the highest efficiency (23%) at the (8,7) slicing configuration.

Another parameter explored is the slicing direction. In addition to slicing along the original (normal incidence) direction, alternative slicing angles are tested. Using the configuration of 6 folding-couplers and 5 out-couplers as a baseline, the slicing direction is varied by rotating the original cut by ±3.25° and ±7.5°. The performance comparison is shown in Fig. 13(b). Interestingly, the best efficiency is not achieved with the original slicing direction (0°), but rather at approximately +3.25°, suggesting that a slight tilt in slicing direction can lead to improved system performance.

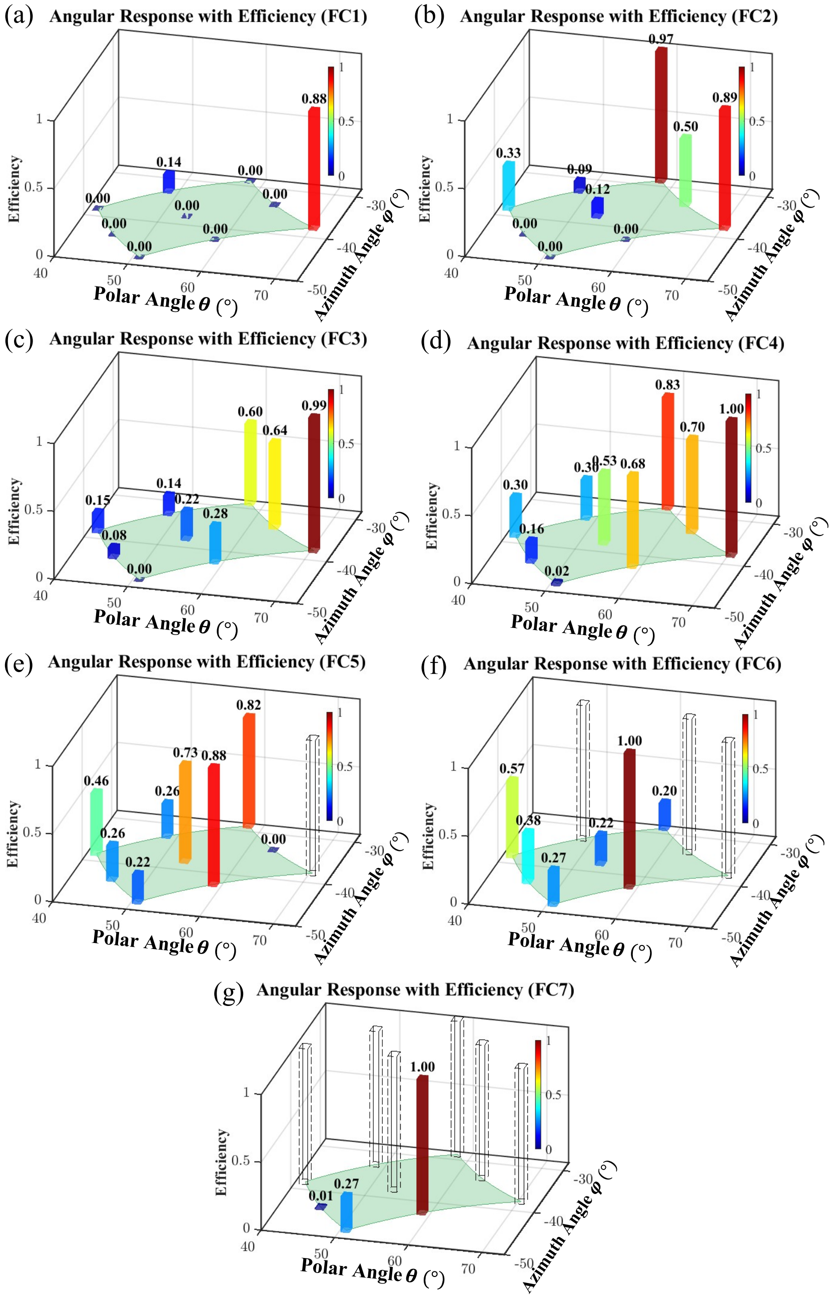

Fig. 11. Optimized +1st order diffraction efficiencies for the folding-couplers FC1 to FC7, shown in subplots (a)–(g). Each histogram displays efficiencies for nine FoV angles.

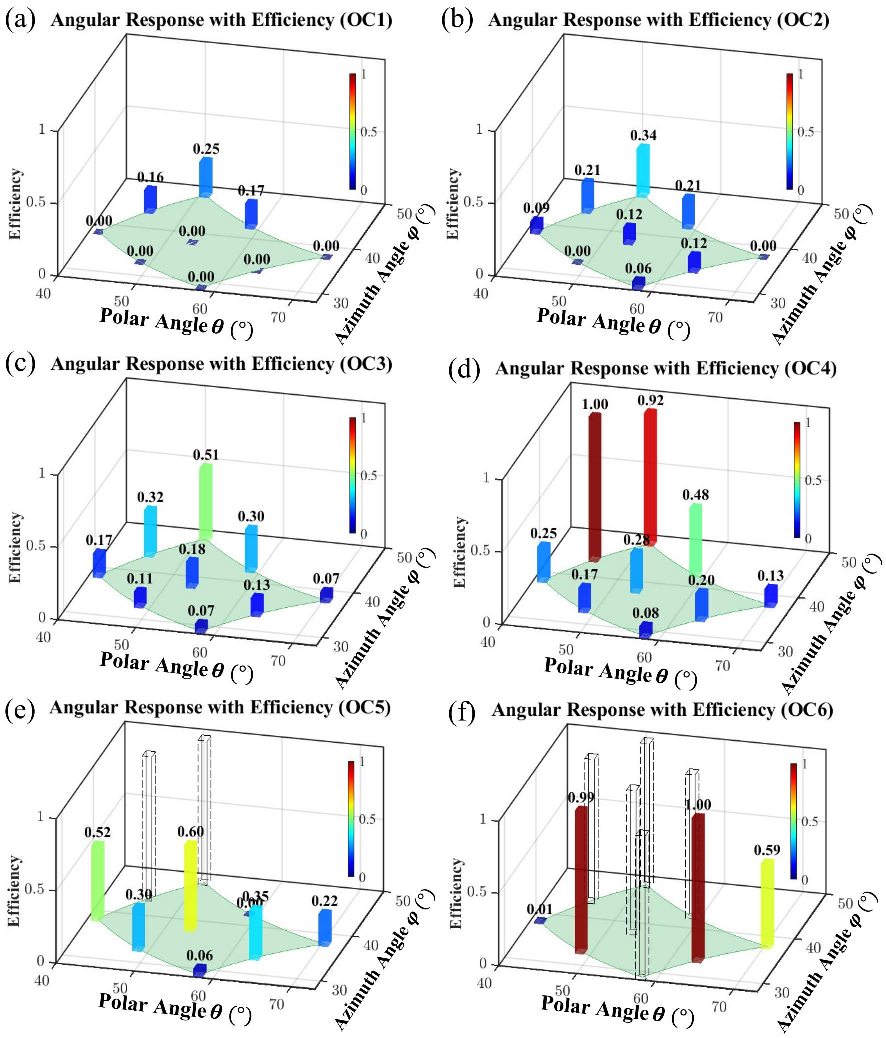

Fig. 12. Optimized +1st order diffraction efficiencies for the out-couplers OC1 to OC6, shown in subplots (a)–(f). Each histogram displays efficiencies for nine FoV angles.

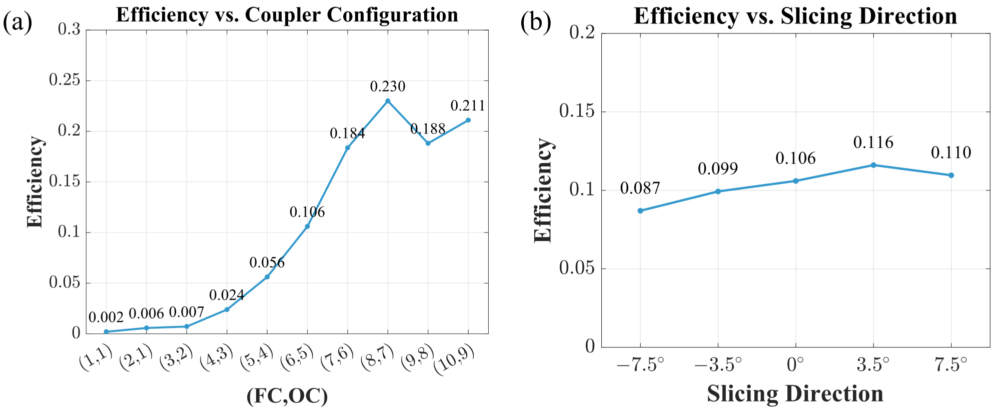

Fig. 13. (a) System efficiency versus the number of folding- and out-coupler slices. (b) Effect of slicing direction on system performance using a (6,5) configuration.

As above-mentioned, during optimization, the diffraction efficiency of each grating is allowed to vary between 0 and 1. According to the optimization results, the efficiency at certain angles approaches 1. While this leads to high optical performance, such a high diffraction efficiency may negatively impact the see-through quality of the AR glasses by reducing the transparency reciprocal diffraction. Additionally, continuously optimizing gratings over a broad angular spectrum at such high efficiency levels is challenging to implement in practice.

Therefore, the above-mentioned results can be considered an upper bound on the system's performance. To evaluate a more practical design, we re-examine the same configuration, using 7 folding-couplers and 6 out-couplers, but this time constrain the grating efficiency range to 0-15%. A minimum uniformity threshold of 80% is still maintained. Under these constraints, the final optical efficiency is reduced to 7.38%, compared to 18.37% in the unconstrained case.

While the current optimization framework uses idealized angular responses as performance targets, the model is structured to allow integration of physically simulated grating efficiencies. Because the beam energy is accumulated afterward through an interaction history vector $[n_1, n_2, \ldots, n_{18}]$, the grating efficiencies can be replaced by those calculated using rigorous methods such as RCWA or the 4×4 matrix method. As a natural next step, the grating structure itself becomes the optimization variable, enabling accurate evaluation of system performance based on realizable coupler designs.

To clarify the differences between conventional optimization approach and our proposed BTM algorithm, a detailed comparison is provided in Table 2. This comparison includes the simulation method, optimization strategy and tools, treatment of coupler responses, and the computational workflow. The calculation time listed corresponds to a single optimization iteration for a representative design.

Optics EXPRESS

Table 2. Comparison between conventional optimization approach and our proposed BTM algorithm

	Conventional approach	BTM
Simulation method	Raytracing	Beam tracing
Optimization tools	Zemax / LightTools	Open-source Python code
Optimization strategy	All FoVs considered simultaneously	Sampled FoVs optimized individually
Optimization workflow	Retracing required after modifying BSDF	Beams stored after tracing for reuse
Coupler response	Requires BSDF generation	Ideal response for each FoV
Calculation Time	~ 1 min	~ 0.42 s

5. Conclusion

In this work, a fast and accurate BTM algorithm is developed to optimize the performance of AR waveguide display systems and derive the target angular response of each coupler. By modeling each beam with its geometric shape, propagation direction, overlap region, and interaction history, the algorithm enables efficient and accurate evaluation of both optical efficiency and uniformity. All effective beams are traced only once and stored, allowing rapid reuse during optimization. Importantly, grating multi-interaction effects, often neglected in conventional simulation tools, are fully incorporated and accelerated using CPU-based parallel computation.

Beyond performance evaluation, this method provides valuable insight into fundamental design questions raised in the introduction. It enables analysis of the maximum efficiency achievable for a given waveguide configuration and reveals the trade-off between efficiency and uniformity. The effects of coupler zoning and slicing direction are also examined, showing that these structural choices significantly affect system performance. Furthermore, we explore scenarios where grating efficiency is limited within specific regions and analyze how these constraints impact overall system performance.

While the current optimization is performed at a single wavelength, the proposed method can be extended to full-color displays with a finite RGB (red, green and blue) spectral bandwidth. This can be achieved by sampling multiple representative wavelengths within the display spectrum. For each wavelength, the effective angular response regions of the folding and out-couplers will differ if a single waveguide is employed. Beam tracing and optimization are then carried out independently for each wavelength, and the result aggregates multiple target angular responses per coupler. This approach enables the algorithm to accommodate RGB display optimization by maintaining wavelength-specific accuracy.

Overall, the proposed method offers a practical, scalable framework for the design and optimization of AR waveguide systems. By combining physical accuracy with high computational efficiency, it serves as a powerful open-source tool for both academic research and industrial development in diffractive AR display technologies.

Funding. AU Optronics.

Acknowledgments. The UCF group is indebted to AUO for partial financial support.

Disclosures. The authors declare no conflicts of interest.

Data Availability. All scripts for waveguide design, beam tracing, and optimization are openly available at [29] The repository will be made publicly accessible upon publication of this paper.

References

1. O. Cakmakci and J. Rolland, "Head-worn displays: a review," J. Display Technol. **2**(3), 199–216 (2006).
2. B. C. Kress and I. Chatterjee, "Waveguide combiners for mixed reality headsets: a nanophotonics design perspective," Nanophotonics **10**(1), 41–74 (2020).
3. J. Xiong, E.-L. Hsiang, Z. He, *et al.*, "Augmented reality and virtual reality displays: emerging technologies and future perspectives," Light:Sci. Appl. **10**(1), 216 (2021).

4. D. Cheng, Q. Wang, Y. Liu, *et al.*, "Design and manufacture AR head-mounted displays: A review and outlook," Light: Advanced Manufacturing **2**(3), 336–369 (2021).
5. A. Bauer and J. P. Rolland, "The Optics of Augmented Reality Displays," in *Springer Handbook of Augmented Reality*, A. Y. C. Nee and S. K. Ong, eds., Springer Handbooks (Springer International Publishing, 2023), pp. 187–209.
6. Y. Ding, Q. Yang, Y. Li, *et al.*, "Waveguide-based augmented reality displays: perspectives and challenges," eLight **3**(1), 24 (2023).
7. S. A. Cholewiak, Z. Başgöze, O. Cakmakci, *et al.*, "A perceptual eyebox for near-eye displays," Opt. Express **28**(25), 38008–38028 (2020).
8. International Committee for Display Metrology, *Information Display Measurements Standard* (Society for Information Display (SID), 2023).
9. Y. Qian, Z. Yang, S.-C. Chen, *et al.*, "Power consumption of light engines for emerging augmented reality glasses: perspectives and challenges," Adv. Photon. **7**(03), 034001 (2025).
10. Y. Amitai, "P-27: A Two-Dimensional Aperture Expander for Ultra-Compact, High-Performance Head-Worn Displays," Dig. Tech. Pap. - Soc. Inf. Disp. Int. Symp. **36**(1), 360–363 (2005).
11. Y. Amitai, "P-21: Extremely Compact High-Performance HMDs Based on Substrate-Guided Optical Element," Dig. Tech. Pap. - Soc. Inf. Disp. Int. Symp. **35**(1), 310–313 (2004).
12. S. J. Walker and J. Jahns, "Optical clock distribution using integrated free-space optics," Opt. Commun. **90**(4-6), 359–371 (1992).
13. J.-R. Yan, Q.-H. Wang, D.-H. Li, *et al.*, "Edge-Lighting Light Guide Plate Based on Micro-Prism for Liquid Crystal Display," J. Display Technol. **5**(9), 355–357 (2009).
14. D. Ni, D. Cheng, Y. Wang, *et al.*, "Design and fabrication method of holographic waveguide near-eye display with 2D eye box expansion," Opt. Express **31**(7), 11019–11040 (2023).
15. Y. Weng, Y. Zhang, W. Wang, *et al.*, "High-efficiency and compact two-dimensional exit pupil expansion design for diffractive waveguide based on polarization volume grating," Opt. Express **31**(4), 6601–6614 (2023).
16. T. Levola and P. Laakkonen, "Replicated slanted gratings with a high refractive index material for in and outcoupling of light," Opt. Express **15**(5), 2067–2074 (2007).
17. M. G. Moharam and T. K. Gaylord, "Diffraction analysis of dielectric surface-relief gratings," J. Opt. Soc. Am. **72**(10), 1385–1392 (1982).
18. J. M. Miller, N. D. Beaucoudrey, P. Chavel, *et al.*, "Design and fabrication of binary slanted surface-relief gratings for a planar optical interconnection," Appl. Opt. **36**(23), 5717–5727 (1997).
19. Y. Weng, D. Xu, Y. Zhang, *et al.*, "Polarization volume grating with high efficiency and large diffraction angle," Opt. Express **24**(16), 17746–17759 (2016).
20. Y.-H. Lee, K. Yin, and S.-T. Wu, "Reflective polarization volume gratings for high efficiency waveguide-coupling augmented reality displays," Opt. Express **25**(22), 27008–27014 (2017).
21. Y. Weng, Y. Zhang, J. Cui, *et al.*, "Liquid-crystal-based polarization volume grating applied for full-color waveguide displays," Opt. Lett. **43**(23), 5773–5776 (2018).
22. N. V. Tabiryan, D. E. Roberts, Z. Liao, *et al.*, "Advances in Transparent Planar Optics: Enabling Large Aperture, Ultrathin Lenses," Adv. Opt. Mater. **9**(5), 2001692 (2021).
23. Y. Ding, Y. Gu, Q. Yang, *et al.*, "Breaking the in-coupling efficiency limit in waveguide-based AR displays with polarization volume gratings," Light:Sci. Appl. **13**(1), 185 (2024).
24. J. P. Rolland and J. Goodsell, "Waveguide-based augmented reality displays: a highlight," Light:Sci. Appl. **13**(1), 22 (2024).
25. J. Goodsell, P. Xiong, D. K. Nikolov, *et al.*, "Metagrating meets the geometry-based efficiency limit for AR waveguide in-couplers," Opt. Express **31**(3), 4599–4614 (2023).
26. J. Goodsell, D. K. Nikolov, A. N. Vamivakas, *et al.*, "Beam interaction and targeted optimization methods for AR waveguide design," Opt. Express **33**(10), 21999–22018 (2025).
27. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," J. Opt. Soc. Am. **71**(7), 811–818 (1981).
28. J. Xiong and S. T. Wu, "Rigorous coupled-wave analyses of liquid crystal polarization gratings," Opt. Express **28**(24), 35960–35971 (2020).
29. Y. Zhang and Y. Ding, "beam_tracing_method Toolkit for AR Waveguides," Github, 2025, https://github.com/yefuzhang/beam_tracing_method.