LENSDREAMER: EXPLORING OPTICAL DESIGN BASINS VIA WORLD-MODEL IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic lens design seeks optical systems that satisfy target imaging specifications in a highly non-convex merit landscape. Existing methods generally explore this landscape through repeated high-fidelity simulator evaluations, so broad search consumes a substantial physical-evaluation budget. We instead use a learned world model to imagine the consequences of design actions, explore more candidate paths, and identify promising routes out of inferior local minima. We introduce LENSDREAMER, an action-conditioned optical world model that encodes layouts, lens prescriptions, design requirements, and physical observations into a multimodal state and predicts its evolution under candidate actions. Its Basin-Map Planner organizes imagined endpoints into candidate regions and selects action sequences for first-action physical verification, after which retained verified states seed local refinement. We further construct LensDreamer-Traj, a trajectory dataset for world-model training, and LensDreamer-2K, a benchmark for automatic lens design. Experiments demonstrate the potential of world-model imagination for navigating non-convex optical design landscapes. More broadly, we hope this work sheds light on applying world models to automatic lens design and other non-convex scientific design problems. Codes and datasets will be released upon acceptance.

1 INTRODUCTION

Automatic lens design uses computational optimization to translate imaging requirements into high-quality lens prescriptions with less manual trial and error (Meiron, 1959; Kingslake & Johnson, 2009). Given a target specification, it balances imaging performance, physical constraints, and other practical requirements. This requires jointly adjusting coupled variables such as surface curvature, element spacing, and glass type. Their interactions create a highly non-convex merit landscape where different initial prescriptions and optimization paths lead to different local optima (Turnhout & Bociort, 2009). Although the ideal objective is a global optimum, it generally cannot be certified in this landscape. The practical goal is therefore to escape inferior local minima and discover better physically feasible optima in more promising attraction basins.

Existing automatic lens-design systems broaden search through specialized optimization. DeepLens progressively increases aperture and field of view from nearly planar surfaces, whereas QGSO combines simulated annealing, genetic selection, and ADAM to produce diverse initial systems (Yang et al., 2024; Gao et al., 2025). Differentiable-optics and learned-initialization methods also reduce manual intervention (Sun et al., 2021; Wang et al., 2022; Côté et al., 2023). However, broad exploration requires many high-fidelity simulator evaluations. Pointwise surrogates estimate candidate quality without full simulation (Sheng et al., 2022), but do not model how current decisions shape later optimization. Reducing physical evaluations enables broader search, while multi-step reasoning can identify sequences that leave the current attraction basin.

World models offer a complementary use of the physical budget. PlaNet and Dreamer learn action-conditioned latent dynamics and imagine trajectories for planning or behavior learning (Ha & Schmidhuber, 2018; Hafner et al., 2019b;a; 2025; Lin et al., 2023; Burchi & Timofte, 2025; Morihira et al., 2026). Applied to automatic lens design, a model can encode a physically evaluated lens, roll out thousands of action sequences, predict their outcomes, and return only a small shortlist to the simulator. High-fidelity simulation remains necessary to verify these predictions, but its role shifts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from exhaustively scoring the search space to selectively validating and refining promising regions. A broader review of automatic lens design and world-model planning is provided in the appendix.

Realizing this idea introduces three challenges. First, *heterogeneous state modeling* is necessary because human designers rely on layouts, prescriptions, specifications, and optical observations to make design decisions. A world model must combine these inputs into a coherent state. Second, *long-horizon prediction* is difficult because errors accumulate over multi-step rollouts and can produce physically misleading endpoints. Third, *escaping local minima* requires more than large-scale imagination because a planner may repeatedly select similar endpoints from one promising region. The system must preserve endpoint diversity and verify different regions to move beyond the current basin and discover better minima.

We address these challenges with LENSDREAMER, an optical world model coupled with a Basin-Map Planner. Its Multimodal Optical State Encoder uses a shared DINOv3 encoder (Siméoni et al., 2025) to extract patch-level tokens from 2D and 3D lens layouts. Qwen3-VL (Bai et al., 2025) fuses them with the prescription, design specification, physical observations, and legal action domain into a unified optical state. Inspired by Transformer-based world models (Micheli et al., 2023; Chen et al., 2022), our hybrid Transformer–RSSM predicts how this state evolves under candidate actions, including future prescriptions and optical properties. During planning, LENSDREAMER imagines many action sequences without repeatedly calling the simulator. The Basin-Map Planner groups imagined endpoints into candidate regions and selects diverse sequences for physical verification. The evaluator executes only the first action of each sequence. Verified outcomes are re-encoded and pruned with incumbents to form the next physical population, while retained designs may seed local refinement.

We train the world model on LensDreamer-Traj and evaluate final designs on all 2,000 LensDreamer-2K tasks. We first compare final design quality with matched-budget physics-only baselines and external lens-design systems under common physical reevaluation. We then test whether Basin-Map broadens verified coverage and yields distinct DLS-stable endpoint groups. Finally, H1–H8 open-loop evaluation and component ablations assess the learned dynamics, state encoder, and planning strategies. In summary, we make the following contributions:

- We formulate automatic lens design as action-conditioned world modeling and introduce LENSDREAMER, to our knowledge the first optical world model for this task. It encodes multimodal lens information and represents lens design as state transitions conditioned on target specifications and design actions, enabling large-scale multi-step imagination of design trajectories.
- We propose the Basin-Map Planner, which organizes imagined endpoints into candidate regions and selects diverse action sequences for first-action physical verification. Retained verified states seed local refinement, broadening search and producing more distinct DLS-stable endpoints than the evaluated alternatives.
- We collect LensDreamer-Traj for learning multi-step optical dynamics and introduce LensDreamer-2K as a benchmark for this new task, where LENSDREAMER demonstrates strong performance in both optical dynamics prediction and final automatic lens design.

2 LENSDREAMER

2.1 PROBLEM FORMULATION

Let $S = (\mathrm{EFL}^*, (F/\#)_{\max}, D, [\lambda_{\min}, \lambda_{\max}], \{(\lambda_i, \alpha_i)\}_{i=1}^{N_\lambda}, \mathrm{MTF}^{\min}, \mathrm{RMS}^*)$ denote the target specification. Here, the components represent the target effective focal length, maximum allowed F-number, detector size, working spectral range, sampled wavelength–weight pairs, task-defined minimum MTF requirements, and target RMS spot radius, respectively. More details are provided in the appendix. Given S, automatic lens design seeks a physically valid prescription $p \in \mathcal{X}_{\mathrm{feas}}(S)$ whose optical response satisfies these requirements. A high-fidelity evaluator $\mathcal{P}$ maps a prescription to its optical response, and we summarize its normalized deviation from the target using the Optical Specification Gap ($\mathrm{SpecGAP}$),

$$Y(p; S) = \mathcal{P}(p; S), \qquad \mathrm{SpecGAP}(p; S) = \mathrm{SpecGAP}\big(Y(p; S), S\big). \tag{1}$$

Here, Y contains the task-relevant optical measurements. To retain invalid prescriptions in the comparison, we use the Failure-Aware Specification Gap ($\mathrm{SpecGAP}_{\mathrm{FA}}$), which assigns them a fixed penalty. These metrics are used in Section 3, and more details are provided in the appendix.

This optimization is highly *non-convex* because ray propagation depends nonlinearly on strongly coupled geometric and material variables, while physical constraints create additional invalid and nonsmooth regions (Kingslake & Johnson, 2009). Consequently, different starting prescriptions may fall into different attraction basins and lead a local optimizer to *different local minima* (Turnhout & Bociort, 2009). Although the ideal objective is the global optimum, it generally cannot be certified. The practical objective is therefore to *explore multiple basins and find a physically feasible local optimum with a smaller* SpecGAP than the one reached from the current start.

In budget-constrained optical design, let $\mathcal{E}_B = ((p_i, Y_i))_{i=1}^{n_B}$ denote the records from $n_B \le B$ calls to the high-fidelity evaluator $\mathcal{P}$, and let $\mathcal{V}_B = \{p_i\}_{i=1}^{n_B}$ be the evaluated prescriptions. We seek the best verified feasible design,

$$p_B^* = \arg \min_{p \in \mathcal{V}_B \cap \mathcal{X}_{\mathrm{feas}}(S)} \mathrm{SpecGAP}(p; S), \qquad n_B \le B. \tag{2}$$

A local optimizer improves a prescription within its current basin, whereas broader exploration requires testing alternative design paths that may reach different basins.

However, an attraction basin cannot be selected directly. Starting from the current prescription, a designer can reach a different region only through a sequence of valid modifications. Each modification changes both the current lens and the designs that remain reachable in subsequent steps. We therefore formulate basin exploration as *action-conditioned state transitions*. At step t, the physical observation $o_t = (I_t^{2D}, I_t^{3D}, p_t, Y_t)$ contains the rendered layouts, current prescription, and measured optical response, conditioned on S. An action a_t selects an allowed operation, target, and magnitude, producing

$$p_{t+1} = \mathcal{A}(p_t, a_t), \qquad Y_{t+1} = \mathcal{P}(p_{t+1}; S). \tag{3}$$

Evaluating such multi-step paths with $\mathcal{P}$ is expensive. This motivates LENSDREAMER, which imagines many candidate paths, verifies selected first actions, and refines retained designs.

2.2 OVERVIEW

Figure 1 summarizes how LENSDREAMER alternates broad exploration through learned imagination with selective verification and refinement through high-fidelity physics. Each task begins with one of Optiland's built-in sample lens designs, which is rescaled to the requested focal length and aperture without further optimization. Before learned exploration begins, the initialized lens is physically evaluated and rendered. The resulting layouts, prescription, design specification, and optical observations are then combined by the Multimodal Optical State Encoder to establish the initial state of the planning process.

Using this encoded state, LENSDREAMER enters an external model predictive control (MPC) loop. At round t, let $\mathcal{V}_t \subseteq \mathcal{V}_B$ denote the retained population of at most M physically verified lenses, with $\mathcal{V}_0$ containing the initialized lens. The action planner allocates N candidate sequences of horizon H across $\mathcal{V}_t$, where $\tau_t^{(n)} = (a_{t,1}^{(n)}, \ldots, a_{t,H}^{(n)})$ for $n = 1, \ldots, N$. Rather than sending every $\tau_t^{(n)}$ to the physical evaluator, the Transformer–RSSM predicts how each sequence changes the lens and its optical response. The Basin-Map Planner organizes the predicted endpoints into candidate regions and selects a representative set $\mathcal{R}_t \subseteq \{\tau_t^{(n)}\}_{n=1}^{N}$, where $|\mathcal{R}_t| \le K$. For each $\tau_t^{(n)} \in \mathcal{R}_t$, the physical evaluator executes and verifies only its first action $a_{t,1}^{(n)}$. The resulting lenses are rendered and encoded again, after which the retained incumbents and verified outcomes are pruned to at most M lenses to form $\mathcal{V}_{t+1}$. Because the next round starts from $\mathcal{V}_{t+1}$, the imagined latent states are discarded rather than propagated across rounds. Repeating this process builds an archive of verified states associated with different candidate regions. Selected states eventually seed local DLS refinement, after which the best physically evaluated design is returned.

Section 2.3 describes multimodal state encoding, action-prefix modeling, recurrent stochastic dynamics, and structured optical readout. Section 2.4 presents imagined basin exploration and physical budget allocation. Section 2.5 describes the physics-supervised trajectories and multi-step learning objective.

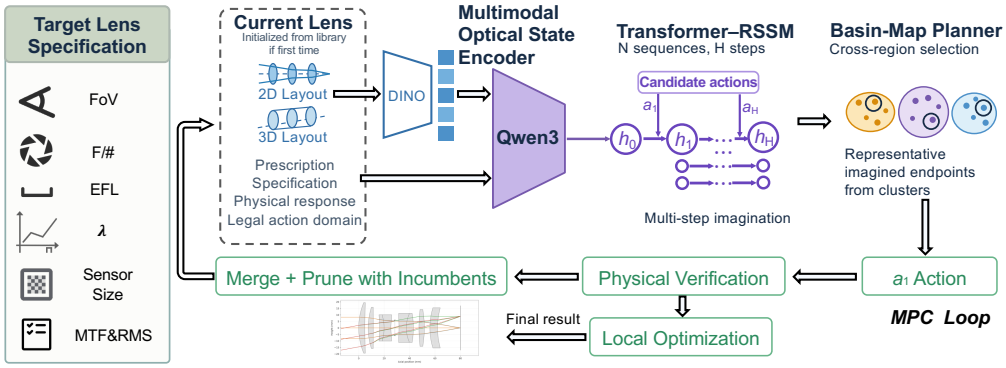

Figure 1: **Overview of LENSDREAMER.** Given a target specification, the pipeline initializes a lens from Optiland's sample library and encodes its multimodal optical state. A model predictive control (MPC) loop then alternates world-model imagination, Basin-Map Planner, and physical feedback before local DLS refinement produces the final design.

2.3 ACTION-CONDITIONED OPTICAL WORLD MODEL

Multimodal Optical State Encoder. Human optical designers make decisions by jointly considering the 2D and 3D lens geometry (I_t^{2D}, I_t^{3D}), current prescription p_t, target specification S, and measured optical response Y_t. A scalar merit value alone cannot preserve this complementary information. A shared DINOv3 backbone E_D (Siméoni et al., 2025) encodes the two layouts into patch-level visual tokens V_t^{2D} and V_t^{3D}, which receive view identifiers before entering Qwen3-VL. We serialize p_t, S, Y_t, and the legal action domain into a structured optical record T_t^{opt} with explicit fields, units, indices, and validity masks. Qwen3-VL E_Q (Bai et al., 2025) directly fuses this record with the visual tokens in a common attention space,

$$V_t^v = E_D(I_t^v), \quad v \in \{2D, 3D\}, \qquad h_0 = \Pi_Q\Big(E_Q\big([V_t^{2D}, V_t^{3D}, T_t^{\text{opt}}, e_{\text{state}}]\big)_{e_{\text{state}}}\Big). \tag{4}$$

Here, e_{state} is an optical-state aggregation token, and Π_Q only maps the Qwen hidden dimension to the world-model state dimension. The resulting h_0 initializes the Transformer–RSSM and supports all imagined trajectories originating from the current verified lens. In our implementation, DINOv3 and Qwen3-VL serve as the visual and multimodal fusion backbones rather than separate prescription, specification, or observation encoders. More details are provided in the appendix.

Action-conditioned Transformer–RSSM dynamics. The effect of a design action depends on the modifications that precede it, while multi-step imagination requires an optical state that evolves after every action. We therefore combine a causal Transformer for action-history modeling with a recurrent state-space model (RSSM) for stochastic latent-state propagation (Hafner et al., 2019b). Following Transformer-based world models (Micheli et al., 2023; Chen et al., 2022), our Transformer conditions the recurrent transition rather than replacing it as in TransDreamer's TSSM.

We describe one candidate rollout within a single MPC round and use k to index its imagined steps. The action encoder, causal Transformer, and recurrent transition are summarized as

$$u_k = E_u(a_k, S), \tag{5}$$

$$c_k = T_{\text{causal}}(h_0, u_{1:k}), \tag{6}$$

$$(h_k, z_k) = F_{\text{RSSM}}(h_{k-1}, c_k, u_k). \tag{7}$$

Here, u_k represents the current action and specification, while c_k summarizes the current and preceding action tokens. The causal mask prevents c_k from accessing future actions, and F_{RSSM} recurrently updates the optical state. During training, z_k is sampled from a posterior conditioned on the observed response features, and a KL objective aligns this posterior with the learned prior. During planning, the next response is unavailable, so the model deterministically uses the prior mean. The stochastic state is incorporated into h_k before the next transition. Consequently, the model can roll out an entire action sequence from h_0 without rendering or re-encoding future layouts.

Structured future-state readout. The latent state alone does not expose the physical quantities needed to compare imagined action sequences. We therefore attach separate prediction heads to

each h_k to decode the future prescription, task-relevant optical responses, feasibility, and predictive uncertainty signals used as auxiliary planner inputs. These structured predictions allow the planner to evaluate how an action sequence changes both the lens and its physical performance instead of reducing its outcome to a single merit value. More details are provided in the appendix.

2.4 World-model Basin Exploration

Broad trajectory imagination. Following the idea of reusing earlier solutions to guide related search problems (Schmidhuber, 2004), the retrieved trajectory provides only a soft proposal prior over actions, while rollouts remain anchored at current physically verified parent lenses and selected transitions are independently checked by the high-fidelity evaluator. Within each external MPC round t, the planner allocates the N candidate sequences $\{\tau_t^{(n)}\}_{n=1}^{N}$ across the verified parent lenses in $\mathcal{V}_t$. Each sequence has horizon H and is sampled from the task-specific action domain. The Transformer–RSSM rolls out these sequences without intermediate physical evaluation, forming the internal imagination loop. After the Basin-Map Planner selects $\mathcal{R}_t$, the physical evaluator applies only $a_{t,1}^{(n)}$ from each $\tau_t^{(n)} \in \mathcal{R}_t$ to its verified parent. The resulting physical states are merged with the retained incumbents and pruned to form $\mathcal{V}_{t+1}$, whose lenses are then rerendered and re-encoded. Thus, each MPC round contains many H-step latent rollouts, while every selected physical branch advances by one verified action before replanning.

Basin-map construction. A large imagination budget does not by itself broaden physical exploration because selecting only the K sequences with the smallest predicted $\mathrm{SpecGAP}$ can concentrate all verification slots in one region. For each imagined sequence $\tau_t^{(n)}$, we form a terminal representation

$$e_t^{(n)} = \mathrm{concat}\Big(h_{t,H}^{(n)}, \hat{p}_{t,H}^{(n)}, \hat{Y}_{t,H}^{(n)}\Big), \qquad \mathcal{C}_t = \mathrm{Cluster}\Big(\{e_t^{(n)}\}_{n=1}^{N}\Big), \tag{8}$$

where $h_{t,H}^{(n)}$, $\hat{p}_{t,H}^{(n)}$, and $\hat{Y}_{t,H}^{(n)}$ denote the terminal latent embedding, predicted prescription, and predicted optical response. The resulting candidate regions $\mathcal{C}_t$ are tracked across imagination, selection, and first-action verification so that near-duplicate endpoints can be suppressed. The planner selects representatives from different regions according to predicted quality, feasibility, an auxiliary predictive uncertainty signal, and novelty. It then globally merges the representatives proposed from all parent lenses into $\mathcal{R}_t$, where $|\mathcal{R}_t| \le K$. More details are provided in the appendix. Allocating the fixed verification slate across candidate regions converts a large imagined set into broader physical coverage and creates more opportunities to leave the current local minimum. These candidate regions lie in imagined state space and are not mathematical attraction basins. We use *certified stable basin* operationally only when associated verified states converge to distinct stable endpoint groups under the same fixed local DLS procedure, without claiming a complete decomposition of the objective landscape.

Physical verification and MPC feedback. Imagined terminal states guide selection but never enter the physical population directly. Let $\pi(n)$ identify the verified parent of $\tau_t^{(n)} \in \mathcal{R}_t$. The high-fidelity transition is

$$p_{t+1}^{(n)} = \mathcal{A}\Big(p_t^{\pi(n)}, a_{t,1}^{(n)}\Big), \qquad Y_{t+1}^{(n)} = \mathcal{P}\Big(p_{t+1}^{(n)}; S\Big). \tag{9}$$

Only these verified outcomes and the retained incumbents are eligible for the next population. We preserve the best feasible designs and prune the remaining outcomes for diversity so that $|\mathcal{V}_{t+1}| \le M$. Every retained lens is then rerendered and passed through the Multimodal Optical State Encoder before the next MPC round. Consequently, neither the imagined terminal prediction nor its recurrent state crosses the physical verification boundary, which limits the accumulation of open-loop model error across rounds.

Cross-round coverage and basin verification. Region-aware selection must persist across MPC rounds because replanning can otherwise return repeatedly to the same promising area. A novelty archive favors endpoints far from previously verified designs, while annealed acceptance permits temporary uphill moves that may cross barriers between regions. When exploration stagnates, the planner enlarges the action scale. More details are provided in the appendix. Meanwhile, the best verified feasible design is preserved independently as

$$p_t^{\mathrm{best}} = \arg \min_{p \in (\bigcup_{i=0}^{t} \mathcal{V}_i) \cap \mathcal{X}_{\mathrm{feas}}(S)} \mathrm{SpecGAP}(p; S), \tag{10}$$

so exploratory acceptance cannot overwrite the best solution found so far. Finally, selected verified states seed this fixed DLS procedure, and their endpoints are grouped by prescription and physical response to determine certified stable basins.

2.5 LensDreamer-Traj and Multi-Step Dynamics Learning

Physics-supervised trajectories. We collect LensDreamer-Traj by applying mixed discrete–continuous design actions to Optiland systems spanning multiple lens families and randomized target specifications and recording the resulting state–action–next-state trajectories. Its training split contains 13,696 trajectories with eight state transitions each, totaling 109,568 high-fidelity transitions for world-model training. A training window consists of an initial multimodal observation o_0, a specification S, and a sequence of physically evaluated transitions $\{(a_k, p_{k+1}, Y_{k+1})\}$. Complete trajectories are kept disjoint across data splits, and contiguous windows provide the supervision needed to learn state transitions that can be rolled forward over multiple actions.

Multi-step learning objective. For each training window, the model is recursively unrolled and supervised at every valid future step without re-encoding an intermediate physical state. The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{MTF}} + \lambda_1 \mathcal{L}_{\mathrm{first}} + \lambda_2 \mathcal{L}_{\mathrm{feas}} + \lambda_3 \mathcal{L}_{\mathrm{surface}} + \lambda_4 \mathcal{L}_{\mathrm{aux}} + \lambda_5 \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{cons}}. \tag{11}$$

The first six terms supervise the predicted MTF, first-order responses, feasibility, surface prescription, auxiliary optical responses, and RSSM prior–posterior alignment. The consistency term ties direct optical readouts to the predicted prescription. Invalid or padded targets are masked from their corresponding losses. Loss weights and optimization settings are provided in the appendix.

3 Experiments

3.1 Experimental Setup

Datasets and splits. We use LensDreamer-Traj to train the optical world model. LensDreamer-2K contains 2,000 evaluation tasks, including 500 in-distribution and 1,500 specification-OOD tasks. A task is specification-OOD when at least one target, such as EFL, the maximum F-number, RMS, or FoV determined by D and EFL, falls outside the range covered by LensDreamer-Traj. This task-level definition reflects coupled lens design because moving one important specification beyond the familiar regime can substantially change design difficulty, as exemplified by aperture and FoV (Kingslake & Johnson, 2009; Yang et al., 2024).

Common evaluation protocol. All final-design and planning experiments use all 2,000 tasks in LensDreamer-2K. We import every final prescription into the same version of Optiland and recompute its physical metrics with identical fields, wavelengths, focal-plane treatment, and validity criteria. The separately reported axis and corner MTF values are evaluated at $100\,\mathrm{lp/mm}$. Controlled comparisons use the same initial prescriptions, legal design variables, random seeds, and high-fidelity Optiland call budget. They control physical evaluations rather than total computational cost, and LensDreamer uses additional model-side computation to improve design quality under this budget. DeepLens and QGSO retain their native procedures and are compared only by final design quality.

Implementation details. Our implementation uses a frozen DINOv3 ViT-L/16 for the 2D/3D layouts and a LoRA-adapted Qwen3-VL-32B-Instruct as the multimodal fusion backbone, producing a 1,024-dimensional state. We train the Transformer–RSSM on eight-step windows for 12 epochs using AdamW (Loshchilov & Hutter, 2017). At inference, each MPC round searches a population of 4,096 length-eight action sequences and physically verifies eight selected first actions, with at most eight rounds and 180 high-fidelity Optiland objective calls per task. The physical budget includes final refinement through four local DLS restarts. Planning evaluates MTF at 50, 100, 200, and $300\,\mathrm{lp/mm}$, while final evaluation uses 256 rays over four fields. Training takes 50 hours on four NVIDIA B200 GPUs. More details are provided in the appendix.

3.2 Automatic Lens Design Comparison on LensDreamer-2K Benchmark

Compared methods. We compare LensDreamer with two groups of methods. The matched-budget baselines include Optiland single-start, Optiland multi-start, physical random search, and the

Table 1: **Comparison on our LensDreamer-2K benchmark.** All final prescriptions are recomputed using the common Optiland evaluator for fair comparison.

Method	Feasible ↑	$\mathrm{SpecGAP}_{\mathrm{FA}}$ ↓	Feasible SpecGAP ↓	RMS (μm) ↓	Axis MTF ↑	Corner MTF ↑	|Distortion| ↓
[Matched-Budget Baselines]							
Optiland single-start	72.20%	3.48	0.97	209.83	4.73%	1.42%	0.37
Optiland multi-start	79.35%	2.83	0.96	104.61	4.71%	1.21%	0.44
Physical random search	77.35%	3.01	0.96	452.93	6.13%	1.43%	0.24
Physical CEM	76.75%	3.06	0.96	332.17	12.23%	1.22%	0.32
[Existing Methods]							
DeepLens (Yang et al., 2024)	68.35%	3.91	1.09	201.31	3.21%	0.91%	0.26
QGSO (Gao et al., 2025)	46.45%	5.81	0.99	378.13	1.83%	1.01%	0.31
LensDreamer (Ours)	85.20%	1.89	0.48	95.67	13.27%	2.37%	0.14

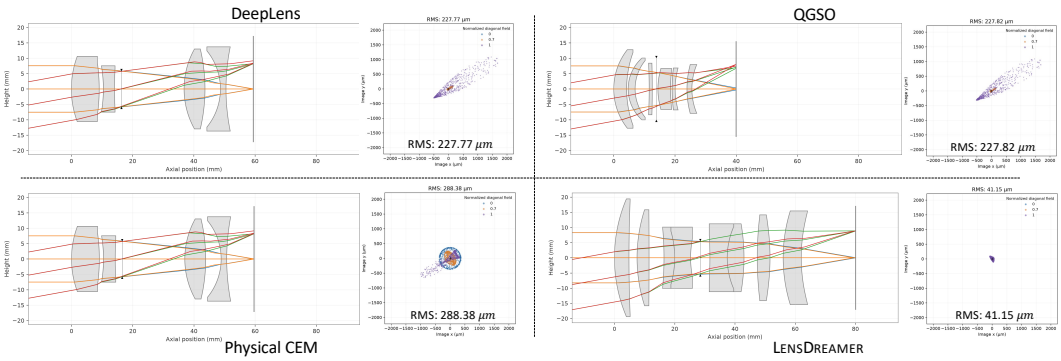

Figure 2: **Visualization of lens layouts and spot diagrams on the LensDreamer-2K benchmark.** The compared methods produce different optical structures, while LensDreamer achieves the smallest RMS spot radius in this example, indicating the best imaging quality among them.

physical cross-entropy method (CEM) (Rubinstein, 1999). Optiland single-start directly applies the built-in lens-optimization routine provided by Optiland to one initial prescription, while Optiland multi-start repeats the same native optimizer from multiple initial prescriptions. All matched-budget baselines use the same initial prescriptions, legal design variables, local DLS refiner, and physical-evaluation budget as our method for fair comparison. We also compare final design quality with existing automatic lens design methods DeepLens (Yang et al., 2024) and QGSO (Gao et al., 2025).

Evaluation metrics. Feasible rate is the fraction of tasks with a physically valid final design. SpecGAP summarizes task-level specification agreement, while $\mathrm{SpecGAP}_{\mathrm{FA}}$ additionally penalizes invalid outputs. Common physical reevaluation also reports direct optical quantities, with RMS and axis and corner MTF measuring imaging performance and distortion measuring geometric image deformation. More details are provided in the appendix.

As shown in Table 1, LensDreamer achieves the highest feasible rate and the lowest specification gaps. Compared with the strongest matched-budget results, it increases feasibility from 79.35% to 85.20%, reduces $\mathrm{SpecGAP}_{\mathrm{FA}}$ from 2.83 to 1.89, and reduces feasible SpecGAP from 0.96 to 0.48. It also improves imaging performance, obtaining the lowest RMS spot radius and the highest axis and corner MTF, while producing the lowest distortion. After common physical reevaluation, LensDreamer surpasses the state-of-the-art automatic lens-design methods DeepLens and QGSO across all reported metrics. Figure 2 shows a representative example in which LensDreamer also achieves the smallest RMS spot radius, visually supporting its stronger imaging performance.

3.3 Basin Exploration Analysis

Compared selection strategies. Scalar top-k/CEM selects candidates by predicted quality, while random diverse endpoints tests whether diversity alone is sufficient. Basin clustering only isolates the effect of organizing terminal states into regions. We further remove the novelty archive, annealed acceptance, and triggered jump individually from the full Basin-Map Planner.

Table 2: **Comparison of basin-exploration strategies under matched budgets.**

Selection strategy	Verified clusters ↑	Certified stable basins ↑	Anchor escape ↑	Best refined SpecGAP ↓	$\text{SpecGAP}_{\text{FA}}$ ↓	Beat physics-only ↑
Scalar top-k/CEM	N/A	1.14	31.62%	1.07	2.57	35.20%
Random diverse endpoints	N/A	1.07	25.47%	1.11	2.73	38.35%
Basin clustering only	4.36	1.27	41.83%	0.91	2.49	49.30%
Full *w/o* novelty archive	3.94	1.29	38.76%	0.97	2.41	47.40%
Full *w/o* annealed acceptance	4.21	1.37	48.29%	0.82	2.29	56.75%
Full *w/o* triggered jump	4.43	1.42	53.14%	0.76	2.17	62.15%
Full Basin-Map Planner	5.24	1.63	62.47%	0.48	1.89	73.25%

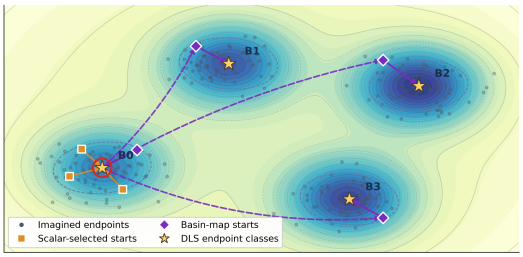

Figure 3: **Basin-Map selection on a real optimization task.** From the same imagined candidates, scalar top-k concentrates DLS starting points near $B0$, whereas Basin-Map selects starting points across multiple candidate regions. Applying the same local DLS procedure then produces the marked endpoint classes.

Evaluation metrics. Verified clusters count imagined regions selected for physical verification, while certified stable basins count distinct endpoints obtained after common local refinement. Anchor escape and Beat physics-only measure whether a strategy reaches a better region than its initial anchor and the matched-budget physics-only search, respectively. Best refined SpecGAP is the lowest specification gap after local refinement, while $\text{SpecGAP}_{\text{FA}}$ additionally penalizes invalid outputs. More details are provided in the appendix.

As shown in Table 2, the full Basin-Map Planner provides the broadest coverage, reaching 5.24 verified clusters and 1.63 certified stable basins. Figure 3 illustrates this behavior in one actual optimization instance by contrasting the local-refinement starting points selected by scalar top-k and Basin-Map from the same imagined candidates and the endpoint classes reached after applying the same DLS procedure. Its 62.47% anchor-escape rate shows that this broader coverage more often reaches lower-SpecGAP DLS-stable endpoint groups beyond the initial anchor. It yields the lowest best-refined SpecGAP and $\text{SpecGAP}_{\text{FA}}$ and beats matched-budget physics-only search on 73.25% of tasks. Basin clustering outperforms scalar and random selection, while removing cross-round exploration mechanisms degrades coverage and final quality.

Table 3: **Ablation of planning strategies under matched budgets.**

Planning Strategy	$\text{SpecGAP}_{\text{FA}}$ ↓	Certified stable basins ↑	Regret ($\times 10^{-3}$) ↓
Random Valid Rollout	2.83	1.07	27.43
Same-Pool Random Verification	2.61	1.12	24.67
Greedy One-Step Planning	2.27	1.27	13.83
Full Basin-Map Planning (Ours)	1.89	1.63	4.17

3.4 Ablation Studies

World-model and component ablations. We evaluate the predictive fidelity of the world model and the contribution of its main components. We compare LensDreamer with persistence, which copies the current response, a linear State–Action–Specification Ridge model, a recursively rolled-out One-Step Dynamics Model, and the earlier DINO-L + GRU model. All methods start from the same state and actions. H1 measures one-step prediction, while later horizons run open loop without physical re-encoding. Mean absolute error (MAE) measures the deviation between predicted and physically evaluated optical quantities, with prescription MAE computed over normalized surface parameters. Feasibility area under the ROC curve (AUROC) measures how well the model distinguishes valid from invalid future states. Figure 4 shows that LensDreamer maintains lower structured-state errors as the horizon increases and reduces H8 MTF error by 18.7% relative to the second-best model, showing that the learned dynamics remain informative over H1–H8 open-loop rollouts. We then ablate the Transformer and RSSM and test encoders that retain one layout, replace DINOv3 with ViT, or remove visual input. All variants use the same data and training schedule. Figure 5 shows that the full model reduces H8 MTF error by 22.2% relative to the strongest ablated variant and maintains the best overall prediction fidelity. These results support the joint Transformer–RSSM dynamics and multimodal state encoder.

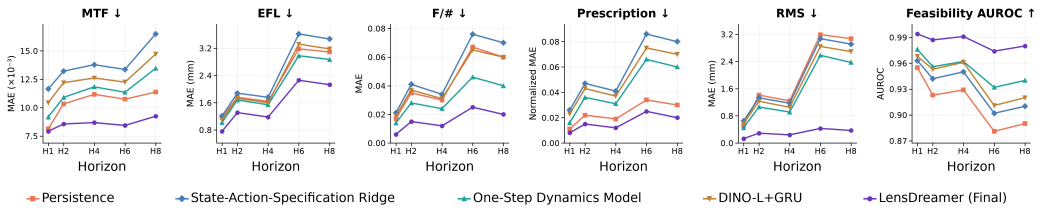

Figure 4: **Open-loop world-model fidelity on held-out LensDreamer-Traj trajectories.** All methods start from the same measured state and receive the same action sequence. After H1, predictions proceed without teacher-forced observations or physical re-encoding.

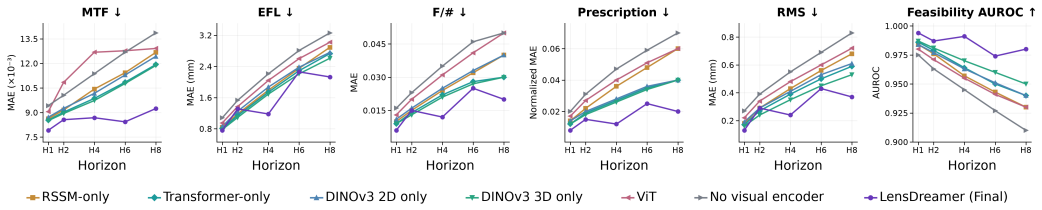

Figure 5: **Ablation of the dynamics model and optical state encoder.** All dynamics-model and encoder variants are trained and evaluated under the same protocol. Exact values and pairwise action-ordering results are provided in the appendix.

Planning-strategy ablations. We isolate the planner under matched budgets. Random Valid Rollout samples legal sequences without model-based ranking, while Same-Pool Random Verification uses the same candidate pool but verifies candidates randomly. Greedy One-Step Planning ranks only the immediate outcome without basin-aware allocation, whereas Full Basin-Map Planning evaluates multi-step futures and distributes verification across candidate regions. Selection regret is the physical $\mathrm{SpecGAP}$ difference between the best selected candidate and the best candidate in the shared pool. Table 3 shows that the full planner discovers 28.3% more certified stable basins than the second-best strategy and obtains the lowest $\mathrm{SpecGAP}_{\mathrm{FA}}$ and selection regret. The same-pool comparison supports the value of basin-aware selection independently of candidate generation. Together, the H1–H8 prediction results and advantage over greedy planning support the usefulness of multi-step imagination within the complete planner.

4 Conclusion

This work formulates automatic lens design as action-conditioned world modeling and introduces LensDreamer. Its Multimodal Optical State Encoder forms a unified state, the Transformer–RSSM imagines multi-step trajectories, and the Basin-Map Planner allocates physical verification across candidate regions. LensDreamer-Traj supports dynamics learning, while LensDreamer-2K supports evaluation. Across 2,000 tasks, LensDreamer achieves the best feasibility, specification agreement, and imaging performance among the compared methods. Basin analysis shows broader verified coverage and more distinct DLS-stable endpoint groups, while the learned dynamics remain informative across H1–H8. These results establish world-model imagination as an effective approach to non-convex optical design under constrained high-fidelity physical-evaluation budgets.

Discussion and Limitations. LensDreamer uses learned imagination to allocate a limited high-fidelity physical-evaluation budget rather than reduce total compute. Evaluation covers specification-OOD tasks, not unseen topologies or lens families, and basin counts operationally denote distinct endpoint groups reached under fixed DLS. H1–H8 fidelity and planner ablations support multi-step imagination within the complete system without isolating horizon as the sole cause. First-action verification and re-encoding limit model-error accumulation. The action space preserves topology, and the results concern better local optima rather than guaranteed global optimality. Topology-changing actions and broader lens families remain future work.

AI USE STATEMENT

In this work, we used generative AI tools to help refine the conceptual framing, research questions, and hypotheses, provide feedback on the methodology and experimental design, write and execute selected parts of the research code, and organize, reformat, and partition the LensDreamer-2K benchmark dataset.

We have not used generative AI tools to generate synthetic datasets, retrieve or discover relevant literature, draft sections of the manuscript, formulate mathematical claims, provide critical ingredients for or write mathematical proofs, translate the manuscript, conduct qualitative or thematic data analysis, or interpret experimental results.

Additionally, we used generative AI tools to polish the manuscript language and paragraph transitions and organize the LaTeX code for tables and figures. We have reviewed all AI-assisted work.

Specifically, the authors manually verified all cited references against their original sources, checked AI-assisted language edits against the implementation, inspected and tested all AI-assisted code execution, and reviewed the edited text and LaTeX code line by line for technical accuracy, unsupported assertions, and formatting issues. The authors also verified the LensDreamer-2K sample assignments and dataset splits, checked for data leakage, and independently validated the experimental code and reported results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. It involves no human or animal subjects, personal data, or sensitive demographic attributes. LensDreamer-Traj and LensDreamer-2K consist of optical-design tasks and trajectories produced through physical simulation. Third-party software and optical resources are used in accordance with their applicable licenses. Because the outputs of LENSDREAMER are candidate optical designs rather than safety-certified devices, any downstream deployment requires independent physical, manufacturing, safety, and application-specific validation. Our method retains a high-fidelity physical evaluator in the loop and does not treat imagined predictions as substitutes for final physical verification.

REPRODUCIBILITY STATEMENT

Section 2 describes the complete LensDreamer pipeline, including the optical state representation, action-conditioned dynamics, Basin-Map Planner, trajectory construction, and training objective. Section 3 specifies the datasets, baselines, common evaluator, metrics, matched-budget protocol, and ablations. The appendix documents the model and planner configurations, action constraints, data splits, physical evaluation settings, and additional results needed to reproduce the reported experiments. We use fixed manifests and retain task-level physical and computational records for paired evaluation. Code, configurations, checkpoints, and datasets will be released upon acceptance.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Maxime Burchi and Radu Timofte. Learning transformer-based world models with contrastive predictive coding. *ArXiv*, abs/2503.04416, 2025. URL `https://api.semanticscholar.org/CorpusID:276813068`.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022.

Geoffroi Côté, Fahim Mannan, Simon Thibault, Jean-François Lalonde, and Felix Heide. The differentiable lens: Compound lens search over glass surfaces and materials for object detection.

In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20803–20812. IEEE, 2023.

Yao Gao, Qi Jiang, Shaohua Gao, Lei Sun, Kailun Yang, and Kaiwei Wang. Exploring quasi-global solutions to compound lens based computational imaging systems. *IEEE Transactions on Computational Imaging*, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Rudolf Kingslake and R Barry Johnson. *Lens design fundamentals*. academic press, 2009.

Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, P. Abbeel, Dan Klein, and Anca D. Dragan. Learning to model the world with language. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:260438420.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Joseph Meiron. Automatic lens design by the least squares method. *Journal of the Optical Society of America*, 49(3):293–298, 1959.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-dreamer: Redundancy-reduced world models without decoders or augmentation. *ArXiv*, abs/2603.18202, 2026. URL https://api.semanticscholar.org/CorpusID:286669844.

Reuven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1:127–190, 1999. URL https://api.semanticscholar.org/CorpusID:18838459.

Jürgen Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54(3):211–254, mar 2004. doi: 10.1023/B:MACH.0000015880.99707.b2.

Lei Sheng, Weichao Zhao, Ying Zhou, Weimeng Lin, Chunyan Du, and Hongwei Lou. A surrogate model based multi-objective optimization method for optical imaging system. *Applied Sciences*, 12(13):6810, 2022. doi: 10.3390/app12136810.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Qilin Sun, Congli Wang, Fu Qiang, Dun Xiong, and Heidrich Wolfgang. End-to-end complex lens design with differentiable ray tracing. *ACM Trans. Graph*, 40(4):1–13, 2021.

Maarten van Turnhout and Florian Bociort. Chaotic behavior in an algorithm to escape from poor local minima in lens design. *Optics express*, 17(8):6436–6450, 2009.

Congli Wang, Ni Chen, and Wolfgang Heidrich. do: A differentiable engine for deep lens design of computational imaging systems. *IEEE Transactions on Computational Imaging*, 8:905–916, 2022.

Xinge Yang, Qiang Fu, and Wolfgang Heidrich. Curriculum learning for ab initio deep learned refractive optics. *Nature communications*, 15(1):6572, 2024.