LENS AS ACTION: WORLD-STATE QUERYING FOR SCENE-CONSISTENT PHOTOGRAPHIC GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera-controllable text-to-image generation aims to synthesize photographs that comply with both semantic prompts and specified camera settings while maintaining scene coherence. Existing methods typically inject lens parameters directly into the denoising process as auxiliary conditions, but their responses to continuously varying parameter values can be weak or non-monotonic. Because scene content and camera-dependent appearance are represented jointly, changing a lens setting may also inadvertently alter scene geometry or semantic layout. In this work, we revisit photographic image generation from the perspective of lens-action world modeling. Rather than treating lens parameters merely as generative conditions, we interpret them as observation actions over a latent scene world. Based on this formulation, we propose a Lens-Action World Model (LAWM) for scene-consistent photographic generation. Given a text prompt, LAWM establishes an anchor observation and encodes its appearance together with geometric cues into a persistent world state shared across lens queries. A structured Lens-action encoder and Lens-Action Scene Querying retrieve action-relevant features from this state and inject them into a temporally coupled diffusion backbone, producing photographs under the requested settings. To alleviate the scarcity of aligned multi-setting photographs, we construct training supervision for lens parameters such as focal length and aperture using field-of-view transformations and depth-guided defocus simulation. Extensive experiments show that our method achieves stronger scene consistency and more reliable responses to continuous lens controls, demonstrating the potential of explicit world-state querying for camera-controllable image generation.

1 INTRODUCTION

Diffusion models Ho et al. (2020); Saharia et al. (2022) have substantially advanced text-to-image generation, while conditional methods Zhang et al. (2023) further enable control over pose, structure, and layout. However, these controls mainly determine what appears in an image, whereas photographic creation also requires control over how the scene is captured. Continuous camera control is valuable for virtual photography, production previsualization, and creative authoring, allowing users to explore lens settings without reconstructing the scene. Changing focal length or aperture should produce faithful and monotonic changes in field of view or depth of field while preserving subject identity and semantic layout. Achieving such camera response without entangling it with scene content remains a central challenge for camera-controllable generation.

Recent studies have made substantial progress in camera-aware image generation and editing. Camera Settings as Tokens Fang et al. (2024) embeds numerical camera metadata into the text feature space of a latent diffusion model, enabling control over several photographic parameters. Generative Photography Yuan et al. (2025) and CamEdit Qin et al. (2026) further demonstrate that diffusion models can learn meaningful associations between camera parameters and visual appearance. Despite their different designs, these methods generally incorporate camera parameters into the denoising process through textual, feature-level, or frame-wise conditioning. Scene content and camera-dependent visual effects are therefore represented jointly within the generative backbone, while consistency across camera settings is maintained primarily by learned generative or temporal priors. Without an explicit scene state shared across camera queries, the model lacks a direct mech-

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.

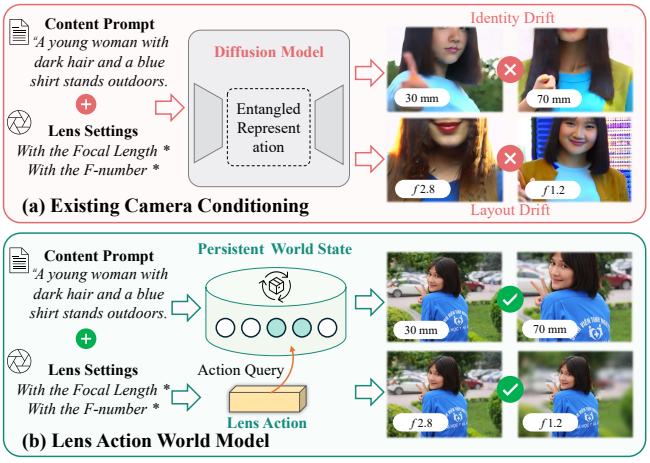

Figure 1: **Comparison with existing methods.** Existing camera-aware generators incorporate camera settings as conditioning signals while representing scene content implicitly within the generative backbone. Consequently, changes in camera settings may be accompanied by unintended variations in subject identity, geometry, or semantic layout. LAWM instead constructs a persistent appearance–geometry world state and queries the same state with different lens actions, producing camera-dependent photographs while preserving the underlying scene.

anism to ensure that only the photographic appearance changes while subject identity and semantic layout remain fixed.

To address this limitation, we revisit camera-aware image generation from the perspective of lens-action world modeling. Our key observation is that changing lens settings does not alter the underlying physical scene itself, but changes how that scene is photographically observed. We therefore represent the shared appearance and geometry of a scene as a persistent photographic world state, and interpret lens settings as lens actions applied to this state. Different lens actions should produce distinct photographic observations, while the identity of the depicted subjects, the scene structure, and the semantic content remain consistently unchanged. This perspective reformulates camera-aware generation from direct parameter-to-appearance conditioning into a lens-action querying process over a shared scene representation, as illustrated in Fig. 1.

To instantiate this formulation, we propose the LAWM, which first fuses an aligned reference observation with VGGT-derived Wang et al. (2025a) geometric cues and encodes them into a persistent multi-scale world state. Each requested lens setting is represented as a structured lens action that combines a physically motivated spatial map with a differential lens representation. At each feature scale, we introduce Lens-Action Scene Querying, through which the action features query compact world-memory tokens, and the retrieved information is injected into the corresponding camera-conditioning pathway of a temporally coupled latent-diffusion backbone. This design keeps the world state shared throughout the output sequence, while allowing only the lens action to vary across photographic observations.

We focus on focal length and aperture because their effects are directly related to scene geometry: focal length changes the field of view and apparent scale, whereas aperture produces depth-dependent defocus. Their explicit geometric effects also make it possible to construct physically motivated supervision when aligned photographs of the same scene under densely sampled lens settings are unavailable. We therefore derive multi-setting targets from individual reference images using field-of-view transformations and depth-guided defocus simulation. This geometry-guided supervision enables LAWM to learn camera-dependent observations from a shared world state without requiring large-scale real paired data. Extensive experiments demonstrate that LAWM achieves strong camera controllability and scene consistency across continuous focal-length and aperture settings.

Our main contributions are summarized as follows:

- We reformulate lens-aware image generation as persistent lens-action querying, where lens parameters specify observation actions over a shared scene state rather than directly conditioning the appearance of generated images.
- We propose a Lens-Action World Model that constructs an explicit appearance-geometry world state from a reference observation and uses Lens-Action Scene Querying to query this state under different lens settings.

- We conduct extensive experiments on multiple photographic controls, including focal length and aperture. The results demonstrate that LAWM substantially improves scene consistency, camera controllability, and visual fidelity over existing camera-aware diffusion models.

2 Related Work

2.1 Camera-Aware Image Generation and Editing

Early studies explored how photographic metadata can be represented and associated with visual appearance. EXIF as Language Zheng et al. (2023) learns cross-modal correspondences between image content and textualized EXIF metadata, establishing that camera properties can be encoded in a shared vision–language space. Camera Settings as Tokens Fang et al. (2024) further introduces numerical camera settings as learnable tokens in latent diffusion models, enabling text-to-image generation conditioned on attributes such as focal length and aperture. PreciseCam Bernal-Berdun et al. (2025) instead focuses on explicit intrinsic and extrinsic camera control, while Generative Photography Yuan et al. (2025) models continuous photographic parameters through frame-wise camera conditioning and temporally coupled generation. These approaches demonstrate that camera metadata provides effective control signals beyond conventional semantic prompts.

More recent methods extend camera control toward specialized photographic effects, reference-based editing, and video generation. Bokeh Diffusion Fortes et al. (2025) and Fine-grained Defocus Blur Control Shrivastava et al. (2026) target aperture-dependent defocus, whereas DiffCamera Wang et al. (2025b) and Generative Refocusing Yuan et al. (2025) support spatially flexible refocusing from a single image. CameraMaster Yang et al. (2025) and CamEdit Qin et al. (2026) introduce unified or continuous camera-parameter control for photographic editing. Less is More Cheng et al. (2026) investigates data-efficient adaptation of text-to-video models to physical camera controls, and DeltaCam Mandal et al. (2026) represents relative changes in intrinsic camera parameters for temporally consistent video generation. Despite their different architectures, these methods primarily inject camera settings as textual, numerical, feature-level, or frame-wise conditions. They do not explicitly maintain a scene state shared across camera queries, leaving camera-dependent effects potentially entangled with subject appearance, scene geometry, and semantic layout.

2.2 World Models and Action-Conditioned Visual Generation

World models Ha & Schmidhuber (2018); Hafner et al. (2023); Xiao et al. (2026); Zhang et al. (2025); Agarwal et al. (2025) learn latent representations of an environment and predict its future states or observations under specified actions. This paradigm supports planning and prediction by maintaining an internal state that can be reused across multiple action queries. Recent generative world models Bruce et al. (2024); Alonso et al. (2024) extend this principle to high-dimensional visual generation by predicting coherent observations conditioned on user actions or motion commands. Related camera-controlled generative models focus on observation changes induced by viewpoint motion. Stable Virtual Camera Zhou et al. (2025) synthesizes consistent novel views under target camera poses, while GEN3C Ren et al. (2025) maintains a geometry-informed 3D cache to support world-consistent generation along specified camera trajectories.

These methods differ from ours in their definition of action. Conventional world models consider actions that change the environment state, while camera-controlled view synthesis mainly models extrinsic camera pose or trajectory. LAWM instead treats intrinsic photographic parameters, specifically focal length and aperture, as observation actions over a persistent photographic world state. By querying a shared appearance-geometry representation, LAWM changes the photographic observation while preserving scene identity, geometry, and semantic layout.

3 Methodology

3.1 Problem Formulation and Overview

Given a semantic description c and a set of target lens settings $\{a_k\}_{k=1}^{K}$, our objective is to generate K photographic observations that follow the camera-parameter variations while preserving the

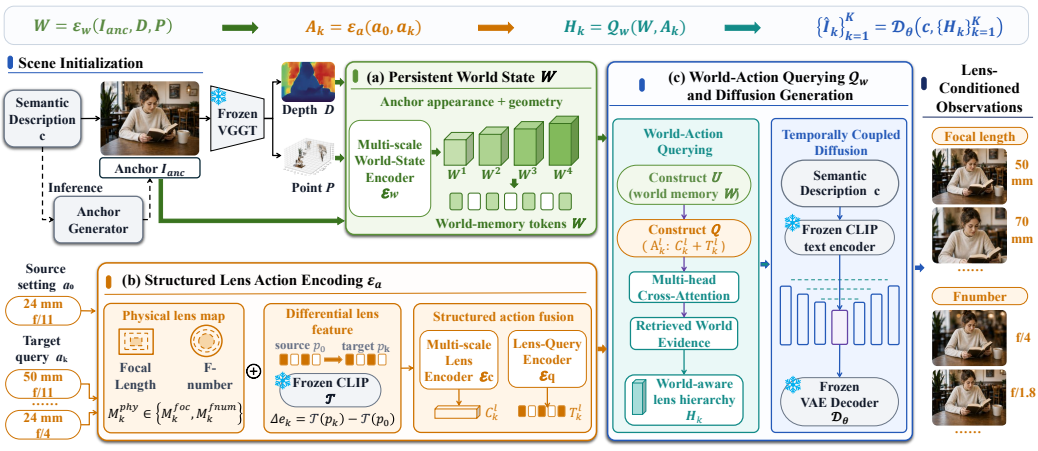

Figure 2: **Overview of our framework.** An anchor observation and VGGT-derived geometry form a persistent multi-scale world state. Source-relative focal-length or f-number queries are encoded as structured lens actions, which retrieve action-specific world evidence and condition a temporally coupled diffusion model to generate scene-consistent observations.

identity characteristics and semantic layout of the scene. We consider continuous control of the 35-mm equivalent focal length and f-number, with only one lens variable changed within each query sequence. During training, the reference image paired with the semantic description c is treated as the anchor observation I_{anc}, which establishes the shared scene, while its source lens setting a_0 serves as the reference point for defining source-relative lens actions. At inference, the generator synthesizes I_{anc} from c under the source lens setting a_0. A frozen VGGT model Wang et al. (2025a) extracts a depth map D and a point-map magnitude P from the anchor observation.

LAWM represents the anchor appearance and geometry as a persistent world state and interprets each source-to-target lens change as an observation action. As illustrated in Fig. 2, the generation process consists of three principal components and is written as:

$$\begin{aligned} W &= \mathcal{E}_w(I_{\text{anc}}, D, P), \quad A_k = \mathcal{E}_a(a_0, a_k), \\ H_k &= \mathcal{Q}_w(W, A_k), \quad \{\hat{I}_k\}_{k=1}^K = \mathcal{D}_\theta\left(c, \{H_k\}_{k=1}^K\right). \end{aligned} \tag{1}$$

The world-state encoder $\mathcal{E}_w$ transforms the aligned anchor observation I_{anc}, depth map D, and point-map magnitude P into a persistent multi-scale world state W. The world state W is constructed once from the anchor scene and shared by all K lens queries. Second, the structured lens-action encoder $\mathcal{E}_a$ represents each source-to-target setting (a_0, a_k) as an action feature A_k using a physics-informed action map and a differential lens feature, where A_k specifies the requested change from a_0 to a_k. Third, Lens-Action Scene Querying $\mathcal{Q}_w$ uses A_k to query the shared world state W, producing the multi-scale world-aware lens feature H_k. The lens-conditioned diffusion generator $\mathcal{D}_\theta$ receives $\{H_k\}_{k=1}^K$ together with the semantic description c and synthesizes $\{\hat{I}_k\}_{k=1}^K$. Its frame dimension indexes different lens queries over the same scene rather than temporal scene evolution.

3.2 Persistent Photographic World State

To ensure that all lens queries share the same scene representation, we construct a persistent appearance-geometry world state from the anchor observation I_{anc}. A frozen VGGT model extracts an aligned depth map D and a three-dimensional point map Π, which is reduced to its per-pixel magnitude P. After normalization, these geometry cues are concatenated with the aligned anchor appearance:

$$X_w = \text{Concat}\left(I_{\text{anc}}, D, P\right), \tag{2}$$

A multi-scale world-state encoder $\mathcal{E}_w$, implemented using a strided convolutional stem and residual blocks, transforms X_w into a hierarchy of scene features:

$$W = \left\{W^l\right\}_{l=1}^L, \qquad W^l = \mathcal{E}_w^l(X_w), \tag{3}$$

where $L = 4$ and l indexes the feature scale. The four levels follow the spatial hierarchy of the diffusion backbone. The anchor appearance provides identity and semantic information, while D and P provide complementary cues about scene structure and depth ordering. Their joint encoding forms the persistent world state W, as shown in Fig. 2(a).

The lens action is deliberately excluded from $\mathcal{E}_w$. Consequently, W is computed once for each scene and reused across all K lens queries to preserve shared scene evidence, while only the action representation varies to model query-dependent photographic changes.

3.3 Structured Lens Action Encoding

LAWM defines a lens action relative to the source setting of the anchor observation rather than encoding a target value in isolation. Let a_0 denote the source setting obtained from the EXIF metadata of I_{anc}, and let a_k denote the setting requested by the k-th query. We consider continuous control of the 35-mm equivalent focal length and f-number, varying only one parameter within each query sequence. During training, target values are sampled from source-relative intervals supported by the supervision process, with $f_k \geq f_0$ for focal length and $N_k \leq N_0$ for f-number. The structured action representation combines a parameter-specific physics-informed action map with a source-relative differential feature, thereby encoding both the expected image-space effect and the magnitude of the requested change from a_0 to a_k.

Focal-length action map. Let f_0 and f_k denote the source and target focal lengths, respectively. Assuming a full-frame sensor with width $s_x = 36\,\text{mm}$ and height $s_y = 24\,\text{mm}$, the field of view along dimension $d \in \{x, y\}$ is:

$$\phi_d(f) = 2\arctan\left(\frac{s_d}{2f}\right), \qquad d \in \{x, y\}. \tag{4}$$

We define the source-relative field-of-view ratio as:

$$r_{d,k} = \text{clip}\left(\frac{\phi_d(f_k)}{\phi_d(f_0)}, 0, 1\right). \tag{5}$$

On a normalized image grid $(u, v) \in [-1, 1]^2$, the corresponding focal map is:

$$M_k^{\text{foc}}(u, v) = \begin{cases} 1, & |u| \leq r_{x,k} \text{ and } |v| \leq r_{y,k}, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

which is replicated across three channels. A longer focal length yields a smaller active region and explicitly represents the expected change in field of view, framing, and apparent subject scale.

F-number action map. Let N_0 and N_k denote the source and target f-numbers. We convert their source-relative change into a bounded f-number control strength:

$$\begin{aligned} \alpha_k &= \text{clip}\left(\frac{N_0/N_k - 1}{N_0/N_{\min} - 1}, 0, 1\right), \\ \kappa_k &= \kappa_{\min} + \alpha_k\left(\kappa_{\max} - \kappa_{\min}\right). \end{aligned} \tag{7}$$

where $N_{\min}$ is the smallest supported f-number and $\kappa_{\min}$ and $\kappa_{\max}$ define the range of the internal f-number representation.

This mapping preserves the photographic ordering that a smaller f-number produces stronger depth-dependent defocus. Let $\overline{\mathcal{G}}_{\sigma_k}$ denote a normalized Gaussian kernel with $\sigma_k = \kappa_k/3$. Its center response $\eta_k = \overline{\mathcal{G}}_{\sigma_k}(0, 0)$ is expanded to form the three-channel f-number map:

$$\left[M_k^{\text{fnum}}\right]_{j,u,v} = \eta_k, \quad (u, v) \in \Omega, \tag{8}$$

where j indexes the feature channel and Ω denotes the image grid.

This construction yields a spatially constant three-channel representation whose magnitude varies continuously with the target f-number.

Differential lens representation. The physical map describes the expected spatial effect but does not fully encode the continuous numerical relation between source and target settings. We therefore represent each lens value using a lens-specific text token:

$$p_k = \begin{cases} \langle \text{focal length: } f_k \rangle, & \text{for focal length,} \\ \langle \text{f-number control: } \kappa_k \rangle, & \text{for f-number,} \end{cases} \tag{9}$$

The tokens are encoded using the frozen CLIP text encoder $\mathcal{E}_{\text{text}}$ Radford et al. (2021), with p_0 defined analogously from the source setting. Given $e_k = \mathcal{E}_{\text{text}}(p_k)$ and $e_0 = \mathcal{E}_{\text{text}}(p_0)$, we compute the source-relative feature $\Delta e_k = e_k - e_0$, standardize it feature-wise, and reshape and interpolate it over the image grid to produce a three-channel differential map M_k^{diff}. The complete structured action map is then:

$$\begin{aligned} M_k &= \text{Concat}\left(M_k^{\text{phy}},\, M_k^{\text{diff}}\right), \\ M_k^{\text{phy}} &\in \left\{M_k^{\text{foc}}, M_k^{\text{fnum}}\right\}, \end{aligned} \tag{10}$$

where $M_k \in \mathbb{R}^{6\times H_{\text{img}}\times W_{\text{img}}}$, with H_{img} and W_{img} denoting image height and width. As shown in Fig. 2(b), a multi-scale lens encoder $\mathcal{E}_c$ extracts spatial lens features $C_k^l = \mathcal{E}_c^l(M_k)$, while a lens-query encoder $\mathcal{E}_q$ produces action tokens $T_k^l = \mathcal{E}_q^l(M_k)$ for $l = 1, \ldots, L$. Together, $A_k = \{(C_k^l, T_k^l)\}_{l=1}^L$ forms the structured lens-action representation in Eq. equation 1.

3.4 World-Action Querying and Diffusion Generation

The world state $W = \{W^l\}_{l=1}^L$ is fixed across lens queries, whereas A_k specifies the requested lens action. As shown in Fig. 2(c), at each scale, Lens-Action Scene Querying uses the spatial feature C_k^l as the query carrier and augments it with the action tokens T_k^l. The world feature W^l is pooled into compact memory tokens and concatenated with T_k^l:

$$\begin{aligned} Q_k^l &= \text{LN}_c^l\big(\text{Tok}(C_k^l)\big) + \text{Avg}\Big(\text{LN}_a^l(T_k^l)\Big), \\ U_k^l &= \text{Concat}\Big[\text{LN}_w^l\big(\text{Tok}\big(\text{Pool}_l(W^l)\big)\big), \text{LN}_a^l(T_k^l)\Big], \end{aligned} \tag{11}$$

where LN denotes LayerNorm, Tok converts a feature map into tokens, Avg averages over the action-token dimension, and Pool_l denotes adaptive pooling at scale l. Q_k^l combines the spatial lens feature with a broadcast action summary, whereas U_k^l concatenates pooled world-state tokens with detailed action tokens and serves as both keys and values. The retrieved world information is merged with the lens feature through a gated residual update:

$$\widetilde{C}_k^l = C_k^l + g_l \ \text{Reshape}\big[\mathcal{O}_l\big(\text{MHA}_l(Q_k^l, U_k^l, U_k^l)\big)\big], \tag{12}$$

where MHA_l denotes multi-head attention, $\mathcal{O}_l$ is an output projection, and g_l is a learnable gate. Thus, all queries reuse the same world memory while retrieving different scene evidence according to their lens actions. Collecting the fused features across scales gives $H_k = \{\widetilde{C}_k^l\}_{l=1}^L = \mathcal{Q}_w(W, A_k)$.

The generator $\mathcal{D}_\theta$ injects $\{H_k\}_{k=1}^K$ through learnable merge projections into the corresponding downsampling, middle, and upsampling blocks of a temporally coupled latent-diffusion U-Net Ronneberger et al. (2015). The semantic description c is supplied through the CLIP text-conditioning pathway. Given a noisy query-indexed latent sequence z_t, its denoising network ϵ_θ predicts:

$$\hat{\epsilon} = \epsilon_\theta\left(z_t, t;\, \mathcal{E}_{\text{text}}(c), \{H_k\}_{k=1}^K\right). \tag{13}$$

The pretrained temporal module jointly processes the K query latents, and the frozen VAE decodes the denoised sequence into $\{\hat{I}_k\}_{k=1}^K$. Here, the frame dimension represents different lens queries over one static scene rather than temporal evolution.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on the full BokehDiffusion dataset Fortes et al. (2025), which contains 15,305 photographs with synthetic captions and camera metadata, including 35-mm equiv-

Table 1: Camera controllability on the unified evaluation protocol. Focal-length response is evaluated using SIFT-estimated image scale, while f-number response is evaluated using the monotonic trend of Laplacian variance.

	Focal Length Control		F-number Control		
Method	Param. Corr. ↑	Scale MAE ↓	Param. Corr. ↑	Rank Corr. ↑	Order Acc. ↑
Stable Diffusion 1.5	-0.173 ± 0.421	0.796 ± 0.302	0.048 ± 0.719	0.067 ± 0.658	0.508 ± 0.271
Stable Diffusion 3.5	0.110 ± 0.622	0.627 ± 0.022	0.229 ± 0.545	0.300 ± 0.543	0.617 ± 0.212
Camera Settings as Tokens	-0.273 ± 0.507	$\underline{0.488 \pm 0.451}$	0.336 ± 0.558	0.275 ± 0.459	0.608 ± 0.183
Generative Photography	$\underline{0.219 \pm 0.742}$	0.572 ± 0.630	$\underline{0.809 \pm 0.317}$	$\underline{0.808 \pm 0.387}$	$\underline{0.900 \pm 0.200}$
LAWM (Ours)	$\mathbf{0.221 \pm 0.589}$	$\mathbf{0.454 \pm 0.335}$	$\mathbf{0.945 \pm 0.046}$	$\mathbf{0.992 \pm 0.029}$	$\mathbf{0.992 \pm 0.029}$

alent focal length and f-number. We randomly partition the dataset into 90% for training and 10% for validation. For each image, frozen VGGT preprocessing provides an aligned RGB observation, a depth map, and point-map magnitude. Camera-dependent targets are generated online after the split using the field-of-view transformation for focal length and depth-guided defocus simulation for f-number. These synthesized observations are used exclusively as training supervision and are not required during evaluation or inference.

Implementation details. We initialize the generative backbone from Stable Diffusion 1.5 Rombach et al. (2022) and load the pretrained temporal module and RealEstate10K image LoRA Zhou et al. (2018); Hu et al. (2022). The VAE, CLIP text encoder, base U-Net, temporal module, and image LoRA are frozen. We optimize the lens encoder, world-state encoder, lens-query encoder, Lens-Action Scene Querying, and lens-feature merge projections. Images are processed at 256×384. The source setting a_0 is obtained from EXIF metadata. Target focal lengths are sampled with $f_k/f_0 \in [1, 2]$, while target f-numbers use $N_k/N_0 \in [0.25, 1]$. We jointly train both controls for 200 epochs on four NVIDIA RTX 5090 GPUs. We use AdamW with a learning rate of 3×10^{-5}, weight decay of 10^{-2}.

Evaluation metrics. Because real pixel-aligned observations of the same scene under different lens settings are unavailable, we adopt a reference-free evaluation protocol. It contains 100 prompts spanning portraits, indoor scenes, streets, and tabletop objects. For focal length, SIFT matching Lowe (2004) and RANSAC-based similarity estimation recover the image-scale response relative to the first observation. We report its Pearson correlation with the analytic full-frame field-of-view trend (Param. Corr.) and the mean absolute log-scale error (Scale MAE). For f-number, Laplacian variance measures image sharpness; we report its Pearson correlation with the queried values, Spearman rank correlation, and pairwise ordering accuracy. Scene consistency is evaluated using adjacent LPIPS Zhang et al. (2018), focal-aligned adjacent LPIPS, and anchor-to-query LPIPS. CLIPScore Radford et al. (2021) measures prompt alignment.

Baselines. We compare LAWM with two prompt-based diffusion baselines, Stable Diffusion 1.5 Rombach et al. (2022) and Stable Diffusion 3.5 Medium Esser et al. (2024), and two camera-aware methods, Camera Settings as Tokens Fang et al. (2024) and Generative Photography Yuan et al. (2025). All methods use the same prompts, target camera schedules, random seed, and output resolution whenever supported by their public implementations.

4.2 Comparison to State-of-the-Art Methods

Lens controllability. Table 1 shows that LAWM provides the strongest overall response to continuous focal-length and f-number queries. For focal length, LAWM achieves the lowest scale MAE of 0.454, although the remaining variance indicates that large spatial changes are still challenging. The improvement is more pronounced for f-number: LAWM reaches a parameter correlation of 0.945 and an ordering accuracy of 0.992, compared with 0.809 and 0.900 for Generative Photography. This advantage arises from the structured lens-action representation: the physics-informed action map explicitly encodes the expected parameter trend, while Lens-Action Scene Querying retrieves the geometry and appearance evidence relevant to each query from a shared world state. Consequently, the model learns an action-dependent photographic transformation instead of relying solely on correlations between numerical conditions and image appearance.

Table 2: Scene consistency and prompt alignment. Adjacent LPIPS measures local changes between consecutive camera queries; aligned LPIPS measures for the intended focal-scale transform; anchor LPIPS measures drift from the anchor.

Method	Control	Adj. LPIPS ↓	Aligned LPIPS ↓	Anchor LPIPS ↓	CLIPScore ↑
Stable Diffusion 1.5	Focal length	0.379 ± 0.149	0.526 ± 0.214	0.403 ± 0.155	0.215 ± 0.029
	F-number	0.324 ± 0.151	0.324 ± 0.151	0.352 ± 0.160	0.220 ± 0.030
Stable Diffusion 3.5	Focal length	0.378 ± 0.057	0.525 ± 0.170	0.590 ± 0.050	0.217 ± 0.021
	F-number	0.159 ± 0.066	0.159 ± 0.066	0.198 ± 0.089	0.241 ± 0.026
Camera Settings as Tokens	Focal length	0.471 ± 0.082	0.576 ± 0.138	0.567 ± 0.107	0.200 ± 0.025
	F-number	0.478 ± 0.079	0.478 ± 0.079	0.569 ± 0.096	0.200 ± 0.024
Generative Photography	Focal length	0.473 ± 0.101	0.443 ± 0.178	0.588 ± 0.091	0.218 ± 0.022
	F-number	0.248 ± 0.041	0.248 ± 0.041	0.477 ± 0.088	0.198 ± 0.028
LAWM (Ours)	Focal length	$\mathbf{0.149 \pm 0.048}$	$\mathbf{0.168 \pm 0.062}$	$\mathbf{0.208 \pm 0.098}$	$\mathbf{0.247 \pm 0.026}$
	F-number	$\mathbf{0.056 \pm 0.026}$	$\mathbf{0.056 \pm 0.026}$	$\mathbf{0.082 \pm 0.037}$	$\mathbf{0.243 \pm 0.026}$



Figure 3: Qualitative comparison of focal-length control. LAWM produces a monotonic field-of-view change while preserving subject identity, object count, and scene layout, whereas the baselines exhibit object, identity, or scene-appearance drift.

Scene consistency and semantic alignment. Table 2 further shows that LAWM achieves the best scene consistency across both lens controls while maintaining the strongest prompt alignment. Its focal-aligned LPIPS decreases from 0.443 for Generative Photography to 0.168, while its f-number adjacent LPIPS decreases from 0.248 to 0.056. LAWM also obtains the highest CLIPScore values of 0.247 and 0.243 for the two controls. Rather than relying only on generative or temporal priors, LAWM constructs the persistent world state once and reuses it across all lens queries; only the action representation changes. Lens-Action Scene Querying further restricts each query to action-relevant world evidence, preventing the requested photographic effect from being absorbed into subject identity, scene geometry, or semantic layout. Since these consistency gains accompany the strongest controllability results, the lower perceptual variation cannot be attributed to an unresponsive generator.

Qualitative comparison. Figures 3 and 4 compare LAWM with a representative baseline for each prompt under the same ordered lens queries. In Fig. 3, LAWM produces a coherent, monotonic change in field of view and apparent scale while retaining the motorcycle appearance, the identities and number of people, and the forest layout. In contrast, the baselines introduce changes in object appearance, collapse the two-person scene into a single-person close-up, or alter the color and structure of the forest. Figure 4 shows a similar advantage for f-number control. As the f-number decreases, LAWM progressively increases background defocus while preserving the identity, pose, and sharpness of the principal subject. The competing methods instead exhibit identity changes, framing drift, or global blur that also degrade the foreground subject. These examples visually confirm that LAWM separates the requested photographic change from the scene content more effectively.

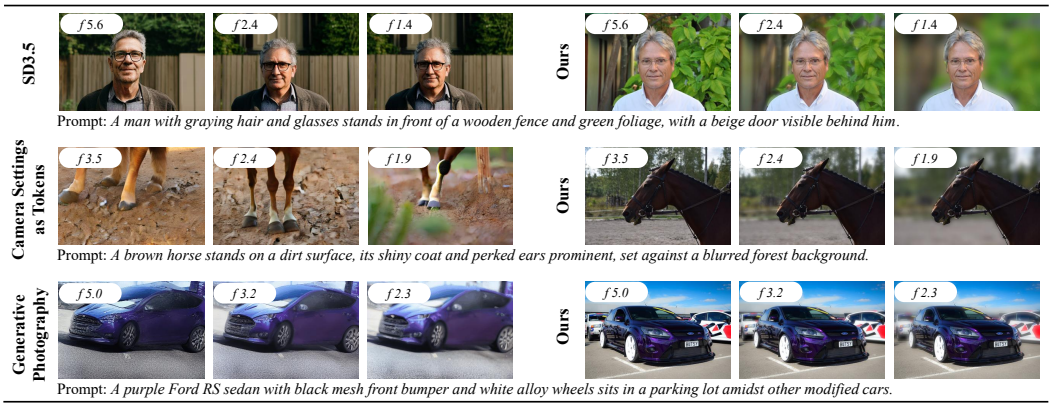

Figure 4: **Qualitative comparison of f-number control.** LAWM progressively increases depth-dependent background defocus while preserving subject identity and foreground sharpness.

Table 3: Component ablation of LAWM for focal-length and f-number control.

	Focal Length Control			F-number Control		
Configuration	Param. Corr. ↑	Scale MAE ↓	Aligned LPIPS↓	Param. Corr. ↑	Order Acc. ↑	Adj. LPIPS↓
Camera conditioning only	-0.450 ± 0.551	2.855 ± 3.896	0.711 ± 0.094	0.313 ± 0.683	0.650 ± 0.306	0.497 ± 0.102
w/o physical lens map	-0.011 ± 0.523	0.997 ± 0.075	0.260 ± 0.186	0.165 ± 0.697	0.536 ± 0.295	0.170 ± 0.091
w/o Lens-Action Scene Querying	-0.244 ± 0.605	1.274 ± 1.040	0.505 ± 0.160	0.368 ± 0.582	0.657 ± 0.250	0.257 ± 0.052
LAWM (full)	$\mathbf{0.221 \pm 0.589}$	$\mathbf{0.454 \pm 0.335}$	$\mathbf{0.168 \pm 0.062}$	$\mathbf{0.945 \pm 0.046}$	$\mathbf{0.992 \pm 0.029}$	$\mathbf{0.056 \pm 0.026}$

4.3 Ablation Study

Table 3 evaluates the structured lens representation and world-action interaction. *Camera conditioning only* removes the persistent world-state pathway, *w/o physical lens map* retains only the differential lens feature, and *w/o Lens-Action Scene Querying* replaces query-based retrieval with additive fusion. For focal length, direct conditioning produces a negative correlation and severe content drift, while removing either the physical map or cross-attention also weakens the focal-scale response. The full model achieves the highest correlation (0.221), lowest scale MAE (0.454), and lowest aligned LPIPS (0.168), although its relatively large correlation variance indicates that focal-length control remains challenging. For f-number, LAWM achieves a correlation of 0.945.Removing the physical map sharply degrades controllability despite its relatively low LPIPS, showing that perceptual similarity alone may result from nearly unchanged outputs. Removing cross-attention degrades both response accuracy and consistency. Overall, the physical map supplies an explicit parameter trend, whereas Lens-Action Scene Querying applies that trend to action-relevant evidence retrieved from the shared world state.

5 Discussion and Conclusion

LAWM reformulates camera-controllable generation as lens-action querying over a persistent photographic world state. Combining anchor appearance, VGGT-derived geometry, and structured lens actions through Lens-Action Scene Querying, LAWM enables continuous focal-length and f-number control without densely paired photographs. Experiments show strong aperture control, improved focal-scale response, and competitive scene consistency.

However, this method still has several limitations. Geometry-guided targets approximate lens responses using centered field-of-view transformations and depth-guided Gaussian filtering rather than calibrated optical supervision. Each query sequence varies only one lens parameter within source-relative training ranges, while evaluation relies largely on reference-free metrics due to the lack of aligned real captures. Future work will explore calibrated paired supervision, stronger multi-view or 3D world representations, and compositional control of additional lens parameters.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2724–2733, 2025.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Shihan Cheng, Nilesh Kulkarni, David Hyde, and Dmitriy Smirnov. Less is more: Data-efficient adaptation for controllable text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14811–14821, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

I-Sheng Fang, Yue-Hua Han, and Jun-Cheng Chen. Camera settings as tokens: Modeling photography on latent diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Armando Fortes, Tianyi Wei, Shangchen Zhou, and Xingang Pan. Bokeh diffusion: Defocus blur control in text-to-image diffusion models. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

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

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Debabrata Mandal, Zhihan Peng, Yujie Wang, and Praneeth Chakravarthula. Deltacam: Differential intrinsic camera modeling for video generation. *arXiv preprint arXiv:2605.25266*, 2026.

Xinran Qin, Zhixin Wang, Fan Li, Haoyu Chen, Renjing Pei, Wenbo Li, and Xiaochun Cao. Camedit: Continuous camera parameter control for photorealistic image editing. *Advances in Neural Information Processing Systems*, 38:114152–114171, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Ayush Shrivastava, Connelly Barnes, Xuaner Zhang, Lingzhi Zhang, Andrew Owens, Sohrab Amirghodsi, and Eli Shechtman. Fine-grained defocus blur control for generative image models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4335–4344, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Yiyang Wang, Xi Chen, Xiaogang Xu, Yu Liu, and Hengshuang Zhao. Diffcamera: Arbitrary refocusing on images. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Qirui Yang, Yang Yang, Ying Zeng, Xiaobin Hu, Bo Li, Huanjing Yue, Jingyu Yang, and Peng-Tao Jiang. Cameramaster: Unified camera semantic-parameter control for photography retouching. *arXiv preprint arXiv:2511.21024*, 2025.

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7920–7930, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Josh Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21685–21695. IEEE Computer Society, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6945–6956, 2023.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12405–12414, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

AI USE STATEMENT

In this work, we used generative AI tools exclusively for grammatical error correction and language polishing. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.