# PUFFIN-WORLD: SCALING A UNIFIED MULTIMODAL MODEL WITH NATIVE 3D WORLD STATES

**Anonymous authors**
Paper under double-blind review

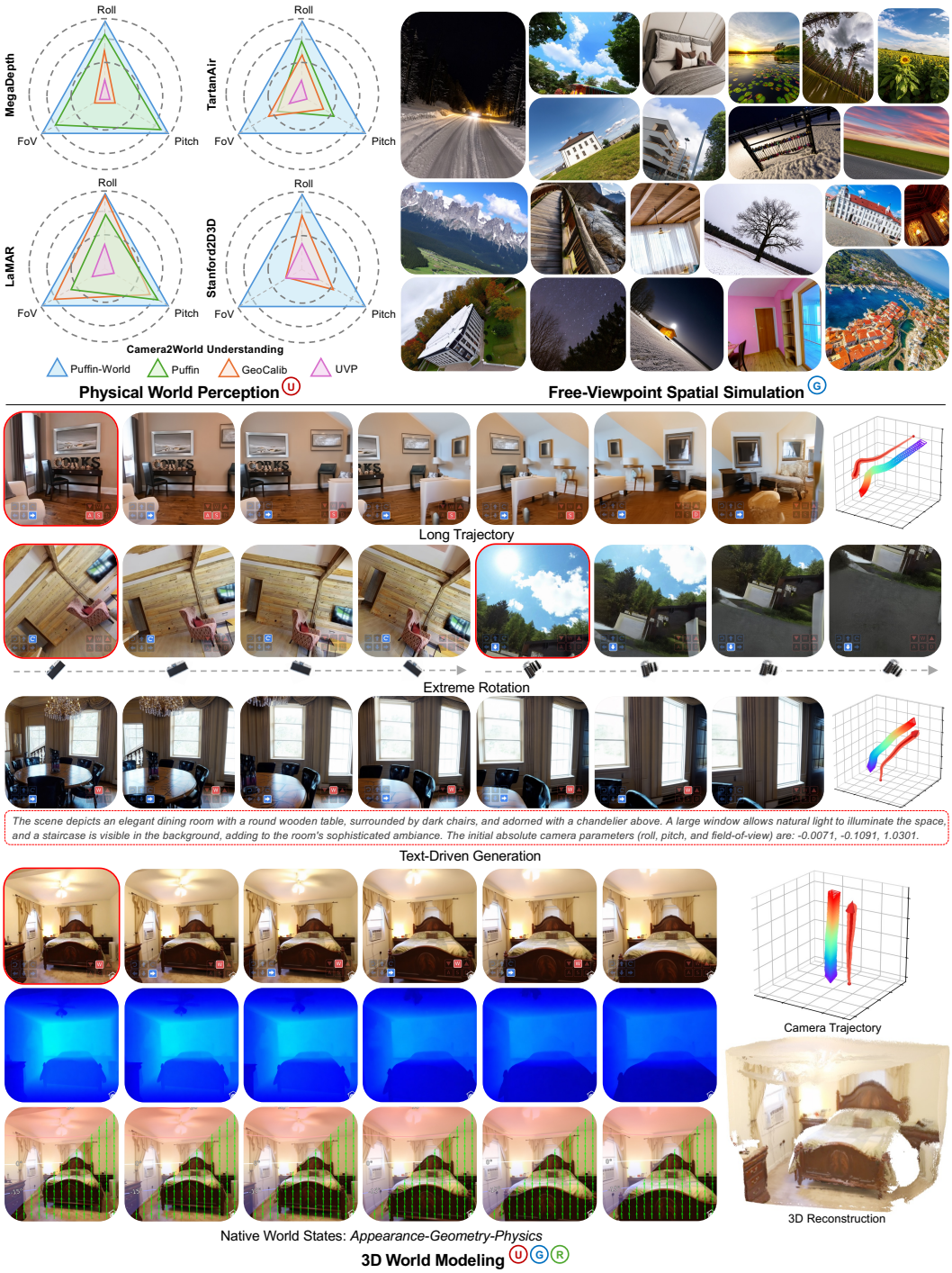


Figure 1: **Versatile capabilities of Puffin-World:** camera-to-world understanding for physical world perception (AUC▷5°), flexible camera-controllable abilities for free-viewpoint spatial simulation, and 3D world generation and reconstruction with flexible actions and native world states.



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.

## ABSTRACT

We propose Puffin-World, a unified multimodal architecture that integrates physical understanding, spatial simulation, and 3D world generation and reconstruction without relying on external offline modules. To reliably construct and interact with 3D worlds, our framework jointly models three native world states: physics (gravity field and latitude), geometry (depth), and appearance (image and sequence), together with a unified Omni-Camera representation that supports diverse tasks and flexible motions. Beyond modeling these states, we introduce a strategy for propagating physical dynamics across future frames. By grounding absolute camera properties in the real world, Puffin-World enables physically consistent and visually stable world generation. We further couple appearance and geometry within a single generative process, jointly synthesizing each future view and reconstructing its underlying geometry. This unified paradigm enables interleaved closed-loop applications requiring synergy across multiple tasks, including mimic and self-calibrated world exploration. To scale Puffin-World to complex scenarios, we construct Puffin-16M, comprising 15 million vision–language–camera triplets and 1 million trajectories featuring various and challenging motions. To foster further research in this area, we will release the code, models, and datasets.


## 1 INTRODUCTION

Building a model that can perceive, generate, and reconstruct the world from arbitrary visual observations is a central goal of multimodal spatial intelligence and physical AI. A capable world model should not merely generate plausible pixels; it should understand where the camera or viewpoint is in the real world, reason about the geometry of the scene, and synthesize what the world looks like from new viewpoints in a way that stays consistent with physical law. The physical world therefore cannot be adequately represented as a collection of 2D images. Instead, effective interaction with the real world calls for a unified framework that jointly models its multiple interconnected states.

Progress so far has advanced along separate fronts that each capture only part of this picture. Generative world models (Agarwal et al., 2025; Parker-Holder et al., 2025; Gao* et al., 2024; Zhu et al., 2026; Huang et al., 2025b) have made striking progress, but they predict the world almost exclusively at the appearance level, treating each frame as RGB content with no explicit notion of the camera's physical orientation or the underlying scene geometry. In parallel, unified multimodal models (Team, 2024; Zhou et al., 2024; Wu et al., 2025a; Chen et al., 2025b) couple understanding and generation within a single network, but do so for only 2D semantics. As a result, the field still lacks a single model that unifies holistic modalities and tasks for 3D world modeling.

Unifying these capabilities is fundamentally non-trivial. A simple combination of existing perception, generation, and reconstruction components is insufficient because a unified world model faces three tightly coupled challenges: (i) establishing a unified action representation that supports both continuous single- and cross-view control while remaining grounded in absolute physical concepts such as gravity, uprightness, and orientation; (ii) maintaining a physically persistent frame that allows knowledge inferred from observed views to consistently guide 3D world modeling at unseen viewpoints; and (iii) scaling these capabilities with data that provides both absolute camera grounding and diverse, challenging motion. Existing relative camera representations, such as Plücker embeddings, offer effective control but lack a global physical anchor, while prevailing 3D datasets (Ling et al., 2024) contain limited rotational diversity and rarely provide absolute camera orientation.

In this work, we present **Puffin-World**, a unified multimodal model that scales 3D world perception, generation, and reconstruction, without relying on any external offline modules. In particular, we formulate three native world states to jointly model the 3D world in an end-to-end manner: the *physics* that grounds an observation in the absolute physical world, the *geometry* that describes underlying 3D spatial structure, and the *appearance* that we finally see from the visual content manifested in images and sequences. To address the aforementioned challenges, we first introduce the Omni-Camera representation, a dense action signal that combines a gravity-anchored absolute field with a relative ray field, which enables physics-grounded text-to-image spatial simulation (single-view) and image-to-3D scene generation (cross-view). Subsequently, we propose physics propagation: by

anchoring the absolute spatial knowledge derived from world perception and propagating it across the relative control signals of future frames, Puffin-World generates appearance that remains stable and gravity-consistent over complex motions. Furthermore, we scale the whole framework with the constructed **Puffin-16M**, a new dataset of 15M vision–language–camera triplets and 1M trajectories featuring diverse and challenging camera motions with precise labels.

Thanks to this unified paradigm, Puffin-World supports diverse multimodal tasks as illustrated in Fig. 1. For physical world perception, it attains state-of-the-art camera-to-world understanding, achieving the best median error and the best AUC at $5^\circ$ across four public benchmarks. For free-viewpoint spatial simulation, it generates camera-controllable images with more faithful distributions than strong general-purpose generators (OpenAI, 2026; Google, 2026). For 3D world modeling, it performs high-DoF action-conditioned, text/image-to-3D generation while jointly reconstructing per-view geometry, enabling consistent and physically grounded world modeling. Furthermore, Puffin-World enables complicated, closed-loop applications that require multi-task synergy, such as mimic and self-calibrated world exploration. We summarize our contributions as follows:

- We propose Puffin-World, a unified multimodal model that formulates native 3D world states and jointly perceives, generates, and reconstructs the 3D world within a single framework.
- We introduce the Omni-Camera representation, a unified dense camera condition that integrates gravity-aware absolute orientation with ray-based relative geometry. A physics propagation mechanism is further proposed to anchor absolute spatial knowledge from perception and propagate it across the subsequent control signals.
- We construct Puffin-16M, comprising Puffin-Cam-15M (15M vision–language–camera triplets) and Puffin-Traj-1M (1M challenging-motion trajectories). Moreover, leveraging Puffin-World's accurate understanding of absolute camera physics, we annotated 28 widely used public datasets. These annotations cover approximately 44.5 million images across diverse data distributions.

## 2 METHOD

Puffin-World unifies physical-world perception, free-viewpoint spatial simulation, and 3D world generation and reconstruction within a single multimodal framework. Its versatility arises from three levels of unification rather than task-specific subnetworks. *(i) Representation unification.* Appearance and geometry share a common latent space, with RGB images and depth maps encoded by the same VAE, while all camera configurations and motions, ranging from in-place rotation to large translation, are represented by the unified Omni-Camera representation $\mathbf{C}$ defined in Section A0.2 of the Appendix. *(ii) Modality unification.* A single backbone both perceives and generates the world: the geometry-aligned vision encoder and LLM yield autoregressive understanding outputs, while the same LLM hidden states are transformed by learnable queries and a lightweight connector into conditioning signals for the diffusion generator, enabling perception and generation as two complementary outputs of one model. *(iii) Task unification.* Each task is determined by the available inputs, including text, target cameras, reference views, together with the role mask $\mathbf{m} \in \mathbb{R}^{4\times H\times W}$. By varying this input composition, the same framework supports diverse tasks. The overview of Puffin-World's framework is illustrated in Figure 2. We detail these capabilities below.

### 2.1 PHYSICS PERCEPTION

Given a single image, Puffin-World estimates the camera's absolute physical state, including its gravity-relative orientation specified by roll and pitch, vertical field-of-view (FoV), and a semantic scene description. Following Puffin (Liao et al., 2026), we formulate this task as autoregressive multi-modal sequence modeling rather than direct regression. The geometry-aligned vision encoder (Liao et al., 2026) extracts visual features from images, which are projected into the LLM embedding space through an MLP projector. The LLM then generates a structured scene and spatial analysis, followed by the numerical camera parameters, and is trained using a next-token cross-entropy loss.

This formulation offers two key advantages for physical-state perception. First, it leverages the LLM's sequence modeling capability and favorable scaling behavior, recasting fine-grained camera understanding as language modeling. Second, because the camera parameters are predicted after reasoning about scene-level cues such as the horizon, vertical structures, and foreground composition,

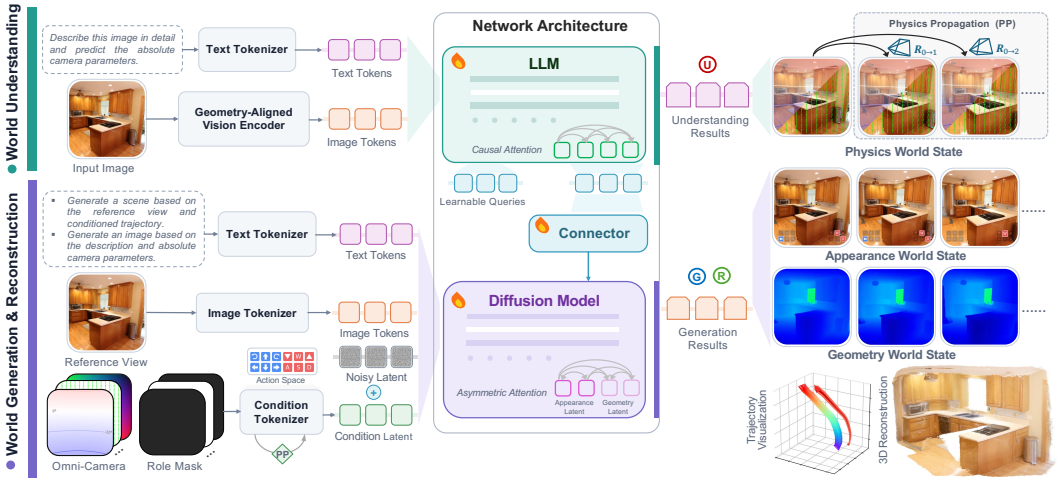


Figure 2: **The network architecture of our Puffin-World.** It mainly comprises 3D world understanding, generation, and reconstruction, formulating native 3D world states within one framework.

the estimation relies on holistic scene understanding rather than low-level visual cues alone. This design enables robust gravity-aware perception across various scenes.

### 2.2 Free-Viewpoint Spatial Simulation

Puffin-World enables camera-controllable image generation: given a text prompt and a desired Omni-Camera map $\mathbf{C}$, it synthesizes an image whose realized viewpoint and intrinsics strictly adhere to $\mathbf{C}$. Because this part focuses on single-view generation, the ray map within the Omni-Camera representation is held constant. The resulting query hidden states pass through a lightweight connector, yielding a pooled conditioning vector and a sequence of joint-attention conditioning embeddings for the multimodal diffusion transformer (MMDiT) (Esser et al., 2024). The MMDiT then denoises a target latent under a flow-matching objective, after which a VAE decoder maps the denoised latent back into pixel space.

Our approach to integrating camera condition is the primary departure from Puffin (Liao et al., 2026). Puffin encodes a 3-channel Perspective Field into a continuous latent via a VAE and injects it into the image latent via cross-attention, limiting its extension to broader tasks and camera motions. Moreover, because a holistic camera representation cannot be formulated with only 3 channels, it cannot be straightforwardly encoded and injected in this manner. Instead, we fuse the Omni-Camera representation directly within the diffusion latent space. A lightweight condition fusion module $\mathcal{F}$ maps $\mathbf{C}$ and the role mask $\mathbf{m}$ into the input latent space of MMDiT. This output is added to the noisy image latent $\mathbf{z}$ prior to the patch embedding operation $\mathcal{P}$:

$$\mathbf{h}^0 = \mathcal{P}\big(\mathbf{z} + \mathcal{F}([\mathbf{C}; \mathbf{m}])\big). \tag{1}$$

To preserve the pretrained generator's behavior at the start of training, $\mathcal{F}$ is initialized such that its contribution starts near zero. Furthermore, to ensure the camera signal remains effective across deeper layers, we re-inject the patchified camera features at a sparse set of Transformer blocks $\mathcal{L}$:

$$\mathbf{h}^l \leftarrow \mathbf{h}^l + \mathbf{W}_l\, \mathcal{P}\big(\mathcal{F}([\mathbf{C}; \mathbf{m}])\big), \quad l \in \mathcal{L}. \tag{2}$$

Compared to using a channel-limited mechanism over a separately VAE-encoded camera branch, this pixel-aligned additive injection supports versatile tasks. It grounds every latent token in its precise camera geometry across multiple layers, providing superior and stable spatial control while introducing negligible parameter overhead.

### 2.3 3D World Modeling

Puffin-World then extends the single-view generator to multi-view trajectories while jointly supporting geometry reconstruction. Instead of denoising a single target image, the model processes a set of

$T$ views within a unified joint-attention sequence. Among them, $K$ reference views are provided as clean latents $\{\mathbf{z}_v^{\mathrm{ref}}\}$ and are also encoded by the geometry-aligned vision encoder and the LLM to provide semantic conditioning. The remaining $T-K$ target views, represented by $\{\mathbf{z}_v^{\mathrm{tgt}}\}$, are corrupted with noise and jointly denoised under the flow-matching objective. Each view, whether reference or target, is associated with its own Omni-Camera condition, which is fused into the corresponding image latent through the additive injection mechanism in Eq. 1 and Eq. 2.

**Physics Propagation.** A trajectory introduces a challenge specific to absolute grounding: relative motion across views can be readily obtained from camera poses, whereas the absolute orientation that anchors the trajectory to the real world is generally unavailable. The Omni-Camera representation addresses this through its complementary components: the relative ray map captures inter-view motion, while the absolute Perspective Field encodes gravity-aligned orientation. We recover the latter for future frames through *physics propagation*. Specifically, the model first perceives the absolute state $(\phi_0, \theta_0, \alpha_0, \dots)$ of the reference view through the Physics Perception pathway (Sec. 2.1), where $\phi_0$, $\theta_0$, and $\alpha_0$ denote its roll, pitch, and vertical FoV, respectively, yielding the gravity direction $\mathbf{g}_0$ in the reference frame. Given the relative rotation $\mathbf{R}^{\mathrm{rel}}_{t\leftarrow 0}$ that transforms vectors from the reference camera coordinate frame to the camera coordinate frame of view $t$, the gravity direction in each view is propagated as:

$$\mathcal{G} = \{\mathbf{g}_t\}_{t=0}^{T-1} = \left\{\mathbf{R}^{\mathrm{rel}}_{t\leftarrow 0}\mathbf{g}_0\right\}_{t=0}^{T-1}. \tag{3}$$

The propagated gravity direction of each view is then used to render its corresponding absolute camera representation, which is concatenated with the relative ray map to form the complete Omni-Camera condition. In this way, all views along the trajectory share a coherent gravity-anchored world frame while retaining their individual relative motions.

**Joint Appearance–Geometry Modeling.** Puffin-World predicts not only the appearance but also the geometry of each generated view, casting reconstruction as part of the same generative process. In particular, we encode depth as an RGB image using an invertible color mapping inspired by the 3D Hilbert curve in Vision Banana (Gabeur et al., 2026), and encode it with the same frozen VAE used for appearance. Instead of directly regressing unbounded depth or disparity, this gives us a bounded three-channel representation that fits the pretrained VAE without changing its input or output interface. The mapping is deterministic and invertible, so the generated result can be decoded back to depth. Its nonlinear transform also allocates more of the color range to nearby surfaces, where geometric accuracy matters most. This allows appearance and geometry to use the same latent representation without requiring a dedicated depth encoder or decoder. For each modeled view, the depth latent is appended as an additional token block that shares the Omni-Camera condition and view-axis index of its RGB counterpart, distinguished only by the geometry channel of the role mask $\mathbf{m}$. A single denoising pass therefore jointly yields the appearance and geometry latents of all target views under the shared flow-matching objective:

$$\mathcal{L}_{\mathrm{fm}}(\mathcal{V}) = \mathbb{E}_{t,\,\boldsymbol{\epsilon}}\left[\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left\|\mathbf{v}_\theta(\mathbf{z}_t^v, t, \mathbf{c}^v) - (\boldsymbol{\epsilon}^v - \mathbf{z}_0^v)\right\|_2^2\right], \tag{4}$$

where $\mathbf{z}_t^v = (1-\sigma_t)\mathbf{z}_0^v + \sigma_t\boldsymbol{\epsilon}^v$, $\mathbf{c}^v$ collects the per-view camera and connector conditioning, and $\mathcal{V}$ is a set of target views. Because the saturated colors of the geometry latents could otherwise corrupt the converged appearance model, we do not simply pool appearance and geometry targets into one objective. Instead, we evaluate Eq. 4 separately over the appearance target views $\mathcal{V}_{\mathrm{rgb}}$ and the geometry target views $\mathcal{V}_{\mathrm{geo}}$, and combine them with a time-dependent weight:

$$\mathcal{L} = \mathcal{L}_{\mathrm{fm}}(\mathcal{V}_{\mathrm{rgb}}) + \omega(\tau)\,\mathcal{L}_{\mathrm{fm}}(\mathcal{V}_{\mathrm{geo}}), \omega(\tau) = \omega_{\max}\,\min\Big(1, \tfrac{\tau}{\tau_0}\Big), \tag{5}$$

where $\tau$ is the training iteration and $\omega(\tau)$ ramps linearly from 0 to $\omega_{\max}$ over the first $\tau_0$ iterations. Normalizing each modality by its own view count keeps the appearance gradient magnitude stable, while the ramp starts the geometry weight at zero so that early depth gradients do not perturb the shared modulation and output layers.

To activate the geometry branch on a converged appearance model without degrading visual quality, we introduce two safeguards that preserve the appearance pathway. First, we adopt *asymmetric* attention as shown in Figure 2: appearance and text queries cannot attend to geometry tokens, while geometry tokens attend to all modalities. This prevents geometry features from leaking into the

appearance stream while retaining full scene context for geometry prediction. Second, we add a zero-initialized learnable geometry-modality embedding after patch embedding. This keeps the model functionally identical to the original multi-view generator at initialization and gradually introduces geometry-specific features during training. At inference, appearance latents are decoded into images, whereas geometry latents are decoded and inverse-mapped from Hilbert color space to scalar depth.

# 3 EXPERIMENTS

## 3.1 IMPLEMENTATION DETAILS

We develop two architectural variants, namely Puffin-World-Base and Puffin-World-Pro, targeting different trade-offs between physical world understanding and world simulation capabilities. Specifically, Puffin-World-Base is initialized with the pretrained C-RADIOv3-H (Heinrich et al., 2025), Qwen2.5-7B-Instruct (Qwen et al., 2024), and SD3.5-Medium-2.5B (Stability AI, 2024) as its geometry-aware vision encoder, large language model, and diffusion backbone, respectively. To further enhance the generative fidelity, Puffin-World-Pro adopts the pretrained C-RADIOv4-H (Ranzinger et al., 2026), Qwen2.5-1.5B-Instruct (Qwen et al., 2024), and SD3.5-Large-8.1B (Stability AI, 2024) for the corresponding components. For both variants, a set of 64 learnable queries and a lightweight six-layer Transformer connector are used to project the LLM hidden states into diffusion conditioning embeddings. In addition to the above two variants, we also release Puffin-World-Caption, an expert model dedicated to annotating physical world states. To support large-scale dataset construction, Puffin-World-Caption is built upon the pretrained C-RADIOv3-H (Heinrich et al., 2025) and Qwen3.5-0.8B-Instruct (Team, 2026), providing an effective balance between annotation accuracy and inference efficiency for large-scale physical-world labeling. More details of training and inference are presented in Sec A0.4.3 of the Appendix.

## 3.2 CAMERA-TO-WORLD UNDERSTANDING RESULTS

Following prior works, we compare Puffin-World against a diverse set of camera-to-world understanding methods, including learning-based approaches such as DeepCalib (Lopez et al., 2019), Perceptual (Hold-Geoffroy et al., 2018), CTRL-C (Lee et al., 2021), MSCC (Song et al., 2024), ParamNet (Jin et al., 2023), GeoCalib (Veicht et al., 2024), and Puffin (Liao et al., 2026), as well as classical geometric methods including SVA (Lochman et al., 2021) and UVP (Pautrat et al., 2023). For each image, gravity estimation performance is measured by the angular errors of roll and pitch, while camera intrinsic estimation is evaluated using the error in vertical (vFoV). Following standard practice, we report both the median error and the Area Under the Recall Curve (AUC) at error thresholds of $1^\circ$, $5^\circ$, and $10^\circ$. Evaluations are conducted on four widely adopted benchmarks: MegaDepth (Li & Snavely, 2018), TartanAir (Wang et al., 2020), LaMAR (Sarlin et al., 2022), and Stanford2D3D (Armeni et al., 2017). As shown in Table 1, Puffin-World consistently outperforms all competing methods in terms of median error and achieves the best performance on the majority of AUC metrics across all four benchmarks. These results demonstrate its strong capability to perceive and ground observations within the physical world. In particular, Puffin-World surpasses Puffin (Liao et al., 2026) with a noticeable margin, showing the meaningful and progressive benefits of scaling the model and data for this task. We note that AnyCalib (Tirado-Garín & Civera, 2025) is specifically designed for camera intrinsic estimation, particularly FoV prediction, and is included only as a specialized reference baseline. Despite its task-specific design, Puffin-World still outperforms AnyCalib on most benchmarks, highlighting the effectiveness and generalizability of our unified framework.

## 3.3 CAMERA-CONTROLLABLE GENERATION RESULTS

We compare Puffin-World against recent state-of-the-art multimodal generation models, including GPT Image2 (OpenAI, 2026), Nano Banana 2 (Google, 2026), Qwen-Image2-Pro (Zhao et al., 2026), FLUX.2-dev (Labs, 2025), and Z-Image (Cai et al., 2025), as well as specialized camera-controllable image generation methods (Bernal-Berdun et al., 2025; Liao et al., 2026). To quantitatively assess camera controllability, we leverage Puffin-World to estimate the camera parameters of each generated image and derive the corresponding pixel-wise perspective fields. We then compare these estimated fields against the GT fields and report both mean and median angular errors for the up vector, gravity direction, and latitude map, all measured in degrees. In addition, we report the Fréchet Inception

Table 1: **Evaluation results on camera-to-world understanding.**

| | Approach | Roll [degrees] | | | | Pitch [degrees] | | | | FoV [degrees] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | error ↓ | AUC ▷ 1/5/10° ↑ | | | error ↓ | AUC ▷ 1/5/10° ↑ | | | error ↓ | AUC ▷ 1/5/10° ↑ | | |
| **Stanford2D3D** | DeepCalib (Lopez et al., 2019) | 1.59 | 33.8 | 63.9 | 79.2 | 2.58 | 21.6 | 46.9 | 65.7 | 6.67 | 8.1 | 20.6 | 37.6 |
| | Perceptual (Hold-Geoffroy et al., 2018) | 2.08 | 26.8 | 53.8 | 70.7 | 3.17 | 21.5 | 41.8 | 57.8 | 13.84 | 2.8 | 7.7 | 16.1 |
| | CTRL-C (Lee et al., 2021) | 3.04 | 23.2 | 43.0 | 56.9 | 3.43 | 18.3 | 38.6 | 53.8 | 8.50 | 7.7 | 18.2 | 31.5 |
| | MSCC (Song et al., 2024) | 3.43 | 13.5 | 36.8 | 57.3 | 2.64 | 22.6 | 45.0 | 60.5 | 5.81 | 9.6 | 23.8 | 41.6 |
| | ParamNet (Jin et al., 2023) | 1.14 | 44.6 | 73.9 | 84.8 | 1.94 | 29.2 | 56.7 | 73.1 | 9.01 | 5.8 | 14.3 | 27.8 |
| | SVA (Lochman et al., 2021) | - | 21.7 | 24.6 | 25.8 | - | 15.4 | 19.9 | 22.4 | - | 6.2 | 11.5 | 15.2 |
| | UVP (Pautrat et al., 2023) | 0.52 | 65.3 | 74.6 | 79.1 | 0.95 | 51.2 | 63.0 | 69.2 | 3.65 | 22.2 | 39.5 | 51.3 |
| | GeoCalib (Veicht et al., 2024) | 0.40 | 83.1 | 91.8 | 94.8 | 0.93 | 52.3 | 74.8 | 84.6 | 3.21 | 17.4 | 40.0 | 59.4 |
| | AnyCalib† (Tirado-Garín & Civera, 2025) | - | - | - | - | - | - | - | - | 2.55 | 21.1 | 46.8 | 64.6 |
| | **Puffin-World** | **0.29** | **93.1** | **97.4** | **98.5** | **0.53** | **73.3** | **88.8** | **94.0** | **1.62** | **34.5** | **61.1** | **76.4** |
| **MegaDepth** | DeepCalib (Lopez et al., 2019) | 1.41 | 34.6 | 65.4 | 79.4 | 5.19 | 11.9 | 27.8 | 44.8 | 11.14 | 5.6 | 12.1 | 22.9 |
| | Perceptual (Hold-Geoffroy et al., 2018) | 1.07 | 47.9 | 72.4 | 83.2 | 3.49 | 19.8 | 39.1 | 54.2 | 13.40 | 2.9 | 8.2 | 16.8 |
| | CTRL-C (Lee et al., 2021) | 0.88 | 54.5 | 75.0 | 84.2 | 4.80 | 16.6 | 33.2 | 46.5 | 18.65 | 2.0 | 5.8 | 12.8 |
| | MSCC (Song et al., 2024) | 0.90 | 53.1 | 72.8 | 82.1 | 5.73 | 19.0 | 33.2 | 44.3 | 10.80 | 6.0 | 14.6 | 26.2 |
| | ParamNet (Jin et al., 2023) | 1.17 | 43.4 | 70.7 | 82.2 | 3.99 | 15.4 | 34.5 | 53.3 | 11.01 | 3.2 | 10.1 | 21.3 |
| | SVA (Lochman et al., 2021) | - | 31.9 | 35.0 | 36.2 | - | 13.6 | 20.6 | 24.9 | - | 9.4 | 16.1 | 21.1 |
| | UVP (Pautrat et al., 2023) | 0.51 | 69.2 | 81.6 | 86.9 | 4.59 | 21.6 | 36.2 | 47.4 | 10.92 | 8.2 | 18.7 | 29.8 |
| | GeoCalib (Veicht et al., 2024) | 0.36 | 82.6 | 90.6 | 94.0 | 1.94 | 32.4 | 53.3 | 67.5 | 4.46 | 13.6 | 31.7 | 48.2 |
| | AnyCalib† (Tirado-Garín & Civera, 2025) | - | - | - | - | - | - | - | - | 3.14 | 19.4 | 40.8 | 59.1 |
| | Puffin (Liao et al., 2026) | 0.32 | 84.9 | 93.4 | 96.2 | 1.08 | 47.6 | 68.2 | 79.4 | 2.42 | 23.9 | 47.8 | 64.1 |
| | **Puffin-World** | **0.28** | **87.7** | **94.2** | **96.7** | **1.01** | **49.7** | **69.3** | **79.6** | **2.41** | **24.8** | **48.8** | **65.5** |
| **TartanAir** | DeepCalib (Lopez et al., 2019) | 1.95 | 24.7 | 55.4 | 71.5 | 3.27 | 16.3 | 38.8 | 58.5 | 8.07 | 1.5 | 8.8 | 27.2 |
| | Perceptual (Hold-Geoffroy et al., 2018) | 2.24 | 23.2 | 48.6 | 66.7 | 2.86 | 23.5 | 44.6 | 61.5 | 15.06 | 5.1 | 8.9 | 17.1 |
| | CTRL-C (Lee et al., 2021) | 1.68 | 32.8 | 59.1 | 74.1 | 2.39 | 24.6 | 48.6 | 65.2 | 5.64 | 10.7 | 25.4 | 43.5 |
| | MSCC (Song et al., 2024) | 3.50 | 15.0 | 37.2 | 57.7 | 3.48 | 18.8 | 38.6 | 54.3 | 11.18 | 4.4 | 11.8 | 23.0 |
| | ParamNet (Jin et al., 2023) | 1.63 | 34.5 | 59.2 | 73.9 | 3.05 | 19.4 | 42.0 | 60.3 | 8.21 | 6.0 | 16.8 | 31.6 |
| | SVA (Lochman et al., 2021) | 9.48 | 32.4 | 39.6 | 44.1 | 18.46 | 21.2 | 28.8 | 34.5 | 43.01 | 8.8 | 16.1 | 21.6 |
| | UVP (Pautrat et al., 2023) | 0.89 | 52.1 | 64.8 | 71.9 | 2.48 | 36.2 | 48.8 | 58.6 | 9.15 | 15.8 | 25.8 | 35.7 |
| | GeoCalib (Veicht et al., 2024) | 0.43 | 71.3 | 83.8 | 89.8 | 1.49 | 38.2 | 62.9 | 76.6 | 4.90 | 14.1 | 30.4 | 47.6 |
| | AnyCalib† (Tirado-Garín & Civera, 2025) | - | - | - | - | - | - | - | - | 3.62 | 15.5 | 36.4 | 55.1 |
| | Puffin (Liao et al., 2026) | 0.40 | 71.7 | 86.2 | 92.1 | 0.95 | 51.0 | 68.2 | 79.3 | 7.48 | 16.3 | 28.5 | 39.0 |
| | **Puffin-World** | **0.31** | **80.1** | **90.0** | **94.2** | **0.67** | **60.2** | **78.1** | **87.2** | **2.34** | **26.6** | **46.4** | **59.5** |
| **LaMAR** | DeepCalib (Lopez et al., 2019) | 1.15 | 44.1 | 73.9 | 84.8 | 4.68 | 10.8 | 28.3 | 49.8 | 10.93 | 0.7 | 13.0 | 24.0 |
| | Perceptual (Hold-Geoffroy et al., 2018) | 1.29 | 40.0 | 68.9 | 81.6 | 2.83 | 21.2 | 44.7 | 62.6 | 17.78 | 3.0 | 5.3 | 10.7 |
| | CTRL-C (Lee et al., 2021) | 1.20 | 43.5 | 70.9 | 82.5 | 1.94 | 27.6 | 54.7 | 70.2 | 5.64 | 9.8 | 24.6 | 43.2 |
| | MSCC (Song et al., 2024) | 1.44 | 39.6 | 60.7 | 72.8 | 3.02 | 20.9 | 41.8 | 55.7 | 14.78 | 3.2 | 8.3 | 16.8 |
| | ParamNet (Jin et al., 2023) | 0.93 | 51.7 | 77.0 | 86.0 | 2.15 | 27.0 | 52.7 | 70.2 | 14.71 | 2.8 | 6.8 | 14.3 |
| | SVA (Lochman et al., 2021) | - | 8.6 | 9.2 | 9.7 | - | 3.4 | 5.7 | 7.0 | - | 1.2 | 2.7 | 4.1 |
| | UVP (Pautrat et al., 2023) | 0.38 | 72.7 | 81.8 | 85.7 | 1.34 | 42.3 | 59.9 | 69.4 | 5.57 | 15.6 | 30.6 | 43.5 |
| | GeoCalib (Veicht et al., 2024) | 0.28 | **86.4** | **92.5** | 95.0 | 0.87 | 55.0 | 76.9 | 86.2 | 3.03 | **19.1** | 41.5 | **60.0** |
| | AnyCalib† (Tirado-Garín & Civera, 2025) | - | - | - | - | - | - | - | - | 2.25 | 24.6 | 51.6 | 70.5 |
| | Puffin (Liao et al., 2026) | 0.38 | 80.6 | 89.8 | 93.5 | **0.71** | 61.7 | 78.9 | 86.4 | 3.62 | 17.0 | 37.3 | 53.1 |
| | **Puffin-World** | **0.26** | 85.8 | **92.5** | **95.3** | **0.71** | **63.5** | **81.2** | **88.2** | **2.73** | 19.0 | **43.0** | 59.4 |

Table 2: **Camera-controllable generation evaluation on Puffin-Cam-Bench.**

| Approach | Up Vector [degrees] | | Latitude [degrees] | | Gravity [degrees] | | Visual Quality |
|---|---|---|---|---|---|---|---|
| | mean error ↓ | median error ↓ | mean error ↓ | median error ↓ | mean error ↓ | median error ↓ | FID ↓ |
| GPT Image2 | 24.13 | 23.44 | 16.42 | 15.98 | 28.83 | 29.38 | 99.36 |
| Nano Banana2 | 24.01 | 23.19 | 15.98 | 15.52 | 28.00 | 26.92 | 90.11 |
| Qwen-Image2-Pro | 24.12 | 23.51 | 16.45 | 15.53 | 28.57 | 28.69 | 98.51 |
| FLUX.2-dev | 24.28 | 23.62 | 15.75 | 15.24 | 28.31 | 28.36 | 97.06 |
| Z-Image | 24.64 | 24.04 | 16.96 | 16.09 | 29.41 | 28.82 | 98.33 |
| PreciseCam | 13.37 | 12.61 | 12.31 | 12.01 | 17.07 | 14.01 | 90.89 |
| Puffin | 3.86 | 3.48 | 4.77 | 4.43 | 4.92 | 2.87 | 80.29 |
| **Puffin-World** | **0.96** | **0.84** | **1.34** | **1.26** | **1.32** | **0.79** | **75.93** |

Distance (FID) to evaluate the overall visual fidelity and realism of the generated images. Since no existing benchmark provides paired text descriptions and precise camera parameters across diverse camera configurations and aspect ratios, we introduce Puffin-Cam-Bench to fill this gap. It contains 600 text–camera specification pairs covering a wide range of scenes, viewpoints, camera poses, fields of view, and aspect ratios, enabling comprehensive evaluation of camera-controllable generation for free-viewpoint spatial simulation.

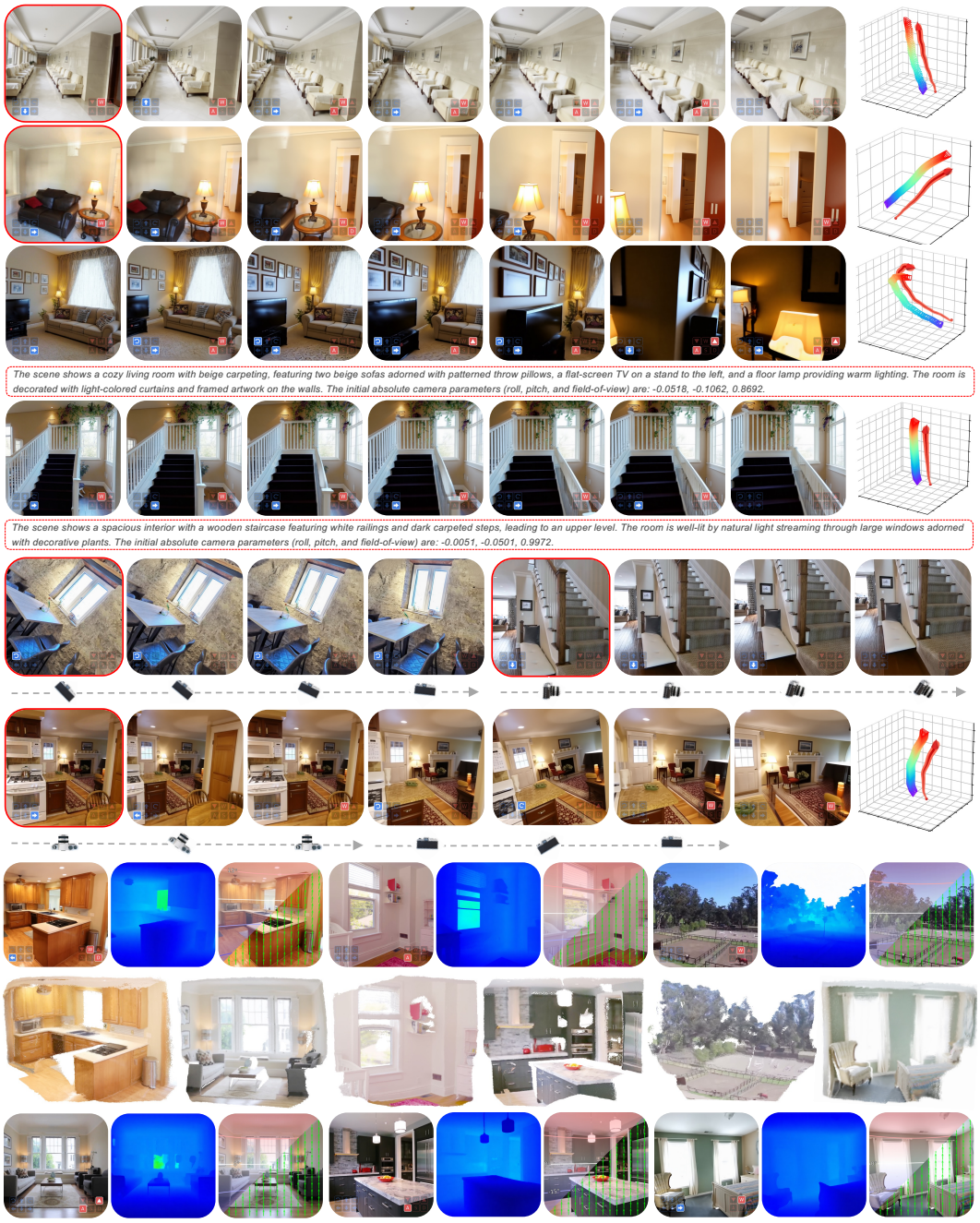


Figure 3: **Qualitative results of 3D world modeling.** Our Puffin-World supports various 3D modeling tasks: image-to-3D generation, text-to-3D generation, flexible rotation control, compound motion control, native 3D world states formulation, and 3D reconstruction.

Table 2 summarizes the quantitative results, while Figure A9 (in Appendix) provides qualitative comparisons across different methods. Our method consistently outperforms existing multimodal generation models by a substantial margin across all evaluation metrics. Although these general-purpose models can produce visually appealing and high-fidelity images, they often fail to preserve spatially consistent scene geometry under specified camera configurations, resulting in noticeable deviations from the desired viewpoints. In contrast, Puffin-World achieves accurate and robust camera control while maintaining high visual quality and scene diversity. Its strong generalization across a wide range of scenes, viewpoints, and camera configurations highlights its practicality for real-world camera-controllable image generation beyond prior specialized models.

Table 3: **Evaluation results on 3D world generation.** Note that camera-control metrics are not reported on RealEstate10K due to the absence of corresponding absolute-camera ground truth. †MVGenMaster's training data include samples from the evaluated RealEstate10K benchmark.

| | Approach | Visual Fidelity | | | Roll [degrees] | | | | Pitch [degrees] | | | | FoV [degrees] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | error↓ | AUC▷1/5/10°↑ | | | error↓ | AUC▷1/5/10°↑ | | | error↓ |
| RealEstate10K | MotionCtrl Wang et al. (2024c) | 13.38 | 0.482 | 0.507 | – | – | – | – | – | – | – | – | – |
| | CameraCtrl He et al. (2024) | 16.93 | 0.589 | 0.352 | – | – | – | – | – | – | – | – | – |
| | ViewCrafter Yu et al. (2025) | 16.04 | 0.569 | 0.331 | – | – | – | – | – | – | – | – | – |
| | SEVA Zhou et al. (2025) | 14.84 | 0.529 | 0.421 | – | – | – | – | – | – | – | – | – |
| | MVGenMaster Cao et al. (2024)† | 17.11 | 0.591 | 0.348 | – | – | – | – | – | – | – | – | – |
| | **Puffin-World** | **17.22** | **0.595** | **0.318** | – | – | – | – | – | – | – | – | – |
| Puffin-Traj-Bench | MotionCtrl Wang et al. (2024c) | 12.53 | 0.438 | 0.604 | 3.20 | 0.25 | 0.43 | 0.54 | 2.15 | 0.28 | 0.49 | 0.59 | 10.04 |
| | CameraCtrl He et al. (2024) | 14.28 | 0.512 | 0.532 | 11.79 | 0.05 | 0.11 | 0.21 | 10.81 | 0.05 | 0.16 | 0.26 | 45.84 |
| | ViewCrafter Yu et al. (2025) | 14.52 | 0.525 | 0.510 | 1.92 | 0.32 | 0.53 | 0.66 | 2.89 | 0.23 | 0.43 | 0.58 | 6.43 |
| | SEVA Zhou et al. (2025) | 17.94 | **0.614** | 0.307 | 1.35 | 0.38 | 0.63 | 0.75 | 1.76 | 0.29 | 0.56 | 0.71 | 12.59 |
| | MVGenMaster Cao et al. (2024) | 13.64 | 0.496 | 0.589 | 9.24 | 0.06 | 0.14 | 0.28 | 9.43 | 0.08 | 0.17 | 0.30 | 26.98 |
| | **Puffin-World** | **18.00** | 0.613 | **0.288** | **0.80** | **0.57** | **0.79** | **0.88** | **1.10** | **0.45** | **0.72** | **0.83** | **2.96** |

### 3.4 3D World Modeling Results

As visualized in Figure 3, Puffin-World supports versatile 3D world modeling. For text-to-3D generation, it first synthesizes a reference view from the text prompt and initial absolute camera parameters, and then generates subsequent views conditioned on the reference image and target camera trajectory; for compound motion, translation and rotational trajectories are sequentially combined to cover diverse roll, pitch, and yaw variations. Across long trajectories, extreme rotations, and complex compound motions, the predicted native 3D world states—appearance, geometry, and physics—remain visually realistic, physically and geometrically plausible, and spatially consistent, while the resulting reconstructions preserve clear scene structure and coherent cross-view geometry. We further quantitatively evaluate Puffin-World on RealEstate10K (50 clips) and Puffin-Traj-Bench (100 clips), following prior work (Cao et al., 2024), against MotionCtrl (Wang et al., 2024c), CameraCtrl (He et al., 2024), ViewCrafter (Yu et al., 2025), SEVA (Zhou et al., 2025), and MVGenMaster (Cao et al., 2024). Visual fidelity is measured by PSNR, SSIM, and LPIPS, while camera-control accuracy on Puffin-Traj-Bench is assessed by re-estimating roll, pitch, and vertical FoV from generated frames using the same Puffin-World camera-understanding branch for all methods, reporting median angular error and AUC@$1/5/10°$. As shown in Table 3, Puffin-World achieves the best PSNR, SSIM, and LPIPS on RealEstate10K, and on the more challenging Puffin-Traj-Bench achieves the best PSNR and LPIPS with competitive SSIM while substantially outperforming all baselines in roll, pitch, and FoV control, demonstrating strong perceptual fidelity together with accurate and robust camera control under challenging spatial transformations.

## 4 Conclusion

We present Puffin-World, a unified multimodal model that perceives, simulates, and generates the physical world through three native 3D world states: physics, geometry, and appearance. In particular, the Omni-Camera representation integrates gravity-aware absolute orientation with ray-based relative motion into a unified camera condition, supporting single-view control, multi-view synthesis, and challenging camera trajectories. Building on this representation, physics propagation transfers the natively perceived absolute state of the reference view across future frames, maintaining a coherent physical frame for stable and self-calibrated world generation. Within a single framework, Puffin-World jointly supports physical-world perception, free-viewpoint spatial simulation, and 3D world generation and reconstruction, while Puffin-16M provides large-scale vision–language–camera supervision and diverse challenging motions for scaling these capabilities. Extensive experiments demonstrate strong and consistent performance across these tasks and further highlight the benefits of multi-task synergy for closed-loop applications.

We believe that grounding multimodal tasks in a shared, physically anchored world representation provides a promising path toward versatile spatial intelligence and physical AI. Future directions include extending the framework to dynamic scenes and longer horizons, and modeling richer physical states beyond gravity and latitude.

## AI USE STATEMENT

Generative AI tools were used solely for limited language editing and polishing to improve the clarity and readability of the manuscript. They were not used for research ideation, methodology or experiment design, implementation, data processing or analysis, result interpretation, or the formulation of scientific claims or conclusions. All AI-assisted text was carefully reviewed and edited by the authors, who take full responsibility for the final content of the paper.

## REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our work. The main paper provides a comprehensive description of the proposed framework, while the appendix includes further details on implementation, dataset construction, training configurations, and evaluation protocols. To facilitate future research in multimodal spatial intelligence, we will release the source code, trained models, dataset construction pipeline, and evaluation benchmark.

## ETHICS STATEMENT

We affirm our adherence to the ICLR Code of Ethics and have conducted this research with integrity and responsibility. This work does not involve human subjects, sensitive personal data, or applications designed for harmful purposes. We aim to contribute positively to the research community by releasing resources that promote openness, reproducibility, and responsible progress in multimodal spatial intelligence.

## 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.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

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.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 2724–2733, 2025.

Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pp. 1–10, 2018.

Ollin Boer Bohan. Megalith-10m. `https://huggingface.co/datasets/madebyollin/megalith-10m`, June 2024. Hugging Face dataset.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Chenjie Cao, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. *arXiv preprint arXiv:2411.16157*, 2024.

Keshigeyan Chandrasegaran, Kyle Sargent, Suchir Agarwal, Michael Jang, Michael Poli, Juan Carlos Niebles, Justin Johnson, Jiajun Wu, and Li Fei-Fei. Gpic: A giant permissive image corpus for visual generation. *arXiv preprint arXiv:2605.30341*, 2026.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567, 2021.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *ICLR*, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Junyuan Deng, Wei Yin, Xiaoyang Guo, Qian Zhang, Xiaotao Hu, Weiqiang Ren, Ping Tan, et al. Boost 3d reconstruction using diffusion-based monocular camera calibration. *arXiv preprint arXiv:2411.17240*, 2024.

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 *International Conference on Machine Learning*, 2024.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024.

Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *CVPR*, 2019.

Google. Nano banana 2, 2026. URL `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`.

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

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, et al. Vision as unified multimodal generation. *arXiv preprint arXiv:2607.06560*, 2026.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Xiankang He, Guangkai Xu, Bo Zhang, Hao Chen, Ying Cui, and Dongyan Guo. Diffcalib: Reformulating monocular camera calibration as diffusion-based dense incident map generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3428–3436, 2025.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22487–22497, 2025.

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2354–2363, 2018.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830, 2018.

Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025a.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025b.

Sebastian Janampa and Marios Pattichis. Sofi: Multi-scale deformable transformer for camera calibration with enhanced line queries. *arXiv preprint arXiv:2409.15553*, 2024.

Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024.

Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Liu Qin, and Lei Zhang. Referring to any person. In *ICCV*, 2025.

Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17307–17316, 2023.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2938–2946, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *ICLR*, 2026.

Jinwoo Lee, Minhyuk Sung, Hyunjoon Lee, and Junho Kim. Neural geometric parser for single image camera calibration. In *European Conference on Computer Vision*, pp. 541–557. Springer, 2020.

Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. Ctrl-c: Camera calibration transformer with line-classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16228–16237, 2021.

Wanhee Lee, Klemen Kotar, Rahul Venkatesh, Jared Watrous, Honglin Chen, Khai Loong Aw, and Daniel Yamins. Unified 3d scene understanding through physical world modeling. In *ICLR*, 2026.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019a.

Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. *arXiv preprint arXiv:1809.00716*, 2018.

Xiaoyu Li, Bo Zhang, Pedro V Sander, and Jing Liao. Blind geometric distortion correction on images through deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4855–4864, 2019b.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2041–2050, 2018.

Kang Liao, Chunyu Lin, Yao Zhao, and Mai Xu. Model-free distortion rectification framework bridged by distortion distribution map. *IEEE Transactions on Image Processing*, 29:3707–3718, 2020.

Kang Liao, Chunyu Lin, and Yao Zhao. A deep ordinal distortion estimation approach for distortion rectification. *IEEE Transactions on Image Processing*, 30:3362–3375, 2021.

Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin, Jing Zhang, Yao Zhao, Moncef Gabbouj, and Dacheng Tao. Deep learning for camera calibration and beyond: A survey. *arXiv preprint arXiv:2303.10559*, 2023.

Kang Liao, Zongsheng Yue, Zhonghua Wu, and Chen Change Loy. Mowa: Multiple-in-one image warping model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Kang Liao, Size Wu, Zhonghua Wu, Linyi Jin, Chao Wang, Yikai Wang, Fei Wang, Wei Li, and Chen Change Loy. Thinking with camera: A unified multimodal model for camera-centric understanding and generation. In *International Conference on Learning Representations*, 2026.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025a.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025b.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, et al. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025c.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

Yaroslava Lochman, Oles Dobosevych, Rostyslav Hryniv, and James Pritts. Minimal solvers for single-view lens-distorted camera auto-calibration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2887–2896, 2021.

Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11817–11825, 2019.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7739–7751, 2025.

Ollin Matsubara and Draw Things AI Team. Megalith-10m: A dataset of 10 million public-domain photographs, 2024. URL `https://huggingface.co/datasets/madebyollin/megalith-10m`. CC0/Flickr-Commons images; Florence-2 captions available in the `megalith-10m-florence2` variant.

OpenAI. Gpt image2, 2026. URL `https://openai.com/zh-Hant-HK/index/introducing-chatgpt-images-2-0/`.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4, 2025.

Rémi Pautrat, Shaohui Liu, Petr Hruby, Marc Pollefeys, and Daniel Barath. Vanishing point estimation in uncalibrated images with prior gravity direction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14118–14127, 2023.

Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2545–2555, 2025.

A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, et al. Qwen2.5 technical report. *arXiv preprint*, 2024.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-radiov4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026.

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 Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10912–10922, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. arxiv 2014. *arXiv preprint arXiv:1409.0575*, 2014.

Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *European Conference on Computer Vision*, pp. 686–704. Springer, 2022.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438, 2019.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

Xu Song, Hao Kang, Atsunori Moteki, Genta Suzuki, Yoshie Kobayashi, and Zhiming Tan. Mscc: Multi-scale transformers for camera calibration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3262–3271, 2024.

Stability AI. Introducing stable diffusion 3.5, 2024. URL `https://stability.ai/news-updates/introducing-stable-diffusion-3-5/`.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*, 2026.

Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Javier Tirado-Garín and Javier Civera. Anycalib: On-manifold learning for model-agnostic single-view camera calibration. *arXiv preprint arXiv:2503.12701*, 2025.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pp. 1–20, 2024.

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.

Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *ICME*, 2021.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024a.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024c.

Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 1369–1373. IEEE, 2015.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025a.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *ICLR*, 2026.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025b.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025c.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025a.

Liuyue Xie, Jiancong Guo, Ozan Cakmakci, Andre Araujo, Laszlo A Jeni, and Zhiheng Jia. Alignddiff: Learning physically-grounded camera alignment via diffusion. *arXiv preprint arXiv:2503.21581*, 2025b.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. In *NeurIPS*, 2022.

Yuxiang Yang, Yingqi Deng, Yufei Xu, and Jing Zhang. Aptv2: Benchmarking animal pose estimation and tracking with a large-scale dataset and beyond. *arXiv preprint arXiv:2312.15612*, 2023.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Xiaoqing Yin, Xinchao Wang, Jun Yu, Maojun Zhang, Pascal Fua, and Dacheng Tao. Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 469–484, 2018.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *NeurIPS*, 2021.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–18, 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.

Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting vanishing points using global image context in a non-manhattan world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5657–5665, 2016.

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025a.

Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024.

Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE transactions on multimedia*, 2019.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025b.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories. *arXiv preprint arXiv:2603.02049*, 2026.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. In *CVPR*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

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 *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.

Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *ICCV*, 2023a.

Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. *Advances in Neural Information Processing Systems*, 36:45137–45149, 2023b.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *ICLR*, 2026.

# APPENDIX

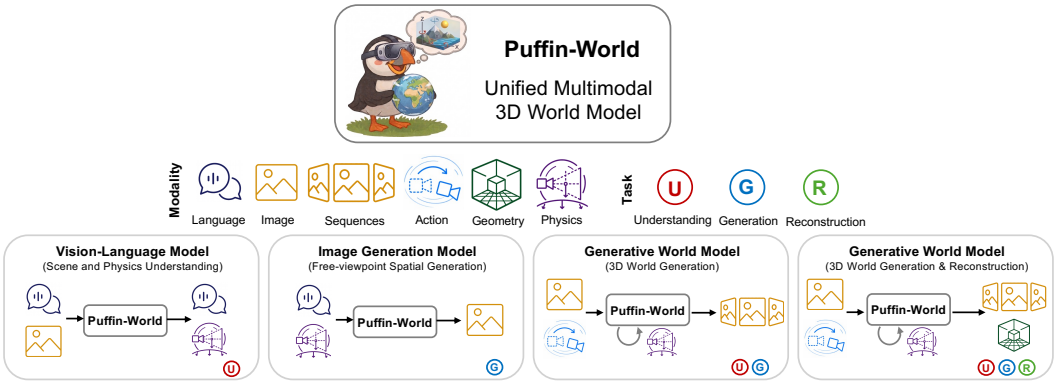


Figure A1: **Puffin-World is a unified multimodal model using native 3D world states (Appearance-Geometry-Physics) for spatial intelligence and physical AI.** It unifies camera physics understanding, free-viewpoint spatial simulation, and 3D world generation and reconstruction within a single framework.


A0.1 Related Work . . . . . 20
A0.1.1 Camera-to-World Understanding . . . . . 20
A0.1.2 Unified Multimodal Models . . . . . 20
A0.1.3 3D World Models . . . . . 21
A0.2 Preliminary . . . . . 21
A0.2.1 3D Native World States . . . . . 21
A0.2.2 Unifying Camera Representations . . . . . 21
A0.3 Dataset . . . . . 22
A0.3.1 Puffin-16M . . . . . 22
A0.3.2 Additional Captions for Public Datasets . . . . . 25
A0.4 Implementation Settings . . . . . 26
A0.4.1 Training Recipe . . . . . 26
A0.4.2 Training Datasets . . . . . 28
A0.4.3 Other Details . . . . . 29
A0.5 Ablation Study . . . . . 30
A0.6 Applications . . . . . 31
A0.7 More Results . . . . . 31
A0.7.1 Camera-to-World Understanding . . . . . 31
A0.7.2 Camera-Controllable Generation . . . . . 31
A0.7.3 3D World Modeling . . . . . 32


This appendix provides additional technical details, analyses, and qualitative results to complement the main paper. We begin with a review of related work, covering camera-to-world understanding, unified multimodal models, and 3D world models, to further position Puffin-World with respect to existing approaches. We then introduce the key preliminaries underlying our formulation, including the hierarchy of native 3D world states (physics, geometry, and appearance) and the unified Omni-Camera representation that combines gravity-aware absolute orientation with ray-based relative

geometry. Next, we provide a detailed description of the datasets used in this work. We first present the construction of Puffin-16M, including Puffin-Cam-15M for large-scale vision–language–camera supervision and Puffin-Traj-1M for diverse and challenging camera trajectories, together with additional physical annotations such as camera height. We further describe the camera and geometric annotations generated for 28 public datasets, analyze their camera-parameter distributions, and explain the construction of dense geometric supervision for real-world datasets with incomplete depth annotations. We then detail the implementation settings of Puffin-World, including the four-stage training recipe, the composition and sampling strategy of the training datasets, module freezing and optimization settings, multi-view training configurations, condition dropout, classifier-free guidance, and inference settings. We subsequently present ablation studies to isolate the contribution of the proposed physics propagation mechanism, evaluating its effect on both visual fidelity and gravity-grounded camera consistency under challenging rotational trajectories. Beyond individual tasks, we further demonstrate closed-loop applications, including mimic world exploration and self-calibrated world exploration, which illustrate the multi-task synergy enabled by the unified framework. Finally, we provide extensive qualitative results for camera-to-world understanding, camera-controllable generation, and 3D world modeling, covering diverse scenes and viewpoints, challenging camera configurations, long trajectories, extreme rotations, compound motions, text-driven world generation, native 3D world-state prediction, and 3D reconstruction.

## A0.1 Related Work

### A0.1.1 Camera-to-World Understanding

Recovering physical camera parameters from images, including camera calibration and pose estimation, is a long-standing problem in 3D vision (Pollefeys et al., 1999; Hartley & Zisserman, 2003; Liao et al., 2023; Veicht et al., 2024; Jin et al., 2023; Lin et al., 2025c). Early learning-based methods directly regress camera parameters from a single image (Hold-Geoffroy et al., 2018; Workman et al., 2015; Bogdan et al., 2018; Zhai et al., 2016; Kendall et al., 2015), whereas more recent approaches narrow the prediction gap by incorporating intermediate geometric structures or semantic cues (Lee et al., 2020; 2021; Song et al., 2024; Janampa & Pattichis, 2024; Yin et al., 2018). A particularly successful direction learns dense, pixel-wise geometric representations, including distortion maps (Liao et al., 2020; 2021), pixel displacement fields (Li et al., 2019b; Liao et al., 2025; Xie et al., 2025b), camera rays (Zhang et al., 2024), perspective fields (Jin et al., 2023; Veicht et al., 2024; Tirado-Garín & Civera, 2025), and incidence fields (Zhu et al., 2023b; He et al., 2025; Deng et al., 2024). These representations provide spatially grounded supervision and generally offer greater robustness than direct global regression. More recently, Puffin (Liao et al., 2026) reframes this task as a language-modeling problem by representing the camera as text and predicting its parameters through spatial reasoning. However, its limited model scale and training data constrain its performance and scalability, causing it to remain inferior to specialized vision-based methods such as GeoCalib (Veicht et al., 2024) on several benchmarks (Sarlin et al., 2022).

### A0.1.2 Unified Multimodal Models

Building upon conventional large multimodal models, unified multimodal models (Team, 2024; Wang et al., 2024b; Tang et al., 2025; Wu et al., 2025c; Lin et al., 2025b; Wu et al., 2024; Tong et al., 2024; Zhou et al., 2024; Deng et al., 2025; Zhang et al., 2025b; Chen et al., 2025b; Song et al., 2026; Agarwal et al., 2026; Su et al., 2026; Han et al., 2026) integrate visual understanding and generation within a single network, either through autoregressive modeling over discrete (Team, 2024; Wu et al., 2024; Wang et al., 2024b; Wu et al., 2025a) or continuous (Fan et al., 2025; Qu et al., 2025) visual tokens, or by connecting pretrained multimodal models with diffusion decoders (Pan et al., 2025; Chen et al., 2025a; Wu et al., 2025b; Lin et al., 2025a; Ma et al., 2025; Huang et al., 2025a; Zhang et al., 2025a; Xie et al., 2024; 2025a). Despite strong progress in general image understanding and generation, these models largely treat images as 2D appearance and semantics under simplified camera assumptions, without explicitly modeling camera physics, scene geometry, or other essential states. Puffin (Liao et al., 2026) takes an initial step toward camera-centric unification, but still focuses on isolated-view perception and lacks a persistent world frame across views. This motivates extending multimodal unification from 2D semantics to 3D world states.

#### A0.1.3 3D WORLD MODELS

Generative world models (Lee et al., 2026) aim to simulate the world by predicting future observations, with recent video- and multi-view diffusion frameworks achieving impressive visual fidelity (Ha & Schmidhuber, 2018; Parker-Holder et al., 2025; Agarwal et al., 2025; Assran et al., 2025; Bar et al., 2025; Yuan et al., 2025; Hansen et al., 2023; Zhu et al., 2026; Huang et al., 2025b; Wu et al., 2026; Dai et al., 2026; Zhang et al., 2026; Zheng et al., 2026). To enable controllable simulation, extensive research conditions generation on camera information: dense camera-pose or Plücker-ray embeddings guide camera-controlled video generation (He et al., 2024; Wang et al., 2024c; Bahmani et al., 2024; Xu et al., 2024; Yang et al., 2024), while multi-view diffusion models synthesize novel views or complete scenes from one or a few reference images (Gao* et al., 2024; Cao et al., 2024; Ren et al., 2025; Bernal-Berdun et al., 2025), sometimes coupled with feed-forward 3D reconstruction models to recover geometry (Wang et al., 2024a; 2025a;b; Lan et al., 2026). Despite this progress, existing 3D world models exhibit three major limitations. First, they predominantly model the world at the *appearance* level, leaving geometry and physical grounding implicit. Second, they rely primarily on *relative* camera motion, which lacks a global physical reference frame. Consequently, the same relative trajectory may correspond to different absolute world orientations, leading to orientation drift and degraded performance under challenging motions, long horizons, and single-view settings. Third, generation is typically treated as an isolated task, decoupled from physical camera understanding and geometry reconstruction, and is often trained on data with limited rotational diversity. These limitations motivate a unified world model that jointly represents physics, geometry, and appearance, anchors generation to an absolute physical frame, and integrates perception, simulation, and reconstruction within a single framework.

### A0.2 PRELIMINARY

#### A0.2.1 3D NATIVE WORLD STATES

Most existing generative world models focus primarily on appearance-level prediction, where future states are represented as RGB images or videos. However, the physical world is natively organized by multiple levels of states, including physics, geometry, and appearance. Physics-level states provide global physical grounding, geometry-level states describe the 3D spatial structure, and appearance-level states represent the observable visual content. Jointly modeling these complementary states is crucial for stable generation, spatially consistent simulation, and grounded world interaction.

In Puffin-World, we formulate a hierarchy of native 3D world states. Specifically, the *physics* state captures absolute physical cues, including the gravity field and latitude map; the *geometry* state represents scene depth; and the *appearance* state corresponds to RGB observations. Since some states, particularly physics-level cues, are difficult to infer directly from visual observations, and their interactions across different levels remain underexplored, Puffin-World jointly perceives, propagates, and generates these states within a unified framework. This design allows physics-aware signals to guide future-view generation toward geometrically consistent structure and realistic appearance.

#### A0.2.2 UNIFYING CAMERA REPRESENTATIONS

The camera serves as a fundamental interface between the 3D world and 2D visual data modalities. In visual perception and reconstruction, camera models explicitly encode the physical rules of perspective projection, thereby supporting tasks such as single-view camera calibration, multi-view geometric reasoning, and joint spatial reconstruction. Beyond perception, cameras also provide effective and practical control signals for generative models, enabling spatially controllable synthesis, continuous world generation, and consistent physical simulation.

Existing camera representations can be broadly categorized as relative or absolute. **Relative camera representations** capture spatial relationships across viewpoints and are widely used for multi-view reconstruction and continuous scene generation. Although readily obtained from structure-from-motion and effective for cross-view geometry, they lack a global physical anchor and cannot encode absolute orientation with respect to the real world. **Absolute camera representations**, in contrast, ground individual observations to physical cues such as the horizon and scene uprightness, but are less suited to continuous multi-view transitions and are substantially harder to obtain from monocular images. These complementary properties motivate a unified camera representation that combines

absolute physical grounding with continuous spatial modeling. To this end, we propose the Omni-Camera representation, a simple but effective unified camera representation for versatile applications. For each pixel $\mathbf{x} = (u, v)$, the Omni-Camera representation combines an absolute camera field and a relative ray field along the channel dimension:

$$\begin{aligned} \mathbf{c}_\mathbf{x} &= \mathrm{Concat}\left(\mathbf{a}_\mathbf{x}, \mathbf{p}_\mathbf{x}\right) \in \mathbb{R}^9, \\ \mathbf{a}_\mathbf{x} &= (\mathbf{u}_\mathbf{x}, \varphi_\mathbf{x}) \in \mathbb{R}^3, \quad \mathbf{p}_\mathbf{x} \in \mathbb{R}^6. \end{aligned} \tag{6}$$

Here, $\mathbf{a}_\mathbf{x}$ denotes the absolute camera representation consisting of the pixel-wise up-vector $\mathbf{u}_\mathbf{x}$ and latitude angle $\varphi_\mathbf{x}$, while $\mathbf{p}_\mathbf{x}$ denotes the relative camera representation given by the ray map composed of the ray origin and direction. Let $\tilde{\mathbf{x}} = [u, v, 1]^\top$ be the homogeneous coordinate of pixel $\mathbf{x}$. Given the camera intrinsic matrix $\mathbf{K}$ and the world-to-camera extrinsic parameters $(\mathbf{R}, \mathbf{t})$, where $\mathbf{X}_c = \mathbf{R}\mathbf{X}_w + \mathbf{t}$, the camera center in the world coordinate system is $\mathbf{o} = -\mathbf{R}^\top\mathbf{t}$. The viewing ray direction from the camera center to pixel $\mathbf{x}$ is

$$\mathbf{d}_\mathbf{x} = \frac{\mathbf{R}^\top\mathbf{K}^{-1}\tilde{\mathbf{x}}}{\|\mathbf{R}^\top\mathbf{K}^{-1}\tilde{\mathbf{x}}\|_2}. \tag{7}$$

The relative ray representation (Gao* et al., 2024) of this ray is then formed by concatenating the ray origin and the ray direction,

$$\mathbf{p}_\mathbf{x} = (\mathbf{o}, \mathbf{d}_\mathbf{x}) \in \mathbb{R}^6, \tag{8}$$

where $\mathbf{o} = -\mathbf{R}^\top\mathbf{t}$ is the ray origin shared by all pixels of the view and $\mathbf{d}_\mathbf{x}$ is the per-pixel ray direction. Relative camera motion between two views is thereby encoded by the change of the ray origin (translation) and the ray direction (rotation).

For the absolute component, we follow the Perspective Field (Jin et al., 2023) to represent the physical orientation of each pixel. Let $\mathbf{g} \in \mathbb{R}^3$ be the unit gravity direction and $\Pi(\cdot)$ be the projection function. For any 3D point $\mathbf{X}$ on the ray of pixel $\mathbf{x}$, satisfying $\Pi(\mathbf{X}) = \mathbf{x}$, the up-vector and latitude angle are defined as

$$\begin{aligned} \mathbf{u}_\mathbf{x} &= \lim_{\epsilon \to 0} \frac{\Pi(\mathbf{X} - \epsilon\mathbf{g}) - \Pi(\mathbf{X})}{\|\Pi(\mathbf{X} - \epsilon\mathbf{g}) - \Pi(\mathbf{X})\|_2}, \\ \varphi_\mathbf{x} &= \arcsin\left(\frac{(-\mathbf{d}_\mathbf{x})^\top\mathbf{g}}{\|\mathbf{d}_\mathbf{x}\|_2}\right). \end{aligned} \tag{9}$$

Here, $-\mathbf{d}_\mathbf{x}$ corresponds to the incoming light ray direction. The up-vector $\mathbf{u}_\mathbf{x}$ describes the image-plane projection of the direction opposite to gravity, and the latitude angle $\varphi_\mathbf{x}$ measures the elevation of the incoming ray with respect to the horizontal plane.

By integrating gravity-aware absolute orientation with ray-based relative geometry, the Omni-Camera representation provides both global physical grounding and continuous spatial modeling. Although conceptually simple, it effectively serves as a unified and flexible camera condition for physical-world understanding, camera-controllable image generation, and continuous world generation, supporting both single-view and cross-view world modeling within a common representation.

### A0.3 Dataset

#### A0.3.1 Puffin-16M

Datasets with precise, high-quality alignment among visual content, semantics, and physical properties remain scarce in the multimodal community. Puffin-4M (Liao et al., 2026) narrows this gap with 4 million vision-language-camera triplets, but still faces three limitations: (i) its scale remains insufficient for larger models; (ii) all images are fixed at $512 \times 512$, restricting multi-resolution generation and requiring non-square inputs to be centrally cropped and resized for camera understanding, which may discard useful content; and (iii) its spatial coverage is dominated by single-view data, with only limited cross-view pairs and no long-range camera trajectories. To overcome these limitations, we introduce Puffin-16M, consisting of 15 million vision-language-camera triplets with diverse resolutions and aspect ratios, together with 1 million trajectories featuring diverse and challenging rational camera motions. We denote these two subsets as Puffin-Cam-15M and Puffin-Traj-1M, respectively, and summarize the comparison with prior datasets in Tab. A1.

**Puffin-Cam-15M** The construction pipeline follows a procedure similar to that of Puffin-4M (Liao et al., 2026), including panoramic data collection and preprocessing, perspective image generation,

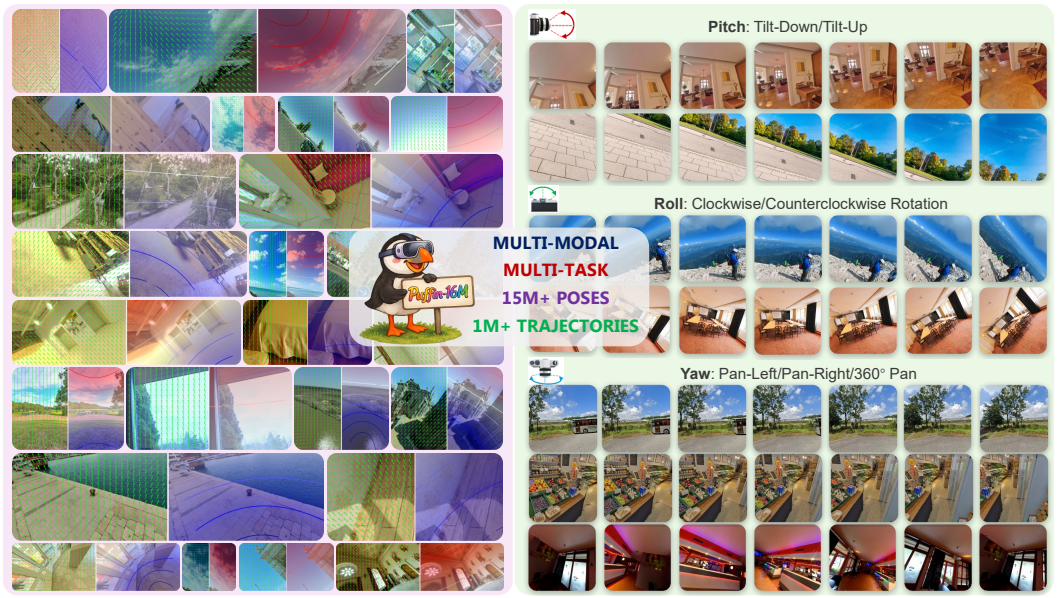


Figure A2: **An overview of the constructed Puffin-16M dataset.** It comprises 15 million vision-language-camera triplets with diverse resolutions and aspect ratios (left), along with 1 million trajectories featuring diverse and challenging rotational camera motions (right).

Table A1: **Dataset Comparisons.** Compared with previous datasets for single-task (Lee et al., 2021; Bogdan et al., 2018; Veicht et al., 2024; Jin et al., 2023; Hold-Geoffroy et al., 2018; Bernal-Berdun et al., 2025) and unified-task (Liao et al., 2026). Our Puffin-16M dataset shows promising advantages across scale, versatility, spatial distribution, and motion diversity. For the camera parameters, we denote the intrinsic parameters: focal length ($f$), radial distortion coefficient ($\xi$); and the extrinsic parameters: roll ($\phi$), pitch ($\theta$), yaw ($\psi$), camera height ($h$).

| Dataset | Task Type | Intrinsics | Extrinsics | # Frames | Details | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Camera | Text | Reasoning | Single-View | Cross-View | Trajectory | Multi-Ratio |
| GeoCalib (Veicht et al., 2024) | Understanding | $f, \xi$ | $\phi, \theta$ | 37K | ✔ | ✗ | ✗ | ✔ | ✗ | ✗ | ✗ |
| CTRL-C (Lee et al., 2021) | Understanding | $f$ | $\phi, \theta$ | 45K | ✔ | ✗ | ✗ | ✔ | ✗ | ✗ | ✗ |
| Deepcalib (Bogdan et al., 2018) | Understanding | $f, \xi$ | - | 67K | ✔ | ✗ | ✗ | ✔ | ✗ | ✗ | ✗ |
| ParamNet (Jin et al., 2023) | Understanding | $f$ | $\phi, \theta$ | 190K | ✔ | ✗ | ✗ | ✔ | ✗ | ✗ | ✗ |
| Perceptual (Hold-Geoffroy et al., 2018) | Understanding | $f$ | $\phi, \theta$ | 390K | ✔ | ✗ | ✗ | ✔ | ✗ | ✗ | ✗ |
| PreciseCam (Bernal-Berdun et al., 2025) | Generation | $f, \xi$ | $\phi, \theta$ | 57K | ✔ | ✔ | ✗ | ✔ | ✗ | ✗ | ✗ |
| Puffin-4M (Liao et al., 2026) | Unified Multimodal | $f$ | $\phi, \theta, \psi$ | 4M | ✔ | ✔ | ✔ | ✔ | ✔ | ✗ | ✗ |
| **Puffin-16M** | Unified Multimodal | $f$ | $\phi, \theta, \psi, h$ | 16M | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |

and scene and reasoning captioning. Specifically, we expand the source panoramic images used in Puffin-4M from 200K to 900K, covering diverse scenarios such as outdoor and indoor scenes, synthetic and real-world environments, and self-driving street views across different countries. Such diversity in the source panoramas provides a strong foundation for scaling large multimodal models, particularly for multimodal spatial intelligence. Due to variations in $360^\circ$ camera calibration and acquisition stability, some panoramas exhibit geometric distortions and misalignment, especially in consumer-grade capture and autonomous-driving onboard settings. We therefore apply geometric correction techniques based on line segmentation and vanishing point estimation, aligning the panoramas with the gravity direction and improving structural consistency. Compared with Puffin-4M, we exclude Stanford2D3D (Armeni et al., 2017) from the source panoramic data to ensure fair evaluation on this benchmark for camera-to-world understanding.

Following the standard protocol adopted in recent studies (Veicht et al., 2024; Liao et al., 2026), we generate diverse perspective views from each panorama using a virtual pinhole camera with varying intrinsic and extrinsic parameters. Specifically, we uniformly sample roll and pitch from $[-45^\circ, 45^\circ]$ and vertical FoV from $[20^\circ, 105^\circ]$, while adaptively determining the number of crops according to the resolution of each source panorama. This process yields approximately 15M perspective images with precise camera parameters and diverse aspect ratios (*e.g.*, 1:1, 2:3, 3:2, 3:4, 4:3, 9:16, and 16:9).

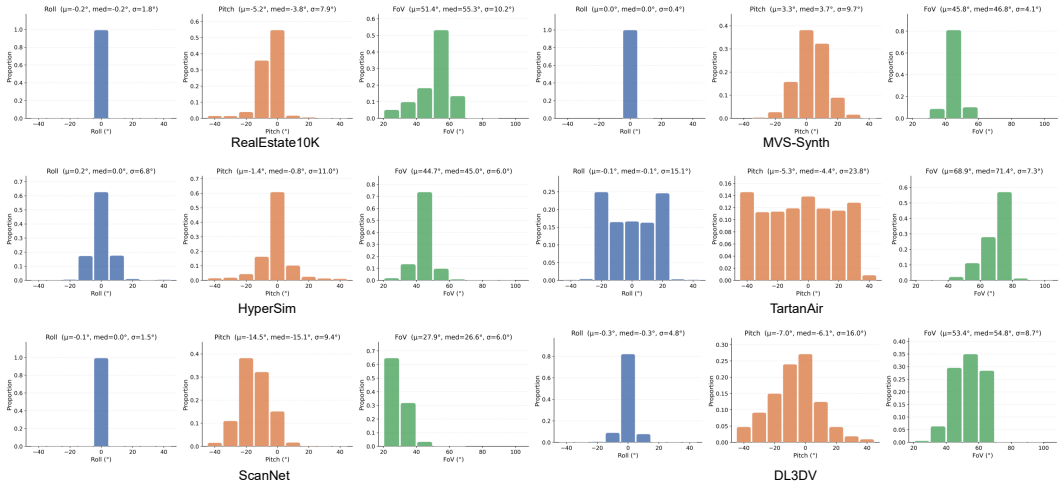


Figure A3: **The distributions of the absolute camera parameters (roll and pitch) and FoV of existing public 3D or sequential datasets** (Ling et al., 2024; Zhou et al., 2018; Dai et al., 2017; Huang et al., 2018; Wang et al., 2020; Roberts et al., 2021).

We retain the full set for camera-to-world understanding, while further curating around 8M images based on the realism and aesthetic quality of the source panoramas for camera-controllable image generation. We then employ advanced multimodal large language models (*e.g.*, Qwen3-VL-8B-Instruct and Qwen3-VL-32B-Instruct (Bai et al., 2025)) to annotate each image with high-quality captions that include detailed scene semantics and structured chain-of-thought reasoning over spatial relationships, camera parameters, and underlying physical principles.

**Puffin-Traj-1M** As shown in Figure A3, most existing 3D world modeling datasets (Ling et al., 2024; Zhou et al., 2018; Dai et al., 2017; Huang et al., 2018; Wang et al., 2020; Roberts et al., 2021) are limited in terms of camera rotational diversity. Specifically, roll is typically restricted to $[-5°, 5°]$ and pitch to $[-10°, 10°]$. This limitation arises because most datasets are captured using handheld cameras, where the primary source of variation comes from camera translation rather than rotation. Consequently, models trained on such data often struggle to understand challenging camera motions and accurately simulate the corresponding future world states, particularly in real-world scenarios.

However, a spatially intelligent agent should be able to actively explore its environment from arbitrary viewpoints. To this end, we construct Puffin-Traj-1M, a large-scale dataset comprising one million trajectories with diverse and challenging camera motions. The dataset includes long-horizon trajectories featuring continuous look-down, look-up, clockwise rotation, counterclockwise rotation, and full $360°$ surrounding-view exploration.

Specifically, we adopt the same data construction pipeline as Puffin-Cam-15M, while introducing an additional camera extrinsic parameter, namely yaw, to enable cross-view exploration and arbitrary-viewpoint camera rotations. All camera extrinsics are sampled from broad ranges: roll and pitch are uniformly sampled from $[-45°, 45°]$, while yaw is sampled from $[0°, 360°)$. For each trajectory, one extrinsic parameter is randomly selected to define the camera motion pattern, and frames are generated either in a single-pass or recursive-pass manner. The single-pass strategy provides long-range scene exploration trajectories, whereas the recursive-pass strategy encourages the emergence of spatial memory and facilitates the preservation of 3D consistency over extended horizons. The vertical field-of-view is randomly sampled from $[60°, 100°]$.

**Camera Height** Beyond the precise camera parameters and camera trajectories provided in Puffin-16M, we further annotate the camera height of each original panoramic image, prior to perspective rendering, using Qwen3-VL-32B (Bai et al., 2025). We categorize camera height into five levels—underwater shot, low-position shot, eye-level shot, high-position shot, and aerial shot—to provide a coarse yet physically meaningful description of the observer's vertical position in the scene. The resulting height label is subsequently inherited by all perspective views rendered from the corresponding panorama, thereby complementing their precise camera parameters with an additional

**###[System Role Instruction]**
Analyze the physical camera height (elevation) assuming a strictly HORIZONTAL camera angle.
Do not consider looking-up or looking-down angles.

Choose exactly ONE label from the following five options: **Underwater shot**, **Low-position shot**, **Eye-level shot**, **High-position shot**, **Aerial shot.**
Classify based on the position of scene objects relative to the horizon line and geometric and semantic cues (perspective geometry, visible ground coverage, and object scale):

**[Underwater shot]**
- Physical Height: Submerged in water (refraction, blue/green cast, turbidity). Distinct from air.
- Key Cue: The entire visual field is immersed in a **volumetric liquid medium**, characterized by optical attenuation and refraction. The concept of a sharp "horizon line" is often replaced by a gradient of water visibility.

**[Low-position shot]**
- Physical Height: Ground level (< 1.0m).
- Horizontal View Geometry: The camera is closer to the floor than the objects.
- Key Cue: Most foreground objects (chairs, pedestrians, cars) tower *above* the image's center line.

**[Eye-level shot]**
- Physical Height: Human standing height (~1.5m - 2m).
- Horizontal View Geometry: Camera height is approximately at human eye level with a near-horizontal viewing direction. The camera is at the similar height as the subjects.
- The image center line intersects the physical "middle band" of the environment, capturing objects (furniture, buildings) in **neutral side-profile** (revealing neither top surfaces nor undersides). It geometrically bisects the scene, creating an equal balance between the ground plane and the ceiling/sky.

**[High-position shot]**
- Physical Height: Elevated but terrestrial (~3m - 10m).
- Horizontal View Geometry: Camera is positioned high (e.g., on a ladder or tall vehicle) looking straight forward.
- Key Cue: Most scene objects (furniture, crowds, cars) sit *below* the image's center line.

**[Aerial shot]**
- Physical Height: Sky and space level (> 20m).
- Horizontal View Geometry: Drone or skyscraper view looking out at the skyline or the space.
- Key Cue: The scene consists of rooftops, distant horizons, clouds, or starscape. Ground objects are insignificant in scale.

Output ONLY the label.



Label: Eye-level Shot

Figure A4: **The designed prompt for captioning the camera height for the panoramic images.** We categorize it into five levels: underwater shot, low-position shot, eye-level shot, high-position shot, and aerial shot. A representative example, consisting of the input panoramic image and the corresponding output label, is shown in the bottom-right corner.



Figure A5: **The overview of the public datasets annotated by Puffin-World, with a total of 28 datasets covering approximately 44.5 million images across diverse scene distributions.**

high-level physical cue. The detailed annotation prompts are provided in Figure A4. We expect these complementary camera annotations to facilitate richer spatial reasoning, viewpoint-aware generation, and future studies of camera-grounded physical world understanding and generation.

### A0.3.2 Additional Captions for Public Datasets

Beyond contributing our self-constructed dataset, we also enrich the commonly used public datasets (Ling et al., 2024; Zhou et al., 2018; Dai et al., 2017; Huang et al., 2018; Wang et al., 2020; Roberts et al., 2021) in terms of the physical and geometric cues.

**Physical Caption** Although these datasets provide accurate per-frame camera poses, they primarily characterize relative motion across views and lack the absolute orientation required to anchor observations to the physical world. Models trained solely on such trajectories can therefore learn scene-centric motion patterns without establishing a persistent physical reference, making gravity direction, horizon location, and absolute camera orientation ambiguous. To address this limitation, we

leverage Puffin-World to estimate absolute camera parameters for every frame in the public training datasets, from which gravity fields and latitude maps are derived as physical priors for our physics propagation mechanism. Beyond the training data, we further annotate a broad collection of widely used public datasets, including ImageNet (Russakovsky et al., 2014), GPIC (Chandrasegaran et al., 2026), Objects365 (Shao et al., 2019), CC12M (Changpinyo et al., 2021), and Megalith-10M (Matsubara & Draw Things AI Team, 2024). In total, our release covers 28 datasets and approximately 44.5M images with absolute camera annotations, providing large-scale supervision for gravity-aware perception, physically grounded world modeling, and future research on spatial intelligence. An overview of the annotated datasets is shown in Figure A5.

As listed in Tab. A2, the collection comprises 22 single-image datasets spanning recognition, detection, segmentation, pose estimation, and generative pre-training, together with six sequential/3D *Absolute-Camera* datasets. Across these datasets, several consistent biases emerge: roll is strongly concentrated around $0^\circ$, indicating a dominant level-camera bias; pitch is generally negative, reflecting a tendency toward downward-looking viewpoints; and FoV varies substantially with capture regime, with web photographs typically exhibiting narrower views than synthetic, scanned, or video-based datasets. Together, these statistics reveal that conventional vision corpora substantially under-sample strongly tilted, rolled, and wide-angle viewpoints.

Several datasets exhibit distinctive camera distributions. For example, *EgoObjects* shows the strongest downward-looking bias, consistent with head-mounted capture of hands and nearby objects, while *DeepFashion* exhibits exceptionally low variation in roll, pitch, and FoV, reflecting tightly controlled studio photography. In contrast, *TartanAir* presents the largest rotational diversity, making it particularly valuable for camera orientations rarely observed in natural photographs; *ScanNet* and *SKU110K* also show pronounced downward pitch due to room-scanning and shelf-facing capture, respectively. These annotations transform widely used vision datasets into camera-grounded resources that support viewpoint-bias analysis, camera-conditioned generation, camera-diverse data curation, and large-scale monocular physics perception. For the *Absolute-Camera* subsets, as analyzed in Figure A3, aligning per-frame predictions with camera trajectories further enables consistency analysis and gravity-grounded world modeling.

**Geometric Caption** For public synthetic datasets (Huang et al., 2018; Wang et al., 2020; Roberts et al., 2021), we directly use the provided dense geometric annotations. For real-world datasets, however, the available depth supervision is often incomplete. Although ScanNet (Dai et al., 2017) and DL3DV (Ling et al., 2024) provide camera poses, their depth maps may be sparse or missing. In ScanNet, depth measurements suffer from sensor holes and invalid observations. For DL3DV, the depth maps obtained through MVS reconstruction (Schonberger & Frahm, 2016), following CUT3R (Wang et al., 2025b) are incomplete. To obtain dense geometric supervision, we run DA3 (Lin et al., 2026) on every frame to predict dense depth maps. We then align these predictions with the available sparse depth annotations, preserving the reliable scale of the original measurements while filling in missing regions. This yields dense and geometrically consistent depth annotations for real-world videos. The comparison of the original sparse depth label, our re-annotated dense depth label, and our predicted depth is visualized in Figure A6.

### A0.4 Implementation Settings

#### A0.4.1 Training Recipe

We adopt a four-stage training strategy to learn the whole framework. The first two stages perform cross-modal alignment and supervised fine-tuning (SFT) for single-view understanding and generation across the vision, language, and camera modalities. These stages cover scene and physical-state perception, as well as free-viewpoint spatial simulation. Building upon the acquired single-view capabilities, the final two stages focus on post-training for 3D world modeling, extending the model to sequential generation and reconstruction. The complete training recipe is summarized in Table A3, and each stage is detailed below.

- **Stage I**. In this stage, we align the vision encoder with the LLM by training only the MLP projector to predict scene descriptions and camera parameters from input images. For generation, text and Omni-Camera representation are used to condition target image synthesis. We train the learnable queries and connector to map LLM hidden states into

Table A2: **Camera parameter statistics of 28 annotated datasets using Puffin-World.** For every source dataset we report its original task, the number of validly annotated images and scenes, and the mean/median/standard deviation of the predicted roll, pitch, and vertical FoV (in degrees). †computed on a ∼4.9M-image sample; ‡pooled from the released train+val statistics.

| | | | | Camera Parameters ($\mu$/med/$\sigma$, °) | | |
|---|---|---|---|---|---|---|
| **Dataset** | **Original task** | **#Images** | **#Scenes** | **Roll** | **Pitch** | **FoV** |
| *Single-image datasets* | | | | | | |
| ADE20K Zhou et al. (2019) | Semantic segmentation | 27.6K | – | 0.0/0.0/1.9 | −1.3/−1.0/8.0 | 32.7/31.2/11.5 |
| AP-10K Yu et al. (2021) | Animal pose estimation | 10.0K | – | 0.4/0.2/8.1 | −9.2/−6.1/13.9 | 26.5/23.2/8.3 |
| APT-36K Yang et al. (2022) | Animal pose tracking | 35.7K | – | 0.2/0.0/3.5 | −4.9/−3.7/11.3 | 24.1/20.3/6.0 |
| APTv2 Yang et al. (2023) | Animal pose tracking | 41.3K | – | 0.2/0.0/3.5 | −5.1/−3.7/11.2 | 24.2/22.9/6.0 |
| CC12M Changpinyo et al. (2021) | Vision–language pre-training | 10.90M | – | 0.2/0.0/5.7 | −2.8/−0.3/11.8 | 32.3/29.0/10.6 |
| COCO Lin et al. (2014) | Detection & segmentation | 122K | – | 0.0/0.0/6.2 | −7.3/−4.0/15.0 | 34.9/32.5/11.7 |
| CrowdPose Li et al. (2019a) | Human pose estimation | 20.0K | – | 0.1/0.0/3.5 | −5.1/−3.7/8.5 | 31.6/28.7/10.7 |
| DeepFashion Liu et al. (2016) | Fashion recognition & retrieval | 13.7K | – | 0.1/0.3/3.5 | 1.1/0.4/3.6 | 32.6/32.1/3.4 |
| EgoObjects Zhu et al. (2023a) | Egocentric object detection | 241.6K | – | −0.1/0.0/8.0 | −24.5/−25.6/15.3 | 39.3/36.7/10.2 |
| GPIC† Chandrasegaran et al. (2026) | Image generation corpus | 4.9M | – | 0.1/0.0/5.6 | −2.6/−1.3/13.0 | 33.8/31.2/11.6 |
| HumanRef Jiang et al. (2025) | Referring expression comprehension | 45.3K | – | 0.1/0.0/3.6 | −2.5/−1.5/8.4 | 32.5/31.2/9.9 |
| IRS Wang et al. (2021) | Stereo matching / depth (syn.) | 188.3K | – | 0.2/0.0/4.9 | −8.3/−5.0/14.8 | 54.8/56.7/6.1 |
| ImageNet-1K‡ Deng et al. (2009) | Image classification | 1.33M | – | 0.1/0.0/7.8 | −7.8/−3.7/15.7 | 29.0/25.6/8.9 |
| InteriorVerse Li et al. (2018) | Inverse rendering (syn. indoor) | 61.2K | – | 0.8/0.0/6.3 | −1.3/−0.8/16.9 | 71.6/68.6/14.9 |
| MPII Andriluka et al. (2014) | Human pose estimation | 25.0K | – | 0.0/0.0/2.7 | −5.0/−3.8/7.8 | 27.6/26.6/6.2 |
| Megalith-10M Bohan (2024) | Image generation corpus | 9.38M | – | 0.0/0.0/6.7 | −3.1/−2.3/12.1 | 34.6/31.3/12.0 |
| Objects365 Shao et al. (2019) | Object detection | 1.72M | – | 0.1/0.0/4.0 | −4.4/−2.7/11.2 | 33.4/31.3/11.0 |
| Open Images Kuznetsova et al. (2020) | Detection & multi-label cls. | 8.32M | – | 0.1/0.0/6.2 | −3.1/−1.1/14.8 | 31.8/31.2/8.2 |
| Rexverse-2M Jiang et al. (2024) | Multimodal grounding & captioning | 414.6K | – | 0.2/0.0/6.8 | −4.0/−0.9/12.9 | 32.2/28.9/11.1 |
| SA-1B Kirillov et al. (2023) | Promptable segmentation | 897.6K | – | 0.2/0.0/4.9 | 2.0/1.5/13.1 | 36.3/34.4/13.2 |
| SKU110K Goldman et al. (2019) | Dense object detection | 19.1K | – | −0.1/0.0/1.9 | −13.9/−14.1/6.6 | 49.6/51.9/12.5 |
| WiderPerson Zhang et al. (2019) | Pedestrian detection | 13.3K | – | 0.1/0.0/1.4 | −1.8/−1.4/5.8 | 29.1/27.2/8.3 |
| *Multi-view datasets (absolute-pose annotations)* | | | | | | |
| DL3DV Ling et al. (2024) | Novel view synthesis / 3D recon. | 2.16M | 6,377 | −0.3/−0.3/4.9 | −7.0/−6.1/16.0 | 53.4/54.8/8.7 |
| HyperSim Roberts et al. (2021) | 3D scene understanding (syn.) | 73.6K | 448 | 0.2/0.0/6.8 | −1.4/−0.8/11.0 | 44.7/45.0/6.0 |
| MVS-Synth Huang et al. (2018) | Multi-view stereo (syn.) | 12.0K | 120 | 0.0/0.0/0.4 | 3.3/3.7/9.7 | 45.8/46.8/4.1 |
| RealEstate10K Zhou et al. (2018) | Novel view synthesis / camera traj. | 595.4K | 5,134 | −0.2/−0.2/1.8 | −5.2/−3.8/7.9 | 51.4/55.3/10.2 |
| ScanNet Dai et al. (2017) | RGB-D reconstruction | 2.32M | 1,468 | −0.1/0.0/1.5 | −14.5/−15.1/9.4 | 27.8/26.6/6.0 |
| TartanAir Wang et al. (2020) | SLAM / visual odometry (syn.) | 306.6K | 18 | −0.1/−0.1/15.1 | −5.3/−4.4/23.8 | 68.9/71.4/7.3 |

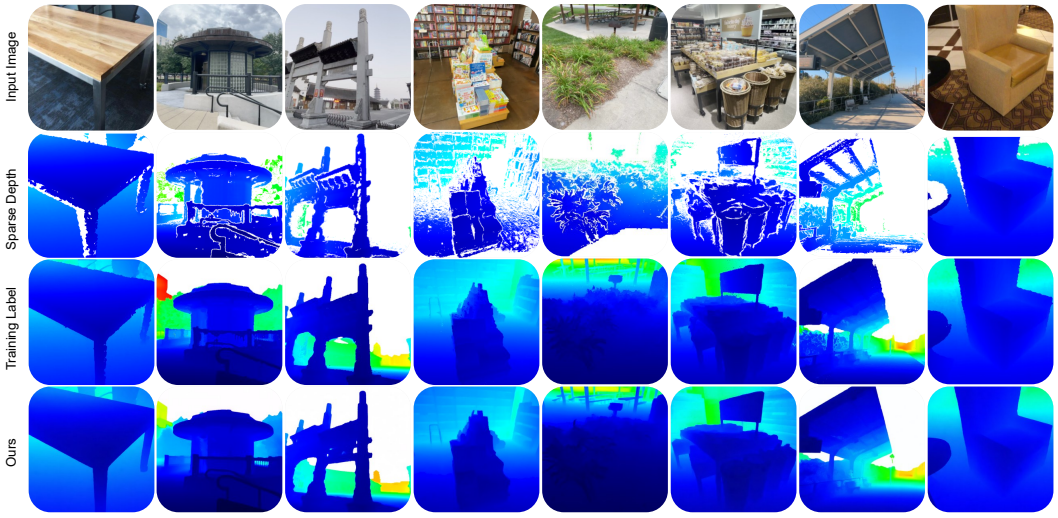


Figure A6: **Depth map visualization on the DL3DV dataset (Ling et al., 2024).** From top to bottom, we show the input image, the sparse label provided in the original dataset, our annotated and aligned dense label (training used), and the predicted result by Puffin-World.

diffusion conditioning signals, while the vision encoder, LLM, and diffusion model remain frozen.

- **Stage II.** After cross-modal alignment, we unfreeze all modules except the VAE and fine-tune the full framework using the same data and objectives as in Stage I. To stabilize optimization and preserve the pretrained visual representations, we scale the gradients of the vision encoder by 0.1.

Table A3: **Training recipe of Puffin-World.** We report the per-stage hyperparameters, the trainable/frozen status of each module, and the data sampling ratio across tasks. *Und.*, *Gen.*, and *Recon.* abbreviate understanding, generation, and reconstruction, respectively.

| | Stage I | Stage II | Stage III | Stage IV |
|---|---|---|---|---|
| **Hyperparameters** | | | | |
| Learning rate | $1 \times 10^{-4}$ | $2 \times 10^{-5}$ | $5 \times 10^{-5}$ | $2 \times 10^{-5}$ |
| LR Scheduler | Cosine | | | |
| Weight Decay | 0.05 | | | |
| Betas | (0.9, 0.95) | | | |
| Optimizer | AdamW | | | |
| Batch Size | 512 | 512 | 256 | 128 |
| **Trainable Modules** | | | | |
| MLP Projector | Trainable | Trainable | Frozen | Frozen |
| Connector | Trainable | Trainable | Trainable | Trainable |
| Condition Tokenizer | Trainable | Trainable | Trainable | Trainable |
| Vision Encoder | Frozen | Trainable | Frozen | Frozen |
| LLM | Frozen | Trainable | Frozen | Frozen |
| Diffusion Model | Frozen | Trainable | Trainable | Trainable |
| **Data Sampling Ratio** | | | | |
| Image→Text-Camera (single-view *Und.*) | 0.5 | 0.5 | – | – |
| Text-Camera→Image (single-view *Gen.*) | 0.5 | 0.5 | 0.14 | 0.09 |
| Image→3D (cross-view *Gen.*) | – | – | 0.86 | – |
| Image→3D (cross-view *Gen.* & *Recon.*) | – | – | – | 0.91 |

Table A4: **Training datasets for our model.** We summarize the source category, number of scenes and frames, depth availability, temporal sampling interval, test-set usage, and the corresponding training stages for each dataset.

| Dataset | Category | Scenes / Frames | Depth | Sample Interval | Testset Split | Training Stage |
|---|---|---|---|---|---|---|
| DL3DV Ling et al. (2024) | Public | 6377 / 2M | Yes* | $[1, 8]$ | No | III, IV |
| RealEstate10K Zhou et al. (2018) | Public | 5134 / 600K | No | $[48, 144]$ | Yes | III, IV |
| HyperSim Roberts et al. (2021) | Public | 756 / 70K | Yes | $[1, 1]$ | No | IV |
| MVS-Synth Huang et al. (2018) | Public | 120 / 12K | Yes | $[1, 2]$ | No | IV |
| TartanAir Wang et al. (2020) | Public | 369 / 300K | Yes | $[1, 10]$ | No | IV |
| ScanNet Dai et al. (2017) | Public | 1468 / 2M | Yes* | $[1, 8]$ | No | IV |
| Puffin-Traj-1M | Self-Constructed | 1M / 100M | No | $[1, 12]$ | Yes | III, IV |
| Puffin-Cam-15M | Self-Constructed | – / 15M | No | – | Yes | I, II, III, IV |

* We refined the sparse depth label into the dense one.

- **Stage III.** We then extend Puffin-World from single-view generation to cross-view 3D world modeling. Conditioned on one or more reference views and their corresponding Omni-Camera maps, the model jointly denoises all target views within a unified sequence. We train the model on trajectories with diverse and challenging camera motions, including large translations, in-place rotations, and full $360^\circ$ exploration. This stage is performed as post-training on the converged single-view model: the vision encoder and LLM are frozen, while the connector, diffusion backbone, and condition fusion module, remain trainable.
- **Stage IV.** Finally, we improve Puffin-World for geometry reconstruction by jointly modeling appearance and depth. We activate a geometry branch that generates the depth maps alongside the target RGB views. Both of them share the same flow-matching formulation but are combined through the separately weighted objective of Eq. 5. This stage keeps the same trainable modules as Stage III and additionally trains the depth-modality embedding.

### A0.4.2 Training Datasets

Table A4 summarizes the datasets used for Puffin-World across different training stages. We combine large-scale public sequential or 3D datasets with our self-constructed Puffin-16M (Puffin-Cam-15M

and Puffin-Traj-1M) to provide complementary supervision in native 3D world states. The public datasets cover diverse real-world and synthetic environments with varying trajectory lengths, motion patterns, and depth availability, thereby providing broad spatial and geometric supervision. When depth annotations are available, they are incorporated to strengthen geometry learning, while datasets without depth primarily contribute to multi-view appearance consistency and camera-conditioned generation. Puffin-Traj-1M complements these sources with substantially richer and more challenging rotational trajectories, extending the motion distribution beyond that of conventional capture datasets. Puffin-Cam-15M further supplies large-scale language-image-camera triplets and is reused across multiple stages to preserve strong camera perception and generation capabilities. Together, these datasets form a balanced training mixture that progressively supports single-view understanding/generation, cross-view generation, trajectory modeling, and 3D world reconstruction within the same framework.

Although ScanNet (Dai et al., 2017) and DL3DV (Ling et al., 2024) provide camera poses, their depth annotations remain incomplete due to sensor holes in ScanNet and sparse MVS reconstruction (Schonberger & Frahm, 2016) in DL3DV. As described in Section A0.3.2, we use DA3 to predict dense depth for each frame and align it with the available measurements, preserving reliable metric scale while filling missing regions. Figure A6 compares the sparse depth provided by the original datasets, our aligned dense annotations used for training, and the depth predicted by Puffin-World. Compared with the original sparse measurements, our annotations provide substantially denser and more complete geometric supervision, reducing holes and missing structures and thereby facilitating unified RGB–depth training without specialized tokenizer designs. Moreover, the depth predicted by Puffin-World exhibits sharper object boundaries and fewer artifacts, while maintaining a spatial layout that is highly consistent with the corresponding input image.

### A0.4.3 Other Details

For single-view camera-controllable generation, we train at the native resolutions and aspect ratios of Puffin-Cam-15M, whereas for 3D world modeling we crop all views to $640 \times 640$. Each sequence contains a total of $8$ views: the number of reference (initial) views is randomly sampled from $1$ to $3$ in Stage III and fixed to $1$ in Stage IV, with the remaining views serving as generation targets. Diffusion timesteps are drawn from a logit-normal distribution ($\mu$=0, $\sigma$=1) with a resolution-dependent shift, and Stage III and Stage IV adopt a larger shift to bias training toward higher-noise timesteps that benefit cross-view consistency. For the Omni-Camera representation, the absolute representation is placed before the relative representation in the channel order, resulting in a $9$-channel dense camera map $\mathbf{C} \in \mathbb{R}^{H\times W\times 9}$ with the same spatial resolution as the image.

For the multi-view post-training Stage III, the model is trained on a mixture of trajectory data to learn diverse camera motions involving both translation (DL3DV (Ling et al., 2024) and RealEstate10K (Zhou et al., 2018)) and rotation (Puffin-Traj-1M), together with single-view camera-controllable generation samples from Puffin-Cam-15M. The single-view batches serve as rehearsal data to preserve the camera controllability acquired in Stage II. We adopt a multi-source sampling strategy, where homogeneous mini-batches are drawn from each data source according to fixed repeat factors of 3:2:1:1. To reduce the training cost of the proposed physics propagation, we pre-compute the absolute camera fields for all training frames offline using the physics perception capability (at Stage-II) of Puffin-World. During training, each view is therefore directly conditioned on its corresponding Omni-Camera map. The condition fusion module is warm-started from the single-view model. Stage IV then inherits the Stage-III weights and activates the joint appearance–geometry branch (Sec. 2.3), widening the multi-view mixture with Hypersim (Roberts et al., 2021), MVS-Synth (Huang et al., 2018), TartanAir (Wang et al., 2020), and ScanNet (Dai et al., 2017) while keeping the same source-balanced sampling and single-view rehearsal. The depth loss weight is ramped from zero to $\omega_{\max}$=1 over the first $3{,}000$ iterations (Eq. 5).

We enable classifier-free guidance (CFG) through condition dropout during training. For single-view camera-controllable generation, we jointly replace the text prompt with an empty string and zero the Omni-Camera condition with probability $0.1$. For multi-view world generation, the same joint dropout is applied to the text and all per-view Omni-Camera maps. We additionally drop the absolute Perspective Field with probability $0.15$ while retaining the relative ray map, reducing over-reliance on absolute cues and enabling component-wise guidance. At inference, camera-to-world understanding

Table A5: **Ablation study on the proposed physics propagation (PP) for 3D world generation with challenging camera motions.**

| Approach | Motion | Image Metrics | | | Camera Metrics | |
|---|---|---|---|---|---|---|
| | | PSNR ↑ | SSIM ↑ | LPIPS ↓ | Error_R ↓ | Error_P ↓ |
| Baseline | Roll | 22.97 | 0.69 | 0.12 | 2.35 | 6.73 |
| | Pitch | 18.53 | 0.60 | 0.26 | 2.17 | 5.60 |
| | Yaw | 18.12 | 0.59 | 0.30 | 1.31 | 3.10 |
| | Average | 19.87 | 0.62 | 0.23 | 1.94 | 5.14 |
| Baseline *w/* PP | Roll | 24.02 | 0.74 | 0.10 | 2.04 | 6.31 |
| | Pitch | 19.18 | 0.65 | 0.24 | 2.03 | 5.28 |
| | Yaw | 18.42 | 0.62 | 0.30 | 1.22 | 2.67 |
| | Average | **20.54** | **0.67** | **0.21** | **1.76** | **4.75** |




Figure A7: **Ablation study of the proposed physics propagation (PP) strategy.** For each sample, we show the generation results marked by the error up-vector and the error value with and without PP.

uses greedy decoding, while generation employs 50 sampling steps. We set the CFG scale to 4.5 and 2.0 for single-view camera-controllable image generation and 3D world modeling, respectively.

### A0.5 Ablation Study

We isolate the contribution of physics propagation, which anchors a generated trajectory to a consistent gravity-aligned frame (Sec. 2.3), and focus on rotation trajectories because they provide the most sensitive probe of absolute grounding: unlike translation, pure rotation directly changes the camera orientation with respect to gravity and the horizon, making inconsistencies in the physical frame immediately observable. We use Puffin-Traj-Bench (spanning roll, pitch, and yaw rotations), generate the corresponding views with and without physics propagation, and recover the camera parameters of each generated frame using Puffin-World's physics-perception capability. The calibrated roll and pitch errors (Error_R and Error_P) are compared against the GT cameras and reported in Table A5, together with PSNR, SSIM, and LPIPS against the target views. Physics propagation consistently improves both visual fidelity and physical world grounding, with the largest gains on roll and pitch trajectories that are directly coupled to gravity. Yaw motion shows smaller appearance improvements because it rotates around the gravity axis, although physics propagation still suppresses spurious horizon tilt and improves physical consistency. These results support our design rationale: propagating the absolute physical state from the reference view through the known relative motion maintains a coherent gravity-anchored representation across the trajectory, leading to more stable horizons, better

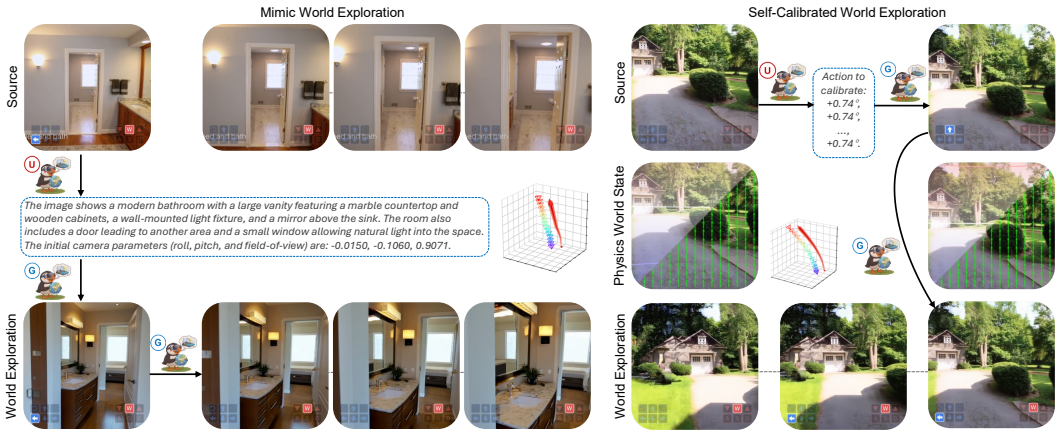


Figure A8: **Closed-loop applications of Puffin-World:** Mimic world exploration (left) and self-calibrated world exploration, both of which require multi-task synergy involving understanding and generation. The type of each sub-task is marked with circled U (understanding) and G (generation).

scene uprightness, improved perceptual quality, and camera states that better follow the intended motion. Qualitative comparisons are provided in Figure A7.

### A0.6 Applications

Beyond the individual tasks above, Puffin-World exhibits promising multi-task synergy for complex, closed-loop applications that jointly involve perception, reasoning, and generation, as illustrated in Figure A8. In *mimic world exploration*, Puffin-World expands a 3D world from a common starting view while reproducing a target exploration process under a prescribed camera trajectory. In *self-calibrated world exploration*, the model reasons about its current physical state, predicts corrective camera actions, and imagines the resulting observations to progressively detect and resolve gravity misalignment, forming an interaction loop reminiscent of World-Action Models (WAMs) (Ye et al., 2026). Both capabilities are realized within the same unified multimodal framework, without relying on auxiliary perception, calibration, or generation modules, highlighting the potential of Puffin-World as a general foundation for interactive 3D environments, virtual reality, and embodied intelligence.

### A0.7 More Results

#### A0.7.1 Camera-to-World Understanding

We provide additional qualitative results across diverse datasets (Dai et al., 2017; Ling et al., 2024; Zhou et al., 2018; Roberts et al., 2021; Huang et al., 2018; Wang et al., 2020; Sarlin et al., 2022; Li & Snavely, 2018) in Figure A10. For each input image, we first estimate its absolute camera parameters using the understanding branch of Puffin-World, and then convert them into the corresponding latitude map and gravity field via Eq. 9 for intuitive visualization. The results demonstrate that Puffin-World generalizes well across diverse scenes and remains robust under challenging camera configurations.

#### A0.7.2 Camera-Controllable Generation

Additional qualitative results are presented in Figure A9 and Figure A11, covering diverse spatial configurations across both indoor and outdoor environments. Puffin-World enables flexible free-viewpoint spatial simulation under a wide range of camera parameters, while preserving realistic appearance, coherent scene layouts, and semantic consistency with the input condition. In particular, the generated views faithfully follow the specified camera orientation and field-of-view, yielding noticeable yet physically plausible changes in perspective, horizon, and spatial composition. These results further demonstrate the robustness and generalization of Puffin-World for camera-controllable generation across diverse scenes and challenging viewpoints.

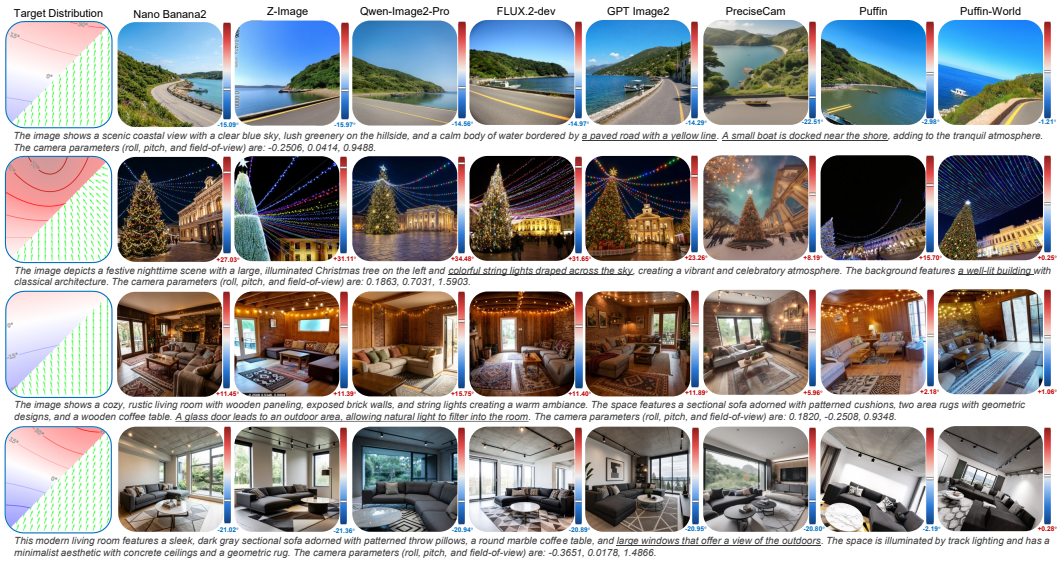


Figure A9: **Comparison results on camera-controllable text-to-image generation for free-viewpoint spatial simulation.** For each result, we visualize the angular error between the simulated gravity and the ground truth; values closer to $0^\circ$ indicate better spatial alignment. Essential elements are underlined in the text prompt.

### A0.7.3 3D World Modeling

As illustrated in Figures A12, A13, and A14, we provide additional qualitative results of 3D world generation under diverse settings, including explicit rotational control, long-horizon trajectories, compound camera actions, text-driven generation, native 3D world-state prediction, and subsequent 3D reconstruction. Across these scenarios, Puffin-World consistently follows the prescribed camera motions while preserving semantic content and coherent spatial structure over extended viewpoint changes. The model remains robust under large rotations and complex combinations of translation and rotation, where maintaining both appearance consistency and physical plausibility is particularly challenging. Moreover, text-driven generation produces geometrically coherent multi-view worlds that remain faithful to the input prompt, while the predicted physics, geometry, and appearance states exhibit strong mutual consistency. The resulting 3D reconstructions further recover clear scene structure and stable spatial layouts, demonstrating that Puffin-World can consolidate its generated multi-view observations into a coherent 3D representation.

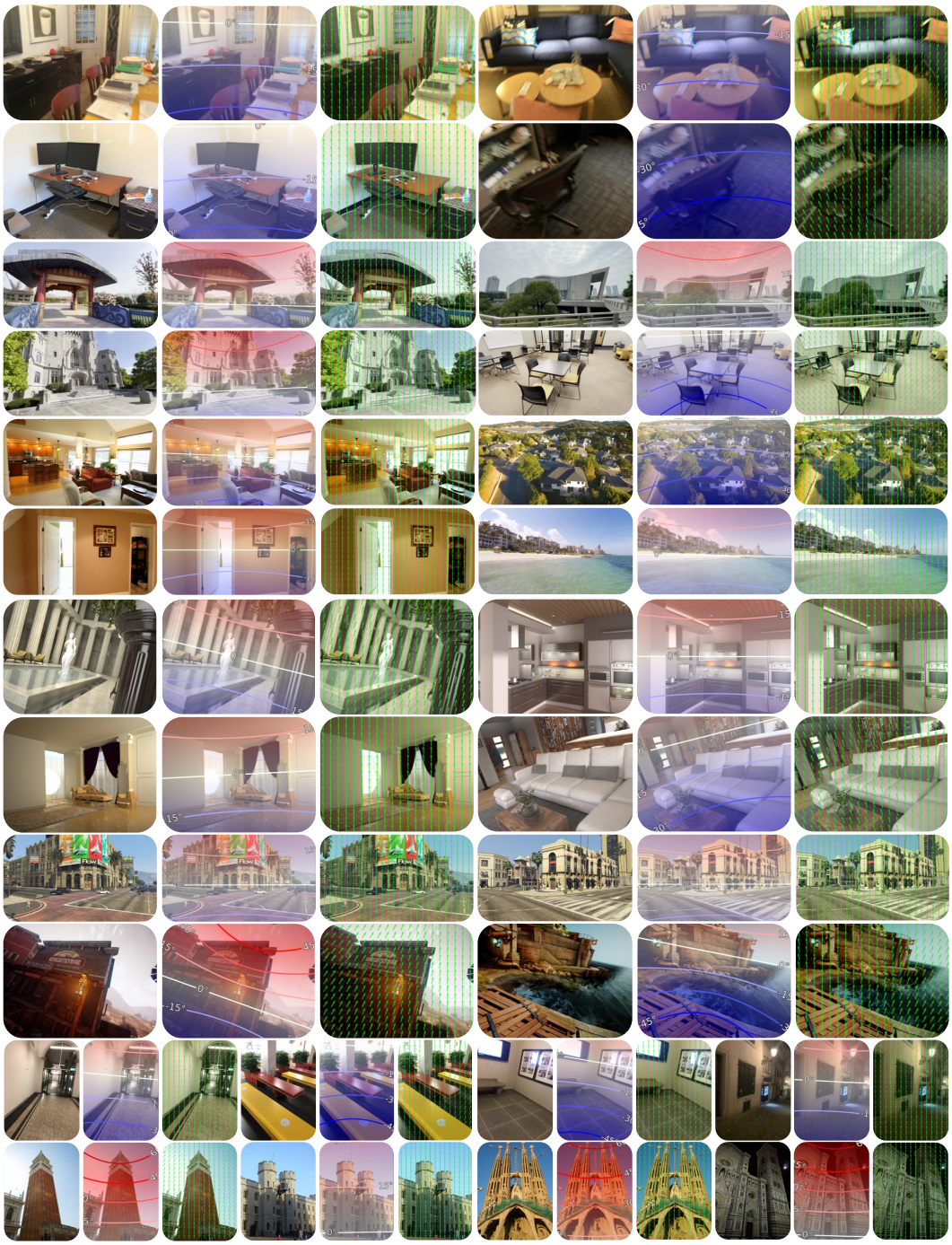

Figure A10: **Visualizations of the estimated physics world state (gravity field and latitude map) on different datasets (Dai et al., 2017; Ling et al., 2024; Zhou et al., 2018; Roberts et al., 2021; Huang et al., 2018; Wang et al., 2020; Sarlin et al., 2022; Li & Snavely, 2018), in which the absolute camera parameters are predicted by Puffin-World.**

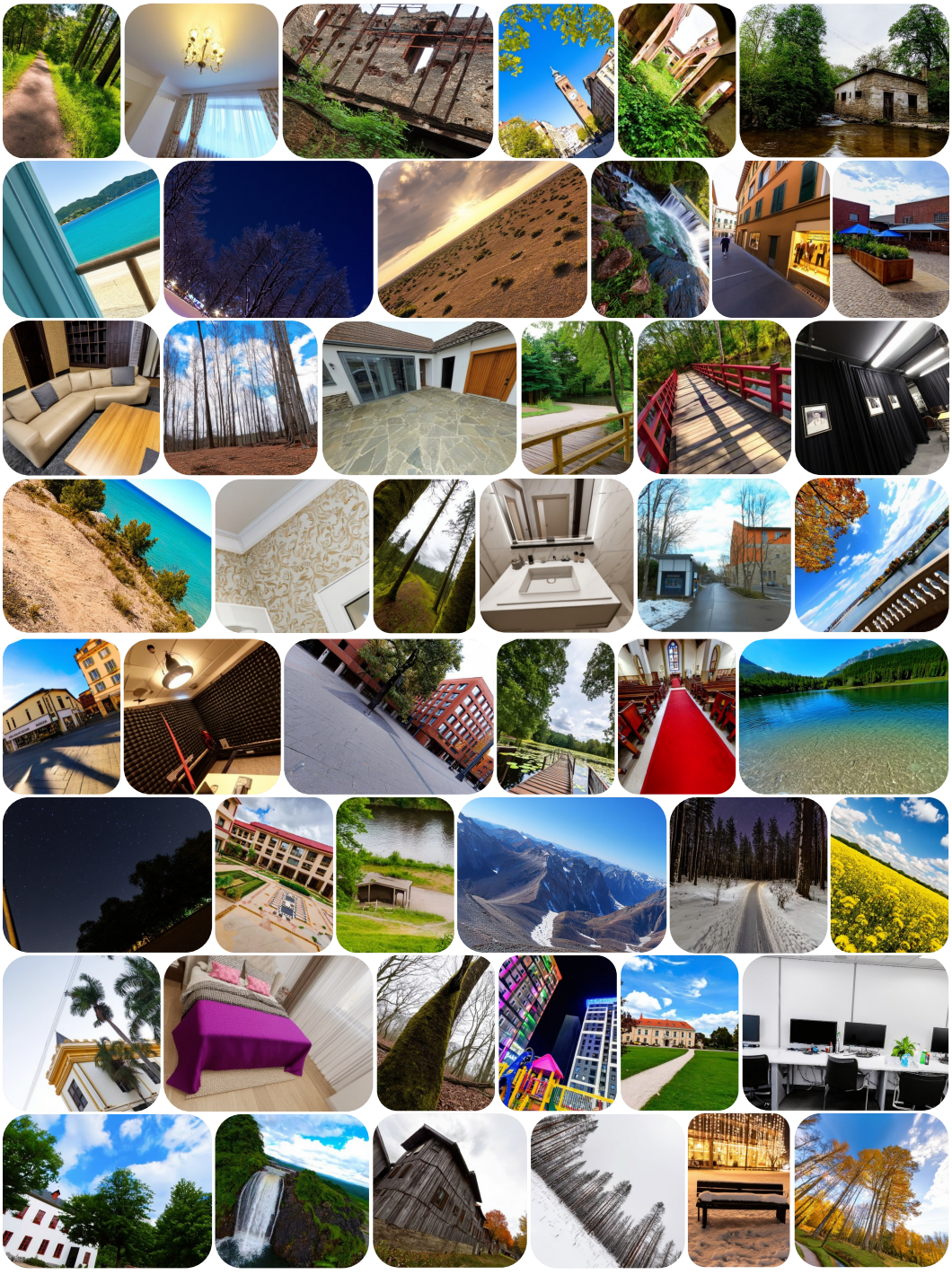

Figure A11: **Camera-controllable text-to-image generation results for free-viewpoint spatial simulation.**



Figure A12: **3D world generation with diverse rotational controls (top: roll, middle: pitch, bottom: yaw).** For each control, we show the generated results from the single-pass path (with visualized physics world state) and recursive path.

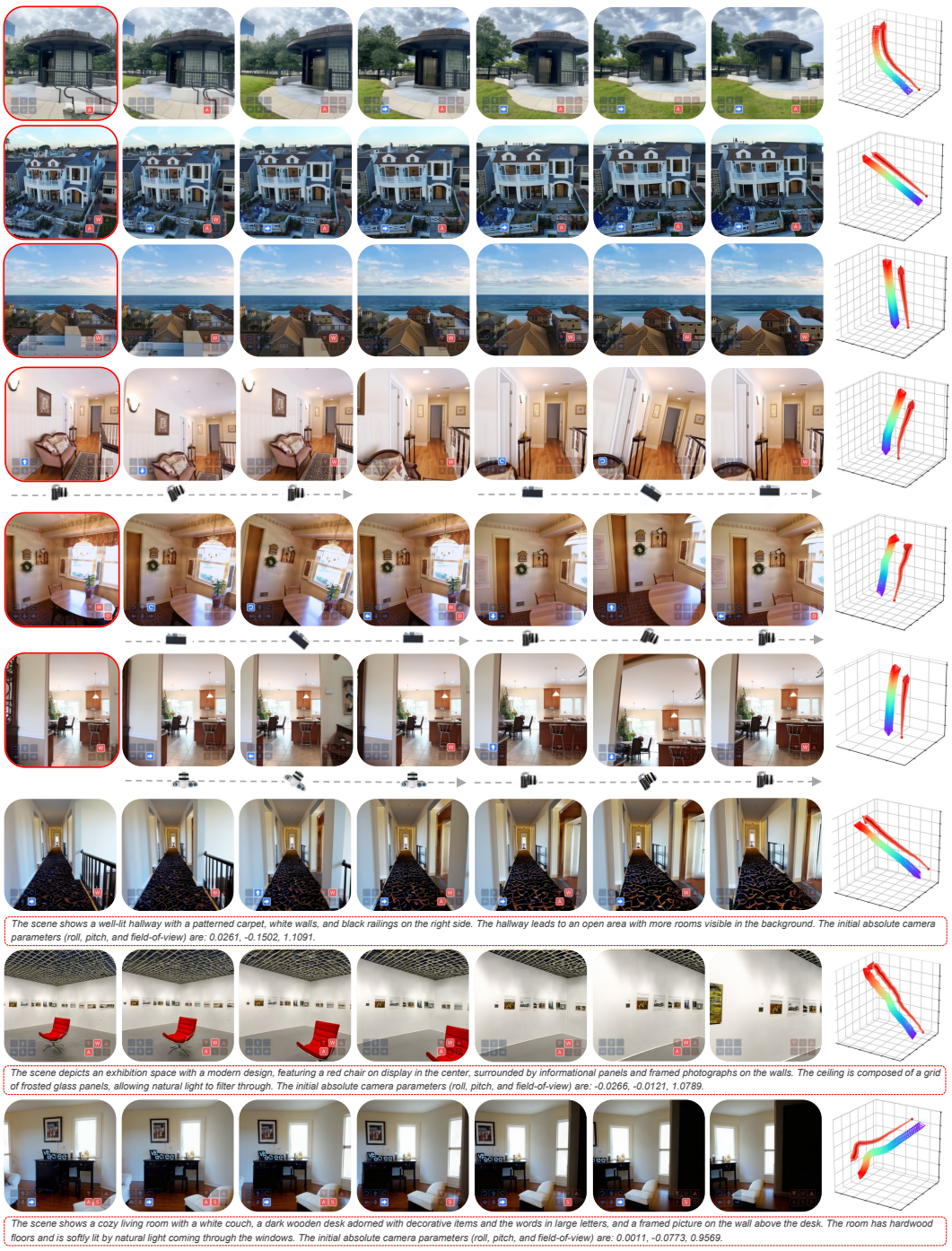


Figure A13: **3D world generation with diverse camera actions and conditions, including long trajectories, compound motions, and text-driven generation.**

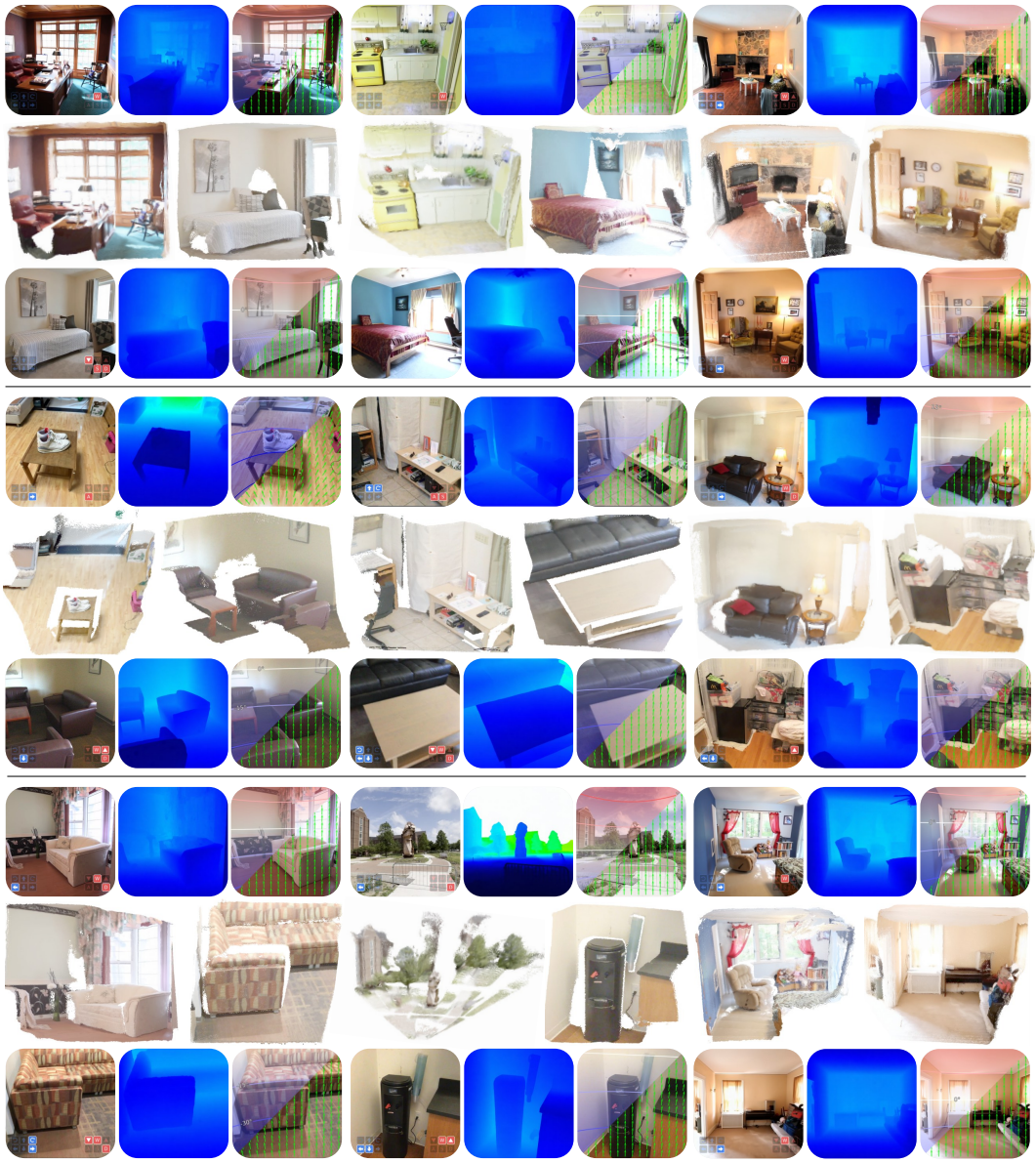

Figure A14: **Additional qualitative results of 3D world modeling.** For each combination, we show the generated native 3D world states: appearance (image), geometry (depth), and physics (gravity field and latitude map), and the 3D reconstructed result. Note that only one group of the generated world states is visualized in each sample for clarity.