Novelty Assessment Report

Paper: Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation
PDF URL: https://openreview.net/pdf?id=5THcDkGGjt
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Camera-centric understanding and generation are two cornerstones of spatial intelligence, yet they are typically studied in isolation. We present Puffin, a unified camera-centric multimodal model that extends spatial awareness along the camera dimension. Puffin integrates language regression and diffusion-based generation to interpret and create scenes from arbitrary viewpoints. To bridge the modality gap between cameras and vision-language, we introduce a novel paradigm that treats camera as language, enabling thinking with camera. This guides the model to align spatially grounded visual cues with photographic terminology while reasoning across geometric context. Puffin is trained on Puffin-4M, a large-scale dataset of 4 million vision-language-camera triplets. We incorporate both global camera parameters and pixel-wise camera maps, yielding flexible and reliable spatial generation. Experiments demonstrate Puffin's superior performance over specialized models for camera-centric generation and understanding. With instruction tuning, Puffin generalizes to diverse cross-view tasks such as spatial imagination, world exploration, and photography guidance. We will release the code, models, dataset pipeline, and benchmark to advance multimodal spatial intelligence research.

Core Task Landscape

This paper addresses: **camera-centric understanding and generation from arbitrary viewpoints**
A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Novel View Synthesis from Sparse or Single Views**
- **Dynamic Scene View Synthesis**
- **Scene-Level Generation and Exploration**
- **Multi-View Consistent Generation**
- **Specialized View Synthesis Modalities**
- **Camera-Aware Perception and Reasoning**
- **Camera Control and Interaction**
- **Foundational Representations and Theory**

Complete Taxonomy Tree

- camera-centric understanding and generation from arbitrary viewpoints Survey Taxonomy
- Novel View Synthesis from Sparse or Single Views
  - Single-Image View Synthesis (4 papers)
  - [2] Synsin: End-to-end view synthesis from a single image (Wiles, 2020) View paper
  - [6] altiro3d: scene representation from single image and novel view synthesis (E. Canessa, 2024) View paper
  - [9] Layer-structured 3d scene inference via view synthesis (Shubham Tulsiani, 2018) View paper
  - [36] PhysGen3D: Crafting a Miniature Interactive World from a Single Image (Boyuan Chen, 2025) View paper
  - Sparse-View Geometry and Appearance Estimation (5 papers)
  - [1] Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views (Shangzhan Zhang, 2025) View paper
  - [5] Free view synthesis (Gernot Riegler, 2020) View paper
  - [22] Free3D: Consistent Novel View Synthesis Without 3D Representation (Chuanxia Zheng, 2024) View paper
  - [38] Generalizable Novel-View Synthesis Using a Stereo Camera (Haechan Lee, 2024) View paper
  - [44] NeuLighting: Neural Lighting for Free Viewpoint Outdoor Scene Relighting with Unconstrained Photo Collections (Quewei Li, 2022) View paper
- Dynamic Scene View Synthesis
  - Monocular Dynamic View Synthesis (3 papers)
  - [8] Generative camera dolly: Extreme monocular dynamic novel view synthesis (Van Hoorick, 2024) View paper
  - [26] Dynamic View Synthesis from Small Camera Motion Videos (Huiqiang Sun, 2025) View paper
  - [35] Pseudo-Generalized Dynamic View Synthesis from a Video (Zhao Xiao-ming, 2023) View paper
  - Multi-Camera Dynamic Reconstruction (3 papers)
  - [20] Fast free-viewpoint video synthesis algorithm for sports scenes (Jun Chen, 2019) View paper
  - [42] ModalNeRF: Neural Modal Analysis and Synthesis for Freeâ□Viewpoint Navigation in Dynamically Vibrating Scenes (Automne Petitjean, 2023) View paper
  - [43] Virtualized reality: Constructing time-varying virtual worlds from real world events (Peter Rander, 1997) View paper
  - Temporal and Spatial Consistency in Dynamic Scenes (2 papers)
  - [21] HybridPlane: A General 4D Representation for Dynamic Scene Reconstruction (Ru Jia, 2025) View paper

    - [27] 3D Scene Prompting for Scene-Consistent Camera-Controllable Video Generation (Jung Jaewoo, 2025) View paper
- Scene-Level Generation and Exploration
    - Large-Scale Scene Generation from Foundation Models (4 papers)
    - [3] Megascenes: Scene-level view synthesis at scale (Chou, 2024) View paper
    - [7] 3d-scenedreamer: Text-driven 3d-consistent scene generation (Songchun Zhang, 2024) View paper
    - [25] Video Perception Models for 3D Scene Synthesis (Huang Rui, 2025) View paper
    - [28] SPATIALGEN: Layout-guided 3D Indoor Scene Generation (Fang Chuan, 2025) View paper
    - Free-Trajectory View Synthesis in Driving Scenes (4 papers)
    - [4] Freevs: Generative view synthesis on free driving trajectory (Wang, 2024) View paper
    - [16] FreeGen: Feed-Forward Reconstruction-Generation Co-Training for Free-Viewpoint Driving Scene Synthesis (Shijie Chen, 2025) View paper
    - [18] ArbiViewGen: Controllable Arbitrary Viewpoint Camera Data Generation for Autonomous Driving via Stable Diffusion Models (Chen Jingfeng, 2025) View paper
    - [34] FreeSim: Toward Free-viewpoint Camera Simulation in Driving Scenes (Lue Fan, 2025) View paper
- Multi-View Consistent Generation
    - Synchronized Multi-Camera Video Generation (2 papers)
    - [11] Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints (Bai Jian-hong, 2024) View paper
    - [33] BEV-VAE: Multi-view Image Generation with Spatial Consistency for Autonomous Driving (Chen Ze-Ming, 2025) View paper
    - Static Multi-View Synthesis (2 papers)
    - [17] S^2VG: 3D Stereoscopic and Spatial Video Generation via Denoising Frame Matrix (Dai Peng, 2025) View paper
    - [32] Improving Viewpoint Consistency in 3D Generation via Structure Feature and CLIP Guidance (Zhang Qing, 2024) View paper
- Specialized View Synthesis Modalities
    - Panoramic and 360-Degree View Synthesis (3 papers)
    - [12] Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image (Giovanni Pintore, 2023) View paper
    - [14] See360: Novel panoramic view interpolation (Zhi-Song Liu, 2022) View paper
    - [37] View synthesis for 360 panoramic spherical images using Multiplane Images (Koirala, 2025) View paper
    - Specialized Sensor and Representation Modalities (3 papers)
    - [13] View-dependent scene appearance synthesis using inverse rendering from light fields (Dahyun Kang, 2021) View paper
    - [15] Controllable Audio-Visual Viewpoint Generation from 360Â° Spatial Information (Marinoni, 2025) View paper
    - [30] An Event Camera Simulator for Arbitrary Viewpoints Based on Neural Radiance Fields (Diego Rodriguez, 2025) View paper
- Camera-Aware Perception and Reasoning
    - Camera-Conditioned Semantic Understanding (2 papers)
    - [19] OccuFly: A 3D Vision Benchmark for Semantic Scene Completion from the Aerial Perspective (Markus Gross, 2025) View paper
    - [23] BEVSegFormer: Bird's Eye View Semantic Segmentation From Arbitrary Camera Rigs (Lang Peng, 2023) View paper
    - Unified Camera-Centric Multimodal Models ★ (3 papers)
    - [0] Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation (Anon et al., 2026) View paper
    - [40] Agent3D-Zero: An Agent for Zero-shot 3D Understanding (Zhang Sha, 2024) View paper
    - [45] MVLLaVA: An Intelligent Agent for Unified and Flexible Novel View Synthesis (Hanyu Jiang, 2024) View paper
- Camera Control and Interaction
    - Interactive Camera Manipulation Interfaces (2 papers)
    - [29] Intuitive and efficient camera control with the toric space (Christophe Lino, 2015) View paper
    - [46] Camera Viewpoint Control with the {Interaction Table} (Hachet, 2024) View paper
    - Camera-Controlled Video Generation (2 papers)
    - [24] Perspectivenet: A scene-consistent image generator for new view synthesis in real indoor environments (David NovotnÃ½, 2019) View paper
    - [39] CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models (He Hao, 2025) View paper
- Foundational Representations and Theory
    - Geometric Representations for View Synthesis (3 papers)
    - [47] Plenoptic image editing (Steven M. Seitz, 2002) View paper
    - [48] From reference frames to reference planes: Multi-view parallax geometry and applications (M. Irani, 1998) View paper
    - [50] Unconstrained Scene Generation with Locally Conditioned Radiance Fields (DeVries, 2021) View paper
    - Image-Based Rendering Foundations (2 papers)
    - [31] Real-time viewpoint image synthesis using strips of multi-camera images (Munekazu Date, 2015) View paper
    - [41] Image-based transformation of viewpoint and scene appearance (Steven M. Seitz, 1997) View paper
    - Spatial Reasoning and Human-Object Interaction (2 papers)
    - [10] Chorus: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images (Sookwan Han, 2023) View paper
    - [49] Hybrid visual and memory search for scenes and objects with variable viewpoints. (Bochao Zou, 2024) View paper

Narrative

Core task: camera-centric understanding and generation from arbitrary viewpoints. The field encompasses methods that synthesize, reason about, or manipulate visual content under varying camera perspectives. At the highest level, the taxonomy divides into eight major branches. Novel View Synthesis from Sparse or Single Views (e.g., Synsin[2], Free View Synthesis[5]) focuses on reconstructing scenes from limited input, while Dynamic Scene View Synthesis extends this to temporal settings. Scene-Level Generation and Exploration (e.g., Megascenes[3], 3D SceneDreamer[7]) emphasizes creating or navigating large-scale environments, and Multi-View Consistent Generation ensures coherence across viewpoints. Specialized View Synthesis Modalities address domain-specific rendering (panoramic, event-based, etc.), whereas Camera-Aware Perception and Reasoning targets tasks like bird's-eye-view segmentation or multimodal understanding that explicitly leverage camera geometry. Camera Control and Interaction (e.g., Generative Camera Dolly[8]) provides user-driven trajectory specification, and Foundational Representations and Theory underpins the geometric and learning frameworks common to all branches.

Several active lines of work highlight contrasting emphases: some pursue end-to-end generative models that produce novel views directly from text or sparse images (Flare[1], ArbiViewGen[18]), while others build explicit 3D representations before rendering. Trade-offs between computational efficiency, geometric fidelity, and generalization remain central. Within Camera-Aware Perception and Reasoning,

a small cluster of Unified Camera-Centric Multimodal Models integrates vision-language understanding with viewpoint reasoning. Thinking with Camera[0] sits squarely in this cluster, emphasizing joint reasoning over camera parameters and scene semantics in a multimodal framework. Compared to Agent3D Zero[40], which focuses on embodied navigation tasks, and MVLLaVA[45], which targets multi-view visual question answering, Thinking with Camera[0] appears to prioritize a broader integration of camera control signals within large-scale vision-language architectures, bridging perception and interactive generation.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Agent3D-Zero: An Agent for Zero-shot 3D Understanding

Authors: Zhang Sha, Huang Di, Sha Zhang, Deng Jia-Jun, Di Huang, et al. (17 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The ability to understand and reason the 3D real world is a crucial milestone towards artificial general intelligence. The current common practice is to finetune Large Language Models (LLMs) with 3D data and texts to enable 3D understanding. Despite their effectiveness, these approaches are inherently limited by the scale and diversity of the available 3D data. Alternatively, in this work, we introduce Agent3D-Zero, an innovative 3D-aware agent framework addressing the 3D scene understanding in ...

Relationship Analysis

Both papers belong to the unified camera-centric multimodal models category, aiming to integrate camera understanding and generation capabilities. While the original paper (Puffin) focuses on unifying single-view camera calibration and text-to-image controllable generation through a joint autoregressive-diffusion framework with explicit camera parameter reasoning, Agent3D-Zero addresses 3D scene understanding through zero-shot multi-view image analysis without requiring 3D data or camera parameter estimation. The key difference is that Puffin performs explicit camera parameter prediction and camera-controllable generation, whereas Agent3D-Zero uses a VLM agent to actively select viewpoints for 3D reasoning tasks without direct camera parameter modeling.

2. MVLLaVA: An Intelligent Agent for Unified and Flexible Novel View Synthesis

Authors: Hanyu Jiang, Jian Xue, Xing Lan, Guohong Hu, Ke Lu | **Year/Venue**: 2024 | **URL**: View paper

Abstract

This paper introduces MVLLaVA, an intelligent agent designed for novel view synthesis tasks. MVLLaVA integrates multiple multi-view diffusion models with a large multimodal model, LLaVA, enabling it to handle a wide range of tasks efficiently. MVLLaVA represents a versatile and unified platform that adapts to diverse input types, including a single image, a descriptive caption, or a specific change in viewing azimuth, guided by language instructions for viewpoint generation. We carefully craft t...

Relationship Analysis

Both papers belong to the Unified Camera-Centric Multimodal Models category, focusing on integrating camera awareness into multimodal architectures. While the original paper (Puffin) unifies camera-centric understanding and generation through a single model that treats camera parameters as language with explicit spatial reasoning, MVLLaVA takes a different approach by using a large multimodal model (LLaVA) as an intelligent agent to orchestrate multiple specialized multi-view diffusion models. The key distinction is that Puffin performs end-to-end camera understanding and generation within one unified framework, whereas MVLLaVA acts as a task router that selects and coordinates separate downstream models based on instruction interpretation.

Contributions Analysis

Overall novelty summary. Puffin proposes a unified camera-centric multimodal model that jointly performs understanding and generation tasks conditioned on camera parameters. Within the taxonomy, it resides in the 'Unified Camera-Centric Multimodal Models' leaf under 'Camera-Aware Perception and Reasoning'. This leaf contains only three papers total, including Puffin itself, indicating a relatively sparse and emerging research direction. The sibling works—Agent3D Zero and MVLLaVA—focus on embodied navigation and multi-view visual question answering respectively, whereas Puffin emphasizes broader integration of camera control with vision-language architectures for both perception and generation.

The taxonomy reveals that most camera-centric work concentrates on synthesis tasks (Novel View Synthesis, Dynamic Scene View Synthesis, Scene-Level Generation) or specialized modalities (panoramic, event-based rendering). Camera-Aware Perception and Reasoning is a smaller branch with only two leaf nodes: Camera-Conditioned Semantic Understanding (two papers on segmentation) and Unified Camera-Centric Multimodal Models (three papers). Puffin's positioning suggests it bridges perception-focused methods and generative approaches, diverging from purely synthesis-oriented branches by incorporating explicit reasoning over camera parameters within a multimodal framework. The taxonomy's scope and exclude notes clarify that Puffin's unified architecture distinguishes it from single-task models in neighboring leaves.

Among 28 candidates examined across three contributions, no clearly refutable prior work was identified. For the unified model contribution, 10 candidates were examined with zero refutations; for the 'thinking with camera' paradigm, 9 candidates yielded no refutations; and for the Puffin-4M dataset, 9 candidates similarly showed no overlapping prior work. This suggests that within the limited search scope—top-K semantic matches plus citation expansion—no existing work directly anticipates Puffin's combination of camera-as-language reasoning, joint understanding-generation architecture, and large-scale vision-language-camera triplet training. The absence of refutations across all contributions indicates potential novelty, though the search was not exhaustive.

Based on the limited literature search of 28 candidates, Puffin appears to occupy a relatively unexplored niche at the intersection of camera-aware reasoning and multimodal generation. The sparse population of its taxonomy leaf and the lack of refutable prior work suggest meaningful novelty, though a broader search might reveal additional related efforts. The analysis covers top semantic matches and does not claim completeness across the entire field.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Puffin: unified camera-centric multimodal model

Description: The authors introduce Puffin, a unified framework that jointly performs camera-centric understanding (estimating camera parameters from images) and generation (controllable image synthesis from camera parameters). This represents the first attempt to unify these two traditionally isolated tasks within a single multimodal model.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Training-free camera control for video generation

URL: View paper

Brief Assessment

Training Free Camera[75] focuses on training-free camera control for video generation using point cloud modeling and noise prior guidance. This is fundamentally different from Puffin's unified framework for camera understanding and image generation within a multimodal model.

---

2. Vggt: Visual geometry grounded transformer

URL: View paper

Brief Assessment

VGGT[69] focuses on 3D reconstruction tasks (camera parameters, depth maps, point maps, point tracks) from multiple views using visual geometry, not on unifying camera-centric understanding with controllable image generation as in Puffin.

---

3. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis

URL: View paper

Brief Assessment

Generative Photography[78] focuses on camera intrinsic settings (aperture, shutter speed, focal length, color temperature) for text-to-image generation, not on unifying camera parameter estimation and controllable synthesis within a single multimodal model.

---

4. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control

URL: View paper

Brief Assessment

RealCam I2V[76] focuses on camera-controlled image-to-video generation with metric-scale alignment and 3D scene reconstruction, not on unifying camera understanding and generation tasks within a single multimodal model framework.

---

5. Vidcraft3: Camera, object, and lighting control for image-to-video generation

URL: View paper

Brief Assessment

VidCraft3[74] focuses on controllable image-to-video generation with camera, object, and lighting control, not on unifying camera parameter estimation and controllable image synthesis within a single multimodal model for static images.

---

6. Gen3c: 3d-informed world-consistent video generation with precise camera control

URL: View paper

Brief Assessment

Gen3C[71] focuses on video generation with 3D-informed camera control using point clouds and depth estimation, not on unifying camera parameter estimation and controllable image synthesis within a single multimodal model like Puffin.

---

7. Imagedream: Image-prompt multi-view diffusion for 3d generation

URL: View paper

Brief Assessment

ImageDream[70] focuses on image-prompt multi-view diffusion for 3D generation, not on unifying camera parameter estimation and controllable image synthesis within a single multimodal model.

---

8. Multimodal image synthesis and editing: A survey and taxonomy

URL: View paper

Brief Assessment

Multimodal Image Synthesis[77] is a survey paper that briefly mentions camera-conditioned generation as one technique among many multimodal synthesis methods. It does not present a unified model integrating camera understanding and generation tasks.

---

9. CamEdit: Continuous Camera Parameter Control for Photorealistic Image Editing

URL: View paper

Brief Assessment

CamEdit[73] focuses on photorealistic image editing through continuous camera parameter control (aperture, focal plane, shutter speed), not on unified camera understanding and generation tasks that Puffin addresses.

---

10. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis

URL: View paper

Brief Assessment

Flovd[72] focuses on camera-controlled video synthesis using optical flow, not on unifying camera parameter estimation and controllable image synthesis within a single multimodal model.

Contribution 2: Thinking with camera paradigm

Description: The authors propose a novel mechanism called thinking with camera that bridges the modality gap between camera parameters and vision-language models. It aligns spatially grounded visual cues with professional photographic terms through structured spatial reasoning across geometric context (roll, pitch, FoV), enabling both accurate parameter prediction and controllable generation.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. TIGeR: Tool-Integrated Geometric Reasoning in Vision-Language Models for Robotics

URL: View paper

Brief Assessment

TIGeR[61] focuses on tool-integrated geometric reasoning for robotics through code generation and external computation libraries, not on aligning camera parameters with vision-language models through spatial reasoning as described in the original contribution.

---

2. RoboRetriever: Single-Camera Robot Object Retrieval via Active and Interactive Perception with Dynamic Scene Graph

URL: View paper

Brief Assessment

RoboRetriever[67] focuses on robotic object retrieval using active and interactive perception with scene graphs, not on aligning camera parameters with vision-language models through spatial reasoning for camera understanding and generation tasks.

3. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation

URL: View paper

Brief Assessment

3DS VLA[63] focuses on robotic manipulation with 3D spatial awareness for action prediction, not on aligning camera parameters with vision-language models through spatial reasoning for camera understanding and generation tasks.

4. CameraBench: Benchmarking Visual Reasoning in MLLMs via Photography

URL: View paper

Brief Assessment

CameraBench[65] focuses on evaluating MLLMs' ability to identify numerical camera settings through visual reasoning tasks, not on aligning camera parameters with vision-language models through spatial reasoning mechanisms for both understanding and generation.

5. Spatialrgpt: Grounded spatial reasoning in vision-language models

URL: View paper

Brief Assessment

SpatialRGPT[53] focuses on grounding spatial reasoning in vision-language models through 3D scene graphs and region-aware representations, not on aligning camera parameters with VLMs through structured spatial reasoning across geometric context (roll, pitch, FoV).

6. SpaceMind: Camera-Guided Modality Fusion for Spatial Reasoning in Vision-Language Models

URL: View paper

Brief Assessment

SpaceMind[66] focuses on camera-guided modality fusion for spatial reasoning in vision-language models, treating camera as a guiding modality for feature fusion. The original paper's 'thinking with camera' paradigm aligns camera parameters with photographic terminology through structured spatial reasoning, which is a distinct approach from SpaceMind's fusion mechanism.

7. Large Language Models and 3D Vision for Intelligent Robotic Perception and Autonomy: A Review

URL: View paper

Brief Assessment

LLMs 3D Vision[68] is a review paper focused on integrating large language models with 3D vision for robotic perception and autonomy. It does not propose a specific mechanism for aligning camera parameters with vision-language models through spatial reasoning, which is the core novelty of the thinking with camera paradigm.

8. Robomm: All-in-one multimodal large model for robotic manipulation

URL: View paper

Brief Assessment

RoboMM[64] focuses on robotic manipulation using camera parameters for 3D spatial perception in robotics tasks, not on aligning camera parameters with vision-language models through spatial reasoning for camera understanding and generation as in the original paper.

9. Grounding actions in camera space: Observation-centric vision-language-action policy

URL: View paper

Brief Assessment

Grounding Camera Space[62] focuses on transforming robot actions from robot base coordinates to camera coordinates for robotic manipulation, not on aligning camera parameters with vision-language models through spatial reasoning for camera understanding and generation tasks.

Contribution 3: Puffin-4M dataset and benchmark

Description: The authors construct Puffin-4M, a large-scale dataset containing 4 million vision-language-camera triplets with precise camera parameters, descriptive captions, pixel-wise camera maps, and spatial reasoning annotations. They also establish comprehensive benchmarks (Puffin-Gen and Puffin-Und) for evaluating camera-centric multimodal models.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding

URL: View paper

Brief Assessment

Llava ST[59] focuses on spatial-temporal video understanding with 4.3M samples for video grounding tasks, not camera-centric vision-language-camera triplets with precise camera parameters and pixel-wise camera maps as in the original paper.

2. Visual language maps for robot navigation

URL: View paper

Brief Assessment

Visual Language Maps[56] focuses on spatial map representations for robot navigation using visual-language features fused with 3D reconstruction, not on constructing large-scale vision-language-camera triplet datasets with camera parameters and spatial reasoning annotations for training multimodal models.

3. A survey on multimodal large language models for autonomous driving

URL: View paper

Brief Assessment

Multimodal Autonomous Driving[51] focuses on multimodal large language models for autonomous driving applications, not on camera-centric vision-language-camera triplet datasets for spatial understanding and generation.

4. Covla: Comprehensive vision-language-action dataset for autonomous driving

URL: View paper

Brief Assessment

Covla[57] focuses on autonomous driving with vision-language-action triplets for trajectory prediction, not camera-centric spatial understanding with camera parameters and pixel-wise camera maps as in Puffin-4M.

---

5. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model

URL: View paper

Brief Assessment

SpatialRGPT[55] focuses on 3D spatial reasoning with region proposals and depth information, not on vision-language-camera triplets with camera parameters for camera-centric understanding and generation.

---

6. Spatialrgpt: Grounded spatial reasoning in vision-language models

URL: View paper

Brief Assessment

SpatialRGPT[53] introduces the Open Spatial Dataset (OSD) with 8.7M spatial concepts and SpatialRGPT-Bench for 3D spatial cognition evaluation, which differs from Puffin-4M's focus on vision-language-camera triplets with precise camera parameters, descriptive captions, pixel-wise camera maps, and spatial reasoning annotations.

---

7. Spatialladder: Progressive training for spatial reasoning in vision-language models

URL: View paper

Brief Assessment

SpatialLadder[60] focuses on progressive training for spatial reasoning with a 26K-sample dataset spanning object localization and multi-modal spatial tasks, while the original paper presents a 4M-sample dataset with vision-language-camera triplets including precise camera parameters and pixel-wise camera maps for camera-centric understanding and generation.

---

8. SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding

URL: View paper

Brief Assessment

SceneVerse[58] focuses on 3D vision-language learning for grounded scene understanding with 68k 3D indoor scenes and 2.5m vision-language pairs, not vision-language-camera triplets with precise camera parameters for camera-centric understanding and generation as in the original paper.

---

9. Thinking in space: How multimodal large language models see, remember, and recall spaces

URL: View paper

Brief Assessment

Thinking in Space[52] focuses on video-based visual-spatial intelligence benchmarks (VSI-Bench) for evaluating how MLLMs understand and recall spatial layouts from videos, not on constructing large-scale vision-language-camera triplet datasets with precise camera parameters for training camera-centric multimodal models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation View paper
- [1] Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views View paper
- [2] Synsin: End-to-end view synthesis from a single image View paper
- [3] Megascenes: Scene-level view synthesis at scale View paper
- [4] Freevs: Generative view synthesis on free driving trajectory View paper
- [5] Free view synthesis View paper
- [6] altiro3d: scene representation from single image and novel view synthesis View paper
- [7] 3d-scenedreamer: Text-driven 3d-consistent scene generation View paper
- [8] Generative camera dolly: Extreme monocular dynamic novel view synthesis View paper
- [9] Layer-structured 3d scene inference via view synthesis View paper
- [10] Chorus: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images View paper
- [11] Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints View paper
- [12] Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image View paper
- [13] View-dependent scene appearance synthesis using inverse rendering from light fields View paper
- [14] See360: Novel panoramic view interpolation View paper
- [15] Controllable Audio-Visual Viewpoint Generation from 360Â° Spatial Information View paper
- [16] FreeGen: Feed-Forward Reconstruction-Generation Co-Training for Free-Viewpoint Driving Scene Synthesis View paper
- [17] S^2VG: 3D Stereoscopic and Spatial Video Generation via Denoising Frame Matrix View paper
- [18] ArbiViewGen: Controllable Arbitrary Viewpoint Camera Data Generation for Autonomous Driving via Stable Diffusion Models View paper
- [19] OccuFly: A 3D Vision Benchmark for Semantic Scene Completion from the Aerial Perspective View paper
- [20] Fast free-viewpoint video synthesis algorithm for sports scenes View paper
- [21] HybridPlane: A General 4D Representation for Dynamic Scene Reconstruction View paper
- [22] Free3D: Consistent Novel View Synthesis Without 3D Representation View paper
- [23] BEVSegFormer: Bird's Eye View Semantic Segmentation From Arbitrary Camera Rigs View paper
- [24] Perspectivenet: A scene-consistent image generator for new view synthesis in real indoor environments View paper
- [25] Video Perception Models for 3D Scene Synthesis View paper
- [26] Dynamic View Synthesis from Small Camera Motion Videos View paper
- [27] 3D Scene Prompting for Scene-Consistent Camera-Controllable Video Generation View paper
- [28] SPATIALGEN: Layout-guided 3D Indoor Scene Generation View paper
- [29] Intuitive and efficient camera control with the toric space View paper
- [30] An Event Camera Simulator for Arbitrary Viewpoints Based on Neural Radiance Fields View paper
- [31] Real-time viewpoint image synthesis using strips of multi-camera images View paper

- [32] Improving Viewpoint Consistency in 3D Generation via Structure Feature and CLIP Guidance View paper
- [33] BEV-VAE: Multi-view Image Generation with Spatial Consistency for Autonomous Driving View paper
- [34] FreeSim: Toward Free-viewpoint Camera Simulation in Driving Scenes View paper
- [35] Pseudo-Generalized Dynamic View Synthesis from a Video View paper
- [36] PhysGen3D: Crafting a Miniature Interactive World from a Single Image View paper
- [37] View synthesis for 360 panoramic spherical images using Multiplane Images View paper
- [38] Generalizable Novel-View Synthesis Using a Stereo Camera View paper
- [39] CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models View paper
- [40] Agent3D-Zero: An Agent for Zero-shot 3D Understanding View paper
- [41] Image-based transformation of viewpoint and scene appearance View paper
- [42] ModalNeRF: Neural Modal Analysis and Synthesis for Freeâ□□Viewpoint Navigation in Dynamically Vibrating Scenes View paper
- [43] Virtualized reality: Constructing time-varying virtual worlds from real world events View paper
- [44] NeuLighting: Neural Lighting for Free Viewpoint Outdoor Scene Relighting with Unconstrained Photo Collections View paper
- [45] MVLLaVA: An Intelligent Agent for Unified and Flexible Novel View Synthesis View paper
- [46] Camera Viewpoint Control with the {Interaction Table} View paper
- [47] Plenoptic image editing View paper
- [48] From reference frames to reference planes: Multi-view parallax geometry and applications View paper
- [49] Hybrid visual and memory search for scenes and objects with variable viewpoints. View paper
- [50] Unconstrained Scene Generation with Locally Conditioned Radiance Fields View paper
- [51] A survey on multimodal large language models for autonomous driving View paper
- [52] Thinking in space: How multimodal large language models see, remember, and recall spaces View paper
- [53] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [54] Multimodal fusion and vision-language models: A survey for robot vision View paper
- [55] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model View paper
- [56] Visual language maps for robot navigation View paper
- [57] Covla: Comprehensive vision-language-action dataset for autonomous driving View paper
- [58] SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding View paper
- [59] Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding View paper
- [60] Spatialladder: Progressive training for spatial reasoning in vision-language models View paper
- [61] TIGeR: Tool-Integrated Geometric Reasoning in Vision-Language Models for Robotics View paper
- [62] Grounding actions in camera space: Observation-centric vision-language-action policy View paper
- [63] 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation View paper
- [64] Robomm: All-in-one multimodal large model for robotic manipulation View paper
- [65] CameraBench: Benchmarking Visual Reasoning in MLLMs via Photography View paper
- [66] SpaceMind: Camera-Guided Modality Fusion for Spatial Reasoning in Vision-Language Models View paper
- [67] RoboRetriever: Single-Camera Robot Object Retrieval via Active and Interactive Perception with Dynamic Scene Graph View paper
- [68] Large Language Models and 3D Vision for Intelligent Robotic Perception and Autonomy: A Review View paper
- [69] Vggt: Visual geometry grounded transformer View paper
- [70] Imagedream: Image-prompt multi-view diffusion for 3d generation View paper
- [71] Gen3c: 3d-informed world-consistent video generation with precise camera control View paper
- [72] Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis View paper
- [73] CamEdit: Continuous Camera Parameter Control for Photorealistic Image Editing View paper
- [74] Vidcraft3: Camera, object, and lighting control for image-to-video generation View paper
- [75] Training-free camera control for video generation View paper
- [76] Realcam-i2v: Real-world image-to-video generation with interactive complex camera control View paper
- [77] Multimodal image synthesis and editing: A survey and taxonomy View paper
- [78] Generative photography: Scene-consistent camera control for realistic text-to-image synthesis View paper