UCLA
UCLA Electronic Theses and Dissertations

Title

Building World Models with Spatial Intelligence

Permalink

https://escholarship.org/uc/item/8r65494d

ISBN

9798277443613

Author

Zhou, Shijie

Publication Date

2026-03-18

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Building World Models with Spatial Intelligence

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Shijie Zhou

2026

© Copyright by
Shijie Zhou
2026

ABSTRACT OF THE DISSERTATION

Building World Models with Spatial Intelligence

by

Shijie Zhou
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2026
Professor Achuta Kadambi, Chair

The rapid ascent of Large Language Models (LLMs) and Vision-Language Models (VLMs) has revolutionized semantic understanding, yet these systems remain "spectators"—observing the world through 2D projections without true physical grounding. This dissertation, Building World Models with Spatial Intelligence, addresses this critical gap by establishing a blueprint for AI systems that possess Spatial Intelligence: the ability to perceive, imagine, interact with, and reason about the 3D world. We posit that achieving this requires a symbiotic architecture that combines the vast semantic knowledge of 2D foundation models with the geometric precision of explicit 3D/4D representations, specifically Gaussian Splatting.

The research is presented through four core contributions corresponding to the essential faculties of a World Model. First, we introduce Feature 3DGS, a framework for Perception that transforms 2D observations into a semantic, editable, and promptable 3D scene representation. By distilling high-dimensional features from 2D foundation models (e.g., SAM, CLIP) into explicit 3D Gaussian fields, this approach enables spatially grounded multimodal world modeling. Second, to address scenarios where direct 2D observations are absent, we present DreamScene360, an engine for Imagination that generates immersive, globally consistent 3D

environments from natural language descriptions by lifting synthesized 360° panoramas into explicit 3D Gaussians. Third, recognizing that the true physical world is inherently dynamic rather than static, we advance towards a more generalized world modeling framework for Interaction. We introduce Feature4X, which transforms monocular videos into a dynamic, multimodal, and interactive 4D world via a unified spatiotemporal latent feature field. This explicit representation equips VLM-driven agents with the spatial and temporal grounding needed to perform complex tasks such as 4D scene editing and spatiotemporal Visual Question Answering (VQA). However, whether frontier AI is truly capable of solving complex, high-level cognitive tasks within the real, dynamic physical world remains an open question. To investigate this, we finally propose VLM4D, a rigorous benchmark for Reasoning that exposes the fundamental limitations of current VLMs in understanding dynamic physical interactions, driving future advancements in physical AI. Together, these works demonstrate a path toward building World Models that do not just recognize patterns in pixels, but understand the geometry and dynamics of the physical world, laying the groundwork for the next generation of Spatial Computing and Embodied AI.

The dissertation of Shijie Zhou is approved.

Bolei Zhou

Jonathan Chau-Yan Kao

Aydogan Ozcan

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2026

To my family, for everything.

TABLE OF CONTENTS

1 Introduction **1**

1.1 The Dawn of Spatial Intelligence 1

1.2 The Concept of the World Model 2

1.3 The Evolution of 3D Representations: The Quest for Structure 5

1.3.1 Explicit Representations: Precision without Semantics 5

1.3.2 Implicit Representations: The NeRF Revolution 5

1.3.3 The Unification: 3D Gaussian Splatting 6

1.4 The Four Pillars of the Dissertation 7

1.4.1 Pillar I: Perception – Distilling Semantic 3D Representations 7

1.4.2 Pillar II: Imagination – From Text to 3D Worlds 8

1.4.3 Pillar III: Interaction – Bridging Space, Time, and Action 8

1.4.4 Pillar IV: Reasoning – Towards Spatiotemporal Intelligence 9

1.5 Dissertation Statement 9

2 Perception - Distilling Semantic 3D Representations **10**

2.1 Introduction 11

2.2 Related Work 13

2.2.1 Implicit Radiance Field Representations 13

2.2.2 Explicit Radiance Field Representations 13

2.2.3 Feature Field Distillation 14

2.3 Method 14

2.3.1 High-dimensional Semantic Feature Rendering 16

2.3.2 Optimization and Speed-up 18
2.3.3 Promptable Explicit Scene Representation 19
2.4 Experiments 20
2.4.1 Novel view semantic segmentation 20
2.4.2 Segment Anything from Any View 23
2.4.3 Language-guided Editing 25
2.5 Discussion and Conclusion 26

3 Imagination - From Text to 3D Worlds 28
3.1 Introduction 29
3.2 Related Works 31
3.3 Methods 32
3.3.1 Text to 360° Panoramas with Self-Refinement 33
3.3.2 Lifting in-the-wild Panorama to 360 Scene 35
3.3.3 Optimizing Monocular Panoramic 3D Gaussians 36
3.3.4 Optimization 40
3.4 Experiments 40
3.4.1 Experiment Setting 40
3.4.2 Main Results 41
3.4.3 Ablation Study 44
3.5 Conclusion 47

4 Interation - Bridging Space, Time, and Action 48
4.1 Introduction 49

4.2 Related Works . . . 51
4.2.1 4D Representations for Reconstruction and Generation . . . 51
4.2.2 4D Reconstruction from Monocular Video . . . 52
4.2.3 Feature Field Distillation . . . 52
4.3 Method . . . 53
4.3.1 Preliminaries: Dynamic 3D Gaussian Splatting . . . 53
4.3.2 Unified Latent Feature Distillation . . . 54
4.3.3 Scaffold-Based Compact Feature . . . 57
4.3.4 Interaction with AI Agent via Feature Fields . . . 57
4.4 Experiments . . . 58
4.4.1 Scene Interaction with Segment Anything . . . 59
4.4.2 Scene Understanding with Semantics . . . 61
4.4.3 Scene Editing with Language and AI Agent . . . 63
4.4.4 Scene Reasoning with 4D Chatbot Agent . . . 64
4.5 Conclusion . . . 66

5 Reasoning - Towards Spatiotemporal Intelligence . . . 68

5.1 Introduction . . . 69
5.2 Related Work . . . 71
5.3 The `VLM4D` Benchmark . . . 72
5.3.1 Benchmark Construction . . . 73
5.3.2 Categorizing Spatiotemporal Performance . . . 77
5.4 Evaluation of `VLM4D` Benchmark . . . 78
5.4.1 Evaluation Setup . . . 78

5.4.2 Benchmark Results . . . 79
5.5 Analysis: Why VLMs Don't Work Well? . . . 80
5.5.1 Limited Spatiotemporal Cognition . . . 80
5.5.2 Deficiencies in Spatiotemporal Labeling . . . 80
5.6 Probing Future Solutions . . . 82
5.7 Conclusion . . . 84

6 Conclusion . . . 88
6.1 The Architecture of a Spatially Intelligent World Model . . . 88
6.2 Theoretical Implications: From "Learning to See" to "Learning to Simulate" . . . 89
6.2.1 The Redefinition of Rendering . . . 90
6.2.2 The Democratization of Simulation (Sim2Real) . . . 90
6.2.3 Agentic Interaction as a Proxy for Understanding . . . 91
6.3 Critical Limitations and Challenges . . . 91
6.3.1 The Distillation Bottleneck . . . 91
6.3.2 The Kinematic vs. Dynamic Gap . . . 92
6.3.3 The Reasoning Gap . . . 92
6.4 Future Directions: Toward Embodied General Intelligence . . . 92
6.4.1 Physics-Informed Gaussian Splatting . . . 92
6.4.2 Closing the Perception-Action Loop . . . 93
6.4.3 Self-Supervised Spatial Learning . . . 93
6.5 Final Remarks . . . 93

References . . . 95

LIST OF FIGURES

1.1 **Building spatially intelligent world models: an overview of this dissertation.** 7

2.1 **Feature 3DGS**. We present a general method that significantly enhances 3D Gaussian Splatting through the integration of large 2D foundation models via feature field distillation. This advancement extends the capabilities of 3D Gaussian Splatting beyond mere novel view synthesis. It now encompasses a range of functionalities, including semantic segmentation, language-guided editing, and promptable segmentations such as "segment anything" or automatic segmentation of everything from any novel view. Scene from [HPP18]. 10

2.2 **An overview of our method.** We adopt the same 3D Gaussian initialization from sparse SfM point clouds as utilized in 3DGS, with the addition of an essential attribute: the *semantic feature*. Our primary innovation lies in the development of a Parallel N-dimensional Gaussian Rasterizer, complemented by a convolutional speed-up module as an optional branch. This configuration is adept at rapidly rendering arbitrarily high-dimensional features without sacrificing downstream performance. 15

2.3 **Novel view semantic segmentation (LSeg) results on scenes from Replica dataset [SWM19] and LLFF dataset [MSO19].** (a) We show examples of original images in training views together with the ground-truth feature visualizations. (b) We compare the qualitative segmentation results using our Feature 3DGS with the NeRF-DFF [KMS22]. Our inference is **1.66**× faster when rendered feature $dim = 128$. Our method demonstrates more fine-grained segmentation results with higher-quality feature maps. 22

2.4 **Comparison of SAM segmentation results obtained by** (a) naively applying the SAM encoder-decoder module to a novel-view rendered image **with** (b) directly decoding a rendered feature. Our method is up to 1.7× faster in total inference speed including rendering and segmentation while preserving the quality of segmentation masks. Scene from [HPP18]. 23

2.5 **Novel view segmentation (SAM) results compared with NeRF-DFF**. (Upper) NeRF-DFF method presents lower-quality segmentation masks - note the failure on segmenting the cup from the bear and the coarse-grained mask boundary on the bear's leg in box-prompted results. (Lower) Our method provides higher-quality masks with more fine-grained segmentation details. Scene from [KKG23]. 24

2.6 **Demonstration of results with various language-guided edit operations by querying the 3D feature field and comparison with NeRF-DFF** (a) We compare our edit results with NeRF-DFF method on the sample dataset provided by NeRF-DFF [KMS22]. Note that our method outperforms NeRF-DFF method by extracting the entire banana hidden by an apple in the original image and with less floaters in the background. (b) We demonstrate results with deletion and appearance modification on different targets. Note that the car is deleted with background preserved, and the appearance of the leaves changes with the appearance of the stop sign remained the same. 25

3.1 **DreamScene360.** We introduce a 3D scene generation pipeline that creates immersive scenes with full 360° coverage from text prompts of any level of specificity. 28

3.2 **Overall Architecture.** Beginning with a concise text prompt, we employ a diffusion model to generate a 360° panoramic image. A self-refinement process is employed to produce the optimal 2D candidate panorama. Subsequently, a 3D geometric field is utilized to initialize the Panoramic 3D Gaussians. Throughout this process, both semantic and geometric correspondences are employed as guiding principles for the optimization of the Gaussians, aiming to address and fill the gaps resulting from the single-view input. 33

3.3 **Diverse Generation.** We demonstrate that our generated 3D scenes are diverse in style, consistent in geometry, and highly matched with the simple text inputs. 38

3.4 **Visual Comparisons.** We showcase 360° 3D scene generation. In each row, from left to right, displays novel views as the camera undergoes clockwise rotation in yaw, accompanied by slight random rotations in pitch and random translations. LucidDreamer [CLN23] hallucinates novel views from a conditioned image (indicated by a red bounding box) but lacks global semantic, stylized, and geometric consistency. In contrast, our method provides complete 360° coverage without any blind spots (black areas in baseline results), and shows globally consistent semantics. 43

3.5 **Ablation of Self-Refinement.** We demonstrate that the self-refinement process greatly enhances the image quality by improving the text prompt. As shown in each row, the image on the left is generated using a simple user prompt, while a prompt augmented by GPT-4V obtains the image on the right. We observe that after the multi-round self-refinement, GPT-4V selects the one panorama with better visual quality, which provides solid support for the immersive 3D scene we further generate. 44

3.6 **Ablation of Optimization Loss.** We demonstrate the impact of Semantic and Geometric losses on the synthesized virtual cameras. (a) Utilizing photometric loss on camera views from a rendered panorama induces artifacts when altering rendered camera positions. (b) Implementing Geometric and (c) Semantic regularizations effectively reduces the artifacts originating from invisible views. (d) Integrating both regularizations yields the most optimal outcomes. 45

3.7 **Ablation Study on 3D Initialization.** We present a comparative visualization of various initialization methods for 3D Panoramic Gaussian Splatting. In the absence of geometric priors of the scene (a), the optimized 3D Gaussian rendering yields plausible results in panoramic camera views. However, artifacts become evident when the rendered camera position is altered. To address this, we employ a monocular depth approach combined with learnable alignment factors and a geometric field (b), which ensures consistent alignment across multiple views. . 46

4.1 **Feature4X: Building 4D Interactive Scenes with Agentic AI from Monocular Videos.** By dynamically distilling model-conditioned features and integrating 2D foundation models with LLMs in feedback loops, Feature4X enables multimodal tasks across 2D, 3D, and 4D with high-level language inputs or direct user interactions, including (but not limited to) segmentation, scene editing, and VQA across novel views and all time steps, unlocking new possibilities for 4D agentic AI. 48

4.2 **Method overview.** Given an input monocular video, we infer 2D priors to segment static background (represented by static 3D Gaussians augmented with latent features) and dynamic foreground (represented by dynamic 3D Gaussians guided by Motion Scaffolds [LWH24a], a set of nodes $\{\mathbf{v}_i\}$ encoding 3D motion trajectories and latent features h_i). Dynamic Gaussian features and motions are computed via interpolation from their K-nearest scaffold nodes. At each timestep, dynamic Gaussians are warped and fused with static Gaussians. A parallel rasterization [ZCJ24] generates RGB images and a unified latent feature map, decoded into task-specific features—illustrated here by SAM2 [RGH24], CLIP-LSeg [LWB22a], and InternVideo2 [WLL24] for representative 2D (novel view segmentation), 3D (scene editing), and 4D (spatiotemporal VQA) tasks. Our framework generalizes to any 2D vision foundation model and is trained end-to-end using input RGB frames and customized features from pretrained 2D models. At inference, rendered feature maps from arbitrary views and timesteps are directly fed into task-specific decoders, seamlessly supporting user prompts and LLM interactions to form a unified 4D agentic AI system. 55

4.3 **Segment Anything in Dynamic 4D Scenes with SAM2 Feature Field.** For any rendered novel view video, we support: (a) Promptless segmentation (segment everything): when no user prompt is provided, segmentation masks are automatically assigned at the first frame ($t = 0$) and then propagated across all frames. (b) Promptable segmentation (segment anything): the user can segment any object—static or dynamic—at any timestep using a point or box prompt, and the corresponding mask is robustly tracked and propagated through subsequent frames. 59

4.4 **Baseline Comparison on SAM2 Inference.** We compare segmentation quality and inference speed between (a) the naive RGB-based approach and (b) our feature-based method. Ours achieves comparable segmentation, accurately tracking the object over time, and avoids RGB artifacts (red box region at $t = 70$), while reducing inference time to about 4× speed-up. 60

4.5 **Semantic 4D Scene Understanding with CLIP Feature Field.** By lifting CLIP-LSeg [LWB22a] features into a 4D feature field, we enable pixel-level semantic segmentation from any view at any timestep. This allows robust 4D scene understanding, even as object appearances change over time—for example, accurately identifying a blooming flower from bud to full bloom across views. . . 61

4.6 **Scene Editing with AI Agent.** Given user prompts, our GPT-powered agent interprets editing intent and autonomously performs scene edits via our 4D CLIP feature field. Examples include both geometric (e.g., "extract" and "delete") and appearance (e.g., "change color") editing in 3D space. While results may not be perfect due to imperfect fine-grained feature alignment and non-optimal editing parameter tuning, the agent adaptively refines parameters and applies edits consistently across views and time—greatly reducing the need for manual tuning—and demonstrates robust, interactive 4D scene manipulation. 63

4.7 **VQA with Chatbot Agent.** (Left) Our model supports free-form VQA across diverse question types—general, spatial, and temporal—by distilling InternVideo2 features. (Right) At each timestep, we reconstruct both a 4D radiance field and a 4D feature field, providing more inference sources beyond the input video frame—including local (moving camera) and global (zoomed-out) novel views and their corresponding feature maps—thereby supporting VQA in 4D and enhancing the model's spatiotemporal reasoning capabilities. 65

5.1 **Spatiotemporal (4D) Awareness**. Humans intuitively reason in 4D (3D space + time), effortlessly reconstructing the dynamic spatial trajectory of moving objects from any perspective. In contrast, current Vision Language Models (VLMs) typically rely on aggregating 2D visual features across time, leading to incorrect predictions when motion understanding and interpretation requires deeper spatiotemporal reasoning. 68

5.2 **Distribution of Dataset Sources and Annotations.** Overview of the dataset composition, illustrating the proportions of real third-person (exo-centric) videos (DAVIS [PPC17], YouTube-VOS [XYF18]), real first-person (ego-centric) videos (Ego4D [GWB22]), and synthetic videos (Cosmos [AAB25]). The real video data is further categorized by annotation types, including translational, rotational, action, counting, and false positive queries (targeting nonexistent events to assess critical reasoning). 73

5.3 **Dataset Generation and Annotation Pipeline.** Our dataset was constructed by collecting real videos and generating synthetic data, followed by human-in-the-loop quality reviews to address ambiguous videos and annotations. After temporal alignment and quality assurance, human-annotated questions and ground-truth answers were created, complemented by multiple-choice (MC) answers generated by large language models (LLMs). The final dataset includes real and synthetic video data with comprehensive VLM scoring metrics. 74

5.4 **Qualitative Examples of Dataset Annotations.** (Top) A third-person (exo-centric) video with translational annotations ("camel turning left from its perspective"). (Middle) A first-person (ego-centric) video with a rotational question ("clockwise rotation of ladle"). (Bottom) A synthetic scene with motion recognition "robotic dog moving left"). 75

5.5 **Comparison of accuracy across types of spatiotemporal questions.** Model accuracy is shown only for the six top-performing VLMs. 77

5.6 **Comparison of CoT and DO Accuracy Across Models.** Accuracy comparison between Chain-of-Thought (CoT) and Direct Output (DO) prompting across VLMs. 81

5.7 **Heatmap of Occurances of Spatial-Temporal Terms in popular video SFT datasets.** 82

LIST OF TABLES

1.1 Comparative Analysis of AI Architectures. This table highlights the fundamental shift from processing 1D/2D data sequences to modeling 3D/4D physical states. 4

2.1 **Performance on Replica Dataset.** (average performance for 5K training iterations, speed-up module rendered feature $dim = 128$). Boldface font represents the preferred results. 21

2.2 **Performance of semantic segmentation on Replica dataset compared to NeRF-DFF.** (speed-up module rendered feature $dim = 128$). Boldface font represents the preferred results. 21

3.1 **Quantitative comparisons** between LucidDreamer and ours. 42

4.1 **Semantic segmentation on the Nvidia dataset [LGM23].** Our method achieves comparable radiance reconstruction (PSNR) and segmentation performance, while significantly reducing memory usage compared to the baselines. . . 62

4.2 **VQA performance across different inference sources with the scene shown in Fig. 4.7.** Feature-based inference supports all question types with lower latency, while view-based methods are limited in spatial and temporal reasoning. 66

4.3 **Spatiotemporal VQA on DAVIS dataset [PPM16].** Compared to 2D video inference, our 4D feature space inference (Global Novel Feature) enhances the Video-LLM's spatiotemporal reasoning while achieving approximately 3× faster inference. 67

5.1 **Evaluation on `VLM4D` Benchmark** across various proprietary and open-source VLMs. Top three performers in each category are highlighted from dark (highest) to light (third highest). Human and random selection baselines are included for reference. 85

5.2 **SFT on Spatial-Temporal Datasets.** MC refers to the performance of multiple-choice, while R, S, and R+S denote real, synthetic, and both real+synthetic usage of data for fine-tuning. 86

5.3 **InternVideo2 Accuracy with 4D Feature Field Reconstruction.** Comparison of InternVideo2's performance given different input modalities from the same subset. 87

ACKNOWLEDGMENTS

Countless times, I have imagined how I would feel when finally submitting my PhD dissertation—whether I would feel nervous, joyful, or relieved. Yet, as this day actually arrives and I look back on the journey that brought me here, I find myself unexpectedly moved to tears. There are simply so many people to whom I owe my deepest thanks.

First and foremost, my deepest gratitude belongs to my family. When I left for the United States in August 2019, I never imagined that a pandemic would keep me away for seven long years. To my beloved grandfather: You passed away during the hardest stretch of my PhD, and it breaks my heart that I could not return for a final goodbye. Failing to keep my promise to rush back to you upon graduation is a regret I will carry forever. To my parents: Although I grew up as the "smart kid," my slow test-taking speed gradually left me behind in China's exam-oriented system. At 18, when my Gaokao results came out and our home fell silent, I truly believed my life was over. But it was your boundless love and support that allowed me to climb out of that dark valley to stand here today. Now, ten years later at 28, I hope to finally honor everything you've done and make you proud.

I am incredibly fortunate to have shared my PhD journey with such amazing colleagues and friends. To my senior labmates, Zhen Wang, Pradyumna Chari, and Yunhao Ba: Thank you for pulling me through the struggles at the very beginning of my PhD. To my peer labmates, Alexander Vilesov, Ellin Zhao, Howard Zhang, and Rishi Upadhyay: Thank you for your endless support; the story of us surviving this PhD together is one I will never forget. To my junior mentees, Haoran Chang, Sicheng Jiang, Shuwang Zhang, Hui Ren, Aditya Nagachandra, Tejas Bharadwaj, and Shreeram Athreya: Thank you for trusting me and for the countless late nights we spent battling deadlines in the lab. My sincere thanks also go to my external collaborators. To Stanford's Jeong Joon Park and Despoina Paschalidou: Your guidance was the key to my very first CVPR paper. To UT Austin's Zhiwen Fan and Dejia Xu: It has been my honor to call you both lifelong colleagues and friends. Finally, to

all my co-authors, especially Yijia Weng, Di Chang, Xuehai He, Ziyu Wan, and Jian Zhang: Your inspiration and help made it possible to publish research I could never have completed on my own.

Beyond research, I want to thank all the incredible friends I met at UCLA: Kaifeng Pang, Ziyang Long, Yu Qiao, Qian He, Mohan Shi, Junyi Yin, Wenbo Zhao, Bingqing Hu, Qi Chen, Hanyu Zhang, Hengjie Liu, and Junyuan Yang. In truth, the PhD journey is an intensely lonely road. Yet, each of you brought the warmth of true friendship into my life, reminding me that I am cared for even in a foreign land. Although 95% of my PhD was confined to the lab, that remaining 5% spent with you was my ultimate sanctuary and spiritual solace. Beyond all else, to my girlfriend, Xiameng Chen: You are the most beautiful, unexpected surprise of this entire journey. Empowered by your love and unwavering support, I have found the courage to conquer any obstacle in my path. The greatest publication of all—the paper of our life together—is one we have only just begun to co-author.

There is an old Chinese proverb: "A teacher for a day is a father for a lifetime." I would like to thank my PhD committee members first: Prof. Bolei Zhou, Prof. Jonathan Kao, and Prof. Aydogan Ozcan. Thank you for your invaluable guidance, insightful feedback, and for helping me shape this dissertation. To Prof. Leonidas Guibas and Prof. Atlas Wang: It has been my absolute privilege to be advised by and collaborate with you both. (Prof. Wang, I also want to express my special thanks for offering me a PhD position at UT Austin years ago; your early belief in me meant the world.) To the late Dr. Suya You: You were the crucial bridge who gave me the opportunity to connect and collaborate with these remarkable professors. I will forever carry the memory of your kindness and mentorship in my heart. Your passing during the final year of my PhD brought me profound sorrow. May you rest in peace. Furthermore, I owe a massive debt of gratitude to the two advisors who shaped my path long before my PhD began. To my undergraduate advisor, Prof. Quan Wen at UESTC: You were my very first research mentor and the guiding light of my academic journey. I simply would not be where I am today without your early guidance. To my

master's advisor at Columbia University, Prof. John Wright: You were my first advisor in the United States, and to this day, I consider you the most pivotal teacher at the turning point of my life. Studying under your guidance at Columbia fundamentally changed my life's trajectory. I learned immeasurably from you, not just about academia, but about character and integrity. Words cannot capture the depth of my gratitude; without your profound influence, the thought of pursuing a PhD would never have even crossed my mind.

Last, but certainly most importantly, I reserve my deepest and most significant thanks for my PhD advisor, Prof. Achuta Kadambi. I thank you for believing in me from the very beginning; you emailed me for an interview the exact night I submitted my UCLA application and ultimately brought me here. I thank you for your tough love, for waking me up and humbling me when I first arrived with too much ego and overconfidence. I thank you for your profound tolerance. During my early years, when I couldn't find a research direction, spiraled into depression, and almost dropped out, you had countless reasons to give up on me. But you never did. Instead, you offered care and support that I will be grateful for the rest of my life. I thank you for your unwavering encouragement. When I failed my first preliminary exam, you stood firmly by my side, giving me the strength to persevere and pass with high marks on my second attempt. I also thank you for your open-mindedness in supporting my internships and my job search. You have been the ultimate guiding light of my PhD journey. We have been through so much over these five years—the ups and downs, the tears and the laughter. If I am being completely honest, there were moments along the way when I felt frustrated or complained. But looking back today, I understand everything you did for me. From this day forward, my greatest hope is to become the student you are most proud of.

VITA

2015–2019 B.Eng. (Electronic Information Engineering), University of Electronic Science and Technology of China.

2019–2021 M.S. (Electrical Engineering), Columbia University.

2024 Student Researcher, Google.

2025 Research Intern, Apple.

2021–2026 Ph.D. Candidate (Electrical and Computer Engineering), UCLA.

PUBLICATIONS

J. Zhang*, **S. Zhou***, B. Liu*, A. Kadambi, Z. Fan. "SpatialStack: Layered Geometry-Semantic Fusion for 3D VLM Spatial Reasoning", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2026. (*equal contribution)

S. Zhou, A. Vilesov, X. He, Z. Wan, S. Zhang, A. Nagachandra, D. Chang, D. Chen, X. Wang, A. Kadambi. "VLM4D: Towards Spatiotemporal Awareness in Vision Language Models", *IEEE International Conference on Computer Vision* (ICCV), 2025.

S. Zhou, H. Ren, Y. Weng, S. Zhang, Z. Wang, D. Xu, Z. Fan, S. You, Z. Wang, L. Guibas, A. Kadambi. "Feature4X: Bridging Any Monocular Video to 4D Agentic AI with Versatile Gaussian Feature Field", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2025.

S. Zhou, Z. Fan, D. Xu, H. Chang, P. Chari, T. Bharadwaj, S. You, Z. Wang, A. Kadambi. "DreamScene360: Unconstrained Text-to-3D Scene Generation with Panoramic Gaussian Splatting", *European Conference on Computer Vision* (ECCV), 2024.

S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, A. Kadambi. "Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2024. (**Top 2.8%, Highlight**)

Z. Wang*, **S. Zhou***, J. J. Park, D. Paschalidou, S. You, G. Wetzstein, L. Guibas, A. Kadambi. "ALTO: Alternating Latent Topologies for Implicit 3D Reconstruction", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2023. (*equal contribution)

Z. Fan, J. Zhang, R. Li, J. Zhang, R. Chen, H. Hu, K. Wang, P. Wang, H. Qu, **S. Zhou**, et al. "VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2026.

R. Li, P. Pan, B. Yang, D. Xu, **S. Zhou**, X. Zhang, Z. Li, A. Kadambi, Z. Wang, Z. Tu, Z. Fan. "4K4DGen: Panoramic 4D Generation at 4K Resolution", *International Conference on Learning Representations* (ICLR), 2025. (**Top 5.1%, Spotlight**)

D. Chang, H. Xu, Y. Xie, Y. Gao, Z. Kuang, S. Cai, C. Zhang, G. Song, C. Wang, Y. Shi, Z. Chen, **S. Zhou**, L. Luo, G. Wetzstein, M. Soleymani. "X-dyna: Expressive dynamic human image animation.", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2025. (**Top 2.98%, Highlight**)

Z. Fan, J. Zhang, W. Cong, P. Wang, R. Li, K. Wen, **S. Zhou**, A. Kadambi, Z. Wang, D. Xu, B. Ivanovic, M. Pavone, Y. Wang. "Large Spatial Model: Real-time Unposed Images to Semantic 3D", *Advances in Neural Information Processing Systems* (NeurIPS), 2024.

CHAPTER 1

Introduction

1.1 The Dawn of Spatial Intelligence

The trajectory of Artificial Intelligence over the past decade has been characterized by the meteoric rise of foundational models trained on vast repositories of internet data. Large Language Models (LLMs) have mastered the syntactic and semantic complexities of human language, demonstrating emergent reasoning capabilities that challenge our definitions of cognition. Similarly, Vision-Language Models (VLMs) have achieved unprecedented success in interpreting two-dimensional imagery, bridging the gap between pixel-level data and semantic description. However, as we stand on the precipice of the next era of AI—often termed "Physical AI" or "Embodied AI"—a critical deficiency in these systems has become glaringly apparent: they lack a fundamental, grounded understanding of the three-dimensional physical reality they inhabit.

Current foundation models operate primarily as spectators. They observe the world through the keyhole of static 2D projections—images and text—and learn to predict statistical correlations between tokens and pixels. While a VLM can describe a photograph of a glass on a table with poetic accuracy, it does not inherently understand the geometry of the glass, the friction of the table surface, the force of gravity holding the glass in place, or the 3D spatial trajectory required to grasp it without causing a catastrophe. It perceives the "what" but lacks the "where" and the "how" in a physical sense. This limitation is the primary bottleneck preventing AI from transitioning from a digital oracle behind a screen to

a physical agent capable of navigating, interacting with, and manipulating the real world.

To bridge this chasm, the field must undergo a paradigm shift from passive 2D perception to active 3D cognition. This new frontier is encapsulated in the concept of **Spatial Intelligence**—a term increasingly championed by leaders in the field such as Fei-Fei Li—which refers to the ability of an AI system to perceive, reason about, and interact with the 3D world in a way that reconciles semantic meaning with geometric structure and physical dynamics. Spatial intelligence is not merely the ability to reconstruct a 3D shape; it is the capacity to transform seeing into doing, understanding into reasoning, and imagination into creation. It requires an internal representation that is continuous, persistent, and manipulable—a **World Model**.

This dissertation, *Building World Models with Spatial Intelligence*, posits that the path to achieving this intelligence lies in constructing systems that leverage the vast semantic knowledge of current 2D foundation models while grounding them in explicit, structured 3D and 4D representations. We argue that while 2D models provide the "mind" (semantic knowledge), explicit representations—specifically 3D Gaussian Splatting—provide the "body" (geometric structure) required for spatial intelligence. By fusing these elements, we can build World Models that possess four critical capabilities: **Perception**, **Imagination**, **Interaction**, and **Reasoning**.

1.2 The Concept of the World Model

The concept of a "World Model" has its roots in control theory and cognitive psychology, suggesting that intelligent agents operate not by reacting directly to sensory stimuli, but by maintaining an internal simulation of the environment. This simulation allows the agent to predict the consequences of its actions, reason about counterfactual scenarios ("what would happen if...?"), and plan over long horizons.

In the context of modern AI, a World Model must satisfy several rigorous requirements

to be considered spatially intelligent. It must be:

1. **Multimodal:** Capable of integrating visual, textual, and potentially physical (depth, tactile) data.
2. **Generative:** Capable of simulating future states and hallucinating unobserved portions of the environment.
3. **Consistent:** Ensuring that objects maintain their identity and geometry across time and varying viewpoints.
4. **Interactable:** Allowing for the manipulation of scene elements, not just passive viewing.

As illustrated in Tab. 1.1, comparing traditional Deep Learning models with the proposed Spatially Intelligent World Models highlights the necessity for a paradigm shift. Traditional vision models rely on 2D feature maps for passive tasks, while Large Language Models (LLMs) excel in semantics but lack physical grounding. In contrast, Spatially Intelligent World Models synthesize multimodal inputs into explicit 3D and 4D internal structures. By grounding their spatial awareness in Euclidean geometry and physics, these models move beyond simple sequence prediction. Instead, they enable robust dynamics modeling, facilitating advanced, active interactions such as physical manipulation, 4D scene editing, and novel 3D asset generation.

The central challenge in building these World Models is the data bottleneck. While we have trillions of tokens of text and billions of 2D images, we lack a comparable scale of annotated 3D data (meshes, point clouds with semantic labels). Therefore, a key theme of this dissertation is **Distillation**: using the abundance of 2D data and the power of pre-trained 2D foundation models to supervise the creation of 3D World Models.

Capability	**Traditional CNNs/ Transformers**	**LLMs**	**Spatially Intelligent World Models**
Primary Input	Images, Videos (Pixel Arrays)	Text (Token Sequences)	Multimodal (Video, Depth, Text, Action)
Internal Representation	Feature Maps (2D)	Attention Weights (High-dim Vector)	Explicit 3D/4D Structure (e.g., Gaussians)
Spatial Awareness	Implicit (Pixel Coordinates)	None (Semantic Only)	Explicit (Euclidean Geometry, Physics)
Temporal Understanding	Sequence Prediction (Frame $t+1$)	Sequence Prediction (Token $t+1$)	Dynamics Modeling ($State_{t+1} = f(State_t, Action)$)
Output	Classification, Segmentation	Text Generation	Novel Views, 3D Assets, Future Trajectories
Interaction	Passive Classification	Chat/Text Response	4D Scene Editing, Physical Manipulation

Table 1.1: Comparative Analysis of AI Architectures. This table highlights the fundamental shift from processing 1D/2D data sequences to modeling 3D/4D physical states.

1.3 The Evolution of 3D Representations: The Quest for Structure

To contextualize the contributions of this research, it is necessary to review the historical evolution of 3D representations in computer vision. The field has long oscillated between two poles: **Explicit Representations** and **Implicit Representations**, each with distinct advantages and fatal flaws regarding spatial intelligence.

1.3.1 Explicit Representations: Precision without Semantics

Explicit representations—such as polygon meshes, point clouds, and voxels—have been the standard in computer graphics for decades.

- **Meshes** explicitly define the surface of an object using vertices and faces. They are efficient to render and compatible with physics engines.
- **Point Clouds** represent the world as a set of discrete coordinates. They are the native output of LiDAR and depth sensors.

However, adapting these representations for Deep Learning has proven difficult. Meshes are topologically rigid; changing the topology (e.g., punching a hole in a donut) during training is non-trivial. Point clouds are unstructured and require massive density to represent fine details. Crucially, these representations historically lacked *semantics*. A mesh of a chair "knows" its geometry but does not "know" it is a chair, nor can it identify its "leg" or "back" without laborious manual annotation.

1.3.2 Implicit Representations: The NeRF Revolution

The introduction of Neural Radiance Fields (NeRFs) [MST20] marked a paradigm shift. NeRFs represent scenes *implicitly* as continuous functions (parameterized by Multi-Layer Perceptrons) that map spatial coordinates (x, y, z) and viewing directions (θ, ϕ) to density and color.

- **Pros:** NeRFs achieved photorealistic novel view synthesis that far surpassed traditional explicit methods, handling complex lighting and view-dependent effects.
- **Cons:** NeRFs function as "black boxes." The scene geometry is encoded in the neural weights, making it notoriously difficult to edit. You cannot simply "delete" an object from a NeRF as you would remove vertices from a mesh. Furthermore, NeRFs are computationally expensive, often requiring seconds or minutes to render a single frame, which renders them unsuitable for real-time interaction in a World Model.

1.3.3 The Unification: 3D Gaussian Splatting

The recent advent of **3D Gaussian Splatting (3DGS)** [KKL23] has provided a middle ground that combines the best of both worlds. 3DGS represents a scene as a collection of explicit 3D Gaussians (ellipsoids), each with learnable parameters: position, rotation, scale, opacity, and spherical harmonic coefficients for color.

- **Explicit yet Differentiable:** Like meshes, Gaussians are explicit primitives that can be moved, deleted, or modified. Like NeRFs, they are "soft" and differentiable, allowing them to be optimized via gradient descent to achieve photorealism.
- **Real-Time Performance:** 3DGS utilizes a rasterization pipeline that allows for real-time rendering (100+ FPS), a prerequisite for interactive World Models.

This dissertation identifies 3D Gaussian Splatting as the optimal substrate for Spatial Intelligence. Because Gaussians are discrete entities, we can attach more than just color to them. We can attach semantic feature vectors, physical properties, or temporal dynamics. This dissertation heavily exploits 3DGS, transforming it from a mere rendering engine into a vessel for high-dimensional semantic and dynamic information.

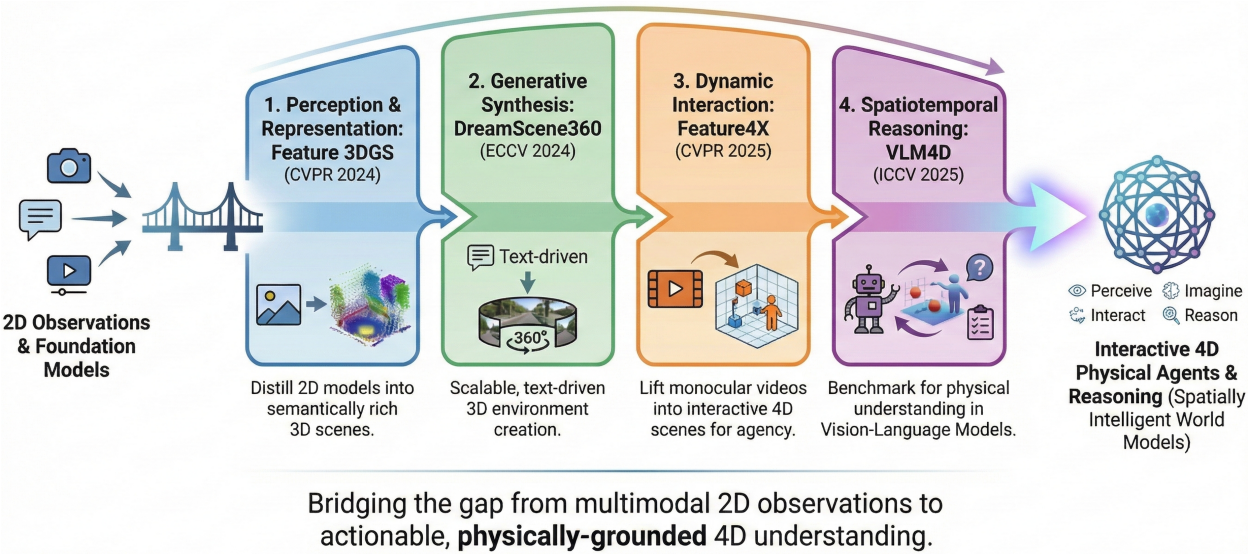


Figure 1.1: **Building spatially intelligent world models: an overview of this dissertation.**

1.4 The Four Pillars of the Dissertation

This research is structured around four pillars that correspond to the essential faculties of a Spatially Intelligent World Model. As illustrated in Fig. 1.1, We begin with **Perception** (Chapter 2), enabling the model to understand what it sees. We proceed to **Imagination** (Section 3), enabling the model to create what it has not seen. We then introduce **Interaction** (Section 4), enabling the model to handle the dynamics of time and intervention. Finally, we assess **Reasoning** (Section 5), benchmarking the model's true understanding of space and time.

1.4.1 Pillar I: Perception – Distilling Semantic 3D Representations

The first requirement is Perception. Perception in this context goes beyond simple object detection; it requires a dense, pixel-perfect understanding of the semantic properties of the 3D environment from any viewing angle. Current 2D foundation models like the Segment Anything Model (SAM) and CLIP have revolutionized 2D perception. They can segment

objects and understand concepts in images with zero-shot capability. However, they are inherently 2D. If we run SAM on a video of a rotating object, the segmentations will flicker and lack consistency across frames. The model has no concept that the "front of the car" and the "side of the car" are part of the same physical object.

Chapter 2 introduces **Feature 3DGS** [ZCJ24], a method to **distill** the knowledge of these 2D foundation models into a 3D Gaussian Field. The core innovation is the **Parallel N-dimensional Gaussian Rasterizer**. Standard 3DGS rasterizes 3-channel RGB colors. Feature 3DGS modifies the rendering pipeline to rasterize arbitrarily high-dimensional feature vectors. This allows us to "bake" high-dimensional semantic features (e.g., CLIP vectors) directly into the 3D Gaussians, giving the World Model a persistent semantic memory.

1.4.2 Pillar II: Imagination – From Text to 3D Worlds

The second requirement is Imagination. A World Model must be able to generate hypothetical scenarios. Current text-to-3D methods often rely on Score Distillation Sampling (SDS) from 2D diffusion models. While innovative, these methods often suffer from the "Janus problem" (multi-faced objects), slow optimization, and poor geometric consistency.

Chapter 3 introduces **DreamScene360** [ZFX24], a pipeline for **text-to-3D scene generation** that prioritizes immersion and consistency. The key insight is using 360° panoramas as a bridge between 2D generation and 3D representation. By lifting generated panoramas into 3D Gaussians and imposing novel geometric constraints, we resolve the "invisible" region problem, creating a "dreaming" engine that allows the World Model to construct diverse training environments from language.

1.4.3 Pillar III: Interaction – Bridging Space, Time, and Action

A static world is insufficient for an agent that needs to act. The third requirement is Interaction, which necessitates handling the **4th dimension: Time**. Monocular video is

the most abundant form of visual data, yet it is fraught with ambiguity.

Chapter 4 presents **Feature4X** [ZRW25], a framework that creates **4D Spatiotemporal Feature Fields**. By decomposing scenes into Static Backgrounds and Dynamic Foregrounds via Motion Scaffolds, Feature4X enables an agent to perform 4D editing (e.g., "Remove the moving person") and Spatiotemporal VQA. This turns the World Model from a passive observer into an active participant capable of causal intervention.

1.4.4 Pillar IV: Reasoning – Towards Spatiotemporal Intelligence

The final pillar is Reasoning. If we build these sophisticated representations, does the AI actually "understand" them? Current benchmarks for VLMs are 2D-centric and fail to test true spatial intelligence.

Chapter 5 introduces **VLM4D** [ZVH25], the first benchmark explicitly designed to stress-test **spatiotemporal reasoning**. The benchmark exposes the critical failure modes of current SOTA models—such as the inability to distinguish between camera motion and object motion—validating the necessity of the explicit 3D/4D representations proposed in this dissertation.

1.5 Dissertation Statement

This PhD dissertation argues that **Spatial Intelligence** in World Models cannot be achieved by scaling 2D generative models alone. It requires a symbiotic architecture where 2D models provide semantic breadth and explicit 3D/4D representations (specifically feature-distilled Gaussian Splatting) provide geometric, semantic, and temporal depth. By bridging the "Pixel-to-Feature" gap in rendering and the "2D-to-4D" gap in representation, we can build World Models that perceive, imagine, interact, and reason with the physical world.

CHAPTER 2

Perception - Distilling Semantic 3D Representations

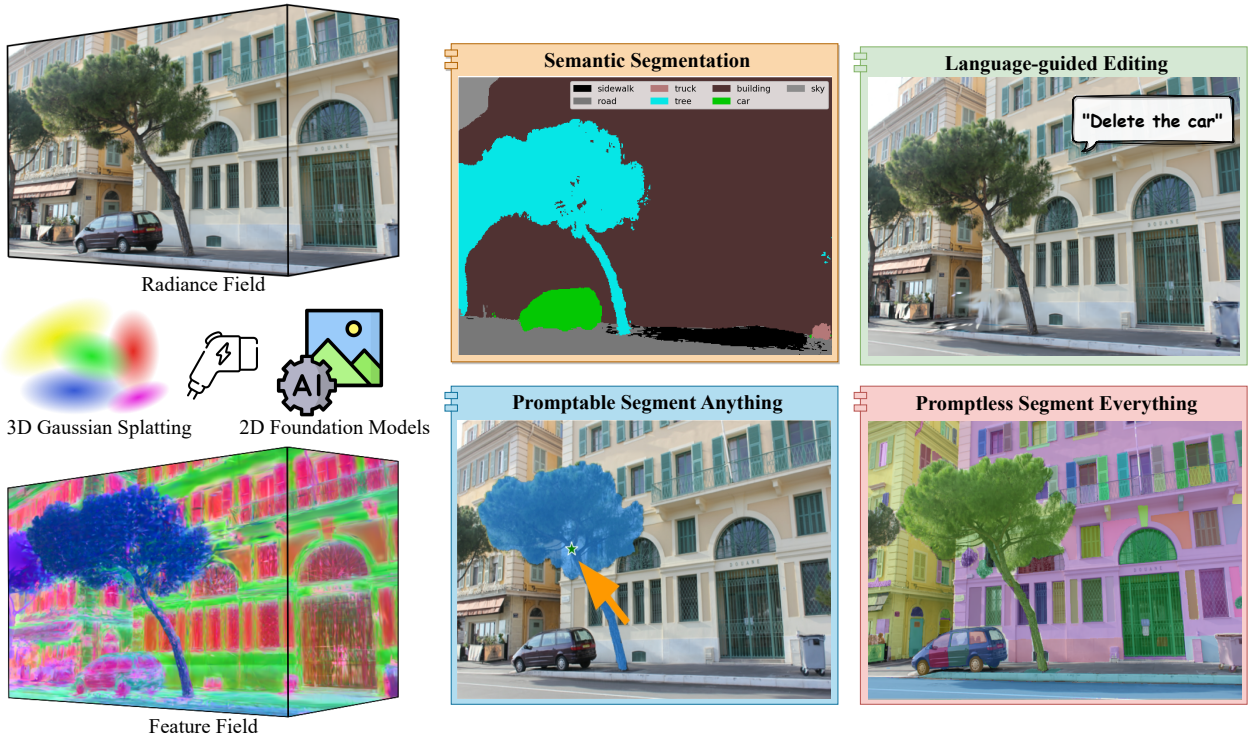


Figure 2.1: **Feature 3DGS**. We present a general method that significantly enhances 3D Gaussian Splatting through the integration of large 2D foundation models via feature field distillation. This advancement extends the capabilities of 3D Gaussian Splatting beyond mere novel view synthesis. It now encompasses a range of functionalities, including semantic segmentation, language-guided editing, and promptable segmentations such as "segment anything" or automatic segmentation of everything from any novel view. Scene from [HPP18].

2.1 Introduction

3D scene representation techniques have been at the forefront of computer vision and graphics advances in recent years. Methods such as Neural Radiance Fields (NeRFs) [MST21a], and works that have followed up on it, have enabled learning implicitly represented 3D fields that are supervised on 2D images using the rendering equation. These methods have shown great promise for tasks such as novel view synthesis. However, since the implicit function is only designed to store local radiance information at every 3D location, the information contained in the field is limited from the perspective of downstream applications.

More recently, NeRF-based methods have attempted to use the 3D field to store additional descriptive features for the scene, in addition to the radiance [KMS22, FWJ22, KKG23, TLL22]. These features, when rendered into feature images, can then provide additional semantic information for the scene, enabling donwstream tasks such as editing, segmentation and so on. However, feature field distillation through such a method is subject to a major disadvantage: NeRF-based methods can be natively slow to train as well as to infer. This is further complicated by model capacity issues: if the implicit representation network is kept fixed, while requiring it to learn an additional feature field (to not make the rendering and inference speeds even slower), the quality of the radiance field, as well as the feature field is likely to be affected unless the weight hyperparameter is meticulously tuned [KMS22].

A recent alternative for implicit radiance field representations is the 3D Gaussian splatting-based radiance field proposed by Kerbl et al. [KKL23]. This explicitly-represented field using 3D Gaussians is found to have superior training speeds and rendering speeds when compared with NeRF-based methods, while retaining comparable or better quality of rendered images. This speed of rendering while retaining high quality has paved the way for real-time rendering applications, such as in VR and AR, that were previously found to be difficult. However, the 3D Gaussian splatting framework suffers the same representation limitation as NeRFs: natively, the framework does not support joint learning of semantic features and radiance

field information at each Gaussian.

In this work, we present Feature 3DGS: the first feature field distillation technique based on the 3D Gaussian Splatting framework. Specifically, we propose learning a semantic feature at each 3D Gaussian, in addition to color information. Then, by splatting and rasterizing the feature vectors differentiably, the distillation of the feature field is possible using guidance from 2D foundation models. While the structure is natural and simple, enabling fast yet high-quality feature field distillation is not trivial: as the dimension of the learnt feature at each Gaussian increases, both training and rendering speeds drop drastically. We therefore propose learning a structured lower-dimensional feature field, which is later upsampled using a lightweight convolutional decoder at the end of the rasterization process. Therefore, this pipeline enables us to achieve improved feature field distillation at faster training and rendering speeds than NeRF-based methods, enabling a range of applications, including semantic segmentation, language-guided editing, promptable/promptless instance segmentation and so on.

In summary, our contributions are as follows:

- A novel 3D Gaussian splatting inspired framework for feature field distillation using guidance from 2D foundation models.
- A general distillation framework capable of working with a variety of feature fields such as CLIP-LSeg, Segment Anything (SAM) and so on.
- Up to **2.7**× faster feature field distillation and feature rendering over NeRF-based method by leveraging low-dimensional distillation followed by learnt convolutional upsampling.
- Up to **23%** improvement on mIoU for tasks such as semantic segmentation.

2.2 Related Work

2.2.1 Implicit Radiance Field Representations

Implicit neural representations have achieved remarkable success in recent years across a variety of areas within the field of computer vision [PFS19, PNM20, MON19, WZP23, MST21a, BMT21]. NeRF [MST21a] demonstrates outstanding performance in novel view synthesis by representing 3D scenes with a coordinate-based neural network. In mip-NeRF [BMT21], point-based ray tracing is replaced using cone tracing to combat aliasing. Zip-NeRF [BMV23] utilized an anti-aliased grid-based technique to boost the radiance field performance. Instant-NGP [MES22] reduces the cost for neural primitives with a versatile new input encoding that permits the use of a smaller network without sacrificing quality, thus significantly reducing the number of floating point and memory access operations. IBRNet [WWG21], MVS-NeRF [CXZ21], and PixelNeRF [YYT21] construct a generalizable 3D representation by leveraging features gathered from various observed viewpoints. However, NeRF-based methods are hindered by slow rendering speeds and substantial memory usage during training, a consequence of their implicit design.

2.2.2 Explicit Radiance Field Representations

Pure implicit radiance fields are slow to operate and usually require millions of times querying a neural network for rendering a large-scale scene. Marrying explicit representations into implicit radiance fields enjoys the best of both worlds. Triplane [CLC22], TensoRF [CXG22], K-Plane [FMW23a], TILED [YZB23] adopt tensor factorization to obtain efficient explicit representation. InstantNGP [MES22] utilizes multi-scale hash grids to work with large-scale scenes. Block-NeRF [TCY22] further extends NeRF to render city-scale scenes spanning multiple blocks. Point NeRF [XXP22] uses neural 3D points for representing and rendering a continuous radiance volume. NU-MCC [LXL23] similarly utilizes latent point features but focuses on shape completion tasks. Unlike NeRF-style volumetric rendering, 3D Gaussian

Splatting introduces point-based α-blending and an efficient point-based rasterizer. Our work follows 3D Gaussians Splatting, where we represent the scene using explicit point-based 3D representation, *i.e.* anisotropic 3D Gaussians.

2.2.3 Feature Field Distillation

Enabling simultaneously novel view synthesis and representing feature fields is well explored under NeRF [MST21a] literature. Pioneering works such as Semantic NeRF [ZLL21] and Panoptic Lifting [SPB22] have successfully embedded semantic data from segmentation networks into 3D spaces. Their research has shown that merging noisy or inconsistent 2D labels in a 3D environment can yield sharp and precise 3D segmentation. Further extending this idea, techniques like those presented in [RAR22] have demonstrated the effectiveness of segmenting objects in 3D with minimal user input, like rudimentary foreground-background masks. Beyond optimizing NeRF with estimated labels, Distilled Feature Fields [KMS22], NeRF-SOS [FWJ22], LERF [KKG23], and Neural Feature Fusion Fields [TLL22] have delved into embedding pixel-aligned feature vectors from technologies such as LSeg or DINO [CTM21] into NeRF frameworks. Additionally, [MSD23, TML23, GSS23, LZZ24, SPP22, SYY23, YWW23] also explore feature fusion and manipulation in 3D. Feature 3DGS shares a similar idea for distilling 2D well-trained models, but also demonstrates an effective way of distilling into explicit point-based 3D representations, for simultaneous photo-realistic view synthesis and label map rendering.

2.3 Method

NeRF-based feature field distillation, as explored in [KMS22], utilizes two distinct branches of MLPs to output the color c and feature f. Subsequently, the RGB image and high-dimensional feature map are rendered individually through volumetric rendering. The transition from NeRF to 3DGS is not as straightforward as simply rasterizing RGB images and

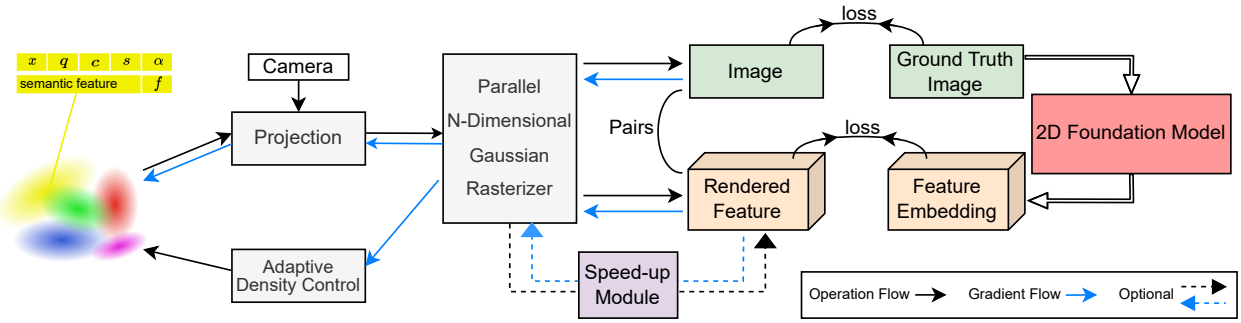


Figure 2.2: **An overview of our method.** We adopt the same 3D Gaussian initialization from sparse SfM point clouds as utilized in 3DGS, with the addition of an essential attribute: the *semantic feature*. Our primary innovation lies in the development of a Parallel N-dimensional Gaussian Rasterizer, complemented by a convolutional speed-up module as an optional branch. This configuration is adept at rapidly rendering arbitrarily high-dimensional features without sacrificing downstream performance.

feature maps independently. Typically, feature maps have fixed dimensions that often differ from those of RGB images. Due to the tile-based rasterization procedure and shared attributes between images and feature maps, rendering them independently can be problematic. A naive approach is to adopt a two-stage training method that rasterizes them separately. However, this approach could result in suboptimal quality for both RGB images and feature maps, given the high-dimensional correlations of semantic features with the shared attributes of RGB.

In this section, we introduce a novel pipeline for high-dimensional feature rendering and feature field distillation, which enables 3D Gaussians to explicitly represent both radiance fields and feature fields. Our proposed parallel N-dimensional Gaussian rasterizer and speed-up module can effectively solve the aforementioned problems and is capable of rendering arbitrary dimensional semantic feature map. An overview of our method is shown in Fig. 4.2. Our proposed method is general and compatible with any 2D foundation model, by distilling the semantic features into a 3D feature field using 3D Gaussian splatting. In our experiments,

we employ SAM [KMR23] and LSeg [LWB22b], facilitating promptable, promptless (zero-shot [BVC19] [GZN20]) and language-driven computer vision tasks in a 3D context.

2.3.1 High-dimensional Semantic Feature Rendering

To develop a general feature field distillation pipeline, our method should be able to render 2D feature maps of arbitrary size and feature dimension, in order to cope with different kinds of 2D foundation models. To achieve this, we use the rendering pipeline based on the differentiable Gaussian splatting framework proposed by [KKL23] as our foundation. We follow the same 3D Gaussians initialization technique using Structure from Motion [SF16]. Given this initial point cloud, each point $x \in \mathbb{R}^3$ within it can be described as the center of a Gaussian. In world coordinates, the 3D Gaussians are defined by a full 3D covariance matrix Σ, which is transformed to Σ' in camera coordinates when 3D Gaussians are projected to 2D image / feature map space [ZPV01]:

$$\Sigma' = JW\Sigma W^T J^T, \tag{2.1}$$

where W is the world-to-camera transformation matrix and J is the Jacobian of the affine approximation of the projective transformation. Σ is physically meaningful only when it is positive semi-definite — a condition that cannot always be guaranteed during optimization. This issue can be addressed by decomposing Σ into rotation matrix R and scaling matrix S:

$$\Sigma = RSS^T R^T, \tag{2.2}$$

Practically, the rotation matrix R and the scaling matrix S are stored as a rotation quaternion $q \in \mathbb{R}^4$ and a scaling factor $s \in \mathbb{R}^3$ respectively. Besides the aforementioned optimizable parameters, an opacity value $\alpha \in \mathbb{R}$ and spherical harmonics (SH) up to the 3rd order are also stored in the 3D Gaussians. In practice, we optimize the zeroth-order SH for the first 1000 iterations, which equates to a simple diffuse color representation $c \in \mathbb{R}^3$, and we introduce 1 band every 1000 iterations until all 4 bands of SH are represented.

Additionally, we incorporate the semantic feature $f \in \mathbb{R}^N$, where N can be any arbitrary number representing the latent dimension of the feature. In summary, for the $i-$th 3D Gaussian, the optimizable attributes are given by $\Theta_i = \{x_i, q_i, s_i, \alpha_i, c_i, f_i\}$.

Upon projecting the 3D Gaussians into a 2D space, the color C of a pixel and the feature value F_s of a feature map pixel are computed by volumetric rendering which is performed using front-to-back depth order [KPL21]:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i T_i, \quad F_s = \sum_{i \in \mathcal{N}} f_i \alpha_i T_i, \tag{2.3}$$

where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$, $\mathcal{N}$ is the set of sorted Gaussians overlapping with the given pixel, T_i is the transmittance, defined as the product of opacity values of previous Gaussians overlapping the same pixel. The subscript s in F_s denotes "student", indicating that this rendered feature is per-pixel supervised by the "teacher" feature F_t. The latter represents the latent embedding obtained by encoding the ground truth image using the encoder of 2D foundation models. This supervisory relationship underscores the instructional dynamic between F_s and F_t in our model. In essence, our approach involves distilling [HVD15] the large 2D teacher model into our small 3D student explicit scene representation model through differentiable volumetric rendering.

In the rasterization stage, we adopted a joint optimization method, as opposed to rasterizing the RGB image and feature map independently. Both image and feature map utilize the same tile-based rasterization procedure, where the screen is divided into 16×16 tiles, and each thread processes one pixel. Subsequently, 3D Gaussians are culled against both the view frustum and each tile. Owing to their shared attributes, both the feature map and RGB image are rasterized to the same resolution but in different dimensions, corresponding to the dimensions of c_i and f_i initialized in the 3D Gaussians. This approach ensures that the fidelity of the feature map is rendered as high as that of the RGB image, thereby preserving per-pixel accuracy.

2.3.2 Optimization and Speed-up

The loss function is the photometric loss combined with the feature loss:

$$\mathcal{L} = \mathcal{L}_{rgb} + \gamma \mathcal{L}_f, \tag{2.4}$$

with

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1(I, \hat{I}) + \lambda \mathcal{L}_{D-SSIM}(I, \hat{I}),$$
$$\mathcal{L}_f = \|F_t(I) - F_s(\hat{I})\|_1.$$

where I is the ground truth image and $\hat{I}$ is our rendered image. The latent embedding $F_t(I)$ is derived from the 2D foundation model by encoding the image I, while $F_s(\hat{I})$ represents our rendered feature map. To ensure identical resolution $H \times W$ for the per-pixel $\mathcal{L}_1$ loss calculation, we apply bilinear interpolation to resize $F_s(\hat{I})$ accordingly. In practice, we set the weight hyperparameters $\gamma = 1.0$ and $\lambda = 0.2$.

It is important to note that in NeRF-based feature field distillation, the scene is implicitly represented as a neural network. In this configuration, as discussed in [KMS22], the branch dedicated to the feature field shares some layers with the radiance field. This overlap could potentially lead to interference, where learning the feature fields might adversely affect the radiance fields. To address this issue, a compromise approach is to set γ to a low value, meaning the weight of the feature field is much smaller than that of the radiance field during the optimization. [KMS22] also mentions that NeRF is highly sensitive to γ. Conversely, our explicit scene representation avoids this issue. Our equal-weighted joint optimization approach has demonstrated that the resulting high-dimensional semantic features significantly contribute to scene understanding and enhance the depiction of physical scene attributes, such as opacity and relative positioning. See the comparison between Ours and Base 3DGS in Tab. 2.1.

To optimize the semantic feature $f \in \mathbb{R}^N$, we minimize the difference between the rendered feature map $F_s(\hat{I}) \in \mathbb{R}^{H \times W \times N}$ and the teacher feature map $F_t(I) \in \mathbb{R}^{H \times W \times M}$, ideally

with $N = M$. However, in practice, M tends to be a very large number due to the high latent dimensions in 2D foundation models (e.g. $M = 512$ for LSeg and $M = 256$ for SAM), making direct rendering of such high-dimensional feature maps time-consuming. To address this issue, we introduce a speed-up module at the end of the rasterization process. This module consists of a lightweight convolutional decoder that upsamples the feature channels with kernel size 1×1. Consequently, it is feasible to initialize $f \in \mathbb{R}^N$ on 3D Gaussians with any arbitrary $N \ll M$ and to use this learnable decoder to match the feature channels. This allows us to not only effectively achieve $F_s(\hat{I}) \in \mathbb{R}^{H \times W \times M}$, but also significantly speed up the optimization process without compromising the performance on downstream tasks.

The advantages of implementing this convolutional speed-up module are threefold: Firstly, the input to the convolution layer, with a kernel size of 1×1, is the resized rendered feature map, which is significantly smaller in size compared to the original image. This makes the 1×1 convolution operation computationally efficient. Secondly, this convolution layer is a learnable component, facilitating channel-wise communication within the high-dimensional rendered feature, enhancing the feature representation. Lastly, the module's design is optional. Whether included or not, it does not impact the performance of downstream tasks, thereby maintaining the flexibility and adaptability of the entire pipeline.

2.3.3 Promptable Explicit Scene Representation

Foundation models provide a base layer of knowledge and skills that can be adapted for a variety of specific tasks and applications. We wish to use our feature field distillation approach to enable practical 3D representations of these features. Specifically, we consider two foundation models, namely Segment Anything [KMR23], and LSeg [LWB22b]. The *Segment Anything Model (SAM)* [KMR23] allows for both promptable and promptless zero-shot segmentation in 2D, without the need for specific task training. LSeg [LWB22b] introduces a language-driven approach to zero-shot semantic segmentation. Utilizing the image feature encoder with the DPT architecture [RBK21] and text encoders from CLIP [RKH21], LSeg

extends text-image associations to a 2D pixel-level granularity. Through the teacher-student distillation, **our distilled feature fields facilitate the extension of all 2D functionalities — prompted by point, box, or text — into the 3D realm.**

Our promptable explicit scene representation works as follows: for a 3D Gaussian x among the N ordered Gaussians overlapping the target pixel, i.e. $x_i \in \mathcal{X}$ where $\mathcal{X} = \{x_1, \ldots, x_N\}$, the activation score of a prompt τ on the 3D Gaussian x is calculated by cosine similarity between the query $q(\tau)$ in the feature space and the semantic feature $f(x)$ of the 3D Gaussian followed by a softmax:

$$s = \frac{f(x) \cdot q(\tau)}{\|f(x)\|\|q(\tau)\|}, \tag{2.5}$$

If we have a set $\mathcal{T}$ of possible labels, such as a text label set for semantic segmentation or a point set of all the possible pixels for point-prompt, the probability of a prompt τ of a 3D Gaussian can be obtained by softmax:

$$\mathbf{p}(\tau|x) = \text{softmax}(s) = \frac{\exp(s)}{\sum_{s_j \in \mathcal{T}} \exp(s_j)}. \tag{2.6}$$

We utilize the computed probabilities to filter out Gaussians with low probability scores. This selective approach enables various operations, such as extraction, deletion, or appearance modification, by updating the color $c(x)$ and opacity $\alpha(x)$ values as needed. With the newly updated color set $\{c_i\}_{i=1}^n$ and opacity set $\{\alpha_i\}_{i=1}^n$, where n is smaller than N, we can implement point-based α-blending to render the edited radiance field from any novel view.

2.4 Experiments

2.4.1 Novel view semantic segmentation

The number of classes of a dataset is usually limited from tens [EEV15] to hundreds [ZZP19], which is insignificant to English words [LWC20]. In light of the limitation, semantic features

Metrics	PSNR(±s.d.)↑	SSIM(±s.d.)↑	LPIPS(±s.d.)↓
Ours (w/ speed-up)	**37.012** (±0.07)	**0.971** (±5.3e-4)	**0.023** (±2.9e-4)
Ours	36.915 (±0.05)	0.970 (±5.7e-4)	0.024 (±1.1e-3)
Base 3DGS	36.133 (±0.06)	0.965 (±1.5e-4)	0.033 (±1.2e-3)

Table 2.1: **Performance on Replica Dataset.** (average performance for 5K training iterations, speed-up module rendered feature $dim = 128$). Boldface font represents the preferred results.

Metrics	mIoU↑	accuracy↑	FPS↑
Ours (w/ speed-up)	0.782	**0.943**	**14.55**
Ours	**0.787**	**0.943**	6.84
NeRF-DFF	0.636	0.864	5.38

Table 2.2: **Performance of semantic segmentation on Replica dataset compared to NeRF-DFF.** (speed-up module rendered feature $dim = 128$). Boldface font represents the preferred results.

empower models to comprehend unseen labels by mapping semantically close labels to similar regions in the embedding space, as articulated by Li et al [LWB22b]. This advancement notably promotes the scalability in information acquisition and scene understanding, facilitating a profound comprehension of intricate scenes. We distill LSeg feature for this novel view semantic segmentation task. Our experiments demonstrate the improvement of incorporating semantic feature over the naive 3D Gaussian rasterization method [KKL23]. In Tab. 2.1, we show that our model surpasses the baseline 3D Gaussian model in performance metrics on Replica dataset [SWM19] with 5000 training iterations for all three models.

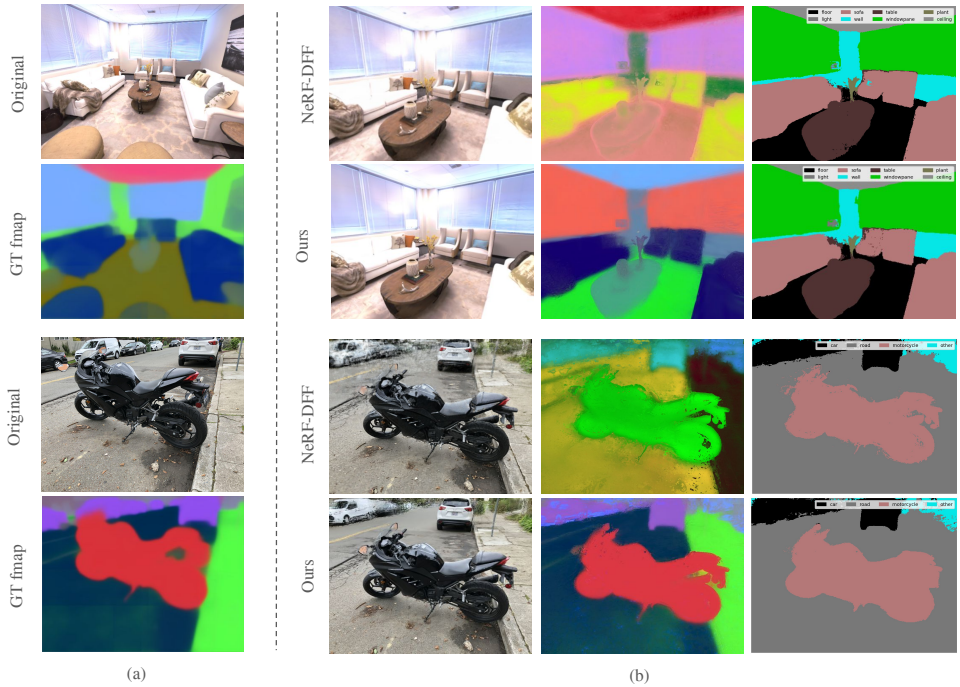


Figure 2.3: **Novel view semantic segmentation (LSeg) results on scenes from Replica dataset [SWM19] and LLFF dataset [MSO19].** (a) We show examples of original images in training views together with the ground-truth feature visualizations. (b) We compare the qualitative segmentation results using our Feature 3DGS with the NeRF-DFF [KMS22]. Our inference is **1.66**× faster when rendered feature $dim = 128$. Our method demonstrates more fine-grained segmentation results with higher-quality feature maps.

Noticeably, the integration of the speed-up module to our model does not compromise the performance.

In our further comparison with NeRF-DFF [KMS22] using the Replica dataset, we address the potential trade-off between the quality of the semantic feature map and RGB images. In Tab. 2.2, our model demonstrates higher accuracy and mean intersection-over-union (mIoU). Additionally, by incorporating our speed-up module, we achieved more than double the frame rate per second (FPS) of our full model while maintaining comparable performance. In Fig. 2.3 (b) the last column, our approach yields better visual quality on

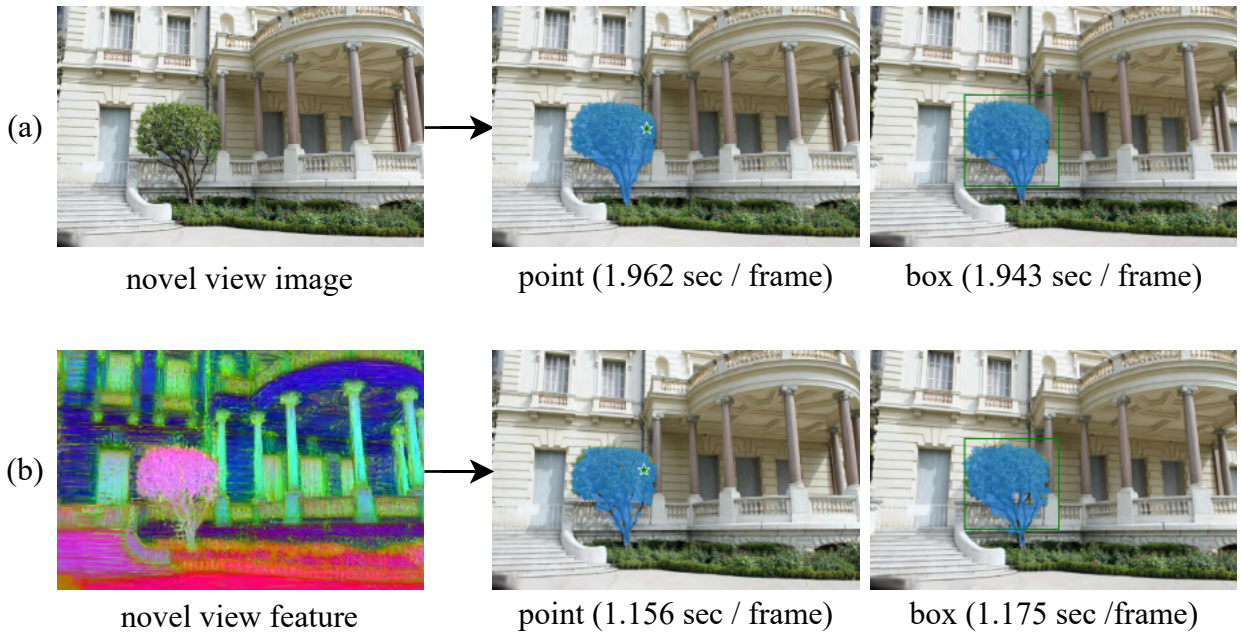


Figure 2.4: **Comparison of SAM segmentation results obtained by** (a) naively applying the SAM encoder-decoder module to a novel-view rendered image **with** (b) directly decoding a rendered feature. Our method is up to 1.7× faster in total inference speed including rendering and segmentation while preserving the quality of segmentation masks. Scene from [HPP18].

novel views and semantic segmentation masks for both synthetic and real scenes compared to NeRF-DFF.

2.4.2 Segment Anything from Any View

SAM excels in performing precise instance segmentation, utilizing interactive points and boxes as prompts to automatically segment objects in any 2D image. In our experiments, we extend this capability to 3D, aiming to achieve fast and accurate segmentation from any viewpoint. Our distilled feature field enables the model to render the SAM feature map directly for any given camera pose. As such, the SAM decoder is the only component needed

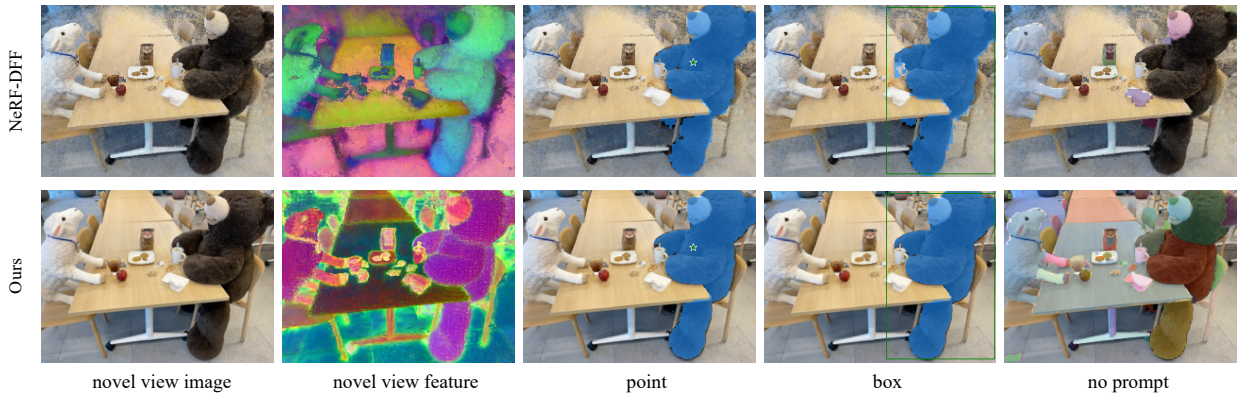


Figure 2.5: **Novel view segmentation (SAM) results compared with NeRF-DFF**. (Upper) NeRF-DFF method presents lower-quality segmentation masks - note the failure on segmenting the cup from the bear and the coarse-grained mask boundary on the bear's leg in box-prompted results. (Lower) Our method provides higher-quality masks with more fine-grained segmentation details. Scene from [KKG23].

to interact with the input prompt and produce the segmentation mask, thereby bypassing the need to synthesize a novel view image first and then process it through the entire SAM encoder-decoder pipeline. Furthermore, to enhance training and inference speed, we use the speed-up module in this experiment. In practice, we set the rendered feature dimension to 128, which is half of SAM's latent dimension of 256, maintaining the comparable quality of segmentation.

In Fig. 4.5, we compare the results of both point and box prompted segmentation on novel views using the naive approach (SAM encoder + decoder) and our proposed feature field approach (SAM decoder only). We achieve nearly equivalent segmentation quality, but our method is up to 1.7× faster. In Fig. 2.5, we contrast our method with NeRF-DFF [KMS22]. Our rendered features not only yield higher quality mask boundaries, as evidenced by the bear's leg, but also deliver more accurate and comprehensive instance segmentation, capable of segmenting finer-grained instances (as illustrated by the more 'detailed' mask on the far right). Additionally, we use PCA-based feature visualization [PVG11] to demonstrate that

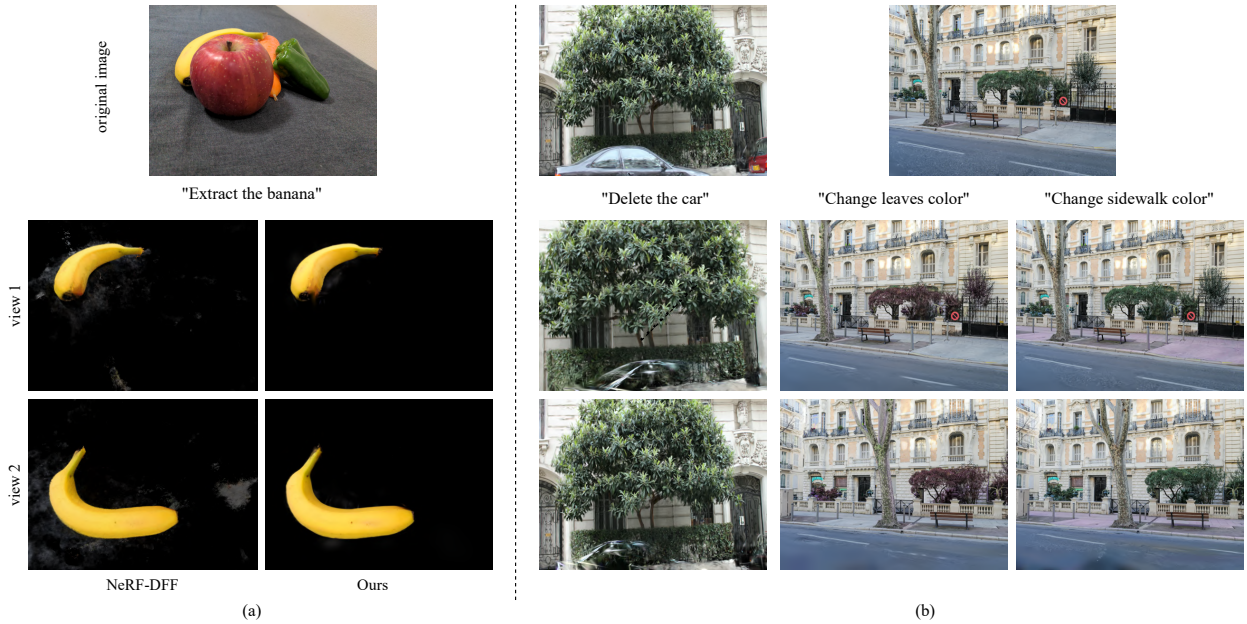


Figure 2.6: **Demonstration of results with various language-guided edit operations by querying the 3D feature field and comparison with NeRF-DFF** (a) We compare our edit results with NeRF-DFF method on the sample dataset provided by NeRF-DFF [KMS22]. Note that our method outperforms NeRF-DFF method by extracting the entire banana hidden by an apple in the original image and with less floaters in the background. (b) We demonstrate results with deletion and appearance modification on different targets. Note that the car is deleted with background preserved, and the appearance of the leaves changes with the appearance of the stop sign remained the same.

our high-quality segmentation masks result from superior feature rendering.

2.4.3 Language-guided Editing

In this section, we showcase the capability of our Feature 3DGS, distilled from LSeg, to perform editable novel view synthesis. The process begins by querying the feature field with a text prompt, typically comprising an edit operation followed by a target object, such as "extract the car". For text encoding, we employ a ViT-B/32 CLIP encoder. Our editing

pipeline capitalizes on the semantic features queried from the 3D feature field, rather than relying on a 2D rendered feature map. We compute semantic scores for each 3D Gaussian, represented by a K-dimensional vector (where K is the number of object categories), using a softmax function. Subsequently, we engage in either soft selection (by setting a threshold value) or hard selection (by filtering based on the highest score across K categories). To identify the region for editing, we generate a "Gaussian mask" through thresholding on the score matrix, which is then utilized for modifications to color c and opacity α on 3D Gaussians.

In Fig. 2.6, we showcase our novel view editing results, achieved through various operations prompted by language inputs. Specifically, we conduct editing tasks such as extraction, deletion, and appearance modification on text-specified targets within diverse scenes. As illustrated in Fig. 2.6 (a), we successfully extract an entire banana from the scene. Notably, by leveraging 3D Gaussians to update the rendering parameters, our Feature 3DGS model gains an understanding of the 3D scene environment from any viewpoint. This enables the model to reconstruct occluded or invisible parts of the scene, as evidenced by the complete extraction of a banana initially hidden by an apple in view 1. Furthermore, compared with our edit results with NeRF-DFF, our method stands out by providing a cleaner extraction with little floaters in the background. Additionally, in Fig. 2.6 (b), our model is able to delete objects like cars while retaining the background elements, such as plants, due to the opacity updates in 3DGS, showcasing its 3D scene awareness. Moreover, we also demonstrate the model's capability in modifying the appearance of specific objects, like 'sidewalk' and 'leaves', without affecting adjacent objects' appearance (e.g., the 'stop sign' remains red).

2.5 Discussion and Conclusion

In this work, we present a notable advancement in explicit 3D scene representation by integrating 3D Gaussian Splatting with feature field distillation from 2D foundation models, a

development that not only broadens the scope of radiance fields beyond traditional uses but also addresses key limitations of previous NeRF-based methods in implicitly represented feature fields. Our work, as showcased in various experiments including complex semantic tasks like editing, segmentation, and language-prompted interactions with models like CLIP-LSeg and SAM, opening the door to a brand new semantic, editable, and promptable explicit 3D scene representation.

However, our Feature 3DGS framework does have its inherent limitations. The student feature's limited access to the ground truth feature restricts the overall performance, and the imperfections of the teacher network further constrain our framework's effectiveness. In addition, our adaptation of the original 3DGS pipeline, which inherently generates noise-inducing floaters, poses another challenge, affecting our model's optimal performance.

CHAPTER 3

Imagination - From Text to 3D Worlds

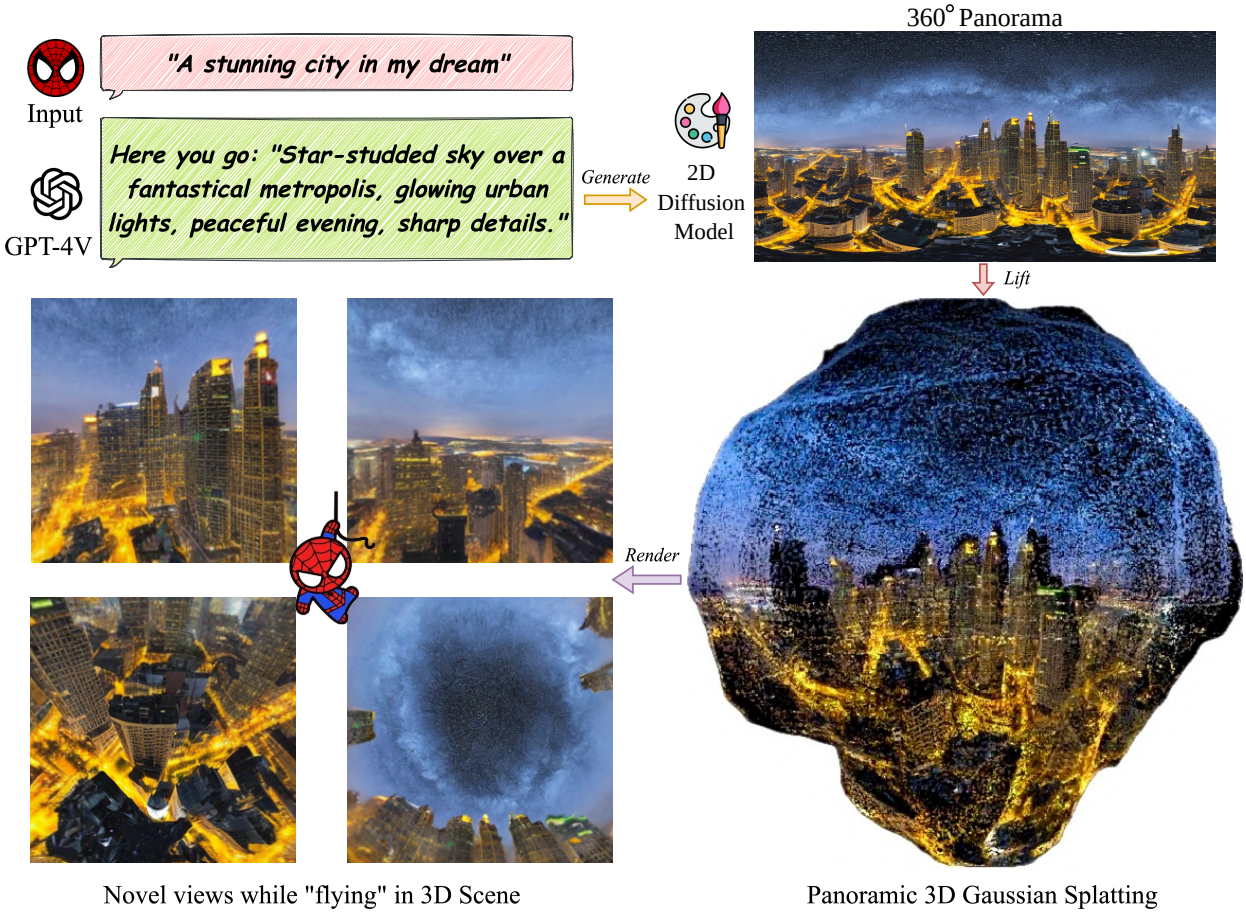


Figure 3.1: **DreamScene360.** We introduce a 3D scene generation pipeline that creates immersive scenes with full 360° coverage from text prompts of any level of specificity.

3.1 Introduction

The vast potential applications of text-to-3D to VR/MR platforms, industrial design, and gaming sectors have significantly propelled research efforts aimed at developing a reliable method for immersive scene content creation at scale. Recent developments in the 2D domain have seen the successful generation or editing of high-quality and adaptable images/videos using large-scale pre-trained diffusion models [RBL22, RDN22] on large-scale datasets, allowing users to generate customized content on demand.

Moving beyond 2D, the generation of 3D content, particularly 3D scenes, is constrained by the limited availability of annotated 3D image-text data pairs. Consequently, efforts in 3D content creation often rely on leveraging large-scale 2D models. This line of approach facilitates the creation of 3D scenes through a time-consuming distillation process. An example of this is DreamFusion [PJB22], which seeks to distill the object-wise 2D priors from diffusion models into a 3D neural radiance field (NeRF) [MST20]. However, these approaches often suffer from low rendering quality, primarily due to the multi-view inconsistency of 2D models, and struggle to extend to scene-scale 3D structure with fine details texture creation, particularly for outdoor scenes [PJB22] with outward-facing viewpoints and unbounded scene scale. Another avenue of 3D generation draws insights from explicit representations, such as point clouds and meshes, as demonstrated in LucidDreamer [CLN23] and Text2Room [HCO23]. These methods attempt to bridge the gap between 2D and 3D generation by initializing with an explicit 3D representation, and then progressively expanding the learned 3D representation to encompass a broader field-of-view. However, the progressive optimization frameworks leveraged by these methods struggle to inpaint substantial missing areas, especially when targeting 360° scenes under unconstrained conditions, resulting in notably distorted and disjointed structures. Moreover, the issue of prompt engineering in text-to-image generation [RBL22, SCS22a], becomes more pronounced in text-to-3D generation frameworks [CLN23, PJB22, AZS23] that rely on either time-consuming score distillation

or complex, multi-step progressive inpainting during the scene generation process, leading to a considerable trial-and-error effort to achieve the desired 3D scene.

To address the above challenges in creating a holistic 360° text-to-3D scene generation pipeline, we introduce **DreamScene360**. Our method initially leverages the generative capabilities of text-to-panorama diffusion models [WXF24] to produce omnidirectional 360° panoramas providing a comprehensive representation of the scene. A self-refining mechanism is used to enhance the image to alleviate prompt engineering, where GPT-4V is integrated to improve the visual quality and the text-image alignment through iterative quality assessment and prompt revision. While the generated panorama images overcome the view consistency issue across different viewpoints, they still lack depth information and any layout priors in unconstrained settings, and contains partial observations due to their single-view nature. To address this, our approach involves initializing scale-consistent scene geometry by employing a pretrained monocular depth estimator alongside an optimizable geometric field, facilitating deformable alignment for each perspective-projected pixel. The gaps, stemming from single-view observations, can be filled by deforming the Gaussians to the unseen regions by creating a set of pseudo-views with a synthesized multi-view effect and the distillation of pseudo geometric and semantic constraints from 2D models (DPT [RBK21] and DINOv2 [ODM23]) to guide the deformation process to alleviate artifacts.

Collectively, our framework, DreamScene360, enables the creation of immersive and realistic 3D environments from a simple user command, offering a novel solution to the pressing demand for high-quality 3D scenes (see the workflow in Fig. 5.1). Our work also paves the way for more accessible and user-friendly 3D scene generation by reducing the reliance on extensive manual effort.

3.2 Related Works

2D Assets Generation. The generation of 2D assets allows for incredible creative liberty, and the use of large-scale learning based priors for content generation. Generative Adversarial Networks [GPM20] were the original state of the art for image generation. Variants such as StyleGAN [KLA19] showed the ability for fine-grained control of attributes such as expression through manipulation of the latent representations as well. More recently, denoising diffusion models [DN21, HJA20, SME20] have been the new state of the art for generative models. Text-guided image diffusion models [SCS22b] have shown the capability of generating high-quality images that are faithful to the conditioning text prompts. Subsequent work has made the generation process more efficient by performing denoising in the latent space [RBL22], and by speeding up the denoising process [SH22, MRG23, GPK24]. Techniques such as classifier free guidance [HS22] have significantly improved faithfulness to text prompts. Additional control of generation has also been shown to be possible, through auxiliary inputs such as layout [ZZL23], pose [ZRA23] and depth maps [BMW23]. More recently, text-to-image diffusion models have been finetuned to generate structured images, such as panoramas [WXF24]. In this work, we utilize text-to-panorama generation as structured guidance for text to 360° scene generation.

Text-to-3D Scene Generation. In recent times, text-to-3D scene generation has been reinvigorated with the advent of 2D diffusion models and guidance techniques. Text-to-image diffusion models, along with techniques such as classifier free guidance, have been found to provide strong priors to guide 3D generation methods [PJB22, WLW23, WDL23, SCX23, XJW22b]. These methods are generally more frequently used than direct text-to-3D diffusion models [JN23, KVN23]. More recent works [LBL23, GLC23, VCK23] successfully generate multi-object compositional 3D scenes. On the other hand, a second class of methods use auxiliary inputs such as layouts [PW23]. The least constrained text-to-3D methods do not rely on any auxiliary input with the only input being the text prompt describing the

3D scene [HCO23, ZLW24, OHL23, FHL23, MCL23, YDH23, CLN23]. Our work requires a text prompt input; however, unlike prior work, we propose using panoramic images as an intermediate input for globally consistent scenes.

Efficient 3D Scene Representation. 3D scene representations include a wide variety of techniques, including point clouds [BTS14], volumetric representations [LSS19, NLT19], and meshes [ZC01]. More recently, however, learning-based scene representations have gained prominence. Implicit representations such as neural radiance fields [MST21a] have shown high quality rendering and novel view synthesis capability. Additional learning-based representations, that span both explicit [FYT22] and mixed [MES22] have been proposed to speed up learning. Most recently, 3D Gaussian splatting [KKL23] has enabled fast learning and rendering of high quality radiance fields using a Gaussian kernel-based explicit 3D representation. Since then, several works have emerged enabling sparse view [ZFJ23, XMU23] and compressed [NSW23, FWW23, NMK23, LRS23, MBH23] 3D Gaussian representations, as well as representation of multidimensional feature fields [ZCJ24]. A variation involves learning a Gaussian-based radiance field representation using panoramic images as input [BHG24]. In this work, we propose a method for text to 360° 3D scene generation, by using panorama images as an intermediate representation.

3.3 Methods

In this section, we detail the proposed DreamScene360 architecture (Fig.3.2). DreamScene360 initially generates a 360° panorama utilizing a self-refinement process, which ensures robust generation and aligns the image with the text semantically (Sec. 3.3.1). The transformation from a flat 2D image to a 3D model begins with the initialization from a panoramic geometric field (Sec. 3.3.2), followed by employing semantic alignment and geometric correspondences as regularizations for the deformation of the 3D Gaussians (Sec. 3.3.3).

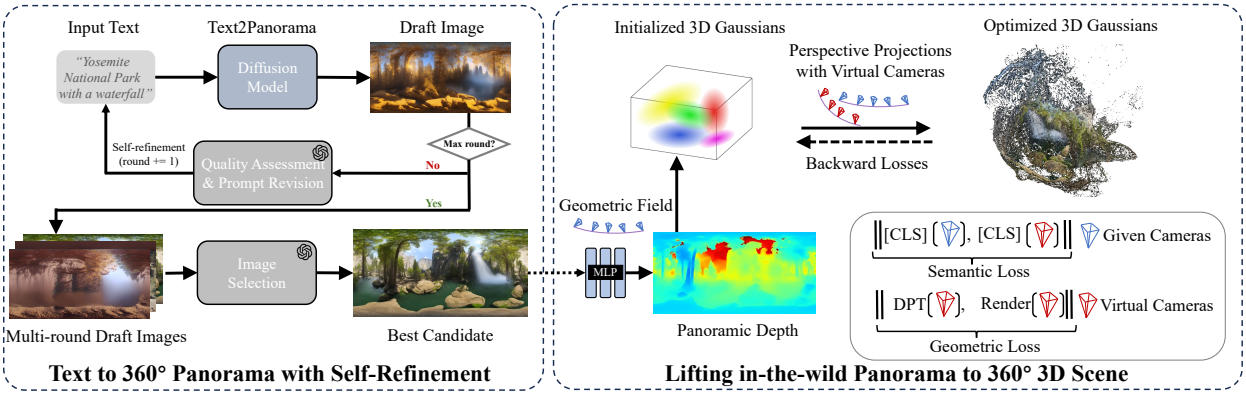


Figure 3.2: **Overall Architecture.** Beginning with a concise text prompt, we employ a diffusion model to generate a 360° panoramic image. A self-refinement process is employed to produce the optimal 2D candidate panorama. Subsequently, a 3D geometric field is utilized to initialize the Panoramic 3D Gaussians. Throughout this process, both semantic and geometric correspondences are employed as guiding principles for the optimization of the Gaussians, aiming to address and fill the gaps resulting from the single-view input.

3.3.1 Text to 360° Panoramas with Self-Refinement

Panoramic images provide an overview of the entire scene in one image, which is essential for generating 360° 3D scenes with global consistency.

360° Panoramic Image Generation. A crucial requirement for the generated panorama is ensuring continuity between the leftmost and rightmost sides of the image. We utilize the diffusion process from MultiDiffuser [BYL23] to generate a panoramic image I_0 of resolution $H \times 2H$ based on a pre-trained diffusion model Φ. Starting from a noisy image I_T, we iteratively denoise the image by solving an optimization problem for each of several patches of the image, selected via a sliding window mechanism. For each patch $P_i(I_t)$, we ensure the distance against their denoised version $\Phi(P_i(I_t))$ is minimized. Though each patch may pull the denoising process in a different direction, the fused result is the weighted average of the

update of each sample, whose closed form is written as follows,

$$\Phi(I_{t-1}) = \sum_{i=1}^{n} \frac{P_i^{-1}(W_i)}{\sum_{j=1}^{n} P_j^{-1}(W_j)} \otimes P_i^{-1}(\Phi(P_i(I_t))), \tag{3.1}$$

where W_i refers to per pixel weight, set to 1 in our experiments.

We use StitchDiffusion [WXF24] as the pretrained 2D diffusion model, where a stitch method is employed in the generation process for synthesizing seamless 360° panoramic images. Trained on a curated paired image-text dataset containing 360° panoramas, a customized LoRA [HSW21] module is incorporated into the MultiDiffusion [BYL23] process. At each denoising timestamp, we not only diffuse at the original resolution $H \times (2H + 2W)$ via MultiDiffuser Φ, but also stitch the leftmost and rightmost image regions ($H \times W$ each) and diffuse the concatenated patch as well to ensure consistency at the border regions. Finally, we consider the center cropped region of $H \times 2H$ as the final 360° panoramic image. It is worth noting that Dreamscene360 is versatile in practice and can also adapt to other text-to-panorama diffusion models.

Multi-Round Self-Refinement. Unlike previous works that generate 3D scenes through a time-consuming score distillation process [WLW23] or progressive inpainting [CLN23], our work uses panorama to achieve user-friendly "one-click" 3D scene generation. We integrate GPT-4V to facilitate iterative self-refinement during the generation process, a feature that was challenging to implement in previous baselines due to the absence of global 2D image representations. Here, we draw inspiration from perspective image generation, Idea2Img [YWL23], and implement a self-improvement framework aiming for better text-image aligned panoramic image generation. Starting from a user-provided rough prompt, we leverage GPT-4V to provide feedback and prompt revision suggestions according to the results generated by the previous step. During each round, GPT-4V is judges the generated image quality in terms of object counts, attributes, entities, relationships, sizes, appearance, and overall similarity with the original user-specified prompt. A score from 0-10 is assigned

to each draft image and the one image with highest score is provided to GPT-4V for additional improvement. GPT-4V will then produce an improved text prompt based on the issues observed in the current generation results, and the new prompt will be used for another round of panorama generation. After a number of rounds, we collect the image with the highest visual quality score judged by GPT-4V in the whole generation process as our final panorama image.

Our 3D scene generation framework, therefore, enjoys a user-friendly self-improvement process, without the need for troublesome prompt engineering for users as in previous methods [PJB22, WLW23, CLN23], but is able to obtain a high-quality, visually pleasing, text-aligned, and 360° consistent panoramic images that can be later converted into an immersive 3D scene via Panoramic Gaussian Splatting, which is detailed in the next sections.

3.3.2 Lifting in-the-wild Panorama to 360 Scene

Transforming a single image, specifically an in-the-wild 360° panoramic image, into a 3D model poses significant challenges due to inadequate observational data to regularize the optimization process, such as those required in 3D Gaussian Splatting (see Fig. 3.7). Rather than beginning with a sparse point cloud (3DGS), we initialize with a dense point cloud utilizing pixel-wise depth information from the panoramic image of resolution H $\times$ W, which are then refined towards more precise spatial configurations, ensuring the creation of globally consistent 3D representations, robust to viewpoint changes.

Monocular Geometric Initialization. Given a single panoramic image $\mathbf{P}$, we project it onto N perspective tangent images with overlaps $\{(\boldsymbol{I}_i \in \mathbb{R}^{H\times W\times 3}, \boldsymbol{P}_i \in \mathbb{R}^{3\times 4})\}_{i=1}^{N}$, following the literature's suggestion that 20 tangent images adequately cover the sphere's surface as projected by an icosahedron [RYR22]. Unlike indoor panoramas, which benefit from structural layout priors in optimization, we employ a monocular depth estimator, DPT [RBK21], to generate a monocular depth map $\boldsymbol{D}_i^{\text{Mono}}$ providing a robust geometric relationship. Never-

theless, these estimators still possess affine ambiguity, lacking a known scale and shift relative to metric depth. Addressing this efficiently is crucial for precise geometric initialization.

Global Structure Alignment. Previous studies in deformable depth alignment [HK18, WWC23, RYR22] and pose-free novel view synthesis [BWL23] have explored aligning scales across multiple view depth maps derived from monocular depth estimations. In this context, we utilize a learnable global geometric field (MLPs), inspired by [WWC23, RYR22], supplemented by per-view scale and per-pixel shift parameters: $\{(\alpha_i \in \mathbb{R}, \boldsymbol{\beta_i} \in \mathbb{R}^{H \times W}\}_{i=1}^{N}$. We define the view direction ($\boldsymbol{v}$) for all pixels on the perspective images deterministically. The parameters of MLPs Θ are initialized with an input dimension of three and an output dimension of one. The parameters $\{(\alpha_i \in \mathbb{R}\}_{i=1}^{N}$ are initialized to ones, and $\{(\boldsymbol{\beta_i} \in \mathbb{R}^{H \times W}\}_{i=1}^{N}$ to zeros. With these optimizable parameters, we define our optimization goal as follows:

$$\min_{\alpha,\beta,\Theta} \left\{ ||\alpha \cdot \boldsymbol{D}^{\text{Mono}} + \boldsymbol{\beta} - \text{MLPs}(\boldsymbol{v}; \Theta)||_2^2 + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}(\boldsymbol{\beta}) + \lambda_\alpha ||\gamma(\alpha) - 1||^2 \right\} \tag{3.2}$$
$$\text{where} \quad \mathcal{L}_{\text{TV}}(\boldsymbol{\beta}) = \sum_{i,j} \left((\beta_{i,j+1} - \beta_{i,j})^2 + (\beta_{i+1,j} - \beta_{i,j})^2 \right)$$

Here, λ_{TV} and λ_α are regularization coefficients to balance the loss weight, $\gamma(\cdot)$ is the softplus function. We set FoV as 80° during the optimization and use the predicted depth from the MLPs for the subsequent Gaussian optimization, as it can provide the depth of any view direction.

3.3.3 Optimizing Monocular Panoramic 3D Gaussians

While 3D Gaussians initialized with geometric priors from monocular depth maps provide a foundational structure, they are inherently limited by the lack of parallax inherent to single-view panoramas. This absence of parallax — critical for depth perception through binocular disparity — along with the lack of multiple observational cues typically provided

by a baseline, poses substantial challenges in accurately determining spatial relationships and depth consistency. Thus, it is imperative to employ more sophisticated and efficient generative priors that function well in unconstrained settings and enable the scaling up and extend to any text prompts.

3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) utilizes the multi-view calibrated images using Structure-from-Motion [SF16], and optimizes a set of Gaussians with center $\boldsymbol{x} \in \mathbb{R}^3$, an opacity value $\alpha \in \mathbb{R}$, spherical harmonics (SH) coefficients $\boldsymbol{c} \in \mathbb{R}^C$, a scaling vector $\boldsymbol{s} \in \mathbb{R}^3$ and a rotation vector $\boldsymbol{q} \in \mathbb{R}^4$ represented by a quaternion. Upon projecting the 3D Gaussians into a 2D space, the color C of a pixel is computed by volumetric rendering, which is performed using front-to-back depth order [KPL21]:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i T_i, \tag{3.3}$$

where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$, $\mathcal{N}$ is the set of sorted Gaussians overlapping with the given pixel, T_i is the transmittance, defined as the product of opacity values of previous Gaussians overlapping the same pixel.

Synthesize Parallax with Virtual Cameras. We emulate parallax by synthesizing virtual cameras that are unseen in training but are close to the input panoramic viewpoint. We methodically create these cameras to simulate larger movements. This procedure is quantitatively described by incrementally introducing perturbation to the panoramic viewpoint coordinates (x, y, z) formalized as:

$$(x', y', z') = (x, y, z) + \delta(d_x, d_y, d_z) \tag{3.4}$$

where (x', y', z') denotes the new virtual camera positions, and $\delta(d_x, d_y, d_z)$ denotes the progressively increasing perturbations in each coordinate direction, under uniform distribution over $[-0.05, +0.05] \times \gamma$, where $\gamma \in \{1, 2, 4\}$ stands for 3-stage progressive perturbations, emulating the camera's movement from the original point.

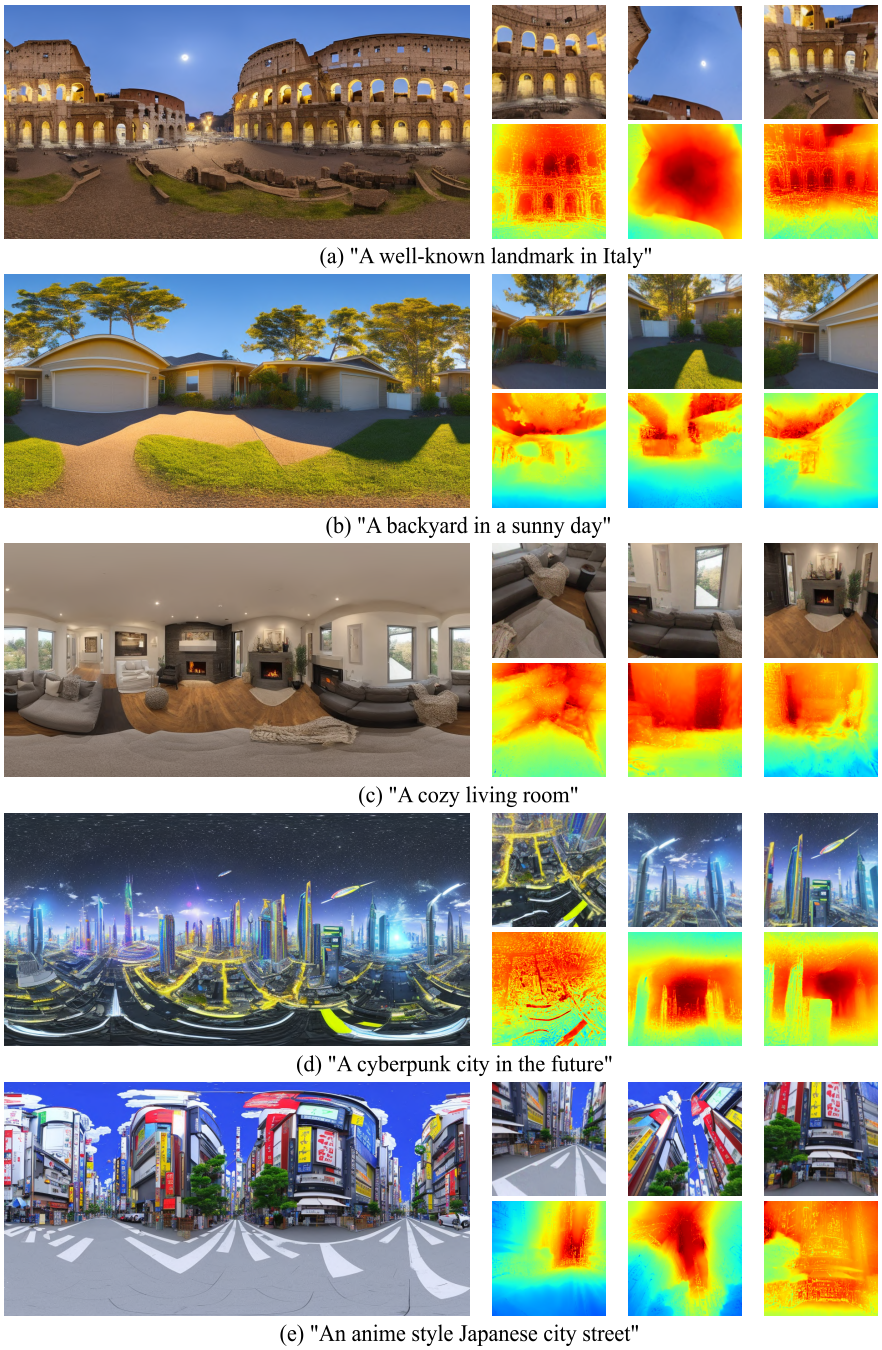


Figure 3.3: **Diverse Generation.** We demonstrate that our generated 3D scenes are diverse in style, consistent in geometry, and highly matched with the simple text inputs.

Distilling Semantic Similarities. Previous research has underscored the significance of capturing the appearance of objects/scenes to establish robust feature correspondences across different views [HKS22, LHR18], which is crucial for tasks like co-segmentation and 2D pre-training. Inspired by these studies, we aim to establish a connection in visual feature correspondence between training views and synthetically generated virtual views. This is achieved by enforcing feature-level similarity, guiding the 3D Gaussians to fill the geometric gaps in the invisible regions effectively. To elaborate, we generate a perspective image $\boldsymbol{I}_i$ from the panoramic 3D Gaussians at a specific training camera viewpoint $\boldsymbol{P}_i$ and create another synthesized image $\boldsymbol{I}'_i$ using Eq 3.4 that is proximate to $\boldsymbol{P}_i$. We employ the [CLS] token from the pre-trained DINOv2 [ODM23] model to encapsulate compact semantic features through the equation:

$$\mathcal{L}_{sem} = 1 - \text{Cos}([\text{CLS}](I_i), [\text{CLS}](I'_i)), \tag{3.5}$$

where Cos denotes the cosine similarity. This approach is inspired by [TBB22, XJW22a], assuming that the [CLS] token from a self-supervised, pre-trained Vision Transformer (ViT) can capture the high-level semantic attributes of an image. Such a mechanism is instrumental in identifying and leveraging similarities between two adjacent rendered images, facilitating our end-to-end 3DGS optimization process.

Regularizing Geometric Correspondences. While leveraging appearance cues is valuable, relying solely on them can lead to spatial inconsistencies. This is because models like DINOv2 might prioritize capturing semantic features over maintaining geometric coherence, potentially resulting in artifacts such as floaters. To mitigate this, we introduce a geometric regularization strategy designed to penalize discontinuities between pixels that exhibit inaccurate depth relationships. We employ a monocular depth estimator DPT [RBK21] for this purpose [DJQ23]. Although it may not provide globally accurate depth, it allows us to ascertain the relative spatial relationships between pixels:

$$\mathcal{L}_{geo}(I_i, D_i) = 1 - \frac{\text{Cov}(D_i, \text{DPT}(I_i))}{\sqrt{\text{Var}(D_i)\text{Var}(\text{DPT}(I_i))}}, \tag{3.6}$$

Here, I_i represents the rendered image at the i-th camera, and D_i signifies the rendered depth. Additionally, we incorporate an unsupervised local smoothness prior (TV loss) on the rendered depth at virtual views.

3.3.4 Optimization

The entire pipeline for transforming a panorama into 3D can be supervised end-to-end through a composite of all the loss functions.

$$\mathcal{L} = \mathcal{L}_{RGB} + \lambda_1 \cdot \mathcal{L}_{sem} + \lambda_2 \cdot \mathcal{L}_{geo} \tag{3.7}$$

Here, $\mathcal{L}_{RGB}$ represents the photometric loss on projected perspective images, which consists of $\mathcal{L}_1$ and a D-SSIM term as described in [KKL23]. $\mathcal{L}_{sem}$ denotes semantic regularization, and $\mathcal{L}_{geo}$ signifies geometric regularization. The weights λ_1 and λ_2 are set to 0.05 each, respectively.

3.4 Experiments

3.4.1 Experiment Setting

Implementation. Given the panorama generated from text, we resize it to 1024×2048 in order to produce a dense enough point cloud by projecting per-pixel panoramic depth values into 3D space along the per-pixel camera ray directions, where the directions can be obtained according to the coordinate transformation in spherical panoramic imaging [RYR22]. Since the panoramic depth map is predicted from the optimized geometric field, the values in local patches are guaranteed to be consistent and smooth, thereby resulting in point cloud in 3D that accurately captures the smooth geometry at the objects' surfaces. This setup enables effective 3D Gaussians initialization within a controllable, bounded area, eliminating redundant points in empty spaces. To circumvent the common floater issue in 3DGS rendering, we disable the densification process, enhancing the overall quality and consistency of the

rendered scenes.

Baseline Methods. Our work tackles the challenging problem of unconstrained 360° generation, including both indoor and outdoor scenarios. However, the works utilizing a bounded NeRF representation using score distillation do not work very well in this case. Thus, the comparisons are conducted between DreamScene360 (ours) and the state-of-the-art LucidDreamer [CLN23]. We use the open-source codebase of LucidDreamer, which starts from a single image and a text prompt. The framework constructs a global point cloud by progressive inpainting into 360° views and then distills a set of 3D Gaussians. In our experiments, we set the input image for the LucidDreamer to be generated from Stable Diffusion [RBL22] v1.5 using the input text for a fair comparison.

Metrics. Since there is no ground truth in the generated 3D scenes, we utilize CLIP [RKH21] embedding distance, following previous works [XJW22b, TRZ23], to measure the text-image alignment as a mechanism to quantify the novel view rendering quality. Specifically, for each method, we render images with camera rotations and translations to mimic the immersive trajectory inside the 3D scenes. Additionally, we utilize multiple non-reference image quality assessment metrics. NIQE [MSB12] and BRISQUE [MMB12] are widely adopted non-reference quality assessment methods in measuring in-the-wild image quality. QAlign [WZZ23] is the state-of-the-art method in quality assessment benchmarks, which adopts a large multi-modal model fine-tuned on available image quality assessment datasets. We utilize the "quality" mode of QAlign, which focuses on the perceptual quality of image contents.

3.4.2 Main Results

360° Scene Generation. As shown in Fig. 3.3, our method can generate diverse 3D scenes in different styles with distinct contents, while preserving high-fidelity novel-view rendering

Table 3.1: **Quantitative comparisons** between LucidDreamer and ours.

	CLIP Distance↓	Q-Align↑	NIQE↓	BRISQUE↓	Runtime
LucidDreamer [CLN23]	0.8900	3.0566	6.2305	51.9764	6min.15sec.
Ours	**0.8732**	**3.1094**	**4.9165**	**38.3911**	7min.20sec.

ability and realistic scene geometry. These results showcase our method's generalization ability to diverse use cases, providing a user-friendly experience in realizing users' imaginations.

Comparisons with Baseline Methods. We show visual comparisons against LucidDreamer [CLN23] in Fig. 3.4. LucidDreamer involves progressive inpainting to find agreement between multiple synthesized images and tries to fuse a point cloud that can be later distilled into 3D Gaussians.

Their pipeline, which inpaints each patch separately based on the same text prompt, tends to produce repetitive results especially when generating complex scenes. In comparison, our method delivers consistent results thanks to the intermediate panorama as a global 2D representation. We provide quantitative comparisons in Tab. 3.1, where the camera undergoes clockwise rotation in yaw, accompanied by slight random rotations in pitch and random translations to capture 4 novel views roughly representing front, back, left, and right, along with two views of up and down by 90° rotation in pitch, to mimic the exploration in an immersive 360° panoramic 3D scene.

For the *bedroom* text prompt, LucidDreamer begins with a view featuring a modern style bedroom with one bed but hallucinates the novel views into multiple bedrooms, while the style of the imagined bedrooms diverges from that of the starting view. In the case of the *Yosemite* text prompt, LucidDreamer merely replicates the waterfall seen in the initial view throughout. In conclusion, our results demonstrate global semantic, stylized, and geometric consistency, offering complete 360° coverage without any blind spots.

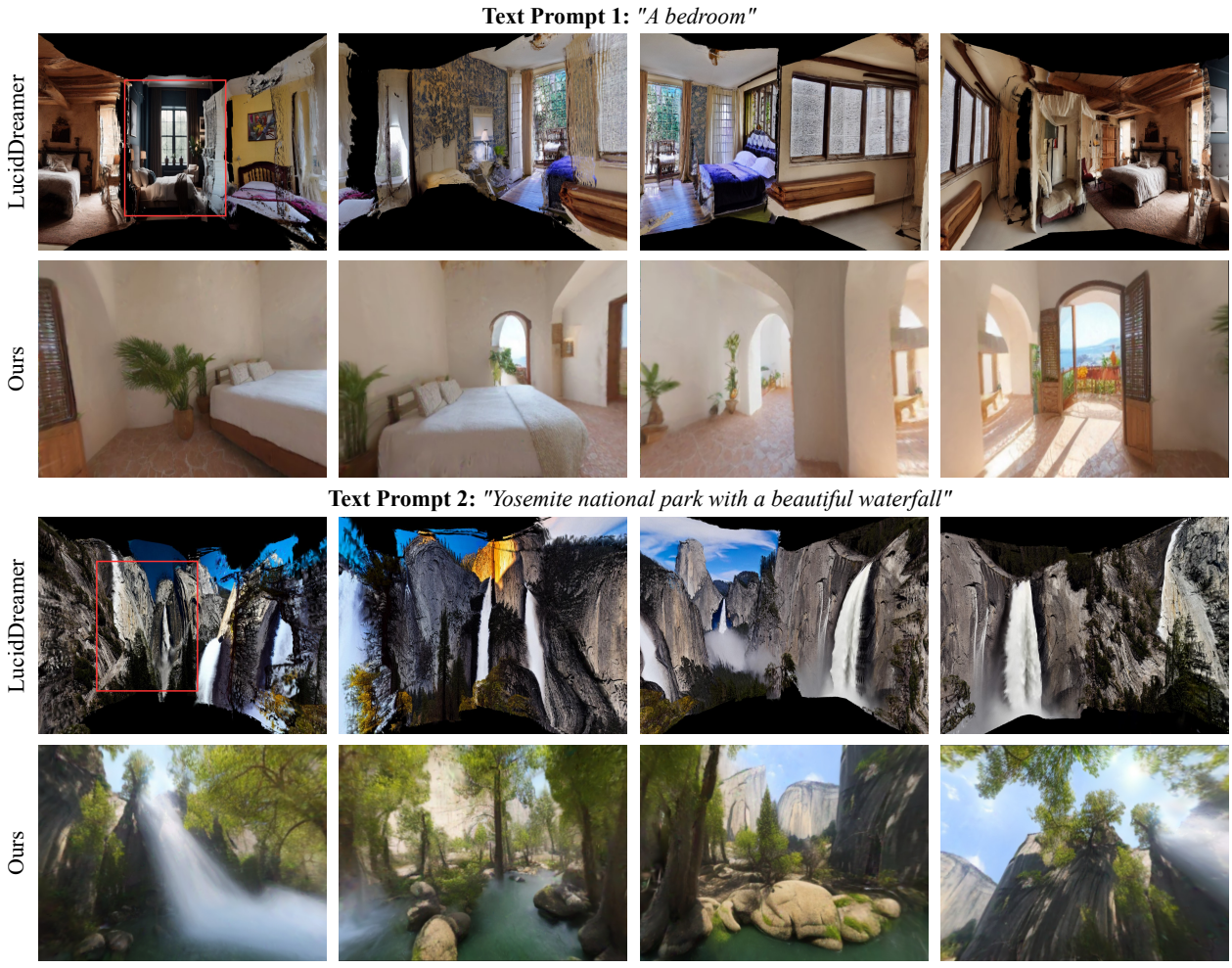


Figure 3.4: **Visual Comparisons.** We showcase 360° 3D scene generation. In each row, from left to right, displays novel views as the camera undergoes clockwise rotation in yaw, accompanied by slight random rotations in pitch and random translations. LucidDreamer [CLN23] hallucinates novel views from a conditioned image (indicated by a red bounding box) but lacks global semantic, stylized, and geometric consistency. In contrast, our method provides complete 360° coverage without any blind spots (black areas in baseline results), and shows globally consistent semantics.

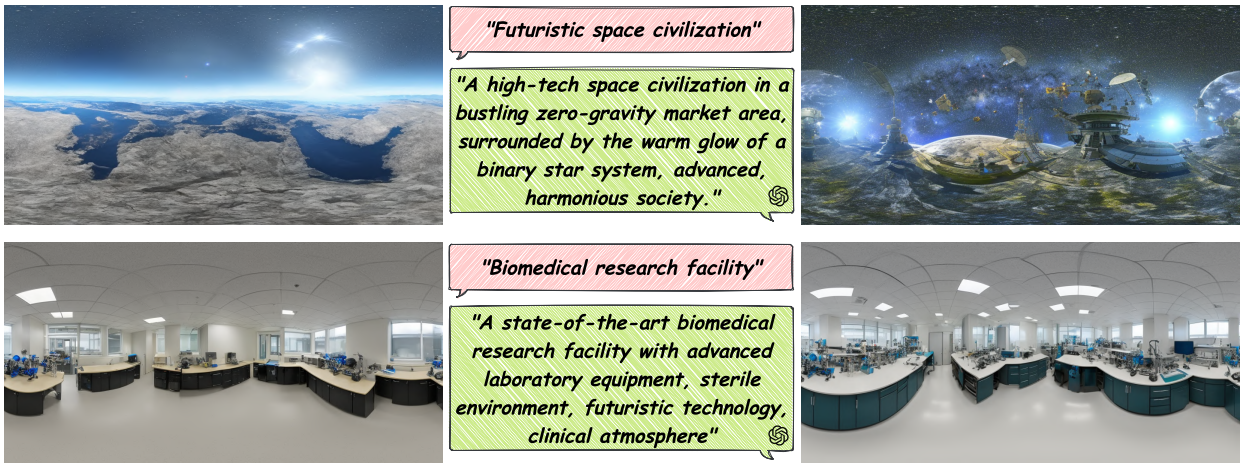


Figure 3.5: **Ablation of Self-Refinement.** We demonstrate that the self-refinement process greatly enhances the image quality by improving the text prompt. As shown in each row, the image on the left is generated using a simple user prompt, while a prompt augmented by GPT-4V obtains the image on the right. We observe that after the multi-round self-refinement, GPT-4V selects the one panorama with better visual quality, which provides solid support for the immersive 3D scene we further generate.

3.4.3 Ablation Study

Self-refinement Process We further evaluate the importance of the self-refinement process. In this part, we mainly focus on the generated panorama since panorama provides a holistic view of the 3D scene and provides an upper bound for the visual quality. As can be seen in Fig. 3.5, we observe that using a simple text prompt usually delivers minimalist results with fewer details. With the help of GPT-4V, our self-refinement process enjoys prompt revision and quality assessment. This ability finally facilitates the selection of a more realistic and detailed panorama image among the draft image candidates. These functionalities are otherwise hard to achieve in previous baselines that do not have global 2D representations, and as a result, our results provide a much better visual appearance than baselines as shown in Fig. 3.4 and Tab. 3.1.



(a) $\mathcal{L}_{RGB}$ (b) $\mathcal{L}_{RGB} + \mathcal{L}_{geo}$ (c) $\mathcal{L}_{RGB} + \mathcal{L}_{sem}$ (d) $\mathcal{L}_{RGB} + \mathcal{L}_{geo} + \mathcal{L}_{sem}$

Figure 3.6: **Ablation of Optimization Loss.** We demonstrate the impact of Semantic and Geometric losses on the synthesized virtual cameras. (a) Utilizing photometric loss on camera views from a rendered panorama induces artifacts when altering rendered camera positions. (b) Implementing Geometric and (c) Semantic regularizations effectively reduces the artifacts originating from invisible views. (d) Integrating both regularizations yields the most optimal outcomes.

Loss Function We investigate the importance of multiple loss functions we adopted in Fig. 3.6. As shown in the figure, especially in the highlighted region, our full configuration of loss functions delivers the best visual quality when rendered from a novel viewpoint. In comparison, removing the geometric regularizations or semantic constraints will result in unpleasant artifacts.

Initialization We showcase the importance of proper initialization in Fig. 3.7. As can be seen in the image, adopting our curated point initialization largely benefits the rendering quality of our 3D Panoramic Gaussian Splatting. An alternate solution involves adopting random initialization of the point cloud locations. This option, however, generates blurry results, mainly due to the challenge of learning proper point locations on the fly without ground truth geometry as supervision. Note that since our full model incorporates this high-quality point initialization, at the training stage, we can safely disable the adaptive density control of 3D Gaussians and, therefore, speed up convergence. Using random initialization



(a) Random initialized geometry for optimizing 3D Gaussians.



(b) Initialized from globally aligned monocular geometry.

Figure 3.7: **Ablation Study on 3D Initialization.** We present a comparative visualization of various initialization methods for 3D Panoramic Gaussian Splatting. In the absence of geometric priors of the scene (a), the optimized 3D Gaussian rendering yields plausible results in panoramic camera views. However, artifacts become evident when the rendered camera position is altered. To address this, we employ a monocular depth approach combined with learnable alignment factors and a geometric field (b), which ensures consistent alignment across multiple views.

of 3D Gaussians, on the other hand, has to rely on adaptive density control introduced in 3D Gaussian Splatting paper [KKL23] (*e.g.* densify, clone, prune) to move the 3D Gaussians to desired locations via backpropagation.

3.5 Conclusion

In this work, we present DreamScene360, a novel framework that unrestrictedly generates high-quality, immersive 3D scenes with complete 360° coverage in various styles from text inputs, regardless of their specificity level. Our pipeline leverages panorama as a middle ground, which provides us with a self-refinement opportunity by asking GPT-4V for quality assessment and prompt revision. Subsequently, we construct a geometric field that initializes 3D Gaussians. With the help of unsupervised geometric and semantic loss on virtual views, we lift the 2D panorama into panoramic 3D Gaussians. Experiments demonstrate our superiority against baseline methods in terms of global consistency and visual quality. However, our generation results are still limited by the default resolution (512×1024) of the pretrained text-to-panorama diffusion model. Moving forward, we will explore generating 3D scenes at higher resolutions and even 4D dynamic scenes [LPY24] for a better, seamless, and immersive user experience.

CHAPTER 4

Interation - Bridging Space, Time, and Action

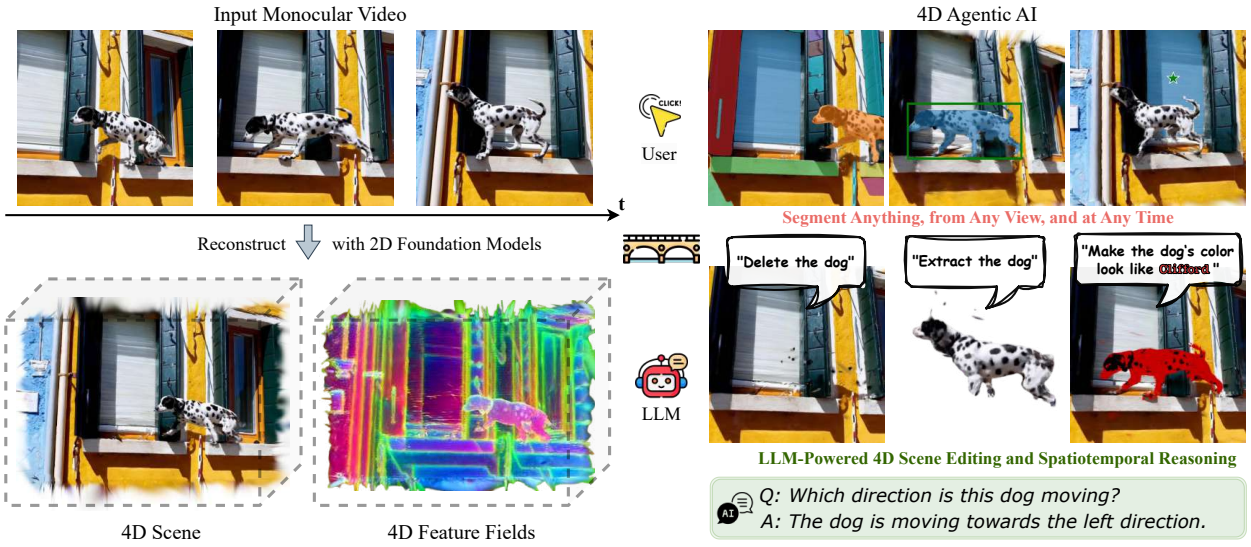


Figure 4.1: **Feature4X: Building 4D Interactive Scenes with Agentic AI from Monocular Videos.** By dynamically distilling model-conditioned features and integrating 2D foundation models with LLMs in feedback loops, Feature4X enables multimodal tasks across 2D, 3D, and 4D with high-level language inputs or direct user interactions, including (but not limited to) segmentation, scene editing, and VQA across novel views and all time steps, unlocking new possibilities for 4D agentic AI.

4.1 Introduction

The rapid evolution of 2D vision and multimodal models has been fueled by access to massive curated datasets with rich annotations, alongside breakthroughs spanning multiple domains. These developments have led to remarkable progress in various tasks, including image segmentation [RKH21], image editing [BHE22], promptable semantic segmentation [RGH24, KMR23], and Visual Question Answering (VQA) [WLL24, LHW23]. Despite these advances, the processing and interpretation of dynamic 3D data—crucial for applications such as autonomous driving, robotics, and 3D asset creation—still lag significantly behind the development of versatile and robust 2D foundation models. A fundamental challenge in 3D vision lies in processing multi-view images and widely available monocular videos, which often suffer from limited camera information and a scarcity of well-annotated, fine-grained per-frame datasets. To bridge this gap, we propose to leverage the latent features of well-trained representations from 2D and multimodal domains [RKH21, RGH24, KMR23, WLL24, LHW23] and adaptively lift them into higher-dimensional feature fields (such as 3D and 4D) in a scalable and efficient manner, enabling the direct transfer of 2D functionalities into 4D with minimal data annotation and training overhead.

Constructing 4D feature fields from casually captured in-the-wild videos is significantly challenging. Existing approaches for static 3D feature fields [YHS24, SWD24, ZCJ24, QLZ24, LZJ24, ZSZ24, LYB24, KMS22, KKG23, TLL22, FZC24] are not directly applicable to 4D scenarios for three primary reasons. First, these methods typically rely on well-calibrated, multi-view input images with precise camera poses, which are difficult to obtain from casual videos. Second, extending static feature fields to include an additional temporal dimension results in substantial memory demands, leading to unstable and costly optimization. Third, previous approaches are limited to task-specific feature fields, necessitating full-parameter re-training for adaptation to new tasks.

In this paper, we set an ambitious goal: to parse in-the-wild monocular videos into a unified dynamic 4D representation that not only reconstructs accurate appearance and geometry but also seamlessly integrates advanced functionalities from a range of 2D foundation models. To tackle the aforementioned challenges, we propose enhancing the dynamic 3D Gaussian Splatting-based 4D scene representation [KKL23, LKL23b] with a unified latent feature capable of distilling diverse 2D foundation features, achieving both flexibility and efficiency. However, directly lifting 2D feature maps to dense, dynamic 3D Gaussians incurs significant computational costs and fails to scale with the spatial and temporal complexity of the scene. Inspired by recent advancements in Gaussian-based 4D reconstruction [LWH24a, WYG24], we propose to leverage the smooth, compact nature of underlying scene semantics and represent the dense 4D feature field using a sparse set of base or "Scaffold" features, enabling efficient representation and scalable adaptation to various downstream tasks.

Our pipeline is fully end-to-end differentiable, supervised by ground truth color and feature maps exported from 2D vision foundation models. Its general and model-agnostic design supports a wide range of vision tasks, spanning 2D (semantic and promptable segmentation), 3D (scene editing), and 4D (spatial and temporal VQA). Furthermore, our compact and versatile 4D feature field serves as a bridge, seamlessly integrating these tasks with LLMs in a continuous feedback loop. This integration enables intuitive and efficient execution through free-form, high-level natural language inputs or direct user interactions. By unifying diverse vision tasks, our framework—seamlessly integrated with LLMs—paves the way for the development of advanced, contextually and spatiotemporally aware 4D agentic AI systems.

We summarize our **contributions** as follows:

- We propose a general 4D feature field distillation technique using Gaussian Splatting to build an interactive 4D scene and lift the functionalities of any 2D vision (image/video) foundation models into the 4D realm, relying solely on monocular video input.
- We introduce the first compact and versatile 4D Gaussian feature field representation,

which leverages the smooth, low-rank nature of scene semantics by modeling the dense 4D feature field as interpolations of a sparse set of base features.

- We develop an LLM-powered agentic AI capable of interpreting natural language prompts, dynamically adjusting configuration parameters, and iteratively refining results through trial and feedback, enabling intelligent 4D scene interaction and manipulation.

4.2 Related Works

4.2.1 4D Representations for Reconstruction and Generation

Various 4D representations are designed to support a wide variety of 4D tasks. Among them, 4D reconstruction is the most widely studied topic that drives the field. Early methods [GTZ21a, PCP21a, PDW21, PSH21, FMW23a, CJ23, KKC21, NSL21, PZX21, SYZ21, WCS22, GTZ21b, LSZ22, DZY21] implement variants of Neural Radiance Fields [MST21b] to represent a dynamic 4D scene through implicit or hybrid representations. More recently, 3D Gaussian [KKL23] has prevailed as a popular explicit 3D scene representation [CLN23, ZFX24, YDH24] and various variants [LKL23b, WYF23, DWD24, YYP23, YGZ24, HSY24, LKL23a, WYG24, LWH24a, MPT25, SHU24, JWF24] are similarly proposed to represent dynamic 4D scenes. Additionally, the field of 4D Generation [YXW23, RPT23, RXM24, LKT23, XLB24] has witnessed great advancements in representations that are more suitable for generative settings. In summary, most 4D representations can be categorized into deformation-based approaches [LKL23b, TTG21, PCP21b, PSB21, PSH21, WZT22, SJL24, LCL24], temporal extended approaches [YYP23, CJ23, FMW23b, DWD24, ZZZ24a], or an ensemble of multiple time-varying representations [XHK21, LNS21, GSK21, LPY24, RXM24]. In our work, we extend the direction of dynamic 3D Gaussian Splatting to a versatile Gaussian feature field that simultaneously embraces multiple features for various

vision tasks beyond 4D reconstruction and generation. To the best of our knowledge, this direction has not been widely explored.

4.2.2 4D Reconstruction from Monocular Video

In the context of 4D reconstruction tasks, most methods [NMO19, LDZ24a] focus on reconstructing dynamic 4D scenes from multiple well-calibrated videos. In contrast, another popular direction is to synthesize generic 4D scenes from monocular videos [XHK21, TTG21, PCP21b, LNS21, PSB21, GSK21, PSH21, WZT22]. For unposed casual videos, non-rigid structure-from-motion workflows [BV99, RKS12, BKL16, KTC15, KTE18, KGF20, YSJ21, YVN22, CL96, NFS15, LSP08, GT19, BPZ20, TD21, ZLG22] are adopted for estimating camera poses for each frame for further processing. Shape-of-motion [WYG24] and MoSca [LWH24a] develop end-to-end workflows that reconstruct 4D scenes from casual captured videos with the help of 3D Gaussians [KKL23, LYX24, LKL23b]. Another series of work [YXW23, LPY24, RPT23, RXM24] assumes the camera is fixed for the video and reconstructs 3D in the camera coordinate space. More recently, the success of DUSt3R [WLC24], a transformer-based novel paradigm for 3D reconstruction of arbitrary image collections, has been adapted for dynamic scenes in MonST3R [ZHH24]. In comparison, we aim to perform monocular 4D reconstruction and feature lifting simultaneously, which cannot be achieved in previous approaches.

4.2.3 Feature Field Distillation

Research in novel view synthesis and feature field representation has been extensively developed within the NeRF framework [MST21a] and 3D Gaussians [KKL23]. Seminal works such as Semantic NeRF [ZLL21] and Panoptic Lifting [SPB22] have integrated semantic information from segmentation networks into 3D spaces, revealing that combining noisy or inconsistent 2D labels within a 3D context can produce clear and accurate 3D segmentations.

Building on this concept, methods such as those in [RAR22] have shown that minimal input, like basic foreground-background masks, can be effective for object segmentation in 3D. Moving beyond label estimation to optimize NeRF, approaches like Distilled Feature Fields [KMS22], NeRF-SOS [FWJ22], LERF [KKG23], and Neural Feature Fusion Fields [TLL22] have embedded pixel-aligned feature vectors from tools such as LSeg or DINO [CTM21] into NeRF structures. More recently, numerous works [YHS24, SWD24, ZCJ24, QLZ24, LZJ24, ZSZ24, LYB24] adopt similar strategies to distill information from well-trained 2D models to 3D Gaussians. However, feature lifting into 4D fields is not properly solved yet. In this work, we make the first attempt to deliver a versatile 4D Gaussian feature field framework that embraces multiple task features simultaneously.

4.3 Method

Given a monocular RGB video $\mathcal{I} = I_1, \ldots, I_t$, we reconstruct the underlying 4D scene as a set of dynamic 3D Gaussians (Sec. 4.3.1), each augmented with a unified latent feature embedding that distills informative features from various 2D foundation models for downstream tasks (Sec. 4.3.2). To address the challenge of handling high-dimensional features across a large number of Gaussians and timesteps, we adopt a compact feature representation (Sec. 4.3.3).

4.3.1 Preliminaries: Dynamic 3D Gaussian Splatting

Following state-of-the-art monocular video-based dynamic 4D scene reconstruction approaches [WYG24, LWH24a, SHU24], we represent the scene with dynamic 3D Gaussians [LKL23b], namely a set of persistent 3D Gaussians [KKL23] that deform over time t. Specifically, we base our approach on MoSca [LWH24a], which can reconstruct the dynamic 4D scene from monocular casual video. Like other methods, MoSca addresses the challenge from single-view partial observation by leveraging priors from 2D foundation models [PYS24, BBW23, TD20,

KRG23, HFF22] and by regularizing Gaussian motion trajectories. At the core of MoSca's method is a structured graph named 4D Motion Scaffold $(\mathcal{V}, \mathcal{E})$ that drives the deformation of individual 3D Gaussians $\mathcal{G} = \{G_j\}_{j=1}^{n}$ (see Fig. 4.2). Each node $\mathbf{v}^{(i)} \in \mathcal{V}$ describes a 3D motion trajectory $[\mathbf{Q}_1^{(i)}, \ldots, \mathbf{Q}_t^{(i)}], \mathbf{Q} = [\mathbf{R}, \mathbf{t}] \in SE(3)$ with a control radius $r^{(i)}$. Edges $\mathcal{E}$ describe a k-nearest neighbor graph over the motion trajectories. Given any 3D Gaussian located at $\mathbf{x}$ at time τ, to compute its deformation to time τ', we first find the trajectory node $\mathbf{v}^{(i^*)}$ with the closest position $\mathbf{p}_\tau^{(i)}$ at time τ, i.e., $i^* = \arg\min_i ||\mathbf{p}_\tau^{(i)} - \mathbf{x}||$, then interpolate the deformation from K nearest trajectory nodes $\{\mathbf{v}^{(i)}\}_{i \in \mathcal{E}(i^*)}$, with weights $\{w_i\}$ computed from the Gaussian-to-trajectory distance and the node control radius. In other words, the dense, per-Gaussian motion trajectories are interpolations of a much smaller ($100\times$) set of trajectory nodes. This design leverages the low-rank nature of the underlying scene motion and effectively regularizes Gaussian motions under limited supervision. We follow [LWH24a]'s protocols to initialize and train 4D Motion Scaffolds and dynamic Gaussians that represent the dynamic foreground elements, as well as another set of static Gaussians that model the static background. Please refer to our supplementary material for full details.

4.3.2 Unified Latent Feature Distillation

We aim to go beyond appearance reconstruction by building a unified, versatile, and dynamic 4D feature field representation, capable of supporting diverse downstream visual tasks $\mathcal{T} = \{T^1, \ldots, T^S\}$ spanning 2D, 3D, and 4D, such as novel-view segmentation (2D), scene editing (3D), and scene-level spatiotemporal VQA (4D). A straightforward way to construct this 4D feature field is to extend existing 3D feature field frameworks [ZCJ24] by replacing their 3D reconstruction modules with dynamic 4D reconstruction. However, this approach addresses each downstream task T^s independently: input frames $\mathcal{I} = \{I_1, \ldots, I_t\}$ are encoded into task-specific 2D feature maps $\mathcal{F}^s = \{F_1^s, \ldots, F_t^s\}$ using dedicated vision foundation model encoders E^s, leading to separate feature fields per task. Consequently, performing three tasks would require three separate reconstructions, making this strategy inefficient and redundant.

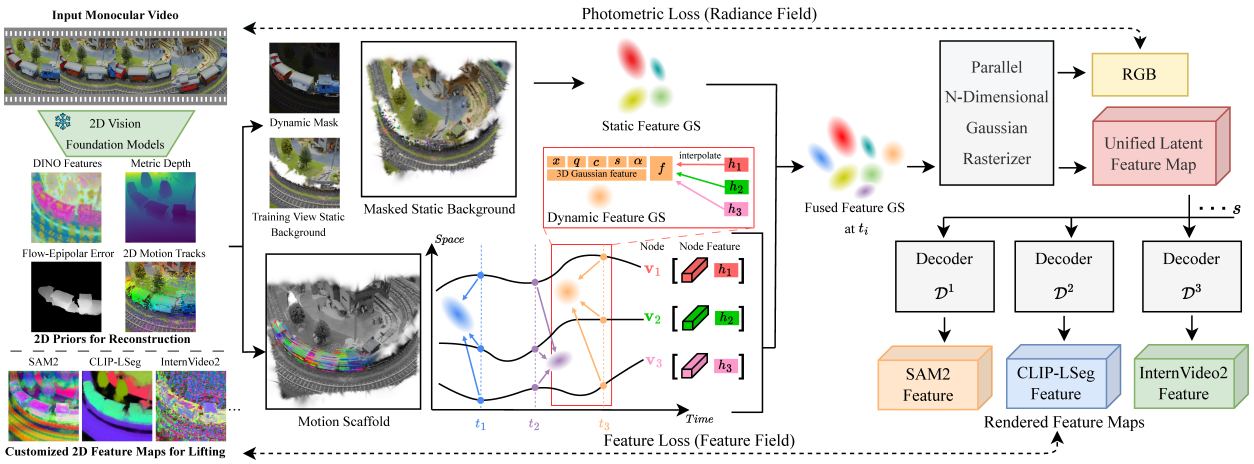


Figure 4.2: **Method overview.** Given an input monocular video, we infer 2D priors to segment static background (represented by static 3D Gaussians augmented with latent features) and dynamic foreground (represented by dynamic 3D Gaussians guided by Motion Scaffolds [LWH24a], a set of nodes $\{\mathbf{v}_i\}$ encoding 3D motion trajectories and latent features h_i). Dynamic Gaussian features and motions are computed via interpolation from their K-nearest scaffold nodes. At each timestep, dynamic Gaussians are warped and fused with static Gaussians. A parallel rasterization [ZCJ24] generates RGB images and a unified latent feature map, decoded into task-specific features—illustrated here by SAM2 [RGH24], CLIP-LSeg [LWB22a], and InternVideo2 [WLL24] for representative 2D (novel view segmentation), 3D (scene editing), and 4D (spatiotemporal VQA) tasks. Our framework generalizes to any 2D vision foundation model and is trained end-to-end using input RGB frames and customized features from pretrained 2D models. At inference, rendered feature maps from arbitrary views and timesteps are directly fed into task-specific decoders, seamlessly supporting user prompts and LLM interactions to form a unified 4D agentic AI system.

To overcome this limitation, we propose distilling a unified 4D feature field to coherently fuse diverse 2D vision foundation model features across views and timesteps, providing consistent feature access in 4D (3D space + time) for various downstream tasks. Specifically,

as shown in Fig. 4.2, we utilize the parallel N-dimensional Gaussian rasterizer [ZCJ24] to simultaneously render RGB images along with a unified latent feature $\mathcal{F}$, which is jointly learned with a set of lightweight decoders $\{\mathcal{D}^1, \ldots, \mathcal{D}^S\}$. Each decoder maps the shared latent feature $\mathcal{F} \in \mathbb{R}^D$ into task-specific target features $\mathcal{F}^s \in \mathbb{R}^{D_s}$, where D and D_s denote the dimensions of the unified latent and task-specific features, respectively, with $D \ll D_s$. Compared to separately optimizing high-dimensional features for each task, our unified latent representation significantly reduces computational overhead during rasterization. During the optimization process, we attach the feature vector $f_j \in \mathbb{R}^D$ to each 3D Gaussian $G_j \in \mathcal{G}$, warp G_j to the target timestep τ following the process introduced in [LWH24a], and rasterize f_j the same way we rasterize Gaussian color c_j as [ZCJ24]. Conceptually, the RGB and feature reconstruction from viewpoint v at timestep τ are computed as:

$$\begin{aligned} \hat{I}_\tau^v &= \text{Rasterize}(v, \{\text{warp}(G_j, \tau), c_j)\}_{G_j \in \mathcal{G}}) \\ \hat{F}_\tau^v &= \text{Rasterize}(v, \{\text{warp}(G_j, \tau), f_j)\}_{G_j \in \mathcal{G}}) \end{aligned} \tag{4.1}$$

The resulting feature map $\hat{F}_\tau$ (with viewpoint v omitted for simplicity) is decoded by $\mathcal{D}^s$ into a task-specific feature map $\hat{F}_\tau^s$, supervised by the ground-truth feature map obtained from a customized vision model encoder E^s (e.g., SAM2). Concretely, we optimize the following feature loss L_{feat} to jointly learn the feature field, alongside the photometric loss from [LWH24a] used for learning the radiance field.

$$L_{\text{feat}} = \sum_{s=1}^{S} \text{MSE}(\hat{F}_\tau^s, F_\tau^s), \tag{4.2}$$

$$\hat{F}_\tau^s = \mathcal{D}^s(\hat{F}_\tau), \qquad F_\tau^s = E^s(I_\tau). \tag{4.3}$$

Note that the feature loss and photometric loss are independent, and introducing the feature field does not degrade the quality of the radiance field reconstruction (see Tab. 4.1).

4.3.3 Scaffold-Based Compact Feature

While our unified feature field effectively reduces feature dimensionality, optimizing a feature vector for every Gaussian remains computationally expensive. However, semantic features tend to be smooth in 3D and exhibit strong local correlations, similar to Gaussian motion trajectories. To leverage this correlation, we propose representing per-Gaussian features $\{f_j\}$ as linear combinations of a smaller (100× fewer) and more compact set of base features $\{h_i\}$, conveniently attached to nodes $\{\mathbf{v}^{(i)}\}$ of the 4D Motion Scaffold (see Fig. 4.2). Recall that we compute per-Gaussian deformations by interpolating from their K-nearest trajectory nodes $\{\mathbf{v}^{(i)}\}_{i\in\mathcal{E}(i^*)}$ with interpolation weights $\{w_i\}$. We reuse these same weights to obtain the per-Gaussian unified features as:

$$f_j = \sum_{i\in\mathcal{E}(i^*)} w_i h_i. \tag{4.4}$$

This compact feature representation significantly reduces the number of parameters required for optimization and provides structural regularization, encouraging smoother learned features.

4.3.4 Interaction with AI Agent via Feature Fields

Agentic AI typically requires cross-modal interactions, notably between language and vision. In this work, we aim to build an AI agent that supports direct language interactions with our dynamic 4D scene representation. Such interaction requires a shared feature space between text and Gaussian features. However, text features are usually high-dimensional. For instance, language-guided scene editing requires CLIP features of 512 dimensions. Directly assigning these high-dimensional features to each Gaussian is computationally expensive. Feature 3DGS [ZCJ24] addresses this by assigning each Gaussian a full 512-dimensional CLIP feature via a parallel N-dimensional Gaussian rasterizer, enabling direct language interaction but resulting in slow rendering speeds and excessive memory usage. Although Feature 3DGS proposes a CNN-based acceleration, it operates only on 2D rendered maps,

failing to directly resolve the feature mismatch in 3D. In contrast, we optimize a compact, lower-dimensional Gaussian feature ($D = 32$), train an MLP-based decoder on rendered 2D features, and apply it directly to 3D Gaussian features during inference. Given the intrinsic flexibility of MLPs, our approach efficiently bridges language features and our compact feature fields, enabling direct interaction between LLMs and our 4D scene through language. Furthermore, by leveraging InternVideo2 features, which can be rendered and decoded from any viewpoint in 3D space over time, we naturally lift the video chatbot LLM from 2D to 4D, enabling free-form language interaction with the AI agent within 4D space.

Beyond direct language interaction, LLM agents can operate in the loop for manipulation tasks by interpreting user prompts, automatically optimizing hyperparameters, and iteratively refining results for downstream applications. For example, given a scene-editing prompt like "Delete the dog," the agent parses the operation command ("delete") and the target object ("dog"), then generates an editing configuration with relevant hyperparameters, such as the softmax threshold for matching Gaussians to the target object. It tests various thresholds, evaluates the quality of rendered image samples, and selects the optimal configuration. This is then consistently applied across the entire 4D scene, enabling efficient, intelligent editing with minimal user input.

This perception-reasoning-action loop empowers the LLM-driven 4D agent to interpret, execute, and refine complex scene manipulations, making it a powerful tool for adaptive 4D scene editing and interaction. The system is particularly valuable for applications like interactive VR content creation and editing, where dynamic scene understanding and precise, context-aware modifications are essential.

4.4 Experiments

In this section, we present experimental results and analyses of the capabilities of our proposed Feature4X framework. Based on the type of desired output, we categorize tasks into

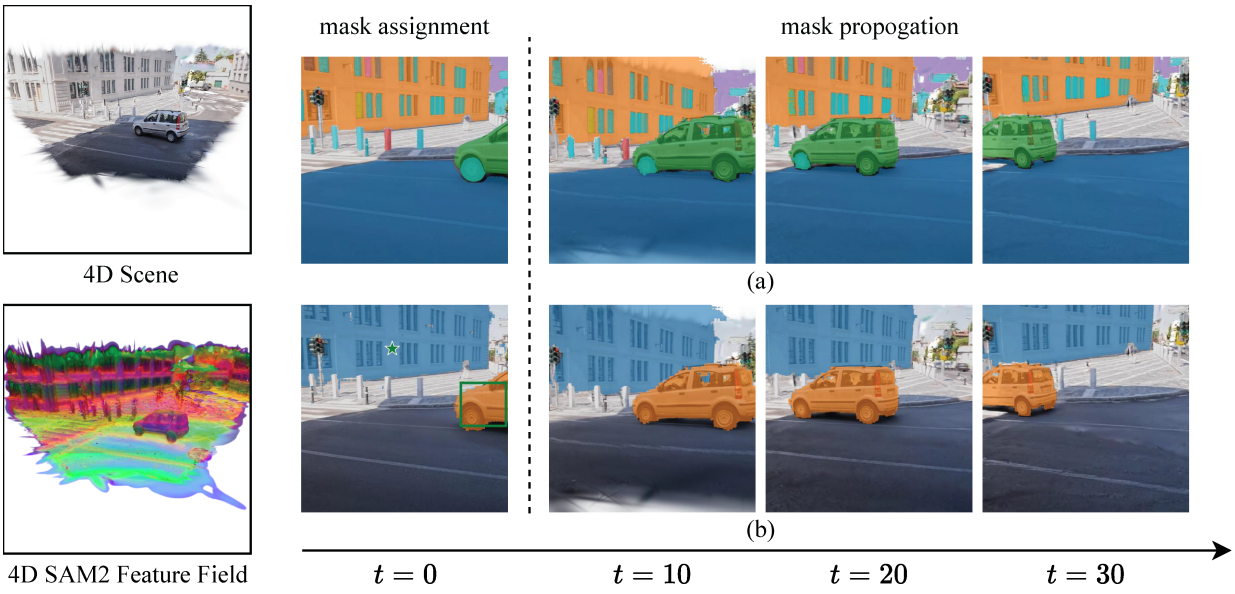


Figure 4.3: **Segment Anything in Dynamic 4D Scenes with SAM2 Feature Field.** For any rendered novel view video, we support: (a) Promptless segmentation (segment everything): when no user prompt is provided, segmentation masks are automatically assigned at the first frame ($t = 0$) and then propagated across all frames. (b) Promptable segmentation (segment anything): the user can segment any object—static or dynamic—at any timestep using a point or box prompt, and the corresponding mask is robustly tracked and propagated through subsequent frames.

2D (segmentation; see Secs. 4.4.1 and 4.4.2), 3D (scene editing; see Sec. 4.4.3), and 4D (spatiotemporal VQA; see Sec. 4.4.4).

4.4.1 Scene Interaction with Segment Anything

Segment Anything Model 2 (SAM2) [RGH24] is an advanced segmentation model for promptable visual segmentation across images and videos, supporting various input types such as points and bounding boxes for interactive and precise results. In our approach, we extract per-frame features using SAM2's image encoder and lift them into a 4D feature field, enabling

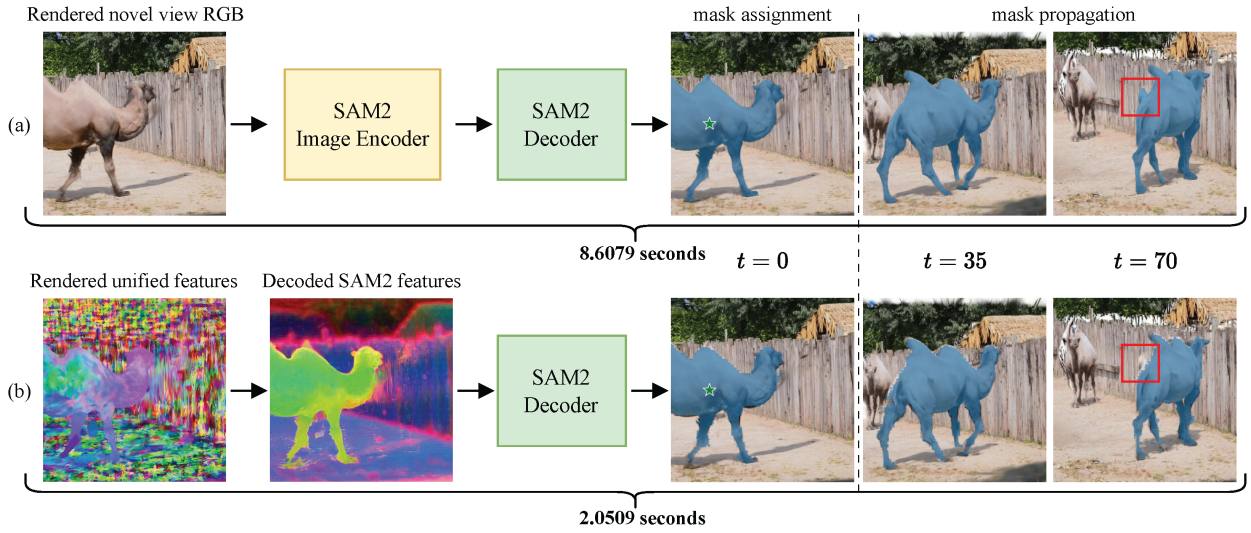


Figure 4.4: **Baseline Comparison on SAM2 Inference.** We compare segmentation quality and inference speed between (a) the naive RGB-based approach and (b) our feature-based method. Ours achieves comparable segmentation, accurately tracking the object over time, and avoids RGB artifacts (red box region at $t = 70$), while reducing inference time to about $4\times$ speed-up.

direct decoding from novel-view rendered feature maps. This allows segmentation that is inherently aware of both viewpoint changes and temporal dynamics, as shown in Fig. 4.3. Masks are seamlessly propagated through space and time, regardless of object motion or camera trajectory, supporting both automatic and user-guided segmentation with temporal coherence.

Notably, our approach bypasses the need to first render RGB videos and then apply the full SAM2 pipeline (see (a) in Fig. 4.4). We observe that artifacts in novel view RGB rendering can mislead SAM2 during encoding, resulting in ambiguous and inaccurate segmentation despite smoother masks. In contrast, our method achieves faster and more robust segmentation in 4D scenes by operating directly in feature space.

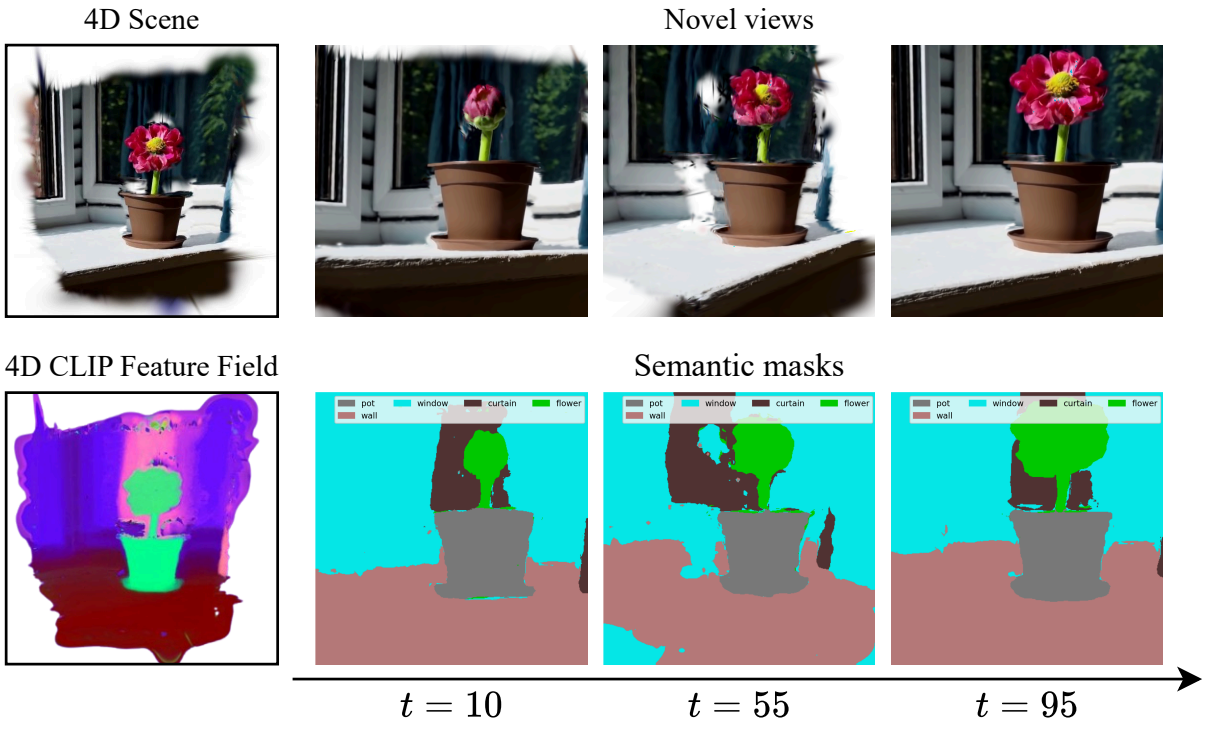


Figure 4.5: **Semantic 4D Scene Understanding with CLIP Feature Field.** By lifting CLIP-LSeg [LWB22a] features into a 4D feature field, we enable pixel-level semantic segmentation from any view at any timestep. This allows robust 4D scene understanding, even as object appearances change over time—for example, accurately identifying a blooming flower from bud to full bloom across views.

4.4.2 Scene Understanding with Semantics

To achieve pixel-level semantic segmentation across any novel view and timestep, we extend the capabilities of CLIP-LSeg [LWB22a] by lifting its features into a 4D CLIP feature field. CLIP-LSeg is a language-driven semantic segmentation model that leverages the Contrastive Language-Image Pre-training (CLIP) framework to align textual descriptions with visual content, enabling zero-shot segmentation by interpreting descriptive input labels. By lifting these 2D features into a 4D representation, our approach captures both spatial and tem-

Method	PSNR↑	mIoU↑	accuracy↑	Size (MB)↓
MoSca	25.166	-	-	**67.726**
MoSca + Feature 3DGS	25.191	0.506	**0.881**	593.907
Ours (single CLIP head)	25.186	**0.510**	0.880	95.294
Ours (full model)	**25.197**	0.503	0.876	95.457

Table 4.1: **Semantic segmentation on the Nvidia dataset [LGM23].** Our method achieves comparable radiance reconstruction (PSNR) and segmentation performance, while significantly reducing memory usage compared to the baselines.

poral information, facilitating consistent semantic understanding in dynamic scenes, even when objects undergo significant appearance changes over time. As depicted in Fig. 4.5, our model accurately identifies a flower throughout its blooming process—from bud to full bloom—across various viewpoints and timesteps. This demonstrates the model's proficiency in maintaining consistent semantic understanding despite substantial visual transformations in dynamic scenes.

Tab. 4.1 presents a quantitative comparison between our method and baselines on the Nvidia dataset [LGM23]. "MoSca [LWH24a]" is the baseline 4D reconstruction method, while "MoSca + Feature 3DGS [ZCJ24]" is a naive, non-compact baseline using RGB + feature rasterization. Our models—using a 32-dimension latent feature and 512-dimension CLIP-LSeg [LWB22a] as the target—maintain comparable PSNR to MoSca, showing no loss in reconstruction quality. The single CLIP head variant achieves the highest mIoU, demonstrating effective feature distillation. Although the naive baseline yields slightly higher accuracy, our models are about 6.2× more space-efficient, with the full model delivering competitive performance while remaining compact and general-purpose.



Figure 4.6: **Scene Editing with AI Agent.** Given user prompts, our GPT-powered agent interprets editing intent and autonomously performs scene edits via our 4D CLIP feature field. Examples include both geometric (e.g., "extract" and "delete") and appearance (e.g., "change color") editing in 3D space. While results may not be perfect due to imperfect fine-grained feature alignment and non-optimal editing parameter tuning, the agent adaptively refines parameters and applies edits consistently across views and time—greatly reducing the need for manual tuning—and demonstrates robust, interactive 4D scene manipulation.

4.4.3 Scene Editing with Language and AI Agent

We employ an LLM agent (GPT-4o [AAA23]) to interpret natural language prompts, optimize editing parameters, execute precise queries, and iteratively refine results for intelligent 3D scene editing. The agent begins by parsing the user's high-level prompt to identify the editing action—such as "extract," "delete," or "change color"—and the target object (e.g., "cow"), as shown in Fig. 4.6. Based on the task, it generates candidate configurations with varying parameters relevant to the prompt. Specifically, we first compute the probability of each 3D Gaussian corresponding to a language description l of the target object as:

$$\mathbf{p}(l \mid j) = \frac{e^{s}}{\sum_{s_i \in \mathcal{L}} e^{s_i}}, \quad s = \frac{q(l) \cdot f_j}{\|q(l)\| \|f_j\|}, \tag{4.5}$$

where $q(l)$ is the text feature, f_j is the CLIP feature of Gaussian j, and $\mathcal{L}$ is the label set of all possible scene objects. The agent filters Gaussians using various thresholds based on

the softmax probability $\mathbf{p}(l \mid j)$, applies the edit configuration, and evaluates the results via rendered image samples. It selects the parameter setting that best aligns with the intended edit and applies it consistently across all video frames, ensuring coherent editing throughout the 4D scene.

This GPT-powered 4D Agentic AI system interprets, executes, and refines complex scene manipulations, serving as an intelligent assistant for adaptive editing and interaction. By autonomously exploring, selecting, and tuning parameters—tasks typically handled by human editors—the agent enhances responsiveness and consistency. GPT acts as a self-refining decision-maker, illustrating the potential of agentic AI in creative and technical workflows where iterative optimization and goal-directed behavior significantly streamline and elevate editing tasks.

4.4.4 Scene Reasoning with 4D Chatbot Agent

While large vision-language models (VLMs) have demonstrated impressive performance in visual question answering (VQA) tasks, their application has predominantly been confined to 2D visual modalities, such as images and videos. To our knowledge, we are the first to extend the capabilities of a VLM (Video-LLM) into 4D—encompassing three spatial dimensions plus time—by integrating VLM knowledge into our 4D feature field. This advancement enables a more comprehensive understanding and interaction with dynamic 4D scenes, capturing both spatial and temporal nuances beyond the static 2D perspective.

We utilize InternVideo2 [WLL24, LHW23]–a state-of-the-art video foundation model–to extract comprehensive video features. Specifically, we sample the input video into several clips, each comprising 8 frames sampled at equal intervals to ensure enough GPU memory available during the whole process. These clips are processed through InternVideo2's vision transformer encoder to obtain segment-level features and then aggregated together. Additionally, for each segment, InternVideo2 generates a class token representing high-level semantic information. We compute the average of these class tokens to form a global video

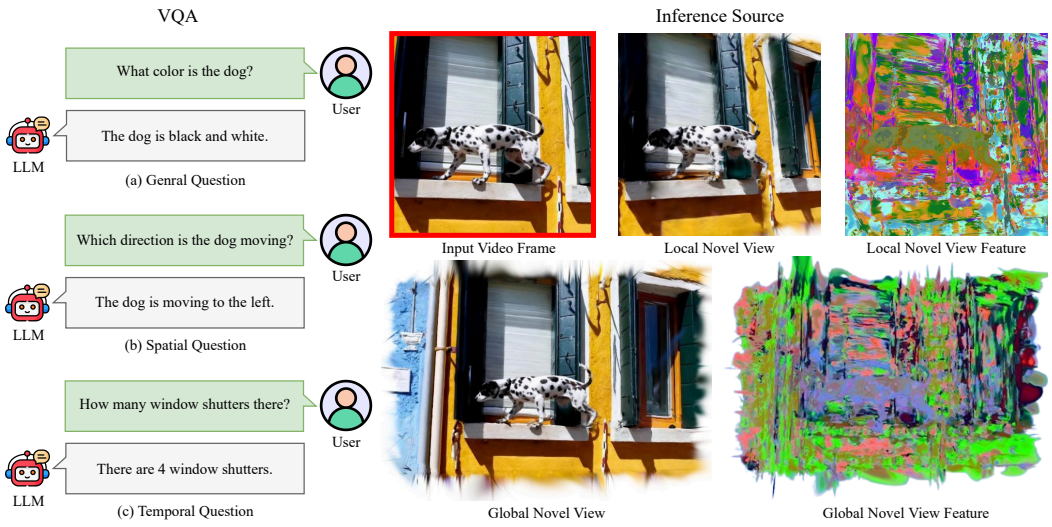


Figure 4.7: **VQA with Chatbot Agent.** (Left) Our model supports free-form VQA across diverse question types—general, spatial, and temporal—by distilling InternVideo2 features. (Right) At each timestep, we reconstruct both a 4D radiance field and a 4D feature field, providing more inference sources beyond the input video frame—including local (moving camera) and global (zoomed-out) novel views and their corresponding feature maps—thereby supporting VQA in 4D and enhancing the model's spatiotemporal reasoning capabilities.

representation. This class feature is then concatenated with the aggregated segment features, effectively integrating temporal information which is necessary for dynamic scene reasoning.

In our system, free-form visual question answering (VQA) is supported through a language model [LHW23], enabling users to interact with a chatbot agent using natural language while exploring 4D dynamic scenes, as illustrated in Fig. 4.7. Unlike prior approaches that rely solely on 2D monocular video—limited to a single view at a single time step—our method leverages the reconstructed 4D radiance and feature fields to support richer inference sources, including both local novel views (via free camera movement) and global novel views (via zoom-out for broader context). In the example shown in Fig. 4.7, we observe that InternVideo2, when operating on the input monocular video, fails to correctly answer spatial and temporal questions—even when the answers are obvious to humans—while the

Inference Source	General	Spatial	Temporal	Time (s)
Input Video View	✓	×	×	33.65
Local Novel View	✓	✓	×	33.86
Local Novel Feature	✓	×	✓	11.99
Global Novel View	✓	×	✓	33.37
Global Novel Feature	✓	✓	✓	12.24

Table 4.2: **VQA performance across different inference sources with the scene shown in Fig. 4.7.** Feature-based inference supports all question types with lower latency, while view-based methods are limited in spatial and temporal reasoning.

same model, using our global novel view feature, succeeds (see Tab. 4.2). This is because monocular video lacks the spatial coverage and contextual cues necessary for spatiotemporal reasoning. In contrast, our framework allows rendering of novel views and their corresponding feature maps at arbitrary time steps, injecting explicit signals that enhance reasoning. To evaluate this more systematically, we construct 400 objective questions across 50 scenes in the DAVIS dataset [PPM16], covering spatial only (e.g., "Is the person facing left?"), temporal only (e.g., "How many times does the dancer spin?"), and spatiotemporal (e.g., "Which direction is the duck moving?") categories. As shown in Tab. 4.3, our method improves VQA accuracy by enriching scene understanding through 4D feature field reconstruction, while also accelerating inference by eliminating the need for video encoding.

4.5 Conclusion

We have introduced Feature4X, a general and scalable framework that bridges casually captured monocular videos to interactive 4D agentic AI systems. By distilling diverse 2D vision foundation model features into a unified, dynamic 4D Gaussian feature field, our method

Inference Source	Spatial Acc.	Temporal Acc.	Overall Acc.	Time (s)
Input Video View	48.50	49.84	49.06	10.02
Local Novel View	49.75	51.13	50.31	9.78
Local Novel Feature	54.50	54.37	56.29	**2.81**
Global Novel View	47.00	49.19	47.48	10.06
Global Novel Feature	**58.75**	**58.25**	**61.32**	3.42

Table 4.3: **Spatiotemporal VQA on DAVIS dataset [PPM16].** Compared to 2D video inference, our 4D feature space inference (Global Novel Feature) enhances the Video-LLM's spatiotemporal reasoning while achieving approximately 3× faster inference.

supports a wide range of tasks across 2D (segmentation), 3D (scene editing), and 4D (spatiotemporal VQA). Through a compact and versatile representation, Feature4X achieves efficient training and inference while maintaining high-quality appearance reconstruction and robust semantic understanding. Furthermore, by integrating LLMs for language-based interaction and autonomous task refinement, our system elevates traditional perception tasks into a perception-reasoning-action loop, enabling intelligent and adaptive manipulation of dynamic 4D scenes. We hope Feature4X opens new avenues in 4D vision and agentic AI research by facilitating immersive, multimodal interaction with dynamic visual content.

CHAPTER 5

Reasoning - Towards Spatiotemporal Intelligence

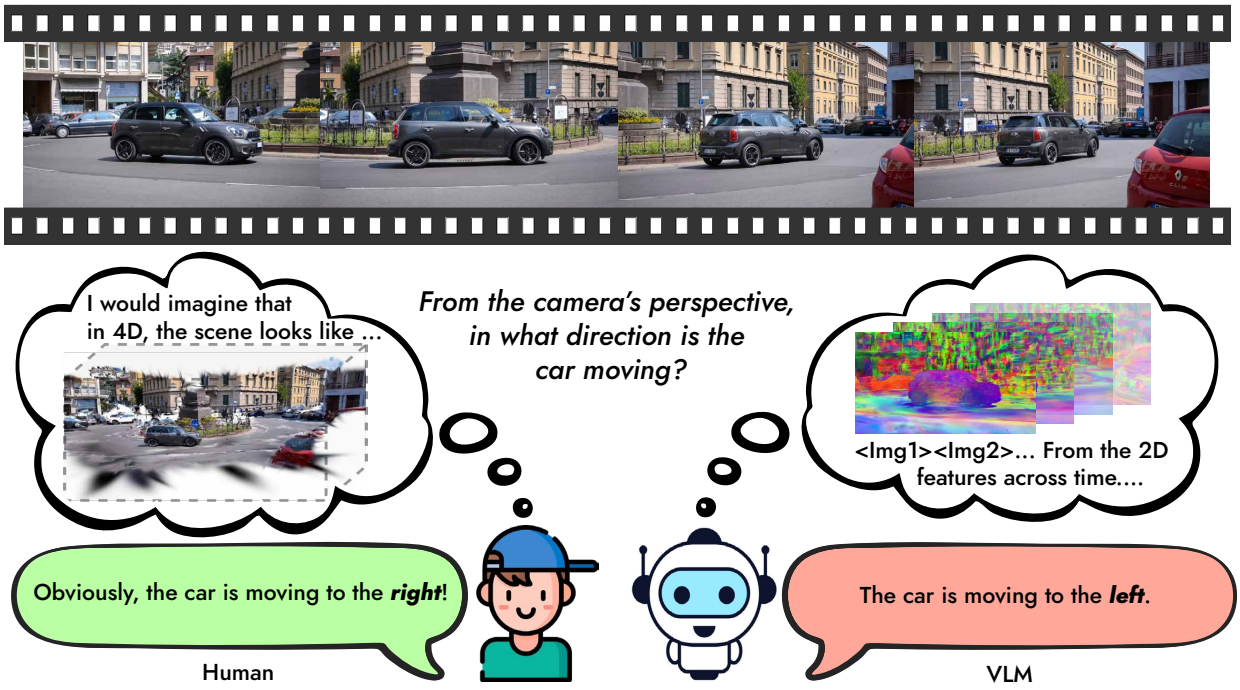


Figure 5.1: **Spatiotemporal (4D) Awareness**. Humans intuitively reason in 4D (3D space + time), effortlessly reconstructing the dynamic spatial trajectory of moving objects from any perspective. In contrast, current Vision Language Models (VLMs) typically rely on aggregating 2D visual features across time, leading to incorrect predictions when motion understanding and interpretation requires deeper spatiotemporal reasoning.

5.1 Introduction

Humans possess an innate ability to perceive, track, and interpret motion as well as spatial and temporal changes, enabling rich interpretations of complex dynamic events from both egocentric and allocentric perspectives [DMN16, MU81, Bur06]. When observing an object move, we can inherently process any changes such as lateral shifts, rotational directions, and periodic or repeated actions unfolding along a specific trajectory [Bur06]. These sophisticated perceptual abilities are a product of our spatiotemporal cognition [FF84], and form an essential foundation that allows us to comprehend and reason about physical phenomena, object interactions, and causal relationships within our environment [Les84, Joh73].

Vision language models (VLMs), which also have the potential to perceive motion and spatiotemporal changes in videos, represent a prominent class of methods aimed at emulating or surpassing human capabilities in integrated visual and linguistic reasoning [LBD89, DBK21]. While previous work on VLMs has primarily focused on static visual understanding—through large-scale training on paired language and image data [RKH21]—or explored video understanding tasks such as captioning [LBP19] and scene understanding [CGY22], we find that their strong performance in these tasks does not naturally translate into robust spatiotemporal reasoning. This limitation is particularly striking given that state-of-the-art VLMs are typically trained on datasets comprising hundreds of billions of tokens [LYZ24b]. In contrast, human infants naturally develop robust spatiotemporal cognition within the first few months of life [SK07].

Another key challenge that limits VLM performance on spatiotemporal tasks is the need to implicitly or explicitly reconstruct a four-dimensional (4D) representation—3D space + time—of dynamic scenes, and subsequently reason over this reconstruction [WMW24]. As illustrated in Fig. 5.1, the car is advancing forwards and turning to the left in its own frame of reference. However, from the camera's perspective, its motion appears as a combination of heading to the right and receding into the distance despite the car being in the center of the

frame due to camera rotation. Human observers can seamlessly disentangle these complex dynamics, accurately interpreting trajectories by synthesizing diverse visual cues including camera rotation compensation, stationary scene landmarks, prior knowledge of 3D and 4D environmental structures, and perspective projections [MU81, Bur06, Les84, FF84]. The inability of current VLMs to integrate spatial, temporal, and semantic cues in a human-like manner stems from their fundamentally different encoding paradigms for visual and language information. This discrepancy highlights a significant gap between human and machine spatiotemporal understanding, suggesting that future VLMs may benefit from insights in cognitive science and neuroscience to develop more advanced mechanisms for perceiving, integrating, reconstructing, and reasoning over dynamic scenes.

In order to effectively characterize and challenge the existing spatiotemporal reasoning abilities of VLMs, we introduce `VLM4D`, a rigorous benchmark specifically designed to probe the spatiotemporal grounding capabilities of current vision language models. Through this contribution, we aim to catalyze research that addresses the critical gap in spatiotemporal understanding and reasoning within VLMs and provide a foundational analysis highlighting key deficiencies in existing models.

We summarize our main **contributions** as follows:

1. We propose the first benchmark `VLM4D` explicitly designed to rigorously evaluate the spatiotemporal (4D) reasoning capabilities of Vision Language Models (VLMs).

2. We introduce a novel, meticulously curated dataset consisting of diverse real-world and synthetic video sequences paired with carefully crafted spatiotemporal question-answer (QA) annotations.

3. We analyze the critical limitations of contemporary VLMs in spatiotemporal reasoning and conduct experiments exploring potential solutions, highlighting fundamental challenges and outlining clear directions for impactful future research.

5.2 Related Work

Spatiotemporal Understanding in Vision Language Models Early methods for video understanding before the advent of large vision language models (VLMs) leveraged trajectory-based representations and separate modeling of spatial and motion cues for tasks like action and motion recognition [Lap05, WKS11, SZ14, CBZ23]. Recently, VLMs have evolved rapidly by fully leveraging the significant achievements of Large Language Models (LLMs) [BMR20, DCL19, WBZ21, BBC23, TLI23, RNS18] and large-scale visual instruction tuning datasets [LLW23, ZCS23, DLL23]. While VLMs [GLZ23, LLW23, ZCS23, AAA25, HLG24, LZG24, TGL24, WBT24] exhibit transformative potential for applications such as embodied AI [SGB24, DXS23, KPK24], robotics [WZD24, PYH25], scene generation [OTK24, ZFX24, LLL25], and world modeling [LYZ24a, ZWL24], most existing methods remain constrained to static images, focusing narrowly on spatial understanding [CYF24, CXK24, RSP24, FZL26, ZZL26] while overlooking the dynamic temporal dimension inherent in real-world interactions. To bridge this gap, emerging research [LHW23, ZLB23, CLZ24, MRK23, ZLC25] has begun exploring video modality integration, aiming to equip VLMs with spatial-temporal awareness critical for tasks like video comprehension, where both contextual details and motion dynamics are essential. For example, VideoLLM-MoD [WCL24] proposes to address the efficiency issue when processing long-term video by mixture-of-depths. [YZL24] introduces VideoRefer to enhance the finer-level (like object-level) spatial-temporal video understanding of VLMs. Grounded-VideoLLM [WXC24] also targets for fine-grained video understanding through incorporating an additional temporal stream. In this work, we aim to rigorously evaluate the 4D spatial-temporal reasoning capabilities of state-of-the-art VLMs, probing how and to what extent these models internalize spatial intelligence and temporal dependencies.

VLM Benchmarks Following the development trends of VLMs, benchmarking VLMs shares the similar trajectory by first evaluating vision QA on static images [LGG24, LDZ24b,

HFZ24, YNZ24], to align with models' early focus on 2D understanding. As VLMs evolved to tackle dynamic scenarios, benchmarks expanded to evaluate general-purpose video comprehension tasks that probe temporal coherence and event understanding [NZX23, KNH24, LHW24, FDL24, LWH24b]. Notably, MMVU [ZXZ25] further proposes a knowledge-intensive benchmark to assess the expert-level reasoning ability of current video-based large models. However, while these works assess perception and semantic understanding, they largely overlook the explicit evaluation of spatial-temporal awareness, a core capability for real-world applications requiring 4D (3D space + time) reasoning. Recent efforts like [YYG24] pioneer benchmarks for 3D visual-spatial intelligence but restrict evaluation to static 3D scene, neglecting the interplay of object motion and temporal dynamics intrinsic to videos. In this work, we introduce `VLM4D`, the first benchmark designed to holistically evaluate the 4D intelligence in VLMs, unifying spatial understanding, temporal continuity, and motion reasoning. By curating tasks that demand precise analysis of dynamic interactions (e.g., direction prediction, perspective anticipation, and motion reasoning), `VLM4D` exposes critical gaps in current models' ability to internalize spatiotemporal relationships. Our work not only advances the granularity of VLM evaluation but also shares insights and potential solutions to improve the model performance.

5.3 The `VLM4D` Benchmark

We introduce `VLM4D`, the first benchmark specifically designed to test the spatiotemporal reasoning ability of VLMs. `VLM4D` consists of 1000 videos paired with over 1800 question-answer pairs, each carefully designed to assess both spatial and temporal understanding jointly. The majority of these videos are sourced from datasets with rich spatiotemporal characteristics, thus ensuring a diverse range of motion-related scenarios. We also augment the dataset with synthetic videos generated by a world-foundation model, Cosmos [AAB25], that has been modified using techniques introduced in [HWY24] to obtain more accurate

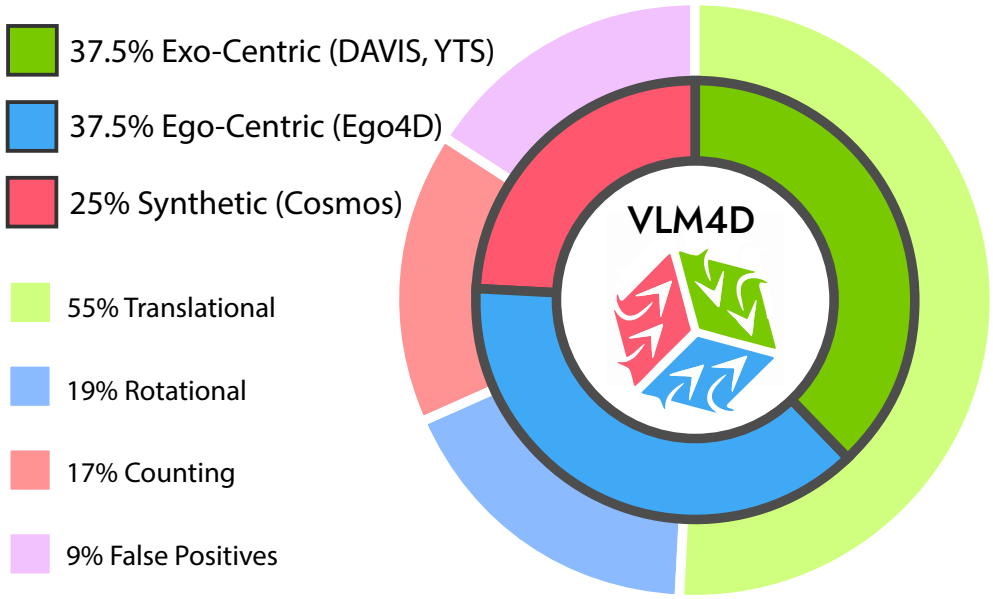


Figure 5.2: **Distribution of Dataset Sources and Annotations.** Overview of the dataset composition, illustrating the proportions of real third-person (exo-centric) videos (DAVIS [PPC17], YouTube-VOS [XYF18]), real first-person (ego-centric) videos (Ego4D [GWB22]), and synthetic videos (Cosmos [AAB25]). The real video data is further categorized by annotation types, including translational, rotational, action, counting, and false positive queries (targeting nonexistent events to assess critical reasoning).

correspondence between motion-oriented prompts and the resulting generated video. Fig. 5.2 illustrates the composition of our dataset.

5.3.1 Benchmark Construction

Unlike prior work that often relies heavily on LLMs and VLMs to generate first iterations of benchmarks and datasets [CWL24] followed by human quality control, we found that existing VLMs and automated methods showed significant limitations in terms of reliability and quality. This shortcoming necessitated direct human annotations that were then followed by augmentation by LLMs to ensure a high-quality benchmark. An overview of the benchmark

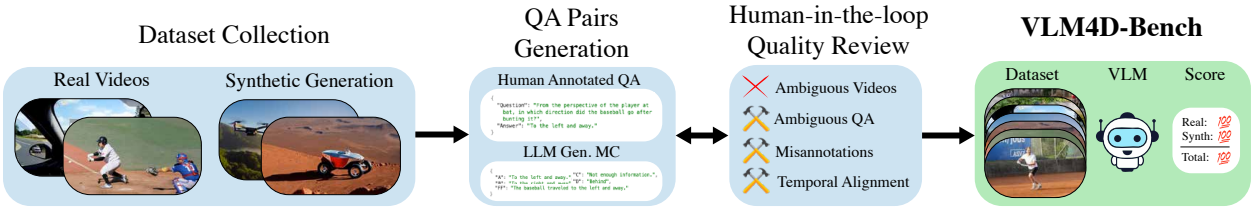


Figure 5.3: **Dataset Generation and Annotation Pipeline.** Our dataset was constructed by collecting real videos and generating synthetic data, followed by human-in-the-loop quality reviews to address ambiguous videos and annotations. After temporal alignment and quality assurance, human-annotated questions and ground-truth answers were created, complemented by multiple-choice (MC) answers generated by large language models (LLMs). The final dataset includes real and synthetic video data with comprehensive VLM scoring metrics.

curation pipeline is shown in Fig. 5.3.

Real Video Data Collection Real-world videos were sourced from datasets with rich spatiotemporal characteristics that ensured diverse motion and perspective variations. For egocentric data, we relied mainly on the Ego4D dataset [GWB22], while most exocentric data points were collected from the DAVIS [PPC17] and YouTube-VOS [XYF18] datasets. To minimize confounders and to focus attention of VLM abilities to only spatiotemporal reasoning, we preprocessed the videos by temporally segmenting and centering them around the most relevant action, thus resulting in videos with an average duration of 3-8 seconds. This ensures that the key event described in the question is clear and reduces ambiguities or confounders that would reduce VLM accuracy.

Synthetic Video Data Generation The rapid advancement of video generation techniques has led to their widespread application in diverse domains like robotics and immersive entertainment [AAB25, BPH24, WWA25, LPY25, WGP25], making 4D reasoning on synthetic content essential for VLMs and underscoring the need to include such videos in our

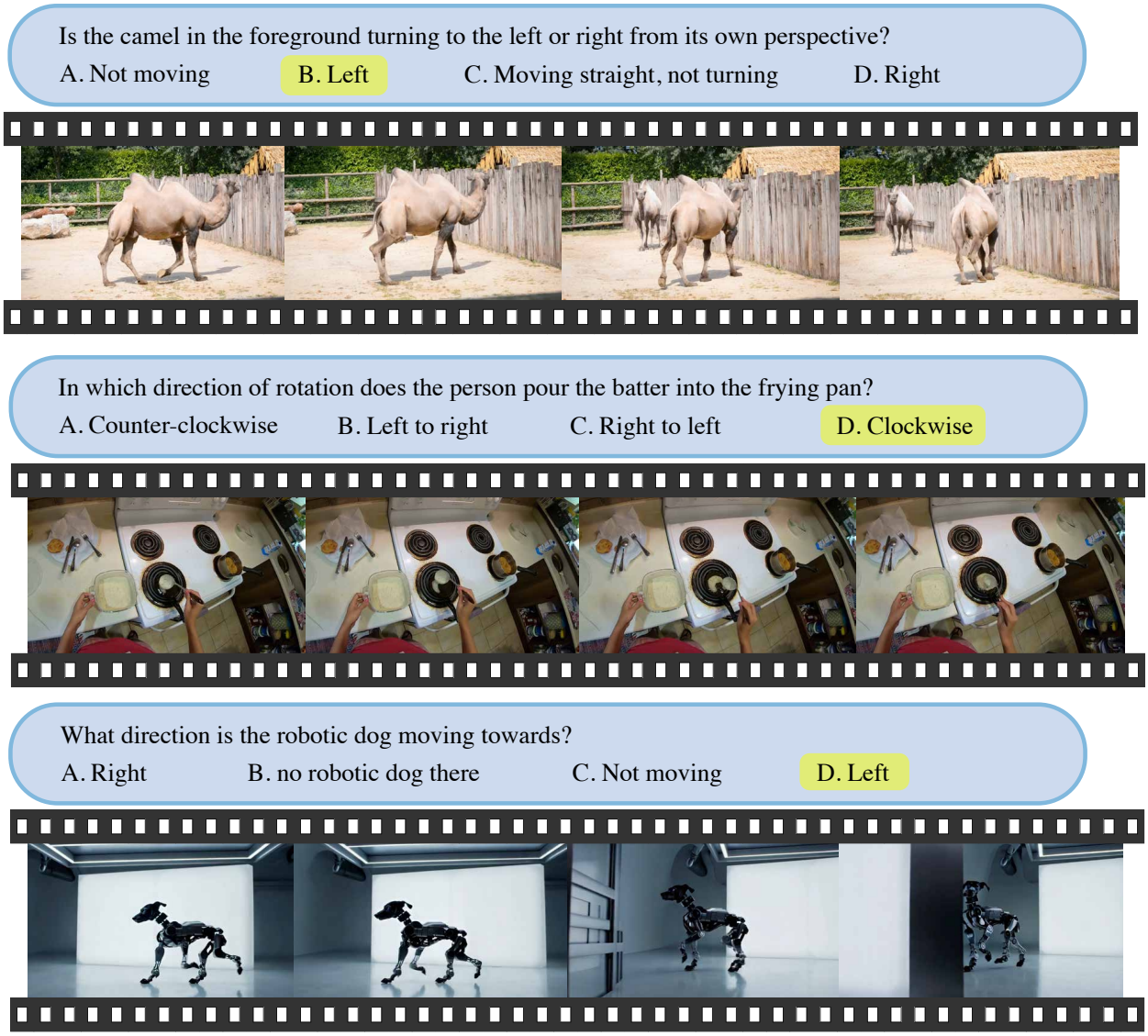


Figure 5.4: **Qualitative Examples of Dataset Annotations.** (Top) A third-person (exo-centric) video with translational annotations ("camel turning left from its perspective"). (Middle) A first-person (ego-centric) video with a rotational question ("clockwise rotation of ladle"). (Bottom) A synthetic scene with motion recognition "robotic dog moving left").

benchmark. For synthetic video generation, we use Cosmos [AAB25] as our video generation backbone. To ensure that the generated videos align with the intended object moving directions, we incorporate input bounding boxes as additional spatial guidance. Specifically,

we follow the approach introduced in [HWY24] modifying the diffusion forward steps to enforce object localization constraints at each timestep, ensuring consistency between the generated object direction and the user-specified trajectory. The average duration of generated synthetic videos is 5 seconds. To maintain high-quality outputs, we perform a manual verification step after generation, filtering out low-quality videos and retaining only those that accurately match the specified directions. Once a video is generated, we use an LLM (GPT-4o) to create two types of evaluation questions: Directional questions, derived from the textual prompt used to generate the video; and False Positive (FP) questions (constituting 10%, matching the ratio in the real dataset), which query non-existent objects in the scene. Both question types follow the format: "What direction is the ⟨Object Name⟩ moving?", where the model must select one of four possible answers: "left", "right", "not moving", or "no ⟨Object Name⟩ there". A final manual review is conducted to filter out or revise ambiguous questions, ensuring the quality of the questions and ground-truth answers.

QA Generation and Quality Control Question-answer pairs are primarily constructed through human annotations. The question answer pairs are then supplemented with alternative answers by an LLM (GPT-4o) for multiple choice (MC) questions. To ensure high-quality annotations, we applied a three-round, multi-person cross-checked verification process, during which ambiguous videos were filtered out and vague, misleading, or incorrect QA pairs were refined to improve spatial and temporal alignment between language and visual content. Fig. 5.4 showcases some qualitative examples of annotations for different types of videos.

Assessing Human Performance To establish a human performance baseline on our benchmark, we conducted an evaluation in which participants independently answered 100 randomly sampled questions from the dataset. The accuracy of human responses was then aggregated to approximate the performance of human spatiotemporal reasoning on the dataset.

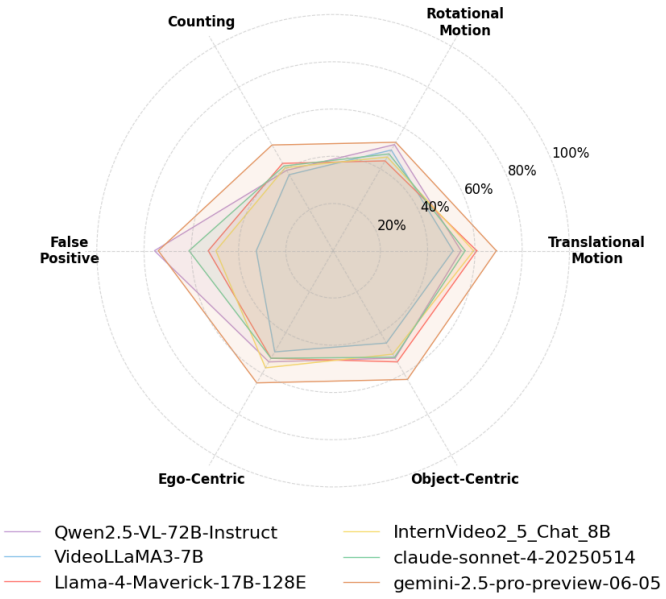


Figure 5.5: **Comparison of accuracy across types of spatiotemporal questions.** Model accuracy is shown only for the six top-performing VLMs.

5.3.2 Categorizing Spatiotemporal Performance

To systematically evaluate spatiotemporal reasoning capabilities, we first categorize videos into two primary groups: egocentric (first-person) videos and exocentric (third-person) videos. Egocentric videos are sourced from the Ego4D [GWB22] dataset, where scenes are captured from a head-mounted camera, thus offering dynamic video data that is inherently coupled with the individual's actions. Exocentric videos encompass a diverse range of recorded scenes, from sports footage to everyday scenes. Beyond this categorization, we also evaluate spatiotemporal performance across four dimensions: translational movement (TM), rotational movement (RM), spatiotemporal counting (STM), and false positives (FP), with their proportions shown in Fig. 5.2. Translational movement assesses a model's ability to track linear motion within scenes, while rotational movement assesses the understanding of changes in orientation and perspective shifts over time. Spatiotemporal counting extends these core motion-based tasks by requiring a more complex reasoning strategy to determine

the number of objects performing a translation or rotational movement. Lastly, the false positives category evaluates the model's critical thinking [FLS25] in determining whether an object or event actually occurred within the spatiotemporal context. By structuring the benchmark along these axes, we aim for a comprehensive framework for assessing spatiotemporal reasoning (Fig. 5.5).

5.4 Evaluation of VLM4D Benchmark

5.4.1 Evaluation Setup

Benchmark Models We evaluate 23 most recently released VLMs thus covering a wide range of model sizes, architectures, and training methodologies. For closed-source VLMs, we evaluate GPT-4o [Ope24], Gemini 2.5 Pro [Gem25], Claude Sonnet 4 [Ant25], and Grok-2-Vision [xAI24]. For open-source models, we include Llama 4 [Met25], DeepSeek-VL [LLZ24], Qwen2.5-VL [YYZ24], Qwen2-VL [WBT24], InternVL2.5 [CWC24], Aria [LLW24], InternVideo 2.5 [WLY25], InternVideo2 [WLL24], Phi-4-multimodal [AAB24], Phi-3.5-vision [AAA24], Pixtral [AAH24], VideoLLama3 [ZLC25], Llava-One-Vision [LZG24], Llava-NeXT-Video [ZLL24]. When available, we evaluate different sizes for each model, resulting in models ranging from 2 to 72 billion parameters.

Evaluation Settings The evaluations were performed in a zero-shot setting with video or a set of sampled frames of video, followed by the prompt forming the input. For each model, we evaluate on two different inference settings. In the first setting, the model prompted to directly output (DO) the answer immediately without any reasoning, and in the second evaluation setting, the model is directed to create intermediate reasoning steps, chain-of-thought (CoT) [WWS22], before inferring the final answer.

Metrics Following prior work [YYG24] and given the nature of our target task, we adopt multiple-choice questions (MCQs) for evaluation, using accuracy as the primary metric. For the two inference settings described earlier, we employ LLM-as-Judge [ZXZ25] to assess the outputs of VLMs. We opt for this method instead of string or template matching, as VLMs—especially under chain of thought (CoT) prompting—often generate all possible answer choices during reasoning, with varying frequencies and slight formatting differences. In some cases, the final answer may even contradict the reasoning. To better evaluate whether the model truly understands the video, we prompt two advanced LLMs (GPT-o3 and o4-mini) to grade based on the full CoT reasoning, not just the final answer. We then perform a cross-check between their judgments and manually resolve any disagreements. The evaluation results are reported in Tab. 5.1.

5.4.2 Benchmark Results

Our comprehensive evaluation on the `VLM4D` benchmark, detailed in Tab. 5.1, systematically assesses the spatiotemporal reasoning capabilities of modern Vision-Language Models (VLMs). The results highlight a clear hierarchy, with proprietary models demonstrating superior performance over their open-source counterparts. Google's Gemini-2.5-Pro emerges as the top-performing model with an overall accuracy of 62.0%, followed by OpenAI's GPT-4o at 57.5%. Within the open-source domain, image-based models like Meta's Llama-4-Scout-17B (54.1%) and video-supported models like Alibaba's Qwen2.5-VL-72B (53.0%) demonstrate highly competitive results, even outperforming some proprietary counterparts. Despite these achievements, a significant performance gap persists when compared to human accuracy (98.8%), underscoring that sophisticated 4D awareness remains a formidable challenge for AI. Notably, performance varies significantly across different categories, such as real versus synthetic data and ego-centric versus exo-centric perspectives, indicating that current models lack generalized spatiotemporal understanding.

5.5 Analysis: Why VLMs Don't Work Well?

5.5.1 Limited Spatiotemporal Cognition

Despite significant advances in VLMs, their abilities to understand and reason about motion, spatial relationships, and temporal coherence remains fundamentally underdeveloped [CLD24, ZWD24]. Chain of thought (CoT) [WWS22] is widely employed as a method to improve accuracy through step-by-step reasoning. We showcase a comparison between CoT and DO in Fig. 5.6. Overall, there is no indication of a large advantage of CoT over all evaluated models. Upon deeper exploration of the CoT reasoning of some models, we observe that the reasoning process was primarily flawed in the following ways: irrelevant information and arriving at conclusions that are inconsistent with the reasoning process. Larger models exhibited strategies that would be similar to how a human processes spatiotemporal information, but the resulting execution falls short of human performance. This demonstrates a disconnect between its visual and linguistic knowledge. We provide examples of this behavior in the supplementary material.

5.5.2 Deficiencies in Spatiotemporal Labeling

Another avenue of exploration we undertook is to understand the richness of spatiotemporal labels in popular supervised fine-tuning (SFT) VLM datasets. Typically, video captioning occurs at the 'scene' level, lacking fine-grained temporal, spatial, and object-level details. We performed an extensive analysis, encompassing over 2 million samples [CWL24, CHM24, LHW23, LWH23, ZLL24]. We performed this analysis through string-matching of spatiotemporal descriptors related to directionality, translational motion, rotation, and perspective shifts and provide the overall results in Fig. 5.7. We then performed a manual finegrained evaluation of the ShareGPT4Video dataset [CWL24] which we found had the highest density of spatiotemporal dataset. We found that from a sample of 100 labels that were detected as spatiotemporal, less than 10% of them were judged as accurate upon human evaluation. This

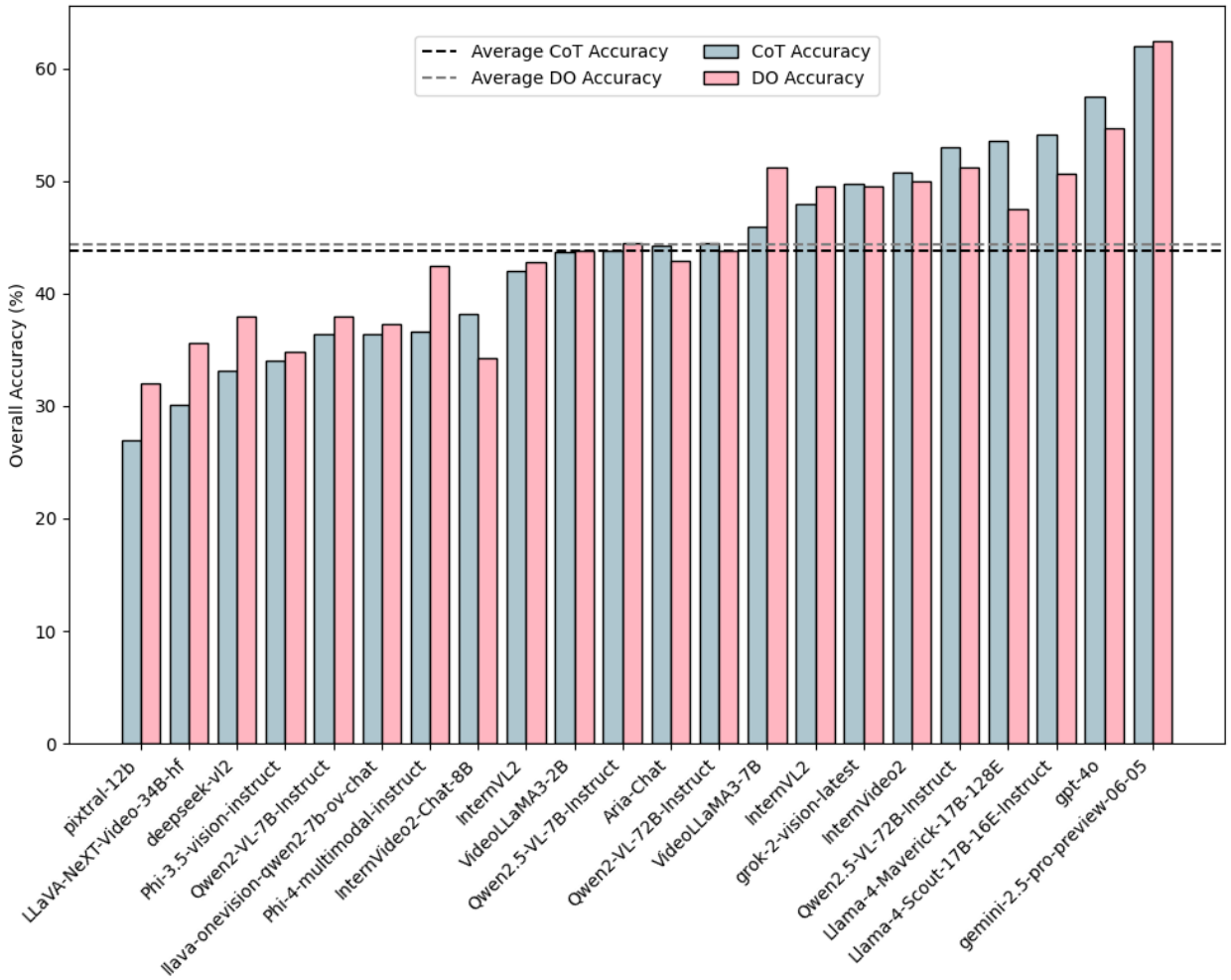


Figure 5.6: **Comparison of CoT and DO Accuracy Across Models.** Accuracy comparison between Chain-of-Thought (CoT) and Direct Output (DO) prompting across VLMs.

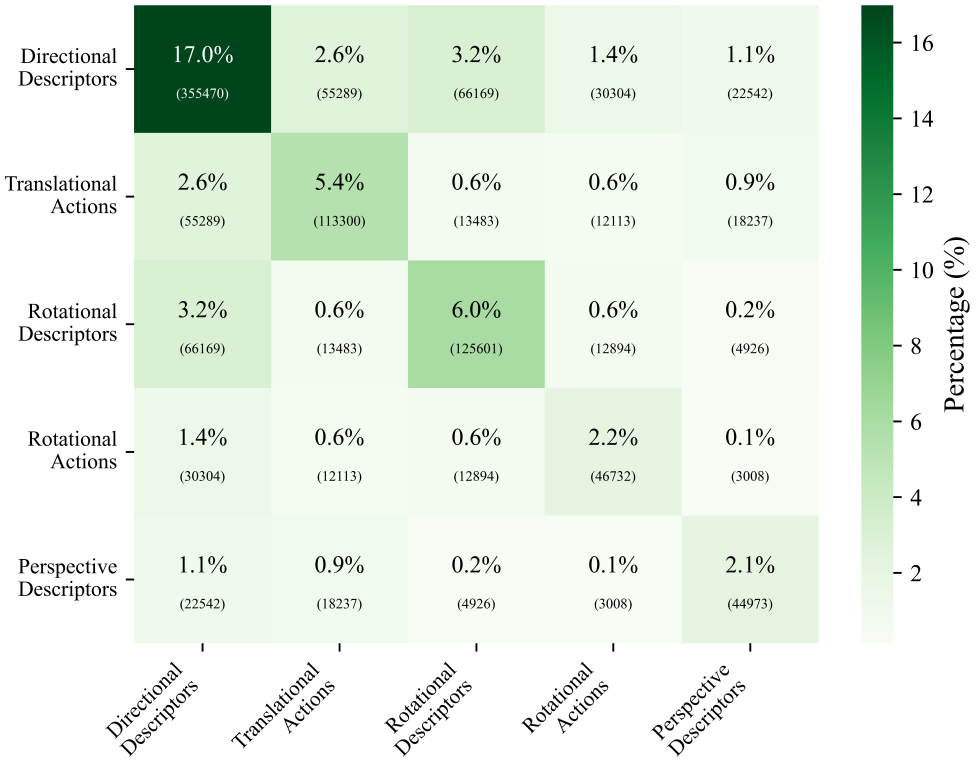


Figure 5.7: **Heatmap of Occurances of Spatial-Temporal Terms in popular video SFT datasets.**

result underscores the inadequacy of current dense captioning approaches, which frequently generate spatiotemporal descriptors without capturing precise motion dynamics. We provide more detailed analysis and explanations in the supplementary material.

5.6 Probing Future Solutions

To probe promising future solutions for enhancing spatiotemporal video understanding, we propose two approaches that address some of the shortcomings of current state-of-the-art VLMs: fine-tuning a VLM on data-rich in spatiotemporal actions and the other leveraging 4D reconstruction and feature fields jointly with a VLM. SFT refines the model's abilities by training on datasets that contain temporally and spatially rich actions and interactions.

By integrating structured visual representations and targeted fine-tuning, these approaches enhance video-language models' ability to interpret motion. The second method lifts the feature space of VLMs into a temporally coherent 4D feature field, providing structured scene representations that improve motion and spatial reasoning in the stage of decoding and inference.

Spatial-Temporal SFT We evaluate on a subset split of the real dataset by randomly splitting the real-world dataset into a training and testing split (80% / 20%) and we try settings using synthetic/real/both for training. We conducted the experiments using Qwen 2VL (7B) and Qwen 2.5VL (7B) through LLama-Factory [ZZZ24b], and compared the performance before and after supervised fine-tuning in Tab. 5.2. The results demonstrated an improvement in accuracy in spatiotemporal reasoning, suggesting that performance gains can be obtained through targeted training. However, the addition of synthetic data does not necessarily increase performance over using real data alone, suggesting the importance of synthetic data quality.

4D Feature Fields Reconstruction Recent advances in 3D/4D feature fields reconstruction methods [KMS22, ZCJ24, YDE24, FZC24, ZRW25] have significantly enhanced the vision foundation model's performance in 3D/4D space by integrating structured latent scene representations into the model's inference stage. Inspired by the promising results of feature lifting, we explore enhancing the InternVideo2-8B model [WLL24] with spatiotemporal awareness by adopting the strategy proposed in Feature4X [ZRW25], which constructs the VLM's 2D feature space along the time dimension into a 4D feature field. To assess this approach, we evaluate performance on a subset of the `VLM4D` benchmark, specifically leveraging all 50 videos from the DAVIS 2016 dataset [PPM16]. Our experimental evaluation compares inference performance across three distinct input modalities: original 2D videos; rendered novel global-view videos (which provide broader contextual information in 2D format); and reconstructed global feature fields (which implicitly incorporate 4D scene-level information

during reasoning). Table 5.3 reveals that reconstructed feature fields achieve the highest accuracy across both reasoning types. This success stems from two key advantages: the inherent structure of 4D representations and the ability of feature field inference to avoid the rendering artifacts of RGB reconstruction (global view video). However, the current approach requires per-scene optimization as a post-processing step, limiting its generalizability and making it computationally intensive.

5.7 Conclusion

Through the construction of the `VLM4D` benchmark, we evaluate the spatiotemporal reasoning capabilities of various vision language models (both open-source and proprietary). While more recently released models demonstrate improved performance over their counterparts, they remain significantly behind human proficiency. Overall, our work questions whether VLMs possess spatiotemporal reasoning abilities that are imperative to have for more sophisticated visual agents in fields ranging from robotics to interactive AI systems that require a deep understanding of dynamic visual environments. We hope to inspire future work to explore novel approaches for integrating spatiotemporal grounding, thereby enhancing their spatiotemporal reasoning capabilities and facilitating robust deployment.

Organization	Model	Release	Real			Synthetic			Overall
			Ego-centric	Exo-centric	Average	Directional	FP	Average	
User Study	Human Performance		99.6	99.7	99.7	95.8	100	96.2	98.8
Random	Random Selection		24.4	23.2	23.6	25.5	24.7	25.4	24.1
Latest Proprietary VLMs									
OpenAI	GPT-4o	2024-11	55.5	62.2	60.0	49.5	53.3	49.9	57.5
Google	Gemini-2.5-Pro	2025-6	64.6	62.9	63.5	54.8	80.0	57.3	62.0
Anthropic	Claude-Sonnet-4	2025-5	52.6	52.1	52.2	44.0	86.7	48.3	51.3
xAI	Grok-2-Vision	2024-12	48.8	49.7	49.4	49.3	66.7	51.0	49.8
Open-source Image VLMs									
Meta	Llama-4-Maverick-17B	2025-4	52.6	54.3	53.8	53.3	51.1	53.0	53.6
	Llama-4-Scout-17B	2025-4	48.6	56.2	53.7	53.3	75.6	55.5	54.1
Microsoft	Phi-4-Multimodal	2025-3	41.0	35.4	37.2	37.5	11.1	34.8	36.6
	Phi-3.5-Vision	2024-7	33.4	38.8	37.1	23.3	37.8	24.7	34.0
DeepSeek	DeepSeek-VL2	2024-12	33.6	32.9	33.1	31.8	46.7	33.3	33.2
Shanghai AI Lab	InternVL2.5-38B	2024-11	46.6	50.1	48.9	43.3	57.8	44.7	47.9
	InternVL2.5-8B	2024-11	39.0	44.0	42.4	40.8	42.2	40.9	42.0
Mistral AI	Pixtral-12B	2024-9	32.3	25.8	27.9	24.3	22.2	24.0	27.0
Rhymes	Aria	2024-11	47.2	44.0	45.1	38.5	71.1	41.8	44.3
Open-source Video VLMs									
Alibaba	Qwen2.5-VL-7B	2025-1	42.3	43.7	43.3	43.5	64.4	45.6	43.8
	Qwen2.5-VL-72B	2025-1	54.3	52.5	53.1	49.5	80.0	52.6	53.0
	Qwen2-VL-7B	2024-8	36.1	34.7	35.2	40.5	35.6	40.0	36.3
	Qwen2-VL-72B	2024-9	48.1	43.0	44.6	40.8	73.3	44.0	44.5
DAMO	VideoLLama3-2B	2025-1	53.2	42.5	46.0	34.3	55.6	36.4	43.7
	VideoLLama3-7B	2025-1	49.4	45.1	46.5	42.8	53.3	43.8	45.9
Shanghai AI Lab	InternVideo2.5-8B	2025-1	57.2	50.5	52.7	44.3	46.7	44.5	50.7
	InternVideo2-8B	2024-8	35.6	39.3	38.1	43.0	0.0	38.7	38.2
LLaVA	LLaVA-One-Vision-7B	2024-9	36.8	35.6	36.0	37.8	35.6	37.5	36.3
	LLaVA-NeXT-Video-34B	2024-6	29.6	31.6	30.9	24.5	55.6	27.6	30.1

Table 5.1: **Evaluation on VLM4D Benchmark** across various proprietary and open-source VLMs. Top three performers in each category are highlighted from dark (highest) to light (third highest). Human and random selection baselines are included for reference.

Model	MC
Original Model	
Qwen 2VL (7B)	38.3
Qwen 2.5VL (7B)	43.4
Finetuned Model	
Qwen 2VL (7B) (R)	54.5
Qwen 2VL (7B) (S)	42.5
Qwen 2VL (7B) (R+S)	**55.5**
Qwen 2.5VL (7B) (R)	55.6
Qwen 2.5VL (7B) (S)	48.6
Qwen 2.5VL (7B) (R+S)	**56.3**

Table 5.2: **SFT on Spatial-Temporal Datasets.** MC refers to the performance of multiple-choice, while R, S, and R+S denote real, synthetic, and both real+synthetic usage of data for fine-tuning.

Input Modality	**Accuracy**
Chain of Thought Response	
Original 2D Video	36.0
Global View Video	32.7
Global Feature Field	**37.4**
Direct Output Response	
Original 2D Video	24.3
Global View Video	23.8
Global Feature Field	**29.0**

Table 5.3: **InternVideo2 Accuracy with 4D Feature Field Reconstruction.** Comparison of InternVideo2's performance given different input modalities from the same subset.

CHAPTER 6

Conclusion

6.1 The Architecture of a Spatially Intelligent World Model

This dissertation set out to explore the frontier of **Spatial Intelligence**—the capacity for AI systems to transcend the limitations of 2D data and engage with the 3D world in a manner that is semantically rich, geometrically consistent, and dynamically interactive. Through the four distinct research threads presented—**Perception (Feature 3DGS)**, **Imagination (DreamScene360)**, **Interaction (Feature4X)**, and **Reasoning (VLM4D)**—we have constructed a cohesive narrative and a technological blueprint for building the next generation of World Models.

The core finding of this research is that the dichotomy between "Semantic AI" (foundation models) and "Geometric AI" (3D graphics) is a false one. The path forward lies in their integration. We have demonstrated that **Gaussian Splatting** serves as the ideal "glue" for this integration, acting as a flexible container that can hold the semantic features of foundation models, the generative output of diffusion models, and the temporal dynamics of video data.

These four components are not isolated projects; they represent the functional organs of a unified system:

- **Feature 3DGS** acts as the **Visual Cortex**, translating raw sensory data (images) into a structured, semantically labeled 3D mental map. It solves the problem of *consistency*, ensuring that an object remains the same entity regardless of the viewing angle.

- **DreamScene360** acts as the **Hippocampus and Prefrontal Cortex**, capable of generating new environments from memory (data) and imagination (text). It solves the problem of *environment creation*, allowing the agent to simulate worlds it has never physically visited.

- **Feature4X** acts as the **Motor and Temporal System**, processing the flow of time and enabling manipulation. It solves the problem of *dynamics*, allowing the agent to understand cause-and-effect relationships in a fluid, changing environment through the use of Motion Scaffolds.

- **VLM4D** acts as the **Cognitive Benchmark**, a rigorous testing ground to ensure that the "intelligence" being developed is grounded in physical reality, not just statistical mimicry.

Synthesized together, these works demonstrate that **Explicit Feature Distillation**—the process of lifting 2D foundation model outputs into 3D/4D Gaussian fields—is a viable and powerful path toward Spatial Intelligence. Unlike implicit methods (NeRFs) which trap knowledge in uninterpretable weights, our approach keeps the world representation **explicit, editable, and interpretable**.

6.2 Theoretical Implications: From "Learning to See" to "Learning to Simulate"

The transition from 2D computer vision to 4D World Models represents a fundamental shift in the ontology of AI. Traditional computer vision asks: *"What is in this image?"* Spatial Intelligence asks: *"What can I do in this world?"*

6.2.1 The Redefinition of Rendering

One of the most profound implications of this research is the redefinition of "rendering." Traditionally, rendering is the process of generating RGB pixels for human consumption. In **Feature 3DGS** and **Feature4X**, we demonstrated the power of **Feature Rendering**. By modifying the rasterization pipeline to output high-dimensional feature maps (e.g., 64-dim or 512-dim vectors) instead of just colors, we transform the virtual camera from a passive capture device into an **active semantic query engine**.

This suggests a future where autonomous agents do not "see" RGB images that then need to be processed by heavy CNNs. Instead, they will "render" semantic maps directly from their internal world model. An autonomous car, for instance, could render a "risk map" or a "drivable surface map" directly from its internal Gaussian representation, bypassing the noisy intermediate step of interpreting raw pixels. This approach aligns with the predictive coding theories in neuroscience, suggesting that perception is largely a process of verifying internal predictions against sensory inputs.

6.2.2 The Democratization of Simulation (Sim2Real)

DreamScene360 highlights a significant trend in the democratization of 3D creativity and simulation. Historically, creating consistent 3D environments required expert knowledge of modeling software and hours of labor. By bridging the gap between natural language and 3D Gaussian Splatting, we empower the creation of infinite training environments. In the context of **Sim2Real** transfer for robotics, this is transformative. The ability to generate diverse, physically consistent training environments from text prompts ("A cluttered kitchen," "A rainy street intersection") allows for the scaling of robotic learning in simulation before deployment in the real world. The resolution of the "Janus problem" via our panoramic approach ensures that these simulations are geometrically plausible, preventing robots from learning policies based on impossible geometries.

6.2.3 Agentic Interaction as a Proxy for Understanding

Feature4X marks a departure from passive video analysis to **Agentic Video Interaction**. Traditional video models (like Video-LLMs) are passive observers. Feature4X enables an agent to *intervene*. The ability to perform "scene editing" or "object removal" in 4D is not just an artistic tool; it is a primitive form of **causal reasoning**. To successfully "delete" a moving person from a video and fill in the background consistently over time, the model must understand the geometry of the background, the trajectory of the person, and the temporal continuity of the scene. This implies that **editing is a proxy for understanding**. If an AI can realistically edit the world, it likely understands the world's underlying structure. This supports the growing consensus that generative modeling is key to understanding.

6.3 Critical Limitations and Challenges

While the contributions of this dissertation are significant, they exist within a rapidly evolving field, and several limitations define the boundaries of our current capabilities.

6.3.1 The Distillation Bottleneck

A recurring theme in Feature 3DGS and Feature4X is **Distillation**. We rely on "teacher" models (SAM, CLIP, DINO) to provide semantic supervision. Consequently, our World Models are strictly bound by the performance of these teachers. If SAM fails to segment an object, our 3D representation will inherit that failure. We are essentially building "shadows" of 2D intelligence cast into 3D. True spatial intelligence may eventually require **end-to-end 3D pre-training**, where features are learned directly from 3D data rather than distilled from 2D priors. However, until 3D datasets scale to the size of the internet, distillation remains the most pragmatic bridge.

6.3.2 The Kinematic vs. Dynamic Gap

While Feature4X introduces dynamics via Motion Scaffolds, it is still fundamentally a **kinematic** representation. It replays observed motion; it does not yet simulate physics (forces, mass, friction). If we drop a virtual ball into a Feature4X scene, it will not bounce unless we explicitly program a physics engine. The current system models *how* things move, but not *why* they move. Bridging this gap—incorporating Newtonian mechanics directly into the Gaussian representation—is the "holy grail" for Physical AI and remains an open challenge.

6.3.3 The Reasoning Gap

The findings from **VLM4D** serve as a critical reality check. The fact that state-of-the-art VLMs struggle with basic spatiotemporal tasks (e.g., distinguishing rotation from translation) indicates that we are still far from AGI. Our proposed benchmarks reveal that while we can *build* high-fidelity 3D/4D representations (Feature 3DGS, Feature4X), the cognitive layer (the VLM) often fails to fully utilize this rich information. There is a disconnect between the precision of the *representation* and the reasoning capability of the *planner*.

6.4 Future Directions: Toward Embodied General Intelligence

The research presented in this dissertation lays the groundwork for several exciting future directions that will occupy the field for the next decade.

6.4.1 Physics-Informed Gaussian Splatting

To move from kinematics to dynamics, future work should explore **Physics-Informed Gaussian Splatting**. By adding physical attributes (mass, elasticity, friction coefficients) to the Gaussian primitives and integrating a differentiable physics engine into the rasterization loop, we could create World Models that not only look real but *act* real. This would

allow agents to predict the outcome of physical interactions (e.g., "If I push this stack of blocks, how will it fall?") with high fidelity.

6.4.2 Closing the Perception-Action Loop

The natural extension of this work is to integrate these 3D/4D World Models into the control loops of **embodied agents** (robots). Instead of training robots on raw pixels, we can train them inside the **DreamScene360** or **Feature4X** environments. The robot would perceive the world through **Feature 3DGS**, using the dense semantic features to plan grasp points or navigation paths. This closes the loop between *Perception*, *Imagination*, and *Action*.

6.4.3 Self-Supervised Spatial Learning

Finally, to overcome the "Reasoning Gap" identified in VLM4D, future models must engage in **Self-Supervised Spatial Learning**. Just as infants learn physics by playing with objects, AI agents should learn spatial intelligence by interacting with their internal world models—predicting the outcome of an action, observing the result, and updating their internal model. Our proposed explicit representations provide the ideal "scratchpad" for this learning process, allowing the agent to hypothesize changes to the 3D geometry and verify them against sensory data.

6.5 Final Remarks

We are witnessing the dawn of a new epoch in Artificial Intelligence. The first epoch was defined by logic and symbols; the second by statistics and 2D patterns. The third epoch, **Spatial Intelligence**, will be defined by the ability to model, understand, and interact with the physical reality of our universe.

This dissertation contributes to this grand endeavor by building the necessary bridges:

between 2D foundations and 3D structure, between language and geometry, and between passive observation and active reasoning. By endowing World Models with the capacity to *Perceive* (Feature 3DGS), *Imagine* (DreamScene360), *Interact* (Feature4X), and *Reason* (VLM4D), we take a decisive step toward AI that not only processes data but truly understands the world. The journey from "Words to Worlds" is just beginning, and the view from here—a future of high-fidelity, semantic, dynamic, and physically accurate spatial world representations—is spectacular.

REFERENCES

[AAA23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774*, 2023.

[AAA24] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. "Phi-3 technical report: A highly capable language model locally on your phone." *arXiv preprint arXiv:2404.14219*, 2024.

[AAA25] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. "Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs.", 2025.

[AAB24] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. "Phi-4 technical report." *arXiv preprint arXiv:2412.08905*, 2024.

[AAB25] 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.

[AAH24] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. "Pixtral 12B." *arXiv preprint arXiv:2410.07073*, 2024.

[Ant25] Anthropic. "System Card: Claude Opus 4 & Claude Sonnet 4." Technical report, 2025.

[AZS23] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. "Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond." *arXiv preprint arXiv:2304.04968*, 2023.

[BBC23] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. "Qwen technical report." *arXiv preprint arXiv:2309.16609*, 2023.

[BBW23] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. "Zoedepth: Zero-shot transfer by combining relative and metric depth." *arXiv preprint arXiv:2302.12288*, 2023.

[BHE22] Tim Brooks, Aleksander Holynski, and Alexei A Efros. "InstructPix2Pix: Learning to Follow Image Editing Instructions." *arXiv preprint arXiv:2211.09800*, 2022.

[BHG24] Jiayang Bai, Letian Huang, Jie Guo, Wen Gong, Yuanqi Li, and Yanwen Guo. "360-GS: Layout-guided Panoramic Gaussian Splatting For Indoor Roaming." *arXiv preprint arXiv:2402.00763*, 2024.

[BKL16] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image." In *ECCV*, 2016.

[BMR20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." *Advances in neural information processing systems*, **33**:1877–1901, 2020.

[BMT21] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. "Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields." *ICCV*, 2021.

[BMV23] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. "Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields." *ICCV*, 2023.

[BMW23] Shariq Farooq Bhat, Niloy J Mitra, and Peter Wonka. "LooseControl: Lifting ControlNet for Generalized Depth Conditioning." *arXiv preprint arXiv:2312.03079*, 2023.

[BPH24] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. "Video generation models as world simulators." 2024.

[BPZ20] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Angela Dai, Justus Thies, and Matthias Niessner. "Neural Non-Rigid Tracking." In *NeurIPS*, 2020.

[BTS14] Matthew Berger, Andrea Tagliasacchi, Lee M Seversky, Pierre Alliez, Joshua A Levine, Andrei Sharf, and Claudio T Silva. "State of the art in surface reconstruction from point clouds." In *35th Annual Conference of the European Association for Computer Graphics, Eurographics 2014-State of the Art Reports*, number CONF. The Eurographics Association, 2014.

[Bur06] Neil Burgess. "Spatial memory: how egocentric and allocentric combine." *Trends in Cognitive Sciences*, **10**(12):551–557, 2006.

[BV99] Volker Blanz and Thomas Vetter. "A Morphable Model for the Synthesis of 3D Faces." In *CVPR*, 1999.

[BVC19] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. "Zero-shot semantic segmentation." *Advances in Neural Information Processing Systems*, **32**, 2019.

[BWL23] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. "Nope-nerf: Optimising neural radiance field with no pose prior." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4160–4169, 2023.

[BYL23] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. "Multidiffusion: Fusing diffusion paths for controlled image generation." 2023.

[CBZ23] Pradyumna Chari, Yunhao Ba, Shijie Zhou, Chinmay Talegaonkar, Shreeram Athreya, and Achuta Kadambi. "On learning mechanical laws of motion from video using neural networks." *IEEE Access*, **11**:30129–30145, 2023.

[CGY22] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. "VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18030–18040, June 2022.

[CHM24] Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. "ShareGPT-4o: Comprehensive Multimodal Annotations With GPT-4o.", 2024.

[CJ23] Ang Cao and Justin Johnson. "HexPlane: A Fast Representation for Dynamic Scenes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 130–141, 2023.

[CL96] Brian Curless and Marc Levoy. "A Volumetric Method for Building Complex Models from Range Images." In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 303–312, 1996.

[CLC22] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. "Efficient geometry-aware 3D generative adversarial networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

[CLD24] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. "Are We on the Right Way for Evaluating Large Vision-Language Models?" *arXiv preprint arXiv:2403.20330*, 2024.

[CLN23] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. "Luciddreamer: Domain-free generation of 3d gaussian splatting scenes." *arXiv preprint arXiv:2311.13384*, 2023.

[CLZ24] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms." *arXiv preprint arXiv:2406.07476*, 2024.

[CTM21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging properties in self-supervised vision transformers." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[CWC24] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling." *arXiv preprint arXiv:2412.05271*, 2024.

[CWL24] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. "ShareGPT4Video: Improving Video Understanding and Generation with Better Captions." *arXiv preprint arXiv:2406.04325*, 2024.

[CXG22] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. "Tensorf: Tensorial radiance fields." In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022.

[CXK24] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

[CXZ21] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14124–14133, 2021.

[CYF24] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. "Spatialrgpt: Grounded spatial reasoning in vision-language models." *Advances in Neural Information Processing Systems*, **37**:135062–135093, 2024.

[DBK21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *International Conference on Learning Representations*, 2021.

[DCL19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[DJQ23] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. "Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20637–20647, 2023.

[DLL23] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning.", 2023.

[DMN16] Julian De Freitas, Nicholas E. Myers, and Anna C. Nobre. "Tracking the changing feature of a moving object." *Journal of Vision*, **16**(3):22, 2016.

[DN21] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in neural information processing systems*, **34**:8780–8794, 2021.

[DWD24] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. "4D Gaussian Splatting: Towards Efficient Novel View Synthesis for Dynamic Scenes." *arXiv preprint arXiv:2402.03307*, 2024.

[DXS23] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. "Palm-e: An embodied multimodal language model." 2023.

[DZY21] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. "Neural Radiance Flow for 4D View Synthesis and Video Processing." In *ICCV*, 2021.

[EEV15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes challenge: A retrospective." *International journal of computer vision*, **111**:98–136, 2015.

[FDL24] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis." *arXiv preprint arXiv:2405.21075*, 2024.

[FF84] Jennifer J. Freyd and Ronald A. Finke. "Representational momentum." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**(1):126–132, 1984.

[FHL23] Chuan Fang, Xiaotao Hu, Kunming Luo, and Ping Tan. "Ctrl-Room: Controllable Text-to-3D Room Meshes Generation with Layout Constraints." *arXiv preprint arXiv:2310.03602*, 2023.

[FLS25] Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. "Missing Premise exacerbates Overthinking: Are Reasoning Models losing Critical Thinking Skill?" *arXiv preprint arXiv:2504.06514*, 2025.

[FMW23a] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. "K-planes: Explicit radiance fields in space, time, and appearance." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12479–12488, 2023.

[FMW23b] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. "K-planes: Explicit radiance fields in space, time, and appearance." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12479–12488, 2023.

[FWJ22] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. "Nerf-sos: Any-view self-supervised object segmentation on complex scenes." *arXiv preprint arXiv:2209.08776*, 2022.

[FWW23] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. "Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps." *arXiv preprint arXiv:2311.17245*, 2023.

[FYT22] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. "Plenoxels: Radiance fields without neural networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5501–5510, 2022.

[FZC24] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. "Large spatial model: End-to-end unposed images to semantic 3d." *Advances in Neural Information Processing Systems*, **37**:40212–40229, 2024.

[FZL26] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, et al. "VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[Gem25] Google Gemini Team. "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities." Technical report, 2025.

[GLC23] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. "GraphDreamer: Compositional 3D Scene Synthesis from Scene Graphs." *arXiv preprint arXiv:2312.00093*, 2023.

[GLZ23] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. "Multimodal-gpt: A vision and language model for dialogue with humans." *arXiv preprint arXiv:2305.04790*, 2023.

[GPK24] Zhengyang Geng, Ashwini Pokle, and J Zico Kolter. "One-step diffusion distillation via deep equilibrium models." *Advances in Neural Information Processing Systems*, **36**, 2024.

[GPM20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks." *Communications of the ACM*, **63**(11):139–144, 2020.

[GSK21] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. "Dynamic View Synthesis from Dynamic Monocular Video." In *ICCV*, 2021.

[GSS23] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. "Interactive segmentation of radiance fields." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4201–4211, 2023.

[GT19] Wei Gao and Russ Tedrake. "SurfelWarp: Efficient Non-Volumetric Single View Dynamic Reconstruction." *arXiv preprint arXiv:1904.13073*, 2019.

[GTZ21a] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Niessner. "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8649–8658, 2021.

[GTZ21b] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction." In *CVPR*, 2021.

[GWB22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

[GZN20] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. "Context-aware feature generation for zero-shot semantic segmentation." In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1921–1929, 2020.

[HCO23] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. "Text2room: Extracting textured 3d meshes from 2d text-to-image models." *arXiv preprint arXiv:2303.11989*, 2023.

[HFF22] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. "Particle video revisited: Tracking through occlusions using point trajectories." In *European Conference on Computer Vision*, pp. 59–75. Springer, 2022.

[HFZ24] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. "Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos." *arXiv preprint arXiv:2406.08407*, 2024.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems*, **33**:6840–6851, 2020.

[HK18] Peter Hedman and Johannes Kopf. "Instant 3d photography." *ACM Transactions on Graphics (TOG)*, **37**(4):1–12, 2018.

[HKS22] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. "Object discovery and representation networks." *arXiv preprint arXiv:2203.08777*, 2022.

[HLG24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. "Gpt-4o system card." *arXiv preprint arXiv:2410.21276*, 2024.

[HPP18] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. "Deep blending for free-viewpoint image-based rendering." *ACM Transactions on Graphics (ToG)*, **37**(6):1–15, 2018.

[HS22] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance." *arXiv preprint arXiv:2207.12598*, 2022.

[HSW21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685*, 2021.

[HSY24] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. "SC-GS: Sparse-Controlled Gaussian Splatting for Editable Dynamic Scenes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4220–4230, 2024.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531*, 2015.

[HWY24] Xuehai He, Shuohang Wang, Jianwei Yang, Xiaoxia Wu, Yiping Wang, Kuan Wang, Zheng Zhan, Olatunji Ruwase, Yelong Shen, and Xin Eric Wang. "Mojito: Motion Trajectory and Intensity Control for Video Generation." *arXiv preprint arXiv: 2412.08948*, 2024.

[JN23] Heewoo Jun and Alex Nichol. "Shap-E: Generating Conditional 3D Implicit Functions." *arXiv preprint arXiv:2305.02463*, 2023.

[Joh73] Gunnar Johansson. "Visual perception of biological motion and a model for its analysis." *Perception & Psychophysics*, **14**(2):201–211, 1973.

[JWF24] Shengxiang Ji, Guanjun Wu, Jiemin Fang, Jiazhong Cen, Taoran Yi, Wenyu Liu, Qi Tian, and Xinggang Wang. "Segment Any 4D Gaussians." *arXiv preprint arXiv:2407.04504*, 2024.

[KGF20] Nilesh Kulkarni, Abhinav Gupta, David Fouhey, and Shubham Tulsiani. "Articulation-aware Canonical Surface Mapping." In *CVPR*, 2020.

[KKC21] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. "Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering." *NeurIPS*, 2021.

[KKG23] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. "Lerf: Language embedded radiance fields." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

[KKL23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. "3d gaussian splatting for real-time radiance field rendering." *ACM Transactions on Graphics (ToG)*, **42**(4):1–14, 2023.

[KLA19] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

[KMR23] 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.

[KMS22] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. "Decomposing nerf for editing via feature field distillation." *Advances in neural information processing systems*, **35**:23311–23330, 2022.

[KNH24] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. "How Good is my Video LMM? Complex Video Reasoning and Robustness Evaluation Suite for Video-LMMs." *arXiv preprint arXiv:2405.03690*, 2024.

[KPK24] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. "Openvla: An open-source vision-language-action model." *arXiv preprint arXiv:2406.09246*, 2024.

[KPL21] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. "Point-Based Neural Rendering with Per-View Optimization." In *Computer Graphics Forum*, volume 40, pp. 29–43. Wiley Online Library, 2021.

[KRG23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. "Cotracker: It is better to track together." *arXiv preprint arXiv:2307.07635*, 2023.

[KTC15] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. "Category-Specific Object Reconstruction from a Single Image." In *CVPR*, 2015.

[KTE18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. "Learning Category-Specific Mesh Reconstruction from Image Collections." In *ECCV*, 2018.

[KVN23] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. "Holodiffusion: Training a 3D diffusion model using 2D images." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18423–18433, 2023.

[Lap05] Ivan Laptev. "On Space-Time Interest Points." *International Journal of Computer Vision*, **64**(2-3):107–123, 2005.

[LBD89] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation*, **1**(4):541–551, 1989.

[LBL23] Yiqi Lin, Haotian Bai, Sijia Li, Haonan Lu, Xiaodong Lin, Hui Xiong, and Lin Wang. "Componerf: Text-guided multi-object compositional nerf with editable 3d scene layout." *arXiv preprint arXiv:2303.13843*, 2023.

[LBP19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks." In *Neural Information Processing Systems*, 2019.

[LCL24] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. "Spacetime gaussian feature splatting for real-time dynamic view synthesis." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8508–8520, 2024.

[LDZ24a] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. "Gaussian-Flow: 4D Reconstruction with Dynamic 3D Gaussian Particle." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21136–21145, 2024.

[LDZ24b] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. "Mmbench: Is your multi-modal model an all-around player?" In *European conference on computer vision*, pp. 216–233. Springer, 2024.

[Les84] Alan M. Leslie. "Spatiotemporal continuity and the perception of causality in infants." *Perception*, **13**(3):287–305, 1984.

[LGG24] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. "Seed-bench: Benchmarking multimodal large language models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

[LGM23] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. "Robust Dynamic Radiance Fields." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[LHR18] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. "Deep object co-segmentation." In *Asian Conference on Computer Vision*, pp. 638–653. Springer, 2018.

[LHW23] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. "VideoChat: Chat-centric Video Understanding." *arXiv preprint arXiv:2305.06355*, 2023.

[LHW24] Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. "Videoeval: Comprehensive benchmark suite for low-cost evaluation of video foundation model." *arXiv preprint arXiv:2407.06491*, 2024.

[LKL23a] Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. "Gaufre: Gaussian Deformation Fields for Real-Time Dynamic Novel View Synthesis." *arXiv preprint arXiv:2312.11458*, 2023.

[LKL23b] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. "Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis." *arXiv preprint arXiv:2308.09713*, 2023.

[LKT23] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. "Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models." *arXiv preprint arXiv:2312.13763*, 2023.

[LLL25] Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Zhaoshuo Li. "Scenethesis: A language and vision agentic framework for 3d scene generation." *arXiv preprint arXiv:2505.02836*, 2025.

[LLW23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning." *Advances in neural information processing systems*, **36**:34892–34916, 2023.

[LLW24] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. "Aria: An open multimodal native mixture-of-experts model." *arXiv preprint arXiv:2410.05993*, 2024.

[LLZ24] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. "Deepseek-vl: towards real-world vision-language understanding." *arXiv preprint arXiv:2403.05525*, 2024.

[LNS21] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. "Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes." In *CVPR*, 2021.

[LPY24] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, and Zhiwen Fan. "4K4DGen: Panoramic 4D Generation at 4K Resolution." *arXiv preprint arXiv:2406.13527*, 2024.

[LPY25] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, and Zhiwen Fan. "4K4DGen: Panoramic 4D Generation at 4K Resolution." In *The Thirteenth International Conference on Learning Representations*, 2025.

[LRS23] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. "Compact 3d gaussian representation for radiance field." *arXiv preprint arXiv:2311.13681*, 2023.

[LSP08] Hao Li, Robert W Sumner, and Mark Pauly. "Global Correspondence Optimization for Non-rigid Registration of Depth Scans." In *Computer Graphics Forum*, volume 27, pp. 1421–1430. Wiley Online Library, 2008.

[LSS19] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. "Neural volumes: Learning dynamic renderable volumes from images." *arXiv preprint arXiv:1906.07751*, 2019.

[LSZ22] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. "Neural 3D Video Synthesis from Multi-view Video." In *CVPR*, 2022.

[LWB22a] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. "Language-driven semantic segmentation." *arXiv preprint arXiv:2201.03546*, 2022.

[LWB22b] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. "Language-driven Semantic Segmentation.", 2022.

[LWC20] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. "Fss-1000: A 1000-class dataset for few-shot segmentation." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2869–2878, 2020.

[LWH23] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. "MVBench: A Comprehensive Multi-modal Video Understanding Benchmark.", 2023.

[LWH24a] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. "MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds." *arXiv preprint arXiv:2405.17421*, 2024.

[LWH24b] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. "Mvbench: A comprehensive multi-modal video understanding benchmark." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

[LXL23] Stefan Lionar, Xiangyu Xu, Min Lin, and Gim Hee Lee. "NU-MCC: Multiview Compressive Coding with Neighborhood Decoder and Repulsive UDF." *arXiv preprint arXiv:2307.09112*, 2023.

[LYB24] Hyunjee Lee, Youngsik Yun, Jeongmin Bae, Seoha Kim, and Youngjung Uh. "Rethinking Open-Vocabulary Segmentation of Radiance Fields in 3D Space." *arXiv preprint arXiv:2408.07416*, 2024.

[LYX24] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. "Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

[LYZ24a] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. "World model on million-length video and language with blockwise ringattention." *arXiv preprint arXiv:2402.08268*, 2024.

[LYZ24b] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. "World Model on Million-Length Video and Language with RingAttention." *arXiv preprint*, 2024.

[LZG24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. "Llava-onevision: Easy visual task transfer." *arXiv preprint arXiv:2408.03326*, 2024.

[LZJ24] Xinyi Liu, Tianyi Zhang, Matthew Johnson-Roberson, and Weiming Zhi. "Spla-Traj: Camera Trajectory Generation with Semantic Gaussian Splatting." *arXiv preprint arXiv:2410.06014*, 2024.

[LZZ24] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. "Weakly supervised 3d open-vocabulary segmentation." *Advances in Neural Information Processing Systems*, **36**, 2024.

[MBH23] Wieland Morgenstern, Florian Barthel, Anna Hilsmann, and Peter Eisert. "Compact 3d scene representation via self-organizing gaussian grids." *arXiv preprint arXiv:2312.13299*, 2023.

[MCL23] Weijia Mao, Yan-Pei Cao, Jia-Wei Liu, Zhongcong Xu, and Mike Zheng Shou. "ShowRoom3D: Text to High-Quality 3D Room Generation Using 3D Priors." *arXiv preprint arXiv:2312.13324*, 2023.

[MES22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. "Instant neural graphics primitives with a multiresolution hash encoding." *ACM Transactions on Graphics (ToG)*, **41**(4):1–15, 2022.

[Met25] Meta. "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation." Technical report, 2025.

[MMB12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain." *IEEE Transactions on image processing*, **21**(12):4695–4708, 2012.

[MON19] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. "Occupancy networks: Learning 3d reconstruction in function space." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

[MPT25] Marko Mihajlovic, Sergey Prokudin, Siyu Tang, Robert Maier, Federica Bogo, Tony Tung, and Edmond Boyer. "Splatfields: Neural Gaussian Splats for Sparse 3D and 4D Reconstruction." In *European Conference on Computer Vision*, pp. 313–332. Springer, 2025.

[MRG23] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. "On distillation of guided diffusion models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

[MRK23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. "Video-chatgpt: Towards detailed video understanding via large vision and language models." *arXiv preprint arXiv:2306.05424*, 2023.

[MSB12] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. "Making a "completely blind" image quality analyzer." *IEEE Signal processing letters*, **20**(3):209–212, 2012.

[MSD23] Kirill Mazur, Edgar Sucar, and Andrew J Davison. "Feature-realistic neural fusion for real-time, open set scene understanding." In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8201–8207. IEEE, 2023.

[MSO19] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines." *ACM Transactions on Graphics (TOG)*, **38**(4):1–14, 2019.

[MST20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis." In *European conference on computer vision*, pp. 405–421. Springer, 2020.

[MST21a] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM*, **65**(1):99–106, 2021.

[MST21b] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." *Communications of the ACM*, **65**(1):99–106, 2021.

[MU81] D. Marr and S. Ullman. "Directional selectivity and its use in early visual processing." *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **211**(1183):151–180, 1981.

[NFS15] Richard Newcombe, Dieter Fox, and Steve Seitz. "DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time." In *CVPR*, 2015.

[NLT19] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. "Hologan: Unsupervised learning of 3d representations from natural images." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7588–7597, 2019.

[NMK23] KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. "Compact3d: Compressing gaussian splat radiance field models with vector quantization." *arXiv preprint arXiv:2311.18159*, 2023.

[NMO19] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. "Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5379–5389, 2019.

[NSL21] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. "Neural Articulated Radiance Field." In *ICCV*, 2021.

[NSW23] Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. "Compressed 3d gaussian splatting for accelerated novel view synthesis." *arXiv preprint arXiv:2401.02436*, 2023.

[NZX23] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. "Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models." *arXiv preprint arXiv:2311.16103*, 2023.

[ODM23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193*, 2023.

[OHL23] Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. "Text2immersion: Generative immersive scene with 3d gaussians." *arXiv preprint arXiv:2312.09242*, 2023.

[Ope24] OpenAI. "Hello GPT-4o." Technical report, 2024.

[OTK24] Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoğlu, and Theo Gevers. "Sceneteller: Language-to-3d scene generation." In *European Conference on Computer Vision*, pp. 362–378. Springer, 2024.

[PCP21a] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. "D-nerf: Neural radiance fields for dynamic scenes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.

[PCP21b] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. "D-NeRF: Neural Radiance Fields for Dynamic Scenes." In *CVPR*, 2021.

[PDW21] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. "Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14314–14323, 2021.

[PFS19] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. "Deepsdf: Learning continuous signed distance functions for shape representation." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

[PJB22] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. "Dreamfusion: Text-to-3d using 2d diffusion." *arXiv preprint arXiv:2209.14988*, 2022.

[PNM20] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. "Convolutional occupancy networks." In *European Conference on Computer Vision*, pp. 523–540. Springer, 2020.

[PPC17] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. "The 2017 davis challenge on video object segmentation." *arXiv preprint arXiv:1704.00675*, 2017.

[PPM16] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. "A benchmark dataset and evaluation methodology for video object segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

[PSB21] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. "Nerfies: Deformable neural radiance fields." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.

[PSH21] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. "HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields." *ACM Transactions on Graphics (TOG)*, 2021.

[PVG11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research*, **12**:2825–2830, 2011.

[PW23] Ryan Po and Gordon Wetzstein. "Compositional 3D Scene Generation using Locally Conditioned Diffusion." *arXiv preprint arXiv:2303.12218*, 2023.

[PYH25] Shivansh Patel, Xinchen Yin, Wenlong Huang, Shubham Garg, Hooshang Nayyeri, Li Fei-Fei, Svetlana Lazebnik, and Yunzhu Li. "A Real-to-Sim-to-Real Approach to Robotic Manipulation with VLM-Generated Iterative Keypoint Rewards." *arXiv preprint arXiv:2502.08643*, 2025.

[PYS24] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. "UniDepth: Universal Monocular Metric Depth Estimation." *arXiv preprint arXiv:2403.18913*, 2024.

[PZX21] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. "Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans." In *CVPR*, 2021.

[QLZ24] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. "LangSplat: 3D Language Gaussian Splatting." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

[RAR22] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G. Schwing, and Oliver Wang. "Neural Volumetric Object Selection." In *CVPR*, 2022.

[RBK21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

[RBL22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[RDN22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents." *arXiv preprint arXiv:2204.06125*, **1**(2):3, 2022.

[RGH24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. "SAM 2: Segment Anything in Images and Videos." *arXiv preprint arXiv:2408.00714*, 2024.

[RKH21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[RKS12] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. "Reconstructing 3D Human Pose from 2D Image Landmarks." In *ECCV*, 2012.

[RNS18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. "Improving language understanding by generative pre-training." 2018.

[RPT23] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. "Dreamgaussian4d: Generative 4d gaussian splatting." *arXiv preprint arXiv:2312.17142*, 2023.

[RSP24] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. "Learning to localize objects improves spatial reasoning in visual-llms." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12977–12987, 2024.

[RXM24] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, et al. "L4GM: Large 4D Gaussian Reconstruction Model." *arXiv preprint arXiv:2406.10324*, 2024.

[RYR22] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. "360monodepth: High-resolution 360deg monocular depth estimation." In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3762–3772, 2022.

[SCS22a] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[SCS22b] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. "Photorealistic text-to-image diffusion models with deep language understanding." *Advances in Neural Information Processing Systems*, **35**:36479–36494, 2022.

[SCX23] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Yang Zhao. "Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture." *arXiv preprint arXiv:2305.11337*, 2023.

[SF16] Johannes L Schonberger and Jan-Michael Frahm. "Structure-from-motion revisited." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

[SGB24] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. "Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding." *arXiv preprint arXiv:2406.13807*, 2024.

[SH22] Tim Salimans and Jonathan Ho. "Progressive distillation for fast sampling of diffusion models." *arXiv preprint arXiv:2202.00512*, 2022.

[SHU24] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. "Dynamic Gaussian Marbles for Novel View Synthesis of Casual Monocular Videos." *arXiv preprint arXiv:2406.18717*, 2024.

[SJL24] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. "3DGStream: On-the-fly Training of 3D Gaussians for Efficient Streaming of Photo-realistic Free-viewpoint Videos." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20675–20685, 2024.

[SK07] Elizabeth S. Spelke and Katherine D. Kinzler. "Core knowledge." *Developmental Science*, **10**(1):89–96, 2007.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." *arXiv preprint arXiv:2010.02502*, 2020.

[SPB22] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. "Panoptic Lifting for 3D Scene Understanding with Neural Fields." *arXiv preprint arXiv:2212.09802*, 2022.

[SPP22] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. "Clip-fields: Weakly supervised semantic fields for robotic memory." *arXiv preprint arXiv:2210.05663*, 2022.

[SWD24] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. "Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5333–5343, 2024.

[SWM19] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. "The Replica Dataset: A Digital Replica of Indoor Spaces.", 2019.

[SYY23] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. "Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation." In *7th Annual Conference on Robot Learning*, 2023.

[SYZ21] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. "A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose." In *NeurIPS*, 2021.

[SZ14] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos." In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pp. 568–576, 2014.

[TBB22] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. "Splicing vit features for semantic appearance transfer." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10748–10757, 2022.

[TCY22] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. "Block-nerf: Scalable large scene neural view synthesis." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8248–8258, 2022.

[TD20] Zachary Teed and Jia Deng. "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow." In *ECCV*, 2020.

[TD21] Zachary Teed and Jia Deng. "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras." *Advances in Neural Information Processing Systems*, **34**:16558–16569, 2021.

[TGL24] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." *arXiv preprint arXiv:2403.05530*, 2024.

[TLI23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971*, 2023.

[TLL22] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. "Neural Feature Fusion Fields: 3D Distillation of Self-Supervised 2D Image Representations." In *3DV*, 2022.

[TML23] Nikolaos Tsagkas, Oisin Mac Aodha, and Chris Xiaoxuan Lu. "VL-Fields: Towards Language-Grounded Neural Implicit Spatial Representations." In *2023 IEEE International Conference on Robotics and Automation*. IEEE, 2023.

[TRZ23] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation." *arXiv preprint arXiv:2309.16653*, 2023.

[TTG21] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. "Non-rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene from Monocular Video." In *ICCV*, 2021.

[VCK23] Alexander Vilesov, Pradyumna Chari, and Achuta Kadambi. "Cg3d: Compositional generation for text-to-3d via gaussian splatting." *arXiv preprint arXiv:2311.17907*, 2023.

[WBT24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution." *arXiv preprint arXiv:2409.12191*, 2024.

[WBZ21] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. "Finetuned language models are zero-shot learners." *arXiv preprint arXiv:2109.01652*, 2021.

[WCL24] Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. "Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation." *Advances in Neural Information Processing Systems*, **37**:109922–109947, 2024.

[WCS22] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. "HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video." In *CVPR*, 2022.

[WDL23] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023.

[WGP25] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. "Cat4d: Create anything in 4d with multi-view video diffusion models." In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26057–26068, 2025.

[WKS11] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. "Action Recognition by Dense Trajectories." In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3169–3176, 2011.

[WLC24] 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, 2024.

[WLL24] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. "InternVideo2: Scaling Foundation Models for Multimodal Video Understanding." In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.

[WLW23] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. "ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation." *arXiv preprint arXiv:2305.16213*, 2023.

[WLY25] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. "InternVideo2. 5: Empowering Video MLLMs with Long and Rich Context Modeling." *arXiv preprint arXiv:2501.12386*, 2025.

[WMW24] Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. "Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering." *arXiv preprint arXiv:2406.00622*, 2024.

[WWA25] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. "Wan: Open and advanced large-scale video generative models." *arXiv preprint arXiv:2503.20314*, 2025.

[WWC23] Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. "PERF: Panoramic Neural Radiance Field from a Single Panorama." *arXiv preprint arXiv:2310.16831*, 2023.

[WWG21] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. "IBRNet: Learning Multi-View Image-Based Rendering." In *CVPR*, 2021.

[WWS22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems*, **35**:24824–24837, 2022.

[WXC24] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. "Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models." *arXiv preprint arXiv:2410.03290*, 2024.

[WXF24] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. "Customizing 360-degree panoramas through text-to-image diffusion models." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4933–4943, 2024.

[WYF23] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. "4d gaussian splatting for real-time dynamic scene rendering." *arXiv preprint arXiv:2310.08528*, 2023.

[WYG24] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. "Shape of Motion: 4D Reconstruction from a Single Video." *arXiv preprint arXiv:2407.13764*, 2024.

[WZD24] Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. "Vlm see, robot do: Human demo video to robot action plan via vision language model." *arXiv preprint arXiv:2410.08792*, 2024.

[WZP23] Zhen Wang, Shijie Zhou, Jeong Joon Park, Despoina Paschalidou, Suya You, Gordon Wetzstein, Leonidas Guibas, and Achuta Kadambi. "Alto: Alternating latent topologies for implicit 3d reconstruction." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 259–270, 2023.

[WZT22] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. "D^2NeRF: Self-Supervised Decoupling of Dynamic and Static Objects from a Monocular Video." In *NeurIPS*, 2022.

[WZZ23] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. "Q-align: Teaching lmms for visual scoring via discrete text-defined levels." *arXiv preprint arXiv:2312.17090*, 2023.

[xAI24] xAI. "Grok-2 Beta Release." Technical report, 2024.

[XHK21] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. "Space-time Neural Irradiance Fields for Free-Viewpoint Video." In *CVPR*, 2021.

[XJW22a] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. "Sinnerf: Training neural radiance fields on complex scenes from a single image." In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pp. 736–753. Springer, 2022.

[XJW22b] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. "NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360deg Views." *arXiv preprint arXiv:2211.16431*, 2022.

[XLB24] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. "Comp4D: LLM-Guided Compositional 4D Scene Generation." *arXiv preprint arXiv:2403.16993*, 2024.

[XMU23] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. "Sparsegs: Real-time 360 {\deg} sparse view synthesis using gaussian splatting." *arXiv preprint arXiv:2312.00206*, 2023.

[XXP22] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. "Point-nerf: Point-based neural radiance fields." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5438–5448, 2022.

[XYF18] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. "Youtube-vos: Sequence-to-sequence video object segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 585–601, 2018.

[YDE24] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. "Improving 2d feature representations by 3d-aware fine-tuning." In *European Conference on Computer Vision*, pp. 57–74. Springer, 2024.

[YDH23] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. "WonderJourney: Going from Anywhere to Everywhere." *arXiv preprint arXiv:2312.03884*, 2023.

[YDH24] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. "Wonderworld: Interactive 3d scene generation from a single image." *arXiv preprint arXiv:2406.09394*, 2024.

[YGZ24] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. "Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20331–20341, 2024.

[YHS24] Justin Yu, Kush Hari, Kishore Srinivas, Karim El-Refai, Adam Rashid, Chung Min Kim, Justin Kerr, Richard Cheng, Muhammad Zubair Irshad, Ashwin Balakrishna, et al. "Language-Embedded Gaussian Splats (LEGS): Incrementally Building Room-Scale Representations with a Mobile Robot." 2024.

[YNZ24] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

[YSJ21] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. "LASR: Learning Articulated Shape Reconstruction from a Monocular Video." In *CVPR*, 2021.

[YVN22] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. "BANMo: Building Animatable 3D Neural Models from Many Casual Videos." In *CVPR*, 2022.

[YWL23] Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. "Idea2img: Iterative self-refinement with gpt-4v (ision) for automatic image design and generation." *arXiv preprint arXiv:2310.08541*, 2023.

[YWW23] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. "Featurenerf: Learning generalizable nerfs by distilling foundation models." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8962–8973, 2023.

[YXW23] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. "4dgen: Grounded 4d content generation with spatial-temporal consistency." *arXiv preprint arXiv:2312.17225*, 2023.

[YYG24] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. "Thinking in space: How multimodal large language models see, remember, and recall spaces." *arXiv preprint arXiv:2412.14171*, 2024.

[YYP23] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. "Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting." *arXiv preprint arXiv:2310.10642*, 2023.

[YYT21] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. "pixelNeRF: Neural Radiance Fields from One or Few Images." In *CVPR*, 2021.

[YYZ24] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. "Qwen2. 5 technical report." *arXiv preprint arXiv:2412.15115*, 2024.

[YZB23] Brent Yi, Weijia Zeng, Sam Buchanan, and Yi Ma. "Canonical Factors for Hybrid Neural Fields." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3414–3426, 2023.

[YZL24] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. "Videorefer suite: Advancing spatial-temporal object understanding with video llm." *arXiv preprint arXiv:2501.00599*, 2024.

[ZC01] Cha Zhang and Tsuhan Chen. "Efficient feature extraction for 2D/3D objects in mesh representation." In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 3, pp. 935–938. IEEE, 2001.

[ZCJ24] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024.

[ZCS23] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. "Minigpt-4: Enhancing vision-language understanding with advanced large language models." *arXiv preprint arXiv:2304.10592*, 2023.

[ZFJ23] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. "FSGS: Real-Time Few-shot View Synthesis using Gaussian Splatting." *arXiv preprint arXiv:2312.00451*, 2023.

[ZFX24] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. "Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian

splatting." In *European Conference on Computer Vision*, pp. 324–342. Springer, 2024.

[ZHH24] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. "MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion." *arXiv preprint arXiv:2410.03825*, 2024.

[ZLB23] Hang Zhang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." *arXiv preprint arXiv:2306.02858*, 2023.

[ZLC25] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. "VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding." *arXiv preprint arXiv:2501.13106*, 2025.

[ZLG22] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. "ParticleSfM: Exploiting Dense Point Trajectories for Localizing Moving Cameras in the Wild." In *European Conference on Computer Vision*, pp. 523–542. Springer, 2022.

[ZLL21] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. "In-Place Scene Labelling and Understanding with Implicit Scene Representation." In *ICCV*, 2021.

[ZLL24] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. "LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.", April 2024.

[ZLW24] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. "Text2nerf: Text-driven 3d scene generation with neural radiance fields." *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[ZPV01] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. "EWA volume splatting." In *Proceedings Visualization, 2001. VIS'01.*, pp. 29–538. IEEE, 2001.

[ZRA23] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[ZRW25] Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan, Suya You, Zhangyang Wang, Leonidas Guibas, et al. "Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature

fields." In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14179–14190, 2025.

[ZSZ24] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. "FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding." *International Journal of Computer Vision*, pp. 1–17, 2024.

[ZVH25] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. "Vlm4d: Towards spatiotemporal awareness in vision language models." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8600–8612, 2025.

[ZWD24] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. "Apollo: An exploration of video understanding in large multimodal models." *arXiv preprint arXiv:2412.10360*, 2024.

[ZWL24] Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Yilun Du, and Chuang Gan. "COMBO: compositional world models for embodied multi-agent cooperation." *arXiv preprint arXiv:2404.10775*, 2024.

[ZXZ25] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. "MMVU: Measuring Expert-Level Multi-Discipline Video Understanding." *arXiv preprint arXiv:2501.12380*, 2025.

[ZZL23] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. "LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22490–22499, 2023.

[ZZL26] Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. "SpatialStack: Layered Geometry-Semantic Fusion for 3D VLM Spatial Reasoning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[ZZP19] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. "Semantic understanding of scenes through the ade20k dataset." *International Journal of Computer Vision*, **127**:302–321, 2019.

[ZZZ24a] Shuai Zhang, Huangxuan Zhao, Zhenghong Zhou, Guanjun Wu, Chuansheng Zheng, Xinggang Wang, and Wenyu Liu. "TOGS: Gaussian Splatting with Temporal Opacity Offset for Real-Time 4D DSA Rendering." *arXiv preprint arXiv:2403.19586*, 2024.

[ZZZ24b] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. "LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.