

000 001 002 003 004 005 LANGUAGE-GUIDED 4D GAUSSIAN SPLATTING FOR 006 REAL-TIME DYNAMIC SCENE RENDERING 007 008 009

010 **Anonymous authors**
 011 Paper under double-blind review
 012
 013
 014
 015
 016
 017
 018
 019
 020
 021
 022
 023
 024

ABSTRACT

025 Dynamic rendering methods often prioritize photometric fidelity while lacking ex-
 026 plicit semantic representations, which constrains their ability to perform semanti-
 027 cally guided rendering. To this end, we introduce Language-Guided 4D Gaussian
 028 Splatting (L4DGS), a lightweight framework for real-time dynamic scene ren-
 029 dering that integrates natural language into semantically structured 4D Gaussian
 030 representations. Central to L4DGS is a Sparse Multi-Scale Attention (SMSA)
 031 mechanism that enables fine-grained, language-driven control by emphasizing se-
 032 mantic regions across space and time. To enforce semantic fidelity and spatial
 033 coherence, we propose a static regularization that aligns language-
 034 guided features with rendered outputs and ensures consistent depth. To further
 035 ensure temporal consistency, a dynamic regularization penalizes abnormal varia-
 036 tions in semantics and depth over consecutive unit time intervals. L4DGS achieves
 037 a 16.1% improvement in PSNR, reduces perceptual error by 58.8%, and increases
 038 rendering speed by over 50%. Experimental results demonstrate the superiority of
 039 our approach in both visual quality and computational efficiency.
 040

1 INTRODUCTION

041 Recent advances in neural rendering have enabled high-fidelity scene synthesis with remarkable
 042 visual realism Mildenhall et al. (2020); Pumarola et al. (2021); Park et al. (2021b); Li et al. (2022b);
 043 Gao et al. (2021). However, most existing approaches are designed for static settings and lack
 044 explicit semantic representations. This absence of semantic structure constrains their applicability
 045 in interactive, semantically guided, and dynamic environments Knapitsch et al. (2017); Hedman
 046 et al. (2018); Barron et al. (2022). Bridging the gap between human intent and real-time visual
 047 content creation is increasingly critical in computer graphics, impacting applications from semantic
 048 scene editing to immersive VR/AR, interactive media, and human-robot interaction. Addressing
 049 this challenge requires joint reasoning over space, time, and semantics, posing intensive demands
 050 on representation learning, cross-modal alignment, and temporally coherent rendering.

051 Recent advances in 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) have demonstrated the ef-
 052 fectiveness of point-based volumetric representations for real-time, photorealistic scene rendering.
 053 These approaches offer efficient rendering pipelines and high visual fidelity, making them well-
 054 suited for interactive graphics applications. In parallel, efforts to incorporate semantic understand-
 055 ing into 3D scene representations, through language, vision models, or segmentation-guided super-
 056 vision, have enabled controllable generation and semantic editing. However, existing methods for
 057 dynamic scene rendering remain largely semantically agnostic, limiting their ability to align visual
 058 outputs with user intent or language-based descriptions. Moreover, current techniques fall short in
 059 addressing the challenge of maintaining semantic consistency over time, which is critical for ren-
 060 dering dynamic environments that evolve coherently. These limitations underscore the need for a
 061 unified framework that integrates language guidance into temporally consistent 4D representations,
 062 enabling semantically grounded, real-time rendering in dynamic and interactive scenarios.

063 Addressing language-guided rendering in dynamic scenes presents a set of fundamental research
 064 challenges. First, it requires learning a joint representation that aligns visual and linguistic modal-
 065 ities across both spatial and temporal dimensions, despite their inherently different structures and
 066 granularities. Achieving effective cross-modal alignment is challenging due to the semantic am-
 067 biguity of natural language and the limited spatial precision of pretrained vision-language models.

054 Second, incorporating language guidance into the rendering pipeline demands attention mechanisms
 055 that are both expressive and computationally efficient. These mechanisms should modulate scene
 056 content selectively and responsively, enabling real-time control without incurring excessive over-
 057 head. Third, the absence of explicit supervision for dynamic semantics complicates training and
 058 generalization, making it difficult to learn robust semantic representations over time. Furthermore,
 059 ensuring temporal coherence in dynamic scenes requires consistent modeling of object semantics
 060 across time. Separate-time-step supervision often leads to semantic drift, identity instability, or
 061 temporal artifacts such as flickering, especially under motion blur or sparse observations.

062 To address these challenges, we propose Language-Guided 4D Gaussian Splatting (L4DGS), a
 063 lightweight framework that integrates natural language understanding into real-time dynamic scene
 064 synthesis via semantically aware 4D Gaussian representations. Our design is motivated by a key
 065 observation: existing rendering pipelines lack the capacity to incorporate language guidance in a
 066 manner that is both spatially and temporally consistent. These methods largely render static scenes
 067 and are unable to capture the continuous evolution of semantics over time. L4DGS is built upon
 068 a sparse, multi-scale cross-modal attention mechanism that dynamically fuses language and visual
 069 features, guiding both the spatial placement and temporal progression of 4D Gaussian primitives.
 070 This core mechanism is complemented by a hierarchical regularization strategy, wherein both static
 071 and dynamic constraints are modulated by language-conditioned attention maps. These compo-
 072 nents enforce semantic consistency, geometric fidelity, and temporal coherence, enabling L4DGS to
 073 generate renderings that are not only photorealistic but also semantically aligned with user intent.
 074 This unified framework integrates high-level language-guided control with low-level dynamic scene
 075 rendering, ensuring real-time performance with scalability and computational efficiency.

076 To enable semantically guided rendering, we integrate natural language understanding with seman-
 077 tically aware 4D Gaussian representations. Leveraging a hierarchical semantic representation, our
 078 language-guided attention mechanism constrains the Gaussian primitives using object-aware fea-
 079 tures, ensuring that the rendered scene accurately reflects the semantics specified by the language
 080 input. In contrast to existing methods that depend on segmentation masks or external generators,
 081 our approach directly extracts fine-grained visual semantics from the input image using a center-
 082 differenced convolutional network. This network is enhanced with dilated convolutions to expand
 083 the receptive field without additional computational overhead, allowing for efficient context aggre-
 084 gation in high-resolution scenes. Building on these hierarchical semantics, we introduce a Sparse
 085 Multi-Scale Attention (SMSA) mechanism that adaptively aligns language with semantically and
 086 structurally salient visual regions. Rather than relying on dense attention across all tokens, SMSA
 087 incorporates a top- k sparse attention strategy to focus the model’s capacity on the most relevant
 088 spatial features, substantially improving both efficiency and interpretability.

089 To ensure spatial consistency, we further introduce a static regularization scheme that aligns
 090 language-guided visual features with rendered scene features in both magnitude and direction. This
 091 facilitates accurate correspondence between semantic features, such as object descriptions or ac-
 092 tion references, and the visual output. Furthermore, we introduce a static depth regularization term,
 093 modulated by language-conditioned attention, to preserve occlusion relationships, relative object
 094 positions, and the overall 3D scene geometry. To address temporal coherence, we further incorpo-
 095 rate a dynamic regularization strategy tailored for dynamic scene rendering. Unlike static settings,
 096 dynamic scenes demand feature continuity over time to prevent artifacts such as flickering, motion
 097 blur, and semantic drift. Our approach extends both semantic and depth consistency across unit time
 098 intervals, rather than separate time steps, enabling robust alignment of features under fast motion,
 099 occlusions, and sparsely observed regions. This design ensures temporally stable and semantically
 100 meaningful rendering in complex, real-time dynamic environments. In summary, our contributions
 101 are as follows:

- 102 • The introduction of a novel framework that integrates language-guided semantics into dy-
 103 namic scene rendering, addressing the gap between visual features and high-level seman-
 104 tics. To our knowledge, L4DGS is the first language-embedded real-time 4D rendering
 105 algorithm.
- 106 • A sparse multi-scale attention mechanism, leveraging language-guided attention to dynam-
 107 ically align language and visual features across multiple granularities, prioritizing seman-
 108 tically salient regions of the scene.

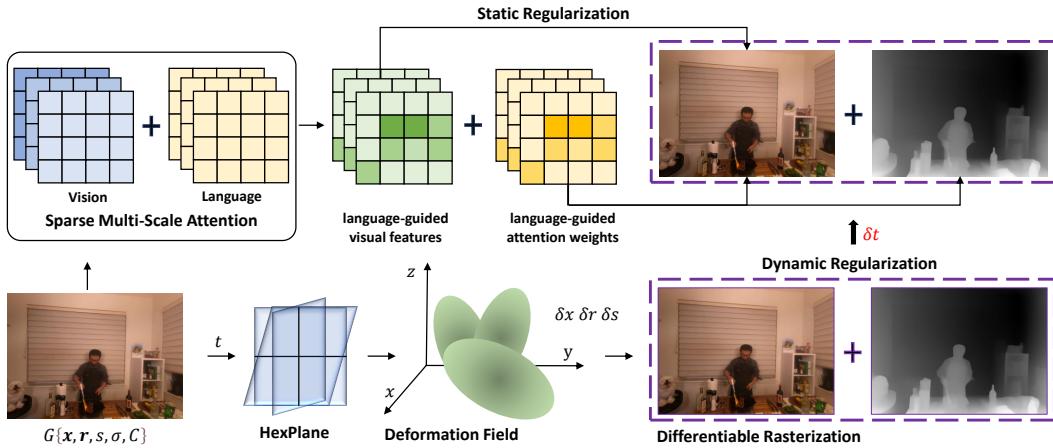


Figure 1: **Framework Overview.** F4DGS leverages a Sparse Multi-Scale Attention (SMSA) mechanism that integrates hierarchical visual features with language embeddings to produce language-guided features and attention weights. Complementary static and dynamic regularization with the attention weights adaptively modulate supervision strength to prioritize semantically salient regions.

- A dynamic regularization that addresses temporal inconsistencies, effectively ensuring smooth transitions of semantic features and depth information across consecutive unit time intervals.
- A static regularization that integrates language-guided semantic and depth features into 4D Gaussians, ensuring efficient optimization while preserving semantic consistency and representational accuracy.

2 RELATED WORK

In this section, we provide an overview of optimization-driven methods for novel view synthesis (NVS), including approaches applicable to both dynamic and static scenes.

Static Novel View Synthesis. Traditional rendering methods, including rasterization, ray tracing, path tracing, and photon mapping, simulate light-object interactions based on physical principles Li et al. (2012); Collet et al. (2015); Kanade et al. (1997); Zitnick et al. (2004). These approaches rely on geometric modelings Riegler & Koltun (2020); Zhou et al. (2018); Flynn et al. (2019); Mildenhall et al. (2019); Srinivasan et al. (2019); Thies et al. (2019); Wood et al. (2023); Kutulakos & Seitz (2000); Penner & Zhang (2017). In contrast, Neural Radiance Field (NeRF) learns a volumetric scene representation through neural networks, eliminating the need for explicit geometric models Du et al. (2021); Gao et al. (2021); Park et al. (2021a;b); Tretschk et al. (2021); Pumarola et al. (2021); Fang et al. (2022); Song et al. (2023). This enables NeRF to efficiently capture complex geometry Li et al. (2022b; 2021); Guo et al. (2023); Tian et al. (2023); Shao et al. (2023). However, NeRFs depend on volumetric rendering integration and repeated forward passes through neural networks Li et al. (2022a); Attal et al. (2023); Fridovich-Keil et al. (2023); Cao & Johnson (2023); Wang et al. (2023b); Gan et al. (2023), resulting in high training cost. 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) improves efficiency by leveraging existing GPU accelerations and 3D Gaussian representations. However, these methods primarily depends on visual features for rendering and lacks direct control over semantic content Luiten et al. (2024); Li et al. (2023). L4DGS addresses this limitation by introducing a language-guided attention mechanism.

Dynamic Novel View Synthesis. Traditional rendering methods convert 3D scenes into 2D images by calculating the interaction between objects and light to achieve realistic visual effects Levoy & Hanrahan (1996);Debevec et al. (1996); Gortler et al. (1996); Seitz & Dyer (1999); Buehler et al. (2001); Waechter et al. (2014). NeRF, in contrast, learns a volumetric scene representation through neural network training Mildenhall et al. (2020); Barron et al. (2021); Verbin et al. (2022);

162 Kopanas et al. (2022); Bemana et al. (2022), eliminating the need for complex modeling. However,
 163 NeRF’s training process is time-consuming Müller et al. (2022); Yan et al. (2023); Fridovich-Keil
 164 et al. (2022); Chen et al. (2022), especially in dynamic scenes, requiring hours to days training.
 165 3DGSS Kerbl et al. (2023) improves rendering efficiency by using efficient GPU optimizations for
 166 real-time performance. However, its extensions struggle with motion blur and scene drift due to its
 167 inability to maintain temporal consistency Yang et al. (2023); Huang et al. (2024); Wu et al. (2023).
 168 L4DGS addresses these limitations by introducing dynamic and static regularization mechanisms,
 169 ensuring geometric consistency.

170

171 3 METHOD

172

173 We introduce Language-Guided 4D Gaussian Splatting (L4DGS), a lightweight framework that in-
 174 tegrates natural language understanding for real-time dynamic scene synthesis using semantically-
 175 aware 4D Gaussian representations. Central to our method is a language-guided attention module
 176 that combines multi-scale CLIP-based visual features with language inputs to guide the spatial and
 177 temporal evolution of 4D Gaussians, incorporating both semantic and depth information.

178

179

180 3.1 LANGUAGE-GUIDED 4D GAUSSIAN SPLATTING

181

182 As illustrated in Fig. 1, L4DGS centers on a language-guided attention mechanism that directs the
 183 model’s focus toward semantically relevant regions of the scene. To enhance visual coherence in
 184 dynamic scene rendering, we first introduce a hierarchical semantic representation that constrains
 185 4D Gaussian distributions using object-aware features. Unlike existing approaches that rely on
 186 segmentation-based masks (e.g., subpart, part, whole) Qin et al. (2024) or external generators such
 187 as SAM Ravi et al. (2024), our method avoids explicit mask supervision. Instead, we adopt a center-
 188 differenced convolutional network Yu et al. (2020) to extract multi-scale CLIP features directly
 189 from the image, capturing fine-grained and illumination-robust semantics. To expand contextual
 190 awareness without incurring additional computational cost, we employ dilated convolutions, which
 191 increase the receptive field and support the processing of high-resolution inputs at varying feature
 192 granularities. This enriched semantic hierarchy informs the placement and refinement of 4D Gaus-
 193 sians, improving both the consistency and fidelity of the rendered output.

194

195 Building on this semantic foundation, L4DGS incorporates a Sparse Multi-Scale Attention (SMSA)
 196 module to adaptively fuse language input (e.g., object references or actions) with visual context (e.g.,
 197 scene layout and object locations). This enables precise localization of user-referenced entities and
 198 allows for fine-grained semantic modulation during rendering. To ensure alignment between seman-
 199 tics and spatial structure, we introduce two complementary static regularization strategies. Static
 200 regularization enforces consistency between language-guided visual features and the rendered scene
 201 content. To further enhance the realism of spatial relationships, we leverage depth features to capture
 202 relative object positions and geometric context, especially in challenging cases involving occlusion
 203 or motion blur. Furthermore, we apply dynamic regularization to enforce temporal coherence. This
 204 term promotes smooth transitions of semantic and depth features across consecutive unit time inter-
 205 vals, addressing issues such as semantic drift and temporal flickering. These two mechanisms ensure
 206 that both spatial and temporal representations evolve coherently, enhancing the realism, stability, and
 207 responsiveness of dynamic scene rendering.

208

209

210 3.2 OPTIMIZATION SCHEME

211

212

213

214

215

Sparse Multi-Scale Attention To address the challenge that dynamic rendering methods lack the
 ability to semantically interpret or interact with the content being rendered, Sparse Multi-Scale At-
 tention (SMSA) facilitates the interaction between language and vision, ensuring that the rendered
 output reflects the user’s intent. To further enhance both efficiency and relevance, we introduce a
 top- k sparse attention mechanism within SMSA that filters out irrelevant tokens and emphasizes
 semantically salient content. The primary function of SMSA in L4DGS is to associate language
 features with visual representations, directing L4DGS’s focus toward spatial regions that are seman-
 tically important. By leveraging multi-scale attention mechanisms, SMSA dynamically adapts to
 different levels of granularity in both the language and visual inputs. This allows L4DGS to pri-

oritize regions that are contextually important, such as objects mentioned in a user’s command or visually salient areas that require attention.

Furthermore, while CLIP encodes images into global semantic features, it lacks fine-grained details and struggles to accurately represent the same object across continuous time intervals. Features extracted from CLIP provide only rough boundaries for different semantic regions, leading to ambiguity and inaccuracies in 4D scene language embeddings. To learn comprehensive semantic features, we begin by extracting the language and vision features into multiple scales, then employ SMSA to compute attention at multiple levels, effectively learning the precise static scene representations. At time t , SMSA selects a subset of salient tokens by computing language and vision modality sparse attention. Given per-head sparse attention for modality $m \in \{v, t\}$, head $h \in \{1, \dots, H\}$, the attention scores can be represented as $A_m^{(h,t)} = \frac{Q_m^{(h,t)} K_m^{(h,t)\top}}{\sqrt{d_h}} \in \mathbb{R}^{L_q \times L_k}$, where Q_m and K_m are the query and key matrices for modality m , respectively. To introduce sparsity, we retain only the top- k key positions for each query token based on the attention scores:

$$\mathcal{I}_i^{(h,t)} = \text{TopKIndices}(A_m^{(h,t)}[i, :]), \quad (1)$$

$$\tilde{A}_m^{(h,t)}[i, j] = \begin{cases} A_m^{(h,t)}[i, j], & \text{if } j \in \mathcal{I}_i^{(h,t)} \\ -\infty, & \text{otherwise} \end{cases}. \quad (2)$$

The sparse attention matrix $\tilde{A}_m^{(h,t)}$ is normalized via softmax as $\bar{A}_m^{(h,t)} = \text{softmax}(\tilde{A}_m^{(h,t)})$ and then multiplied with $V_m^{(h,t)}$ to yield $H_m^{(h,t)} = \bar{A}_m^{(h,t)} V_m^{(h,t)}$. We then concatenate and project back to obtain a visual modality feature $SMSA_v^{(t)} = [H_v^{(1,t)} \parallel \dots \parallel H_v^{(H,t)}] W^O$, where $SMSA_v^{(t)} \in \mathbb{R}^{L_f \times d_f}$ be the rendered features from L4DGS. By incorporating top- k sparsity, SMSA dynamically attends to only the most relevant tokens across multiple spatial scales and modalities, enabling precise localization of language entities, such as the red chair. This approach not only improves the interpretability of attention but also reduces computational overhead, allowing L4DGS to operate efficiently in real-time scenarios.

Attention-Salient Static Regularization Central to L4DGS framework is the use of language-modulated attention weights, derived from the vision branch of the SMSA mechanism. These weights are designed to selectively emphasize spatial regions that are both semantically and structurally important, enabling L4DGS to align features in a targeted and content-aware manner. Unlike uniform regularization strategies, our method utilizes attention scores computed within SMSA to determine which visual regions warrant stronger supervision. At time t , we compute attention weights $w_q^{(t)} \in [0, 1]$ by first aggregating attention scores across all SMSA heads in the vision modality: $\bar{A}_v^{(t)} = \frac{1}{H} \sum_{h=1}^H \tilde{A}_v^{(h,t)}$. We then compute the semantic importance score of each visual token q by measuring the average attention it receives across all queries: $w_q^{(t)} = \frac{1}{L_f} \sum_{i=1}^{L_f} \bar{A}_v^{(t)}[i, q]$. These weights reflect how strongly the visual region q is attended to under language guidance and are further normalized across valid positions: $w_q^{(t)} = \frac{w_q^{(t)}}{\sum_{j=1}^{L_f} M_j^{(t)} \cdot w_j^{(t)}}$. These language-modulated attention weights naturally highlight semantically meaningful structures within the scene, which are then used to modulate the static semantic regularization. To ensure that supervision is concentrated where it matters most for perception or interaction, language-guided static semantic regularization is defined as:

$$\mathcal{L}_{StaticSem}^* = \sum_{q=1}^{L_f} M_q^{(t)} \cdot w_q^{(t)} \left[\|SMSA_{v,q}^{(t)} - F_{\text{rendered},q}^{(t)}\|_2^2 + \lambda \left(1 - \cos(SMSA_{v,q}^{(t)}, F_{\text{rendered},q}^{(t)}) \right) \right]. \quad (3)$$

This regularization enables supervision is concentrated on regions that are semantically salient under language intent, such as object boundaries, interactable elements, and foreground structures.

Furthermore, we extend this attention mechanism to the depth domain by introducing an attention-weighted static depth regularization. The attention weights $w_q^{(t)}$, derived from the language-conditioned SMSA vision stream, are used to emphasize structurally important regions during depth supervision:

$$\mathcal{L}_{StaticDepth}^* = \sum_{q=1}^{L_f} M_q^{(t)} \cdot w_q^{(t)} \left[\|D_{\text{rendered},q}^{(t)} - D_{GT,q}^{(t)}\|_2^2 + \lambda \left(1 - \cos(D_{\text{rendered},q}^{(t)}, D_{GT,q}^{(t)}) \right) \right]. \quad (4)$$

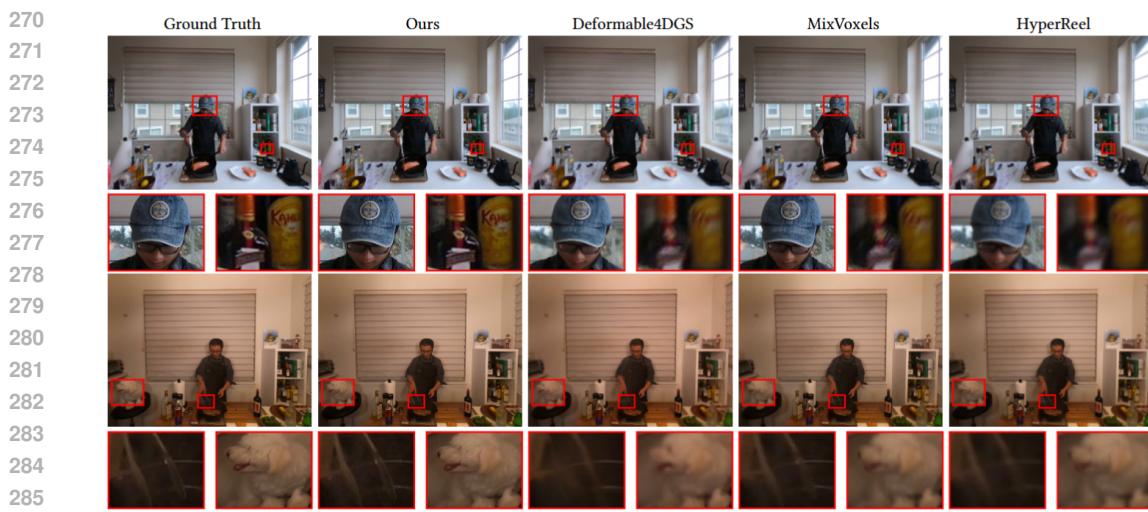


Figure 2: **Qualitative Comparison on Plenoptic Video Dataset.** L4DGS outperforms leading methods in rendering hierarchical visual details, e.g., the letters on a bottle inside a distant cabinet, the frequently moving faces, and the kitchenware.

This formulation enables the model to better resolve spatial relationships and occlusion patterns, focusing depth alignment on regions with higher semantic and structural relevance.

The final attention-salient static regularization is given by:

$$\mathcal{L}_{Static}^* = \mathcal{L}_{StaticSem}^* + \lambda_{Static} \mathcal{L}_{StaticDepth}^*. \quad (5)$$

This language-guided regularization allows L4DGS to perform semantic and geometric supervision in a spatially selective and content-aware manner.

Lifting Representations into the 4D Space
 Ensuring temporal consistency is critical for dynamic scene rendering, especially when modeling deformable or moving objects using 4D Gaussian representations. Unlike 3D Gaussians, which capture spatial structure at a single time step, 4D Gaussians encode both spatial and temporal information. This added temporal dimension introduces challenges: semantic features associated with the same object may drift or become inconsistent across time, especially in cases of rapid motion, occlusion, or limited visual evidence. Separate-time-step supervision alone is insufficient to address these issues, as it does not constrain inter-frame coherence.

To overcome this limitation, we introduce dynamic regularization, a temporal consistency constraint that operates over continuous unit time intervals rather than isolated time steps. Specifically, we enforce smoothness in the evolution of semantic features associated with each Gaussian primitives within unit time intervals.

This regularization penalizes abrupt temporal deviations, encouraging stable and coherent semantic trajectories in both spatial and temporal dimensions. It is applied directly to the learned feature embeddings, promoting continuity in appearance without modifying point cloud density.

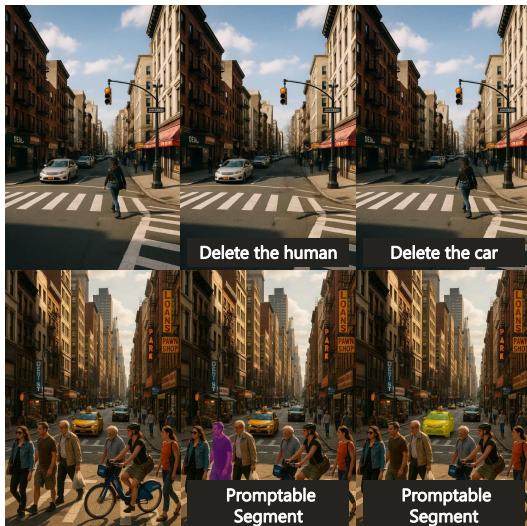


Figure 3: **Language-guided editing and promptable segmentation.**

This temporal smoothing is especially beneficial in challenging regions, such as those affected by motion blur, sparse observations, or disocclusions, where supervision from individual time steps is noisy or unreliable. By preserving the consistency of semantic distributions across unit time intervals, dynamic regularization maintains accurate object identity and geometry, reduces temporal flickering, and enhances the realism of dynamic appearance modeling. Furthermore, as this approach operates over existing Gaussian distributions, it ensures visual fidelity without increasing point cloud density.

Dynamic Regularization To operationalize temporal consistency across 4D Gaussians, we introduce a dynamic regularization objective that directly penalizes inconsistent feature trajectories over time. Specifically, we compute the temporal variations of both the language-guided vision features $SMSA_{v,q}^{(t)}$ and the rendered fused features $F_{\text{rendered},q}^{(t)}$ across unit time intervals δt . For each visual token q , the temporal variations are defined as:

$$\begin{aligned}\delta SMSA_{v,q}^{(\delta t)} &= SMSA_{v,q}^{(t+\delta t)} - SMSA_{v,q}^{(t)}, \\ \delta F_{\text{rendered},q}^{(\delta t)} &= F_{\text{rendered},q}^{(t+\delta t)} - F_{\text{rendered},q}^{(t)}, \\ \delta \text{Direction} &= 1 - \cos(\delta SMSA_{v,q}^{(\delta t)}, \delta F_{\text{rendered},q}^{(\delta t)}).\end{aligned}\quad (6)$$

To promote semantic stability, we penalize abnormal deviations in temporal gradients through a weighted combination of feature magnitude differences and their directional misalignment:

$$\mathcal{L}_{\text{DynamicSem}}^* = \frac{1}{L_f} \sum_{\delta t} \sum_{q=1}^{L_f} M_q^{(t)} w_q^{(t)} \left(\lambda_D \cdot \delta \text{Direction} + \|\delta SMSA_{v,q}^{(\delta t)}\|_2^2 + \|\delta F_{\text{rendered},q}^{(\delta t)}\|_2^2 \right). \quad (7)$$

where $M_q^{(t)} \in \{0, 1\}$ masks out invalid tokens and $w_q^{(t)}$ are the attention-salient weights described previously. This loss encourages Gaussian primitives' temporal coherence in appearance, enabling smooth semantic transitions even in the presence of fast object motion, partial occlusion, or sparse frame sampling. By focusing on temporal feature gradients rather than static states, the regularization captures transferable appearance evolution, which is crucial for rendering high-frequency textures and temporally consistent reflections on dynamic surfaces.

To further ensure physically plausible motion and accurate structural evolution, we extend our formulation with a depth-based regularization that constrains temporal changes in the predicted geometry. We penalize excessive depth fluctuations in Gaussian primitives across consecutive unit time intervals:

$$\mathcal{L}_{\text{DynamicDepth}}^* = \frac{1}{L_f} \sum_{\delta t} \sum_{q=1}^{L_f} M_q^{(t)} w_q^{(t)} \|\delta D_{\text{rendered},q}^{(\delta t)}\|_2^2. \quad (8)$$

where $\delta D_{\text{rendered},q}^{(\delta t)} = D_{\text{rendered},q}^{(t+\delta t)} - D_{\text{rendered},q}^{(t)}$ represents the temporal change in rendered depth.

The full dynamic regularization objective integrates both semantic and geometric components:

$$\mathcal{L}_{\text{Dynamic}}^* = \mathcal{L}_{\text{DynamicSem}}^* + \lambda_{\text{Dynamic}} \cdot \mathcal{L}_{\text{DynamicDepth}}^*. \quad (9)$$

This dynamic regularization enhances the stability and responsiveness of L4DGS in time-varying scenes. It guides the deformation of Gaussians in a physically consistent manner, enabling the model to track non-rigid motions, handle occlusions, and interpolate missing frames in low-frame-rate or sparse input scenarios. By capturing coherent motion trajectories without increasing point cloud density, our method enables high-quality real-time rendering with reduced computational overhead.

Semantic Consistency Finally, to ensure consistent and semantically accurate rendering across both spatial and temporal domains, we define the comprehensive semantic consistency regularization that integrates both static and dynamic regularization components:

$$\mathcal{L}_{\text{4DSemantic}}^* = \mathcal{L}_{\text{Dynamic}}^* + \lambda_{\mathcal{O}} \mathcal{L}_{\text{Static}}^*. \quad (10)$$

where $\lambda_{\mathcal{O}}$ is the learnable hyperparameter. Our approach encourages consistent representation of object semantics and depth throughout the 4D Gaussian field, effectively addressing semantic drift, flickering, and identity instability.

378 **4 EXPERIMENTS**
 379

380 **4.1 DATASETS AND IMPLEMENTATION DETAILS**
 381

382 We assess our method using two
 383 widely recognized datasets, each
 384 presenting distinct challenges in
 385 dynamic scene modeling. **Plenoptic**
 386 **Video Dataset** Li et al. (2022b)
 387 includes six real-world scenes, with
 388 17 to 20 views per scene for training,
 389 and one central view reserved for
 390 evaluation. All images have a resolution
 391 of 1352×1014 Li et al. (2022b).
 392 **D-NeRF Dataset** Pumarola et al.
 393 (2021) consists of monocular video
 394 sequences from eight different
 395 scenes. Each scene contains between
 396 50 and 200 training images, 10 to
 397 20 validation images, and 20 test
 398 images, all resized to a resolution of
 399 800×800 Pumarola et al. (2021). Experiments are run on a single RTX 3090 GPU. The optimization
 400 parameters are fine-tuned using the configuration settings from 3DGS Kerbl et al. (2023).

401 **4.2 RESULTS**

402 **Evaluation on Plenoptic Video Dataset.** We
 403 compare L4DGS with several state-of-the-art
 404 dynamic rendering baselines. As shown in
 405 Table 1, our method achieves the highest render-
 406 ing quality by a notable margin, with a PSNR
 407 of 34.00 and an LPIPS of 0.05, outperforming
 408 leading methods in both fidelity and perceptual
 409 similarity (see Figure 2). L4DGS also demon-
 410 strates superior efficiency in training and infer-
 411 ence: it completes training in just 30 minutes,
 412 over 3x faster than MixVoxels and more than
 413 60x faster than K-Planes, while enabling real-
 414 time rendering at 50 FPS, exceeding existing
 415 baselines. Experimental results highlight that L4DGS not only achieves state-of-the-art visual qual-
 416 ity but also enables substantial gains in computa-
 417 tional efficiency, ensuring high-fidelity, real-time
 418 dynamic scene rendering with language-guided control. Figure 3 further confirms the key advan-
 419 tages of L4DGS. It enables language-driven control for accurate and localized scene editing.

420 **Evaluation on D-NeRF Dataset** We eval-
 421 uate L4DGS against existing dynamic render-
 422 ing methods. As shown in Table 2, L4DGS
 423 achieves the highest overall rendering quality,
 424 with a PSNR of 37.00, SSIM of 0.98, and
 425 LPIPS of 0.02. These results represent a sub-
 426 stantial improvement over all baselines, ex-
 427 ceeding the next-best method by over 4 dB
 428 in PSNR. Beyond accuracy, L4DGS demon-
 429 strates exceptional efficiency, requiring only 5
 430 minutes of training, substantially faster than
 431 other real-time-capable methods such as De-
 432 formable4DGS and TiNeuVox, and enables
 433 real-time rendering at 150 FPS. These findings confirm the effectiveness of L4DGS in jointly ensur-
 434 ing high visual fidelity, rapid training, and real-time performance.



Figure 4: **Qualitative Ablation Study of Different Components in L4DGS.**

Table 1: **Quantitative Comparison on Plenoptic Video Dataset.** We compare L4DGS against lead-
 ing methods. L4DGS obtains the highest PSNR
 while boosting training efficiency. *: trained on 8
 GPUs and tested only on the Flame Salmon scene.

Method	PSNR↑	SSIM↑	LPIPS↓	Train↓	FPS↑
DyNeRF Li et al. (2022b)*	29.58	-	0.08	1344 h	0.015
StreamRF Li et al. (2022a)	28.16	0.85	0.31	79 min	8.50
HyperReel Attal et al. (2023)	30.36	0.92	0.17	9 h	2.00
NeRFPlayer Song et al. (2023)	30.69	-	0.11	6 h	0.05
K-Planes Fridovich-Keil et al. (2023)	30.73	0.93	0.07	190 min	0.10
MixVoxels Wang et al. (2023b)	30.85	0.96	0.21	91 min	16.70
Deformable4DGS Wu et al. (2023)	28.42	0.92	0.17	72 min	39.93
Ours	34.00	0.95	0.05	30 min	50.00

Table 2: **Quantitative Comparison on D-NeRF Dataset.** We compare our approach with lead-
 ing dynamic scene rendering methods. L4DGS
 effectively balances visual quality and training ef-
 ficiency in dynamic scene rendering.

Method	PSNR↑	SSIM↑	LPIPS↓	Train↓	FPS↑
D-NeRF Pumarola et al. (2021)	29.17	0.95	0.07	24 h	0.13
TiNeuVox Fang et al. (2022)	32.87	0.97	0.04	28 min	1.60
K-Planes Fridovich-Keil et al. (2023)	31.07	0.97	0.02	54 min	1.20
FFDNeRF Gu et al. (2023)	31.70	0.96	0.05	-	<1.20
MSTH Wang et al. (2023a)	30.40	0.97	0.05	9.80 min	-
V4D Gan et al. (2023)	32.67	0.97	0.05	10.21 h	2.64
Deformable4DGS Wu et al. (2023)	32.99	0.97	0.05	13 min	104.00
Ours	37.00	0.98	0.02	5 min	150.00

432

433 Table 3: **Ablation Study with Quantitative Comparison on D-NeRF Dataset.** We validate different
434 components in L4DGS on rendering quality PSNR.

ID	Dyn. Sem.	Stat. Sem.	Dyn. Depth	Stat. Depth	Attn. Focus	Jumping Jacks	Mutant	Stand Up
<i>a</i>						34.33	35.87	35.91
<i>b</i>	✓					35.85	37.76	38.03
<i>c</i>		✓				35.47	37.29	37.50
<i>d</i>			✓			35.09	36.81	36.97
<i>e</i>				✓		34.71	36.34	36.44
<i>f</i>	✓	✓	✓	✓		37.96	40.36	41.14
<i>Full</i>	✓	✓	✓	✓	✓	38.44	40.89	41.50

435

436

437 4.3 ABLATION STUDIES

438

439 **Language-Guided Semantics Consistency.**

440 To evaluate the impact of language-guided semantic consistency, we conduct an ablation
 441 study that retains only the static semantic regularization module. As shown in Table 3 (*c*), this
 442 configuration consistently outperforms the non-regularized baseline (Table 3 (*a*)), highlighting
 443 the importance of aligning rendered content with language-conditioned visual features. The
 444 results demonstrate that language-guided spatial attention serves as strong supervision for
 445 rendering salient structures and object boundaries, which is a crucial foundation for achieving
 446 spatiotemporal consistency.

447

448 Furthermore, to evaluate the effect of attention-salient weighting, we compare
 449 the full model (Table 3 *Full*) with its ablated variant (Table 3 (*f*)). The inclusion of the attention-
 450 salient module yields consistent improvements across all scenes, with especially pronounced gains in
 451 detail-rich scenes such as Mutant. These results indicate that attention-aware weighting enhances the
 452 effectiveness and spatial selectivity of supervision, improving cross-modal alignment and rendering
 453 quality in structurally complex regions.

454

455

456 **Dynamic Consistency.** To isolate the impact of temporal semantic consistency, we evaluate a
 457 model variant (Table 3 (*b*)), which shows a notable performance gain over the baseline (Table 3 (*a*)).
 458 This demonstrates the effectiveness of enforcing temporal coherence in semantic space. Among all
 459 components, dynamic semantic supervision yields the highest average PSNR improvement across
 460 scenes, underscoring its central role in addressing temporal flickering and semantic drift. In Figure 4,
 461 qualitative results further reveal that omitting dynamic regularization introduces artifacts such as
 462 motion blur. These findings confirm that temporal feature alignment is essential for robust, high-
 463 fidelity dynamic scene rendering.

464

465

466 5 CONCLUSION

467

468

469

470

471

472

473

474

475 We have presented L4DGS, a language-guided framework for real-time dynamic scene rendering
 476 based on semantically enriched 4D Gaussian representations. By integrating natural language
 477 understanding with hierarchical visual features through a Sparse Multi-Scale Attention (SMSA) mech-
 478 anism, our approach enables language-guided, fine-grained rendering in complex dynamic environ-
 479 ments. To ensure spatial and temporal consistency, we introduce static and dynamic regularization
 480 strategies that align semantic and depth features across both space and time, effectively addressing
 481 temporal semantic drift and inconsistency. Extensive experiments demonstrate that incorporating
 482 language semantics into the rendering pipeline substantially ensures realistic rendering and enhances
 483 scene interpretability, while maintaining comparable training efficiency.

484 Figure 5: **Optical Flow Visualization.**

486 REFERENCES
487

- 488 Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew
489 O’Toole, and Changil Kim. Hyperreal: High-fidelity 6-dof video with ray-conditioned sampling.
490 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
491 16610–16620, 2023.
- 492 Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and
493 Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields.
494 In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864,
495 2021.
- 496 Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf
497 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference
498 on Computer Vision and Pattern Recognition*, pp. 5470–5479, 2022.
- 499 Mojtaba Bemana, Karol Myszkowski, Jeppe Revall Frisvad, Hans-Peter Seidel, and Tobias Ritschel.
500 Eikonal fields for refractive novel-view synthesis. In *ACM SIGGRAPH 2022 Conference Proceed-
501 ings*, pp. 1–9, 2022.
- 502 Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstruc-
503 tured lumigraph rendering. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2001.
- 504 Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings
505 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 130–141, 2023.
- 506 Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance
507 fields. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022.
- 508 Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues
509 Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM
510 Transactions on Graphics (TOG)*, 34(4):1–13, 2015.
- 511 Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from
512 photographs: A hybrid geometry-and image-based approach. *ACM Transactions on Graphics
513 (Proc. SIGGRAPH)*, 1996.
- 514 Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance
515 flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on
516 Computer Vision (ICCV)*, pp. 14304–14314. IEEE Computer Society, 2021.
- 517 Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias
518 Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH
519 Asia 2022 Conference Papers*, pp. 1–9, 2022.
- 520 John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck,
521 Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent.
522 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
523 2367–2376, 2019.
- 524 Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo
525 Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF
526 Conference on Computer Vision and Pattern Recognition*, pp. 5501–5510, 2022.
- 527 Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo
528 Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of
529 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12479–12488, 2023.
- 530 Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4d: Voxel for 4d novel
531 view synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- 532 Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic
533 monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,
534 pp. 5712–5721, 2021.

- 540 Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. *ACM
541 Transactions on Graphics (Proc. SIGGRAPH)*, 1996.
- 542
- 543 Xiang Guo, Jiadai Sun, Yuchao Dai, Guanying Chen, Xiaoqing Ye, Xiao Tan, Errui Ding, Yu-
544 meng Zhang, and Jingdong Wang. Forward flow for novel view synthesis of dynamic scenes. In
545 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16022–16033,
546 2023.
- 547 Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Bros-
548 tow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics
549 (TOG)*, 37(6):1–15, 2018.
- 550
- 551 Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-
552 gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the
553 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4220–4230, 2024.
- 554 Takeo Kanade, Peter Rander, and PJ Narayanan. Virtualized reality: Constructing virtual worlds
555 from real scenes. *IEEE multimedia*, 4(1):34–47, 1997.
- 556
- 557 Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splat-
558 tting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14,
559 2023.
- 560 Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking
561 large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.
- 562
- 563 Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis.
564 Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics
565 (TOG)*, 41(6):1–15, 2022.
- 566
- 567 Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal
568 of computer vision*, 38:199–218, 2000.
- 569
- 570 Marc Levoy and Pat Hanrahan. Light field rendering. *ACM Transactions on Graphics (Proc. SIG-
571 GRAPH)*, 1996.
- 572
- 573 Hao Li, Linjie Luo, Daniel Vlasic, Pieter Peers, Jovan Popović, Mark Pauly, and Szymon
574 Rusinkiewicz. Temporally coherent completion of dynamic shapes. *ACM Transactions on Graph-
575 ics (TOG)*, 31(1):1–11, 2012.
- 576
- 577 Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d
578 video synthesis. *Advances in Neural Information Processing Systems*, 35:13485–13498, 2022a.
- 579
- 580 Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim,
581 Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video
582 synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer
583 Vision and Pattern Recognition*, pp. 5521–5531, 2022b.
- 584
- 585 Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time
586 dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023.
- 587
- 588 Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-
589 time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer
590 Vision and Pattern Recognition*, pp. 6498–6508, 2021.
- 591
- 592 Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians:
593 Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision
3DV*, 2024.
- 594
- 595 B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing
596 scenes as neural radiance fields for view synthesis. In *European conference on computer vision*,
597 2020.

- 594 Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ra-
 595 mamoothi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with
 596 prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.
 597
- 598 Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics prim-
 599 itives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):1–15,
 600 2022.
- 601 Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M
 602 Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of*
 603 *the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021a.
- 604 Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman,
 605 Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for
 606 topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG)*, 40(6):1–12,
 607 2021b.
- 608 Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics*
 609 (*TOG*), 36(6):1–11, 2017.
- 610 Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural
 611 radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer*
 612 *Vision and Pattern Recognition*, pp. 10318–10327, 2021.
- 613 Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d lan-
 614 guage gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
 615 *Pattern Recognition*, pp. 20051–20060, 2024.
- 616 Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham
 617 Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images
 618 and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- 619 Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer*
 620 *Vision*, pp. 623–640, 2020.
- 621 Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *Inter-*
 622 *national Journal of Computer Vision*, 35:151–173, 1999.
- 623 Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Ten-
 624 sor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering.
 625 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
 626 16632–16642, 2023.
- 627 Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and An-
 628 dreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural
 629 radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742,
 630 2023.
- 631 Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah
 632 Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings*
 633 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 175–184, 2019.
- 634 Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis
 635 using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.
- 636 Fengrui Tian, Shaoyi Du, and Yueqi Duan. Mononerf: Learning a generalizable dynamic radi-
 637 ance field from monocular videos. In *Proceedings of the IEEE/CVF International Conference on*
 638 *Computer Vision*, pp. 17903–17913, 2023.
- 639 Edith Treitschke, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and
 640 Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis
 641 of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Con-*
 642 *ference on Computer Vision*, pp. 12959–12970, 2021.

- 648 Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings*
 649 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5481–5490. IEEE,
 650 2022.
- 652 Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of
 653 3d reconstructions. In *European Conference on Computer Vision*, pp. 836–850. Springer, 2014.
- 655 Feng Wang, Zilong Chen, Guokang Wang, Yafei Song, and Huaping Liu. Masked space-time hash
 656 encoding for efficient dynamic scene reconstruction. *Advances in Neural Information Processing*
 657 *Systems*, 2023a.
- 658 Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural vox-
 659 ells for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference*
 660 *on Computer Vision*, pp. 19706–19716, 2023b.
- 662 Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and
 663 Werner Stuetzle. Surface light fields for 3d photography. In *Seminal Graphics Papers: Pushing*
 664 *the Boundaries, Volume 2*, pp. 487–496. 2023.
- 665 Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian,
 666 and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint*
 667 *arXiv:2310.08528*, 2023.
- 669 Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular ob-
 670 jects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*
 671 (*CVPR*), pp. 8285–8295, June 2023.
- 672 Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. De-
 673 formable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint*
 674 *arXiv:2309.13101*, 2023.
- 675 Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoy-
 676 ing Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceed-
 677 ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5295–5305,
 678 2020.
- 680 Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification.
 681 *ACM Transactions on Graphics*, 37(4):1–12, 2018.
- 682 C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski.
 683 High-quality video view interpolation using a layered representation. *ACM transactions on graph-
 684 ics (TOG)*, 23(3):600–608, 2004.
- 686
 687
 688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701