The Structure-Equivalent Prior: Unifying Temporal Dynamics and 3D Evolution in 4D Latent Space

Jingyuan Gao, Tianyu Shen*, Ruosen Hao, Te Guo, Zhiwei Li, Kunfeng Wang*

College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China
{jygao, tianyu.shen, rshao, te.guo, lizw, wangkf}@buct.edu.cn

Abstract

Recent advances in deep learning-based 3D representation have achieved remarkable success, particularly in modeling static high-fidelity geometries. However, the extension of these techniques to dynamic 3D scenes introduces a critical challenge of effectively representing spatio-temporal dependencies, i.e., jointly modeling detailed spatial structures within frames and temporal dynamics across frames. To address this challenge, this paper proposes that the temporal evolution observed in dynamic 3D scenes is fundamentally attributable to the deformation of underlying spatial structures. To capture this relationship, we introduce a unified continuous 4D latent space representation incorporating a structure-equivalence prior, named SEP-4D. The core of SEP-4D is an efficient 4D tensor decomposition-fusion approach. This method fuses decomposed learnable 2D feature planes via a plane-wise spatio-temporal fusion mechanism of planar distributions, explicitly enforcing the principle that temporal evolution originates from geometric deformations of the 3D structure. To mitigate the associated computational demands, we sample the 3D probability volumes generated by VAE-based fusion into a spatio-temporally consistent 4D latent representation. The efficacy of our approach is validated through experiments on the fundamental task of 4D occupancy reconstruction. Extensive results demonstrate that, by leveraging the inherent equivalence of temporal dynamics and structural deformation, our method achieves high-quality reconstruction across various sequence lengths. Notably, for 4-frame scenes, we attain an impressive 91.68% mIoU, significantly outperforming state-of-the-art baselines on standard benchmarks.

Code — https://github.com/BUCT-IUSRC/SEP-4D

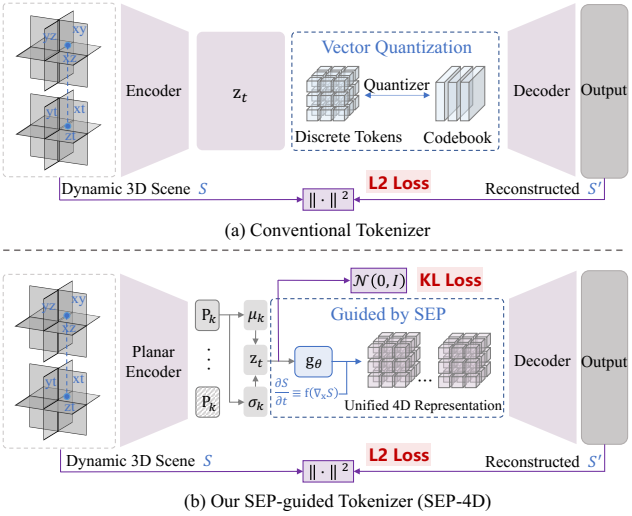

Figure 1: Comparison between (a) conventional tokenizer and (b) our proposed SEP-guided 4D representation tokenizer (SEP-4D). In (a), an encoder transforms the input dynamic 3D scene into latent embedding z_t, which are directly quantized into discrete tokens. In (b), SEP-4D employs a planar encoder to fuses learnable feature planes P_k via a plane-wise spatio-temporal fusion mechanism. By SEP guided, SEP-4D represents dynamic 3D scene into a unified continuous 4D latent space representation, which enforcing that temporal evolution originates from geometric deformations of the 3D structure.

Introduction

Understanding and representing dynamic 3D scenes is a fundamental challenge in computer vision, with profound implications for autonomous driving (Wang et al. 2024; Bian et al. 2024), robotics (Mittal et al. 2023), medicine (Lai et al. 2025), and AR/VR applications (Engel et al. 2023). While neural fields (Mildenhall et al. 2021; Yariv et al. 2021) and discrete representations (Choy et al. 2016; Groueix et al. 2018; Achlioptas et al. 2018; Fan, Su, and Guibas 2017) have advanced static 3D reconstruction, extending these to dynamic scenes introduces prohibitive computational costs and unstructured optimization challenges (Cao and Johnson 2023). Critically, dynamic 3D scenes demand rigorous spatio-temporal consistency, where physical plausibility requires continuous geometric deformation rather than discrete frame-by-frame approximations.

Recent approaches have extensively investigated solutions to achieve spatio-temporal consistency. As shown in Figure 1 (a), exploiting the natural autoregressive property of videos, some methods (Zheng et al. 2024; Zhu et al.

*Corresponding Authors

2025; Zhuo et al. 2025; Pumarola et al. 2021) leverage vector quantization with learnable codebooks (Van Den Oord, Vinyals et al. 2017) to discretize 3D representations into tokens. These approaches recursively derive current token representations from previous ones. Although achieving impressive results, their recursive pipelines incur unpredictable computational overhead and inference latency when generating fine-grained 4D representations, along with severe error accumulation in long-sequence reconstruction (Liao et al. 2025). More critically, the autoregressive paradigm neglects a fundamental constraint of dynamic 3D reconstruction: spatial structures evolve under stricter temporal continuity than sequential video data.

To tackle the limitation of weak integration of motion and structure in dynamic 3D space, recent works have explored structured representations such as K-Plane (Fridovich-Keil et al. 2023) demonstrates that a d-dimensional scene can be decomposed into k planar representations, while HexPlane (Cao and Johnson 2023) further validates the effectiveness of compressing spatio-temporal data into six 2D feature planes aligned with spatial and spatio-temporal axes. However, simple combinations (e.g., concat, hadamard product) of 2D planes often fail to capture the intricate interdependencies between spatial and temporal dimensions, resulting in sub-optimal reconstruction and limited generalization to unseen scenarios.

In this context, this work posits that a unified 4D latent representation can more effectively model structural dependencies than simple combinations of 2D planes. This advantage stems from a strong prior: **the temporal evolution in dynamic 3D scenes fundamentally corresponds to the deformation of underlying spatial structures.** This insight naturally raises the question: *Can we construct a unified, continuous 4D latent space that intrinsically encodes such structured spatio-temporal regularity, enabling efficient dynamic 3D reconstruction?* Therefore, addressing this core question, we argue that an effective unified 4D representation must satisfy the following conditions: 1) maintain the fine-grained structure of the high-dimensional 4D tensor (3D + temporal 1D); 2) ensure intrinsic temporal coherence through a unified spatio-temporal representation. Unfortunately, none have yet achieved all of these objectives simultaneously.

To this end, this paper proposes a unified, continuous 4D latent space representation for dynamic 3D scene reconstruction, by constraining the inherent spatial regularity of these scenes within the latent space. As shown in Figure 1 (b), our core innovation is that the dynamic 3D scene is encoded into a highly compressed latent 3D probability space following a unified distribution, which is achieved by a feature-plane-based representation combined with a VAE-based compressor. Subsequently, the results sampled from this 3D probability space are transformed into a unified structured 4D latent space representation with a plane-wise spatio-temporal fusion mechanism of planar distributions. In contrast to simple combinations of 2D planes, which fail to model complex spatio-temporal dependencies, our approach enforces the key prior that the temporal evolution of dynamic 3D scenes is governed by smooth geometric transformations rather than arbitrary per-frame changes. Extensive experiments demonstrate that our method outperforms other state-of-the-art methods in terms of reconstruction accuracy and computational efficiency in 4D occupancy reconstruction, particularly excelling in long-sequence scenarios.

Our main contributions can be summarized as follows:

- A structure-equivalent prior (SEP) is proposed that formally establishes the intrinsic equivalence between temporal dynamics and spatial structural evolution in dynamic 3D scenes. This foundational principle addresses the critical limitation of weak spatio-temporal integration observed in existing approaches.
- Building on SEP, we develop a continuous 4D latent space representation (SEP-4D) constructed through probabilistic plane decomposition and plane-wise spatio-temporal fusion. Our representation uniquely enforces that temporal transitions originate from smooth geometric deformations.
- Superior performance is demonstrated through validation of the intuitive benchmark task of 4D occupancy reconstruction. SOTA alternatives across varying sequence lengths while requiring only two NVIDIA A6000 GPUs.

Related Work

Dynamic 3D Reconstruction

Dynamic 3D reconstruction aims to recover the evolving 3D geometry and appearance of a scene or object from multiview observations captured over time. Early neural approaches modeling scenes in the 3D domain with time to address dynamic characteristics. NeRF-T (Wang et al. 2021; Xian et al. 2021) consider time as an additional input dimension to NeRF representation (Shao et al. 2023), while D-NeRF (Pumarola et al. 2021) maps deformations of non-rigid scenes at different time instances to a shared, static canonical space via a deformation network, followed by standard NeRF-style rendering in this space by a canonical network. TiNeuVox (Fang et al. 2022) introduces a highly efficient framework by leveraging time-aware neural voxels to overcome the limitations of explicit representations in modeling temporal dynamics. However, these NeRF variants (Müller et al. 2022; Sun, Sun, and Chen 2022; Liu et al. 2020) still bear a non-negligible latency in the rendering process (Wu et al. 2024).

Recently, some studies have explored the use of VAE tokenizers to represent dynamic 3D scenes for high-quality representations in downstream tasks. OccWorld (Zheng et al. 2024) tokenizes 3D occupancy into discrete tokens using a reconstruction-based scene tokenizer, then employs an autoregressive transformer to sequentially generate future scene and ego tokens to decode upcoming occupancy states and trajectories. I^2-World (Liao et al. 2025) introduces dual spatio-temporal tokenizers and a controlled encoder-decoder architecture to autoregressively generate consistent 4D occupancy forecasts. To avoid the appearance drift introduced by the autoregressive architecture, AR4D (Zhu et al. 2025) incorporates a refinement stage based on a global deformation field and the geometry of the 3D representation of each

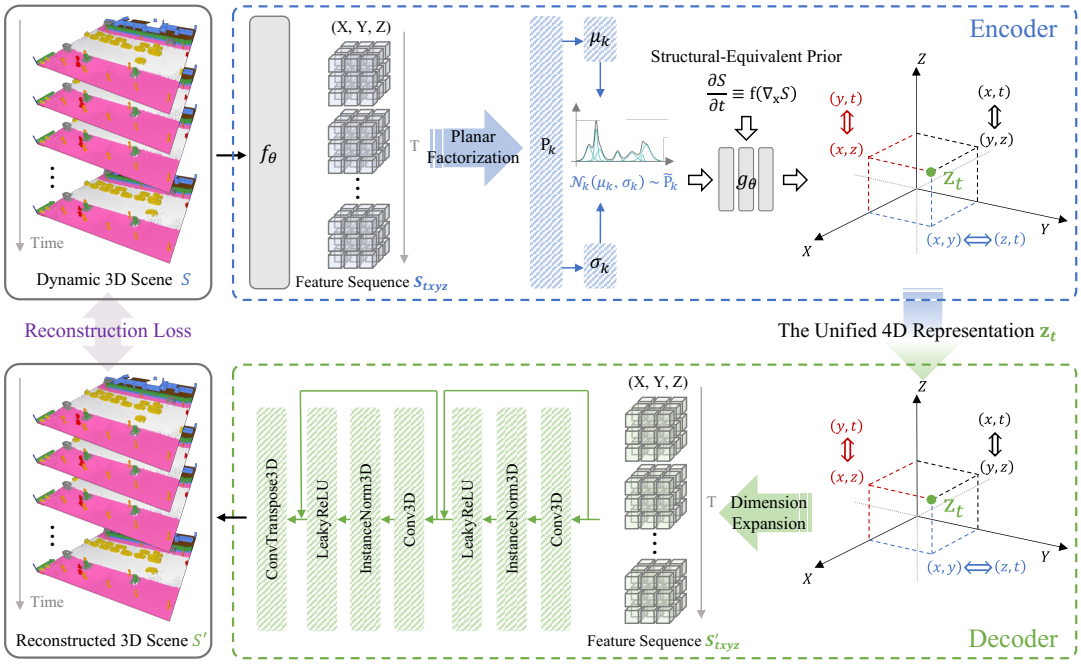

Figure 2: The overall framework of proposed **SEP-4D**. **Encoder**: Convolutional network f_θ encodes dynamic 3D scenes into feature planes P_k via planar factorization. **Fusion**: Unified 4D representation z_t is obtained by sampling P_k distributions with plane-wise spatio-temporal fusion mechanism g_θ. **Decoder**: After expansion in time dimension, z_t is decoded through multi-layer residual networks for scene reconstruction.

frame. Meanwhile, VAT (Zhang, Xiong, and Xu 2024) investigates the feature collapse problem that arises when extending 2D tokenizers to 3D representations, where the inherent lack of sequential order in 3D data impedes compression into fewer tokens while preserving structural details. We observe that all these recently methods typically leverage VAE-based tokenizers to quantize 3D representations into discrete tokens, enabling autoregressive or masked generative modeling within this latent space. However, discrete tokenization of continuous 3D geometry inherently compromises the representation of time-varying fine structural details (Zhang, Xiong, and Xu 2024). In contrast to these token-based approaches, our work establishes a continuous latent space for dynamic 3D scene reconstruction, breaking the limitations of discrete representations.

Structured 4D Representations

Representing dynamic 3D scene remains a challenging topic due to the curse of dimensionality. An increasingly popular direction focuses on factorizing 4D spatio-temporal representations into structured and learnable components. This class of methods is motivated by the observation that dynamic 3D scenes often exhibit strong geometric regularities and smooth temporal evolution, which can be exploited through lower-dimensional decompositions to improve both efficiency and generalization (Chan et al. 2022). D-TensoRF (Jang and Kim 2022) considers the radiance field of a dynamic scene as a 5D tensor and decomposes the grid into rank one vector components or low-rank matrix components. Tensor4D (Shao et al. 2023) projects a unified 4D tensor into three temporal-aware volumes, further decomposed into nine compact feature planes for dynamic scene reconstruction and rendering. HexPlane (Cao and Johnson 2023) shows that dynamic 3D scenes can be explicitly represented by six planes of learned features, which is less redundant compared to Tensor4D. While K-Plane (Fridovich-Keil et al. 2023) presents the first interpretable model capable of representing radiance fields in arbitrary dimensions by decomposing any d-dimensional scene into k planar representations. Inspired by HexPlane, 4DGS (Wu et al. 2024) efficiently decomposes 4D neural voxels and constructs Gaussian features, followed by a MLP that predicts Gaussian deformations at new timestamps. Despite their effectiveness, these methods mostly rely on oversimplified fusion mechanisms, such as concatenation or element-wise operations, which fail to capture the complex nonlinear dependencies between space and time, especially in chaotic motion.

In this paper, we introduce a unified and continuous 4D latent representation guided by SEP, which extends the benefits of structured decomposition and enforces temporal evolution as a continuous deformation of the underlying spatial structure.

Method

Preliminaries: VAE Tokenizer for Representation

Recent approaches that employ VAE-based tokenizers to represent image features have achieved remarkable results.

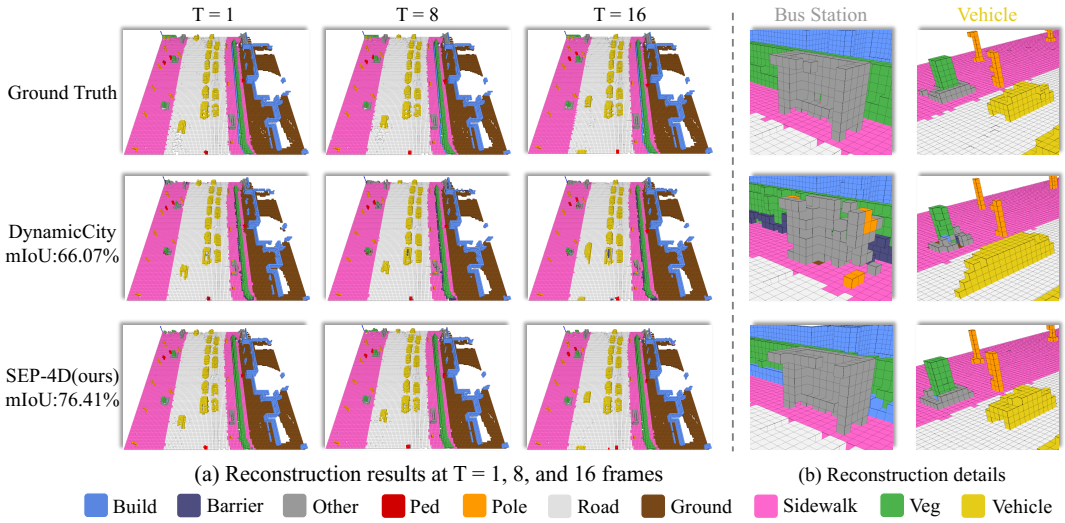

Figure 3: Qualitative visualization of reconstruction results on CarlaSC. (a) is the reconstruction scenes from the 1st, 8th, and 16th frames. (b) visualizes detailed reconstructions of the 16th frame. It can be observed that, owing to the structure-equivalent prior, SEP-4D effectively captures structural features in static and dynamic scenarios, whereas DynamicCity produces inferior reconstructions.

Notably, NAR (He et al. 2025) investigates how the inherent spatial and temporal locality of tokens and their neighbors influences 2D image and 3D video representations, explicitly modeling the Manhattan distance between tokens. This naturally raises the question of whether such a temporal–structure prior could similarly enhance 4D representations. Furthermore, VAT (Zhang, Xiong, and Xu 2024) examines the issue of feature collapse that arises when extending discrete VAE tokenizers from 2D representation to 3D. These observations motivated us to propose the structure-equivalent prior (SEP). By embedding SEP directly into the 4D latent space representation, we aim to better support downstream tasks.

SEP-guided 4D Latent Space Representation

Dynamic 3D scenes inherently couple temporal evolution with spatial structural changes—a fundamental physical prior overlooked by existing 4D representations. Formally, let a dynamic 3D scene be represented as a time-varying function $\mathcal{S}(\mathbf{x}, t) : \mathbb{R}^3 \times \mathbb{R}^+ \mapsto \mathbb{R}^n$, where $\mathbf{x} \in \mathbb{R}^3$ denotes spatial coordinates and $t \in \mathbb{R}^+$ time. Therefore, this physical prior can be formulated as:

$$\frac{\partial \mathcal{S}}{\partial t} \equiv \mathbf{f}\left(\nabla_{\mathrm{x}} \mathcal{S}\right) \tag{1}$$

while $\frac{\partial \mathcal{S}}{\partial t}$ represents the temporal rate of change of the scene, $\nabla_{\mathrm{x}} \mathcal{S}$ denotes the spatial gradient, $\mathbf{f}$ is an unknown deformation functional governing how spatial gradients drive temporal changes. This axiom implies that infinitesimal temporal evolution $\Delta \mathcal{T}$ must derive from smooth spatial deformations $\Delta \mathcal{X}$.

Conventional approaches such as HexPlane (Cao and Johnson 2023) and K-Plane (Fridovich-Keil et al. 2023) decouple temporal and spatial factors through simple compositions (e.g., concatenation or Hadamard products), which can be formulated as:

$$\mathrm{z}_t^{(\text{plane})} = \bigoplus_{k=1}^{6} \mathrm{P}^{(k)} \tag{2}$$

while $\mathrm{P}^{(k)} \in \mathbb{R}^2$ are 2D feature planes, $\oplus$ denotes simple composition, z_t represents a d-dimensional latent vector. However, this formulation violates the SEP because it fails to enforce the intrinsic relationship between the temporal derivative $\frac{\partial}{\partial t}$ and the spatial gradient ∇_{x} in Equation (3), making it impossible to ensure spatio-temporal continuity and physical plausibility.

However, Equation (2) violates the core prior proposed in this paper, SEP. This is because, after simply combining multiple 2D feature planes, there is no inherent relationship between the temporal derivative $\frac{\partial}{\partial t}$ and the spatial gradient ∇_{x} in Equation (3), making it impossible to ensure spatio-temporal continuity and physical plausibility, leading to reconstruction artifacts and poor generalization. As shown in Figure 3, the encoder inherited from HexPlane in DynamicCity (Bian et al. 2024) exhibits significant detail distortion.

$$\frac{\partial}{\partial t}\left(\mathrm{P}^{(xy)} \oplus \mathrm{P}^{(tx)}\right) \not\propto \nabla_{\mathrm{x}}\left(\mathrm{P}^{(xy)} \oplus \mathrm{P}^{(tx)}\right) \tag{3}$$

As shown in Figure 1, to bridge this gap, we introduce a VAE encoder $\mathcal{E}$ that projects dynamic 3D scenes into a unified 4D latent space:

$$\mathrm{z}_t = \mathrm{g}_\theta(\mathcal{E}(\mathcal{S}(\cdot, t))); \mathrm{z}_t \in \mathbb{R}^d \tag{4}$$

while g_θ is a learnable fusion network that represents our fusion mechanism. Therefore, this design ensures the SEP in Equation (5). The operator $\mathbf{h}$ represents the spatio-temporal Jacobian transformation intrinsically defined by the fusion

Dataset	Resolution	Frames	Classes	OccSora	DynamicCity		**SEP-4D (ours)**	
				mIoU	mIoU	Training Time	mIoU	Training Time
CarlaSC	128×128×8	4	10	41.01%†	76.25%	1912 s	**91.68% (+15.43%)**	1474 s
	128×128×8	8	10	39.91%†	70.61%	2509 s	**85.33% (+14.72%)**	2712 s
	128×128×8	16	10	33.40%†	66.07%	2037 s	**76.41% (+10.34%)**	2585 s
	128×128×8	32	10	28.91%†	59.31%†	5891 s	**68.47% (+9.16%)**	6639 s
	128×128×8	64	10	-	47.17%	11156 s	**56.80% (+9.63%)**	13208 s

Table 1: We compare the 4D scene reconstruction performance between OccSora, DynamicCity and our SEP-4D framework across the CarlaSC datasets. Results are reported under varying voxel resolutions and sequence lengths. Scores marked with † are taken directly from the DynamicCity paper, other results are reproduced using the official codebase.

mechanism g_θ, which establishes a differentiable correspondence between spatial structural deformations $\nabla_{\mathrm{x}}\mathrm{z}_t$ and temporal evolution rates $\frac{\partial \mathrm{z}_t}{\partial t}$.

$$\frac{\partial \mathrm{z}_t}{\partial t} = \mathbf{h}\left(\nabla_{\mathrm{x}}\mathrm{z}_t\right) \tag{5}$$

The Unified 4D Reconstruction Framework

Planar Factorization Encoder. As shown in Figure 2, given input dynamic 3D scene $\mathcal{S}(\mathrm{x}, t)$, we first utilize a shared 3D convolutional feature extractor $f_\theta(\cdot)$ to extract and downsample features from each occupancy frame, resulting in a feature sequence $\mathcal{S}_{txyz} \in \mathbb{R}^{T\times C\times X\times Y\times Z}$.

To maintain the fine-grained structure of the high-dimensional 4D tensor, we decompose $\mathcal{S}_{txyz}$ into six orthogonal planes:

$$\mathbf{P} = \{\mathbf{P}_{xy}, \mathbf{P}_{xz}, \mathbf{P}_{yz}, \mathbf{P}_{tx}, \mathbf{P}_{ty}, \mathbf{P}_{tz}\} = \mathcal{PF}(\mathcal{S}_{txyz}) \tag{6}$$

where $\mathcal{PF}$ denotes the planar factorization operator. Each plane is encoded into a probability distribution via VAE in Equation (7) and then sample from these distributions in Equation (8).

$$\mathcal{N}_k(\mu_k, \sigma_k) = \mathcal{E}_k(\mathbf{P}_k), \quad k \in \{xy, xz, yz, tx, ty, tz\} \tag{7}$$

$$\tilde{\mathbf{P}}_k \sim \mathcal{N}_k(\mu_k, \sigma_k) \tag{8}$$

The sampled planes are fused through a plane-wise spatio-temporal fusion mechanism of planar distributions:

$$\mathrm{z}_t = \mathrm{g}_\theta\left(\tilde{\mathbf{P}}_{xy}, \tilde{\mathbf{P}}_{xz}, \tilde{\mathbf{P}}_{yz}, \tilde{\mathbf{P}}_{tx}, \tilde{\mathbf{P}}_{ty}, \tilde{\mathbf{P}}_{tz}\right) \tag{9}$$

We employ an MLP within the plane-wise spatio-temporal fusion mechanism $\mathbf{g}_\theta$ to establish a mapping between spatial structure and temporal evolution. The fusion mechanism $\mathbf{g}_\theta$ enforces the SEP through its architectural design:

$$\mathbf{g}_\theta(\cdot) = \mathrm{MLP}\left(\sum_{k\in\mathcal{S}_{\text{spatial}}} \phi_k \otimes \tilde{\mathbf{P}}_k, \sum_{k\in\mathcal{S}_{\text{spatio-temporal}}} \psi_k \odot \tilde{\mathbf{P}}_k\right) \tag{10}$$

where $\mathcal{S}_{\text{spatial}} = \{xy, xz, yz\}$, $\mathcal{S}_{\text{spatio-temporal}} = \{tx, ty, tz\}$, ϕ_k, ψ_k represent Geometry-aware attention weights, $\otimes$ and $\odot$ respectively enforce spatial structural consistency and temporal coherence.

Through this learnable fusion network g_θ, we obtain a 4D latent space representation by sampling from a unified distribution. The latent vector z_t in Equation (9) therefore takes the form of a 3D tensor, $\mathrm{z}_t \in \mathbb{R}^{C\times X\times Y\times ZT}$.

Multi-layer Residual Decoder. To reconstruct $\mathcal{S}^{'}$ from the learned 4D latent space representation z_t, we first expand the channel dimension of $\mathrm{z}_t^{(\text{exp})}$ to match the shape of $\mathcal{S}_{txyz}$ in Equation (6):

$$\mathrm{z}_t^{(\text{exp})} = \mathrm{g}_\theta(\mathcal{E}(\mathrm{z}_t)) \in \mathbb{R}^{T\times C\times X\times Y\times Z} \tag{11}$$

This expanded latent representation is then processed by a hierarchical 3D transposed convolutional network that progressively upsamples the features while maintaining structural fidelity. To stabilize training and improve the representational capacity of the decoder, we adopt residual connections within each convolutional block. Specifically, each block learns the residual mapping over the input feature, allowing the model to preserve low-level spatial details while refining high-level semantic representations. This design helps mitigate feature collapse and oversmoothing in high-dimensional data reconstruction.

During training, both encoder parameters $\{\mathcal{E}_k, \mathrm{g}_\theta\}$ and decoder parameters $\mathcal{D}_\theta$ are jointly optimized, enabling simultaneous compression of dynamic 3D scenes into unified 4D latent representations and SEP-guided reconstruction that intrinsically couples spatial deformations with temporal evolution.

Experiments

Experimental Setups

Datasets. CarlaSC (Wilson et al. 2022) dataset contains 6 training scenes, each duplicated into Light, Medium, and Heavy based on traffic density. Each scene lasts approximately 180 seconds and is sampled at a frequency of 10 Hz. This dataset contains 22 semantic categories, and the scene resolution is $128 \times 128 \times 8$, covering a region 25.6 meters around the ego vehicle, with a height of 3 meters.

Evaluations Metrics. We evaluate the reconstruction quality of our VAE-based model using standard 3D segmentation metrics. Specifically, we report the overall Intersection over Union (IoU) across the entire voxel space to assess holistic accuracy, and the mean Intersection over Union (mIoU) across semantic categories to measure category-wise fidelity.

Method	mIoU (%)	Building	Barrier	Other	Pedestrian	Pole	Road	Ground	Sidewalk	Vegetation	Vehicle
Resolution: $128 \times 128 \times 8$		**Sequence Length: 8**†									
OccSora	39.910	33.001	3.260	5.659	19.224	19.357	93.038	57.335	85.551	30.899	51.776
DynmicCity	76.181	70.874	50.025	52.433	87.958	85.866	97.513	83.074	93.944	58.626	81.498
SEP-4D (ours)	**85.330**	**76.858**	**52.433**	**81.788**	**96.255**	**92.638**	**98.542**	**93.703**	**97.318**	**71.955**	**91.810**
Improv.	+9.149	+5.984	+2.408	+29.355	+8.270	+6.772	+1.029	+10.629	+3.374	+13.329	+10.321
Resolution: $128 \times 128 \times 8$		**Sequence Length: 16**†									
OccSora	33.404	19.264	2.205	3.454	11.781	9.165	92.054	50.077	82.594	18.078	45.363
DynmicCity	74.223	66.852	51.901	**49.844**	79.410	82.369	96.937	84.484	94.082	58.217	78.134
SEP-4D (ours)	**76.408**	**68.433**	**55.092**	44.034	**80.445**	**84.691**	**97.843**	**85.101**	**95.020**	**60.428**	**92.993**
Improv.	+2.175	+1.581	+3.182	-5.810	+1.035	+2.322	+0.906	+0.617	+0.938	+2.211	+14.859
Resolution: $128 \times 128 \times 8$		**Sequence Length: 32**†									
OccSora	28.911	16.565	1.413	0.944	6.200	4.150	91.466	43.399	78.614	11.007	35.353
DynmicCity	59.308	52.036	25.521	29.382	56.811	57.876	94.792	78.390	89.955	46.080	62.234
SEP-4D (ours)	**68.473**	**57.428**	**33.551**	**37.587**	**71.618**	**68.744**	**96.104**	**81.383**	**97.245**	**57.472**	**83.598**
Improv.	+9.165	+5.392	+8.03	+8.205	+14.807	+10.868	+1.312	+2.993	+7.29	+11.392	+21.364
Resolution: $128 \times 128 \times 8$		**Sequence Length: 64**									
DynmicCity	47.167	44.869	10.533	10.294	39.476	32.211	86.832	68.131	83.453	42.178	53.693
SEP-4D (ours)	**56.798**	**50.137**	**13.110**	**19.186**	**44.793**	**43.224**	**93.782**	**83.883**	**89.179**	**48.252**	**82.434**
Improv.	+9.631	+5.268	+2.577	+8.892	+5.317	+11.013	+6.95	+15.752	+5.726	+6.074	+28.7

Table 2: Comparisons of Per-Class IoU Scores. We compared the performance of OccSora (Wang et al. 2024), DynamicCity (Bian et al. 2024) and our SEP-4D framework on CarlaSC (Wilson et al. 2022) across 10 semantic classes. Scores marked with † are taken directly from the DynamicCity paper; other results are reproduced using the official codebase.

Implementation Details. Our experiments are conducted using two NVIDIA A6000 GPUs. During the VAE training stage, the batch size is set to 2 for sequence lengths of 4 and 8, and 1 for sequence lengths of 16, 32, and 64. The loss weights for Lovász-softmax and cross-entropy are set to 1, while the weight for the KL term is set to 0.005. The initial learning rate for VAE is set to 10^{-3}, and we adopt the MultiStepLR scheduler.

Dynamic 3D Scene Reconstruction

Comprehensive Validation Across Various Sequence Lengths. To further verify the superiority of our method on different sequence length occupancy reconstruction tasks, we conducted comprehensive experiments on the CarlaSC (Wilson et al. 2022) dataset with varying numbers of input frames: 4, 8, 16, 32, and 64. As summarized in Table 1, our method consistently outperforms DynamicCity (Bian et al. 2024), achieving significant mIoU improvements of 15.43, 14.72, 10.34, 9.16, and 9.63. Whether using short or long input sequences, SEP-4D demonstrates stronger reconstruction capability, indicating the robustness of our approach across different temporal scales. Notably, even when given an ultra-long input sequence of 64 frames, our model still maintains high mIoU and IoU scores, showcasing strong robustness and excellent scalability.

To better illustrate the model's temporal modeling capacity, Figure 3 presents visualizations of predicted frames at different time steps ($t = 1, 8, 16$) under the 16-frame input setting. Figure 5 presents visualizations of SEP-4D vs. DynamicCity (Bian et al. 2024) across different sequence lengths. SEP-4D achieves the best reconstruction accuracy across all sequence lengths under comparable training time. In summary, our method consistently produces structurally more accurate and temporally coherent reconstructions across all time steps, further demonstrating its superior temporal modeling capability and consistent semantic representation over time.

The mIoU of specific categories. SEP-4D achieves higher reconstruction accuracy across different semantic categories, we report the mIoU results for each category, as shown in Table 2. SEP-4D achieves significant performance improvements across all semantic categories. This demonstrates that SEP-4D does not rely on a few high-frequency categories to boost the overall mIoU, but possesses strong generalization and modeling capabilities across all categories, thereby achieving a more balanced and robust reconstruction performance.

Convergence rate. SEP-4D framework exhibits superior training convergence efficiency in the occupancy reconstruction task. As shown in Figure 4, SEP-4D converges signifi-

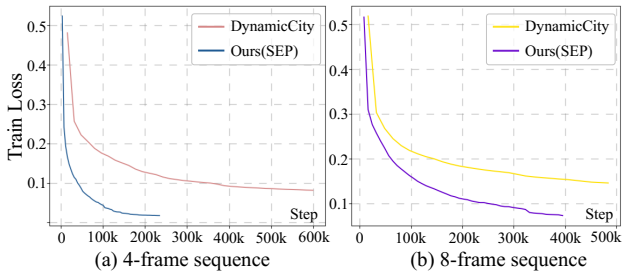

Figure 4: Training dynamics of SEP-4D vs. DynamicCity on CarlaSC. The x-axis is the training steps and the y-axis is the training loss. SEP-4D converges faster than DynamicCity in the early training phase and performs better in the long run.

Fusion Method	mIoU (%)	IoU (%)
Concate	63.71	91.07
Conv Fusion	74.78	91.66
Averaging	75.39	92.31
Dynamic Weighting	75.61	91.98
Ours	**76.41**	**92.32**

Table 3: Comparison of fusion mechanisms on CarlaSC.

cantly faster than DynamicCity (Bian et al. 2024) under both 4-frame and 8-frame input settings, requiring fewer training epochs to reach stable performance and achieving lower final errors. This demonstrates that our architectural design not only improves reconstruction quality but also accelerates the training process, effectively reducing time and resource consumption. Such efficiency is particularly valuable for large-scale scene reconstruction and time-sensitive applications.

Ablation Study

Fusion Strategy Comparison. The first ablation study investigates the impact of different fusion strategies on the performance of the model, as shown in Table 3. We compare four fusion methods: Conv1D fusion, MLP fusion, simple averaging, and dynamic weighting. The results demonstrate that our plane-wise spatio-temporal fusion mechanism consistently achieves the best performance across all metrics. This approach effectively enhances the collaborative modeling among the six input planes and maximally preserves the underlying spatio-temporal dependencies.

Spatio-temporal Modeling Strategy. The second ablation study focuses on comparing different approaches for capturing spatio-temporal structure. We compared the planar factorization modeling with a baseline method that directly applies Conv3D and SE modules on the full voxel input. Table 4 shows that the planar factorization modeling strategy significantly outperforms the baseline in both mIoU and IoU, validating its effectiveness in capturing complex spatio-temporal structures and improving the precision and fidelity of scene reconstruction.

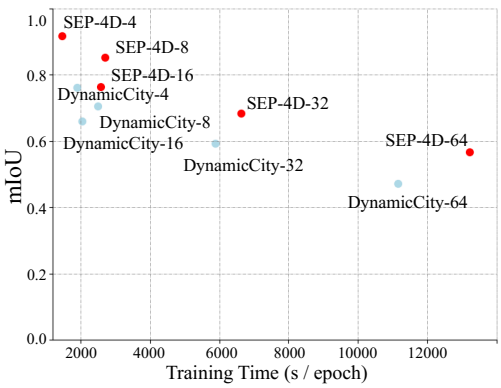

Figure 5: The computational efficiency of ours SEP-4D vs. DynamicCity across different sequence lengths. The number following each method name in the figure indicates the number of frames. For the training time, the less the better.

Method	mIoU (%)	IoU (%)
3D Conv + SEBlock	46.74	84.0
Ours	**76.41**	**92.32**

Table 4: Ablation study about encoder. We compare the original planar factorization design with a 3D Conv-only variant.

Conclusion

In this paper, we have introduced a novel continuous 4D latent-space representation guided by the structure-equivalent prior (SEP-4D), which fundamentally bridges temporal dynamics and spatial structural evolution. By decomposing dynamic scenes into learnable 2D feature planes and constraining their fusion via plane-wise spatio-temporal mechanism to originate, our model encodes structured latent 4D representations of dynamic 3D scenes. This approach eliminates the suboptimal reconstruction issues inherent in prior methods that rely on naive compositions or discrete tokenization. To validate our method's efficacy, we adopted 4D occupancy reconstruction, an intuitive benchmark task. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art methods like DynamicCity and OccSora, achieving up to 15.42% higher mIoU for 4-frame sequences and maintaining superior performance across varying sequence lengths. Critically, the SEP-guided latent space guarantees inherent consistency between temporal evolution and spatial structures in dynamic scenes. The proposed residual-enhanced decoder further mitigates feature collapse, yielding balanced performance across all semantic categories including challenging sparse objects.

Acknowledgments

This work is supported by the Beijing Natural Science Foundation (Grant L241017 and Grant 4252048), the National Natural Science Foundation of China (Grant 62302047), and the National Key Laboratory of Hybrid Human-Machine Augmented Intelligence, Xi'an Jiaotong University under Grant HMHAI-202416.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, 40–49. PMLR.

Bian, H.; Kong, L.; Xie, H.; Pan, L.; Qiao, Y.; and Liu, Z. 2024. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. *arXiv preprint arXiv:2410.18084*.

Cao, A.; and Johnson, J. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 130–141.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, 628–644. Springer.

Engel, J.; Somasundaram, K.; Goesele, M.; Sun, A.; Gamino, A.; Turner, A.; Talattof, A.; Yuan, A.; Souti, B.; Meredith, B.; et al. 2023. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*.

Fan, H.; Su, H.; and Guibas, L. J. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 605–613.

Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 216–224.

He, Y.; He, Y.; He, S.; Chen, F.; Zhou, H.; Zhang, K.; and Zhuang, B. 2025. Neighboring autoregressive modeling for efficient visual generation. *arXiv preprint arXiv:2503.10696*.

Jang, H.; and Kim, D. 2022. D-tensorf: Tensorial radiance fields for dynamic scenes. *arXiv preprint arXiv:2212.02375*.

Lai, Y.; Zhong, J.; Su, V.; and Yang, X. 2025. Patient-Specific Autoregressive Models for Organ Motion Prediction in Radiotherapy. *arXiv preprint arXiv:2505.11832*.

Liao, Z.; Wei, P.; Zhang, R.; Chen, S.; Wang, H.; and Ren, Z. 2025. I^2-World: Intra-Inter Tokenization for Efficient Dynamic 4D Scene Forecasting. *arXiv preprint arXiv:2507.09144*.

Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.-S.; and Theobalt, C. 2020. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33: 15651–15663.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mittal, M.; Yu, C.; Yu, Q.; Liu, J.; Rudin, N.; Hoeller, D.; Yuan, J. L.; Singh, R.; Guo, Y.; Mazhar, H.; et al. 2023. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6): 3740–3747.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10318–10327.

Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.

Sun, C.; Sun, M.; and Chen, H.-T. 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5459–5469.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, C.; Eckart, B.; Lucey, S.; and Gallo, O. 2021. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*.

Wang, L.; Zheng, W.; Ren, Y.; Jiang, H.; Cui, Z.; Yu, H.; and Lu, J. 2024. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*.

Wilson, J.; Song, J.; Fu, Y.; Zhang, A.; Capodieci, A.; Jayakumar, P.; Barton, K.; and Ghaffari, M. 2022. MotionSC: Data set and network for real-time semantic mapping in dynamic environments. *IEEE Robotics and Automation Letters*, 7(3): 8439–8446.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20310–20320.

Xian, W.; Huang, J.-B.; Kopf, J.; and Kim, C. 2021. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9421–9431.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34: 4805–4815.

Zhang, J.; Xiong, F.; and Xu, M. 2024. 3D representation in 512-Byte: Variational tokenizer is the key for autoregressive 3D generation. *arXiv preprint arXiv:2412.02202*.

Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; Duan, Y.; and Lu, J. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, 55–72. Springer.

Zhu, H.; He, T.; Yu, X.; Guo, J.; Chen, Z.; and Bian, J. 2025. Ar4d: Autoregressive 4d generation from monocular videos. *arXiv preprint arXiv:2501.01722*.

Zhuo, D.; Zheng, W.; Guo, J.; Wu, Y.; Zhou, J.; and Lu, J. 2025. Streaming 4D Visual Geometry Transformer. *arXiv preprint arXiv:2507.11539*.