APF-GS: Adaptive Prior Fusion and Compositional Reconstruction for Dynamic Humans in the Wild

Tao Cheng[1], Chengliang Wang[2,1,*], Chaoyue Chen[1], Jianbin Chen[3]
[1]National Elite Institute of Engineering, Chongqing University, Chongqing, China
[2]College of Computer Science, Chongqing University, Chongqing, China
[3]Chongqing Changan Automobile Co., Ltd., Chongqing, China
wangcl@cqu.edu.cn

***Abstract*—High-fidelity reconstruction of dynamic pedestrians in 3DGS-based driving scenes remains challenging, as existing approaches often produce results that lack high-frequency geometric details and involve structural conflation of humans with their accessories. To address these issues, we propose APF-GS, a unified framework for high-fidelity, structured reconstruction. The Adaptive Prior Fusion (APF) module recovers high-frequency geometric details through a neural residual deformation field, reconciling macro-level geometric correction with micro-level detail generation via dynamic regularization. The Compositional Reconstruction (CR) module mitigates structural conflation by self-supervised decoupling of supervision signals, enabling cleaner human-accessory separation. It further introduces a parent-child kinematic linkage to ensure physically consistent interactions during motion.Extensive evaluations on challenging dynamic scenarios from the Waymo Open Dataset demonstrate that APF-GS outperforms existing methods in reconstruction fidelity and robustness, yielding sharper geometric details and structurally consistent interactions. Specifically, it achieves a PSNR of 33.46 on human-centric regions, a significant improvement of +6.73 dB over existing baselines.**

***Index Terms*—3D Gaussian Splatting, Dynamic Human Reconstruction, Adaptive Prior Fusion, Compositional Reconstruction**

I. Introduction

High-fidelity reconstruction of dynamic pedestrians with complex non-rigid deformations is a fundamental challenge in frontier applications such as autonomous driving [1]–[5]. Beyond accurate geometric recovery, faithful reproduction of high-frequency details—such as clothing wrinkles—is essential for achieving visual realism in simulation environments [6]–[8]. Recent studies have focused on reconstructing dynamic pedestrians directly from vehicle-mounted sensor logs, offering a data-driven pathway toward scalable and photorealistic scene modeling for autonomous driving validation [9]–[12].

To guide and regularize the complex reconstruction of non-rigid humans [13]–[17], the mainstream approach is to introduce a parametric model (e.g., SMPL [18]) as a prior, combined with a 3D Gaussian splatting [19] representation. This "prior-driven" paradigm has become a cornerstone of the field.

*Corresponding author. This work was supported by the Chongqing Technology Innovation & Application Development Major Project under Grant CSTB2025TIAD-STX0029 and the Southwest Hospital Cultivation Project of Clinical Innovative Technologies under Grant 2025CXJS26.

Fig. 1. Comparison: OmniRe (top) vs. Ours (bottom). Our method preserves finer details for both humans and accessories, achieving more realistic reconstructions.

In the area of animatable avatars, works like GauHuman [20] and Human Gaussian Splatting [21] construct a Gaussian field in a canonical space under the guidance of an SMPL prior, which is then driven by skinning techniques to achieve high-quality dynamic rendering. For complex autonomous driving scenarios, works such as OmniRe [22] also adopt this approach, leveraging SMPL priors and a general deformable field to uniformly model dynamic pedestrians within the scene.

However, existing holistic frameworks still face key challenges in reconstructing dynamic pedestrians with object interactions. Parametric priors like SMPL provide only a smooth geometric base, lacking high-frequency details such as clothing wrinkles [23]–[25]. Under occlusion, these priors become unreliable, as the need for macro-level geometric correction conflicts with micro-level detail generation, leading to structural and textural distortions.Meanwhile, coarse dataset annotations often merge humans with attached objects (e.g., backpacks), introducing semantic ambiguity that leads models to erroneously fuse accessories into the human body [26], [27].

To this end, we propose APF-GS, a novel framework for high-fidelity, compositional reconstruction of dynamic pedestrians in autonomous driving scenarios. The core of our approach is the Adaptive Prior Fusion (APF) module, which employs dynamic regularization to reconcile macro-level geometric correction with micro-level detail generation, thereby overcoming the limitations of template priors. Additionally, we introduce a self-supervised decoupling strategy that separates

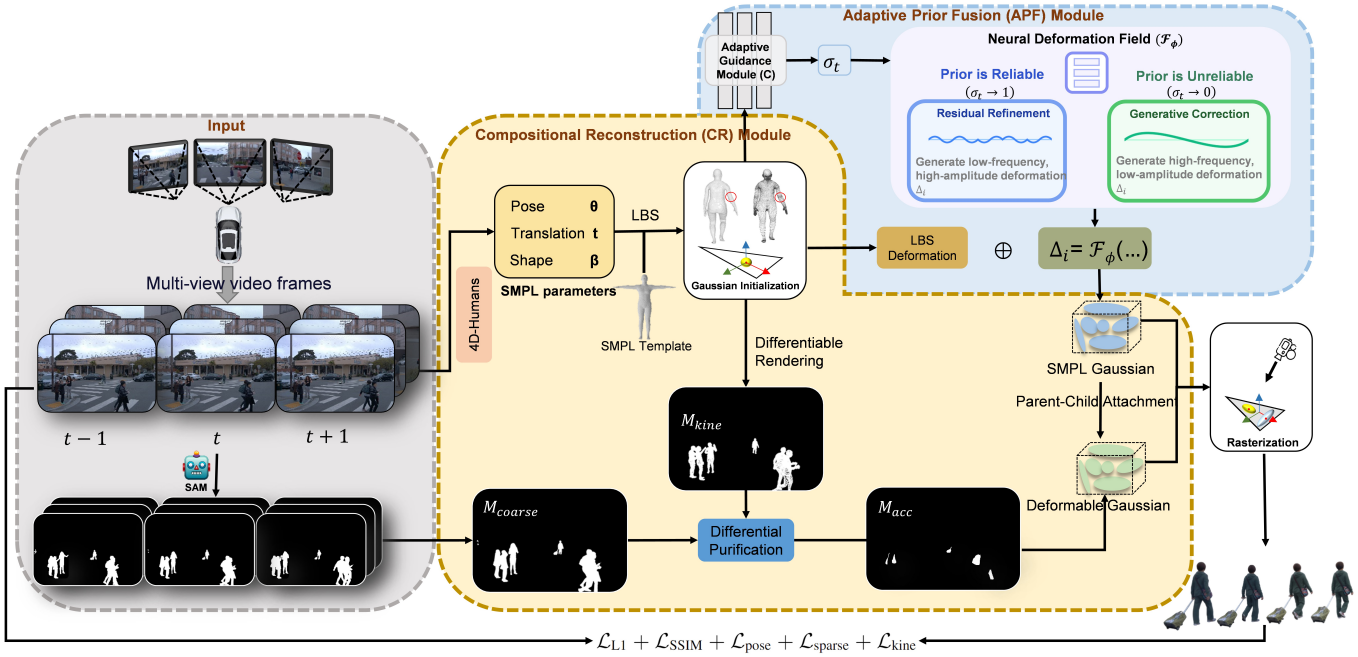

Fig. 2. Overview of the APF-GS framework. Taking multi-view video frames as input, the Compositional Reconstruction (CR) Module leverages SMPL-based priors to disentangle human (M_{kine}) and accessory (M_{acc}) regions via self-supervised decoupling. These masks supervise SMPL Gaussians and Deformable Gaussians, respectively, bound by a parent-child kinematic constraint to ensure physical consistency while allowing non-rigid flexibility. Meanwhile, the Adaptive Prior Fusion (APF) Module derives a confidence score σ_t from SMPL priors to dynamically modulate the Neural Deformation Field ($\mathcal{F}_\phi$). This modulation determines the generation focus: prioritizing macro-geometric repair when priors are unreliable, while shifting to micro-detail generation when macro-geometry is reliable, thereby balancing macro-correction and micro-detail synthesis.

humans from accessories and reconstructs their interactions in a physically plausible manner, effectively mitigating structural conflation caused by coarse annotations.

The main contributions of this work are as follows:

- Adaptive Prior Fusion (APF) Module: We introduce a dynamic regularization mechanism within the 3D Gaussian deformation field. It robustly corrects inaccurate kinematic priors while simultaneously recovering high-fidelity micro-details (e.g., clothing wrinkles).
- Compositional Reconstruction (CR) Module: We propose a self-supervised decoupling strategy to resolve human-accessory structural conflation. Coupled with a parent-child kinematic linkage, it ensures physically consistent interactions for complex dynamic pedestrians.
- Unified APF-GS Framework: We present a holistic framework for dynamic human reconstruction. Validated on the Waymo Open Dataset, it significantly improves upon existing baselines, achieving a +6.73 dB PSNR gain in human-centric regions.

II. METHOD

As illustrated in Fig. 2, our APF-GS framework is designed for the high-fidelity, structured reconstruction of dynamic pedestrians and their accessories. The pipeline begins by addressing structural conflation and purifying supervision signals via a self-supervised decoupling step (§II-B). Subsequently, our core adaptive prior fusion module (§II-C) balances macro-level geometric correction with micro-level detail generation through a dynamic regularization mechanism. Finally, the compositional reconstruction of the human-accessory interaction is achieved using a parent-child kinematic linkage (§II-D).

A. Preliminaries

3D Gaussian splatting. 3DGS models scenes using 3D Gaussian primitives $\mathcal{G} = \{g_i\}$, each defined by position μ, opacity α, covariance Σ, and SH color c. For rendering, primitives are projected and alpha-blended via differentiable rasterization. The pixel color C is computed as:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)\, \sigma_t. \tag{1}$$

where $\mathcal{N}$ denotes primitives overlapping the pixel.

Parametric Human Model. SMPL manipulates a fixed-topology mesh $\mathcal{M}_h = (\mathcal{V}, \mathcal{F})$ via shape β and pose θ parameters. The vertex deformation $\mathcal{V}_S$ is formulated as:

$$\mathcal{V}_S = \mathcal{V} + B_S(\beta) + B_P(\theta). \tag{2}$$

where B_S and B_P represent displacements from shape and pose. Driven by LBS, SMPL serves as a critical geometric prior in our framework.

B. Self-Supervised Decoupling for Supervision Signal Purification

To address structural conflation in coarse dataset annotations, we propose a self-supervised decoupling module that separates the human and accessory without requiring additional labels.

Fig. 3. (a) Input frames. (b) Dataset-provided masks (M_{coarse}) with structural conflation from merged accessories. (c) Our kinematic masks (M_{kine}) capture only the human body, enabling clean supervision for dynamic human reconstruction.

We first generate a clean kinematic prior from the initial SMPL parameters (θ_0, β) and render it onto the image plane via a differentiable renderer Π, producing the kinematic human mask M_{kine}. As shown in Fig. 3, M_{kine} effectively avoids the human-object entanglement present in the original coarse mask M_{coarse}, providing a semantically cleaner supervision target. Note that kinematic prior construction relies on instance tracking IDs; thus, untracked pedestrians are excluded from M_{kine} and handled by the background model. The accessory mask is then explicitly obtained via a set difference operation:

$$M_{\text{acc}} = M_{\text{coarse}} \setminus M_{\text{kine}}. \tag{3}$$

Based on these two semantically clear masks, we instantiate two composable entity nodes employing differentiated modeling strategies for subsequent training. Specifically, the human node is strictly supervised by the purified mask M_{kine} and adopts an SMPL-driven Gaussian model to precisely capture articulated skeletal motion. In contrast, the accessory node is supervised by M_{acc} and utilizes a general Deformable Gaussian model to robustly adapt to non-rigid deformations that lack a fixed topological structure.

C. Adaptive Prior Fusion via Dynamic Regularization Mechanism

Following the acquisition of a pure human node, high-fidelity reconstruction is confronted with a core conflict. On one hand, the model must perform "macro-level geometric correction" to compensate for an unreliable initial SMPL prior. On the other hand, it must generate "micro-level details," such as clothing wrinkles, which are absent from the inherently smooth prior surface. To resolve the challenge of a single neural deformation field struggling to reconcile these two tasks at conflicting scales, we propose the Adaptive Prior Fusion module. This module dynamically assesses the prior's reliability, enabling an intelligent trade-off between geometric correction and detail generation.

We term this core mechanism *dynamic regularization*, which adaptively balances geometric correction and detail generation by weighting $\mathcal{L}_{\text{kine}}$ and $\mathcal{L}_{\text{sparse}}$ via a reliability weight σ_t. To achieve this, we first design a unified deformation field composed of the classic LBS transformation and a proposed neural residual deformation field $\mathcal{F}_\phi$. For each Gaussian primitive $\bar{\mathbf{g}}_i$ in the human node, its final state $\mathbf{g}_i(t)$ at time t is determined by:

$$\mathbf{g}_i(t) = T_h(t) \otimes (\text{LBS}(\bar{\mathbf{g}}_i, \theta(t)) \oplus \Delta_i)\,. \tag{4}$$

Here, $T_h(t)$ is the global rigid transformation. The crucial residual deformation $\Delta_i = (\delta\mu_i, \delta q_i, \delta s_i)$ is predicted by the neural residual deformation field $\mathcal{F}_\phi$. This process of predicting and generating residual deformations is precisely modulated by an Adaptive Guidance Module (C). Conditioned on multi-modal inputs: pedestrian visual features extracted from the current image frame, a learnable instance embedding vector e_h representing the specific pedestrian identity, and the current timestamp t. From these inputs, it directly predicts two key signals: a high-dimensional conditional feature vector $\mathbf{c}_t$, which encapsulates all necessary information about the current observational state to guide the deformation; and a scalar reliability weight σ_t, strictly constrained to the range $(0, 1)$ via a Sigmoid activation function, acting as a dynamic gate to modulate the output scale of the residual deformation. To prevent mode collapse (e.g., $\sigma_t \to 0$), we initialize the MLP bias for high prior confidence ($\sigma_t \approx 0.8$), encouraging adherence to geometric priors unless contradicted by strong photometric evidence.

Subsequently, the conditional feature $\mathbf{c}_t$ is concatenated with the positional encoding of the Gaussian primitive $\gamma(\bar{\mu}_i)$ to jointly guide the final deformation prediction:

$$\Delta_i = \mathcal{F}_\phi\big(\gamma(\bar{\mu}_i), \mathbf{c}_t\big). \tag{5}$$

The APF-GS framework is driven by a unified joint optimization objective, with the total loss function defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} \\ + \sigma_t\lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}} + (1-\sigma_t)\lambda_{\text{kine}}\mathcal{L}_{\text{kine}}.\end{aligned} \tag{6}$$

This function includes the photometric loss ($\mathcal{L}_{\text{photo}}$), a combination of L1 and SSIM losses to ensure visual fidelity, and the pose loss ($\mathcal{L}_{\text{pose}}$) to guarantee smooth motion. The core of our optimization lies in the adaptive regularization mechanism, which is dynamically modulated by the reliability weight σ_t and implicitly supervised by the competitive interplay between data fidelity and prior constraints. Specifically, in regions with high photometric error (indicating an unreliable prior), the network automatically drives $\sigma_t \to 0$. This relaxes the sparsity constraint on $\mathcal{F}_\phi$ and activates the kinematic consistency loss ($\mathcal{L}_{\text{kine}}$), allowing the model to generate large-magnitude macro-geometric corrections. Conversely, in regions where the prior is well-aligned, the network raises $\sigma_t \to 1$. This

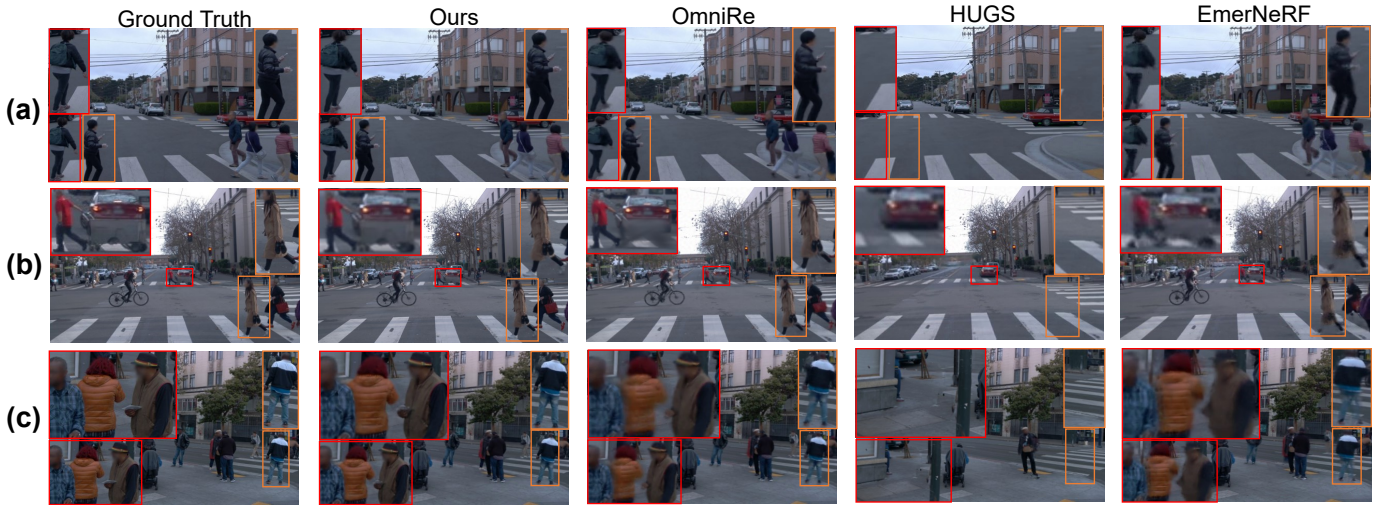

Fig. 4. Qualitative comparison with baseline methods. Results are shown for: (a) scene 23, (b) scene 24, and (c) scene 327. Our method (Ours) achieves better reconstruction of complex poses and texture details.

TABLE I: Quantitative comparison on the Waymo Open Dataset. We report metrics on both the full image and human-centric regions. "Human" indicates metrics computed on human-related regions, and the "GT Box" column indicates whether a method utilizes ground truth 3D bounding boxes for dynamic object modeling. The best results are highlighted in bold.

		Image Reconstruction					Novel View Synthesis				
		Full Image			Human		Full Image			Human	
Methods	GT Box	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
EmerNeRF [28]		31.24	0.882	0.120	22.59	0.557	29.41	0.871	0.134	21.29	0.479
3DGS [19]		26.73	0.891	0.104	15.76	0.374	26.39	0.886	0.145	15.47	0.368
PVG [29]		32.40	0.931	0.107	24.40	0.700	28.80	0.880	0.129	21.70	0.558
DeSiRe-GS [30]		32.49	0.943	0.105	24.52	0.723	30.59	0.936	0.120	22.05	0.596
HUGS [31]	✓	28.22	0.920	0.093	16.18	0.398	27.60	0.907	0.098	15.85	0.377
StreetGS [32]	✓	29.09	0.933	0.083	16.69	0.427	28.59	0.925	0.107	16.48	0.389
OmniRe [22]	✓	34.55	0.962	0.059	26.73	0.841	33.12	0.943	**0.056**	24.05	0.712
Ours	✓	**35.87**	**0.965**	**0.054**	**33.46**	**0.931**	**33.26**	**0.945**	0.061	**28.42**	**0.796**

enforces the sparsity loss ($\mathcal{L}_{\text{sparse}}$) to suppress large deformations, strictly restricting the residual field to focus solely on generating high-frequency micro-details.

The activated kinematic consistency loss is defined as:

$$\mathcal{L}_{\text{kine}} = \lambda_{\text{bone}}\mathcal{L}_{\text{bone}} + \lambda_{\text{limit}}\mathcal{L}_{\text{limit}}. \tag{7}$$

where $\mathcal{L}_{\text{bone}}$ (bone length loss) maintains constant bone lengths, and $\mathcal{L}_{\text{limit}}$ (joint angle limit loss) constrains joint rotations to a plausible physiological range. Through this unified optimization objective, APF-GS effectively resolves the intrinsic conflict between macro-level correction and micro-level detail generation, achieving robust and high-fidelity reconstruction of dynamic humans.

D. Compositional Scene Construction from Decoupled Entities

To achieve physically plausible modeling of dynamic human-object interactions, we recompose the decoupled human and accessory entities through a parent-child kinematic chain. Based on high-fidelity human reconstruction, we obtain continuous and accurate skeletal joint transformation matrices $\mathbf{T}_{\text{joint}}(t) \in \mathbb{R}^{4\times4}$. According to semantic rules (e.g., backpack linked to torso, suitcase to wrist), we select the most relevant joint as the parent node, denoted by $\mathbf{T}_{\text{parent}}(t)$.

The global pose of the accessory is attached to the parent node via a learnable relative transformation:

$$\mathbf{T}_{\text{accessory}}(t) = \mathbf{T}_{\text{parent}}(t) \cdot \Delta\mathbf{T}_{\text{attach}}. \tag{8}$$

where $\Delta\mathbf{T}_{\text{attach}}$ is a learnable parameter that is optimized end-to-end during training, adaptively capturing the stable spatial offset between the human and the object (e.g., grasping or carrying posture).

This structured composition ensures that accessories move naturally with the human body, avoiding artifacts such as floating or penetration. Moreover, it establishes a hierarchical, editable scene graph, enabling independent manipulation of individual entities and providing a structured foundation for downstream applications such as simulation and interactive rendering.

III. EXPERIMENTS

A. Experimental Setup

Datasets and Metrics. We conduct experiments on the Waymo Open Dataset [33]. To rigorously evaluate robustness,

we curated a benchmark of 24 highly challenging sequences. These were specifically selected to cover diverse complex scenarios. This selection targets "long-tail" perception challenges in autonomous driving, ensuring our metrics reflect performance on difficult dynamic cases with statistical significance. For Novel View Synthesis (NVS), every 15th frame is held out for testing, while the remaining frames are used for training. We report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

Implementation Details and Efficiency. We implement APF-GS using the PyTorch framework. All experiments are conducted on a single NVIDIA RTX 4090 GPU (24GB). specifically, the training process requires approximately 1.2 hours per sequence, and the inference speed reaches 60 FPS, ensuring real-time rendering capabilities.

Comparison Methods. We compare APF-GS with a range of state-of-the-art dynamic scene reconstruction methods, categorized into two groups: (i) unsupervised methods without instance annotations, including EmerNeRF [28], 3DGS [19], PVG [29], and DeSiRe-GS [30]; and (ii) scene graph-based methods that rely on instance-level annotations (e.g., 3D bounding boxes), represented by StreetGS [32], HUGS [31], and OmniRe [22]. This comprehensive evaluation demonstrates the superiority of our approach in complex real-world scenarios. While unsupervised baselines avoid annotations, they often compromise object-level controllability and high-frequency detail. By leveraging standard 3D bounding box priors—widely available in autonomous driving pipelines—our approach secures the structured editing capabilities essential for realistic simulation. This strategy significantly enhances reconstruction fidelity and downstream utility compared to fully implicit decomposition.

B. Quantitative and Qualitative Analysis

Quantitative Analysis. As shown in Table I, APF-GS achieves state-of-the-art performance in both full-scene and human-centric reconstruction. Notably, our method attains a PSNR of 33.46 on human regions, significantly outperforming OmniRe (26.73), demonstrating superior modeling of complex non-rigid motions. This gain stems from two key components: (1) the self-supervised decoupling module effectively separates humans from accessories, eliminating supervision contamination; and (2) the adaptive prior fusion enables high-frequency detail generation (e.g., clothing wrinkles). APF-GS also surpasses rigid-motion-specific methods (StreetGS, HUGS) and advanced unsupervised approaches (PVG, DeSiRe-GS).

Qualitative Analysis. As shown in Fig. 4, APF-GS achieves high-fidelity reconstructions consistent with the ground truth across all tested scenes. For instance, in Scene 23, our model accurately reconstructs a pedestrian's backpack and clothing textures while clearly separating human and accessory geometries; in Scene 24, it successfully restores the complex interaction of a pedestrian dragging a trolley, precisely capturing both human form and the spatial layout of large accessories; in Scene 327, it reproduces rich texture details on an orange down jacket. These visual comparisons validate the effectiveness of our core mechanisms—self-supervised decoupling and adaptive prior fusion—in enhancing realism and fidelity in dynamic human reconstructions.

TABLE II: Ablation Study on the Core Components of APF-GS.

Settings	Full PSNR ↑		Human PSNR ↑	
	Recon.	NVS	Recon.	NVS
w/o Decoupling	34.75	32.68	28.76	24.18
w/ Strong Prior	35.21	33.16	32.54	27.08
w/ Weak Prior	34.92	32.90	31.98	26.20
Full model	**35.87**	**33.26**	**33.46**	**28.42**

C. Ablation Studies

As shown in Table II, we conduct ablation studies to validate the effectiveness of key components in our framework. Removing the self-supervised decoupling module (w/o Decoupling) leads to a significant performance drop, with a 4.70 dB decrease in PSNR on human regions, demonstrating its critical role in preventing structural conflation between the human body and accessories. For the adaptive prior fusion mechanism, enforcing a strong reliance on the SMPL prior (Strong Prior) limits geometric correction of inaccurate initial poses, resulting in a 0.92 dB degradation. In contrast, adopting a purely data-driven regime without prior guidance (Weak Prior) fails to capture fine-grained motion patterns, causing a larger drop of 1.48 dB. These results confirm that our adaptive fusion mechanism effectively balances prior knowledge and observational evidence, ensuring both robustness and high-fidelity reconstruction.

IV. Conclusion

In this paper, we present APF-GS, a unified framework for high-fidelity reconstruction of dynamic pedestrians in autonomous driving scenes using 3D Gaussian Splatting. Our method addresses two key challenges in parametric prior-based approaches: the lack of high-frequency details (e.g., clothing wrinkles) and structural conflation between humans and accessories. By introducing the Adaptive Prior Fusion (APF) and Compositional Reconstruction (CR) modules, we enhance geometric accuracy, detail realism, and structural integrity. Experiments on the Waymo dataset demonstrate that APF-GS outperforms state-of-the-art methods in fidelity and robustness. The framework establishes a compositional representation of pedestrians, enabling future applications in editable modeling and interactive simulation for autonomous driving and virtual reality.

REFERENCES

[1] Jiahao Wu, Rui Peng, Zhiyan Wang, Lu Xiao, Luyang Tang, Jinbo Yan, Kaiqiang Xiong, and Ronggang Wang, "Swift4d: Adaptive divide-and-conquer gaussian splatting for compact and efficient reconstruction of dynamic scene," *arXiv preprint arXiv:2503.12307*, 2025.

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

[3] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman, "Humannerf: Free-viewpoint rendering of moving people from monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16210–16220.

[4] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli, "The language of motion: Unifying verbal and non-verbal language of 3d human motion," *CVPR*, 2025.

[5] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan, "Neuman: Neural human radiance field from a single video," in *European Conference on Computer Vision*. Springer, 2022, pp. 402–418.

[6] Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu, "Disco4d: Disentangled 4d human generation and animation from a single image," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26331–26344.

[7] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao, "Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment," in *European Conference on Computer Vision*. Springer, 2024, pp. 124–142.

[8] Bing He, Yunuo Chen, Guo Lu, Qi Wang, Qunshan Gu, Rong Xie, Li Song, and Wenjun Zhang, "H3d-dgs: Exploring heterogeneous 3d motion representation for deformable 3d gaussian splatting," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[9] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20310–20320.

[10] Jiageng Mao, Boyi Li, Boris Ivanovic, Yuxiao Chen, Yan Wang, Yurong You, Chaowei Xiao, Danfei Xu, Marco Pavone, and Yue Wang, "Dreamdrive: Generative 4d scene modeling from street view images," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 367–374.

[11] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21634–21643.

[12] Georg Hess, Carl Lindström, Maryam Fatemi, Christoffer Petersson, and Lennart Svensson, "Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11982–11992.

[13] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis, "Gart: Gaussian articulated template models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19876–19887.

[14] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie, "Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 634–644.

[15] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu, "Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19711–19722.

[16] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan, "Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20763–20774.

[17] Jianing Chen, Zehao Li, Yujun Cai, Hao Jiang, Chengxuan Qian, Juyuan Kang, Shuqin Gao, Honglong Zhao, Tianlu Mao, and Yucheng Zhang, "Haif-gs: Hierarchical and induced flow-guided gaussian splatting for dynamic scene," *arXiv preprint arXiv:2506.09518*, 2025.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[20] Shoukang Hu, Tao Hu, and Ziwei Liu, "Gauhuman: Articulated gaussian splatting from monocular human videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20418–20431.

[21] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero, "Human gaussian splatting: Real-time rendering of animatable avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 788–798.

[22] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang, "Omnire: Omni urban scene reconstruction," in *The Thirteenth International Conference on Learning Representations*, 2025.

[23] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black, "Scanimate: Weakly supervised learning of skinned clothed avatar networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2886–2897.

[24] Yichen Wang, Yiyi Zhang, Xinhao Hu, Li Niu, Jianfu Zhang, Yasushi Makihara, Yasushi Yagi, Pai Peng, Wenlong Liao, Tao He, Junchi Yan, and Liqing Zhang, "Pedestrian motion reconstruction: A large-scale benchmark via mixed reality rendering with multiple perspectives and modalities," in *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang, "Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4568–4577.

[26] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *European Conference on Computer Vision*. Springer, 2024, pp. 162–179.

[27] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein, "Gaussian shell maps for efficient 3d human generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9441–9451.

[28] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang, "Emernerf: Emergent spatial-temporal scene decomposition via self-supervision," *arXiv preprint arXiv:2311.02077*, 2023.

[29] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang, "Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering," *arXiv:2311.18561*, 2023.

[30] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan, "Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6782–6791.

[31] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21336–21345.

[32] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng, "Street gaussians: Modeling dynamic urban scenes with gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 156–173.

[33] Pei Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2446–2454.