Consistent Instance Field for Dynamic Scene Understanding

Supplementary Material

Overview. In the supplementary material, we first present our proposed multi-view instance segmentation to get cross-view consistent instance masks on the Neu3D dataset (Sec. A.1). We then present additional qualitative video results (Sec. B.1) and quantitative results (Sec. B.2). Finally, we discuss the limitation of our method (Sec. C.1) and the societal impact (Sec. C.2).

A. Implementation Details

A.1. Multi-View Consistent Instance Segmentation

Pre-processing (merging multi-view videos). As discussed in Sec. 4.1 of the main paper, the multi-view benchmark Neu3D [4] provides synchronized videos but does not include ground-truth instance annotations. Therefore, we follow prior works [2, 5] and use DEVA [1], a video object tracking model, to generate input masks. However, because DEVA processes each video independently, the resulting masks are inconsistent across views. To address this limitation, we reinterpret the spatial multi-view dimension as an inter-temporal one and merge all views into a single pseudo-monocular sequence. More precisely, we concatenate each video with the reversed video of its adjacent view. Formally, given N spatially adjacent camera views and T frames per video corresponding to each view, we first reorder frames to maximize temporal continuity:

$$\mathcal{S}_n^{\text{ordered}} = \begin{cases} (I_1, I_2, \ldots, I_T), & \text{if } n \text{ is odd,} \\ (I_T, I_{T-1}, \ldots, I_1), & \text{if } n \text{ is even,} \end{cases} \tag{S1}$$

where I_t denotes the t-th frame from that n-th view. The input to DEVA is then constructed by concatenating all reordered view sequences:

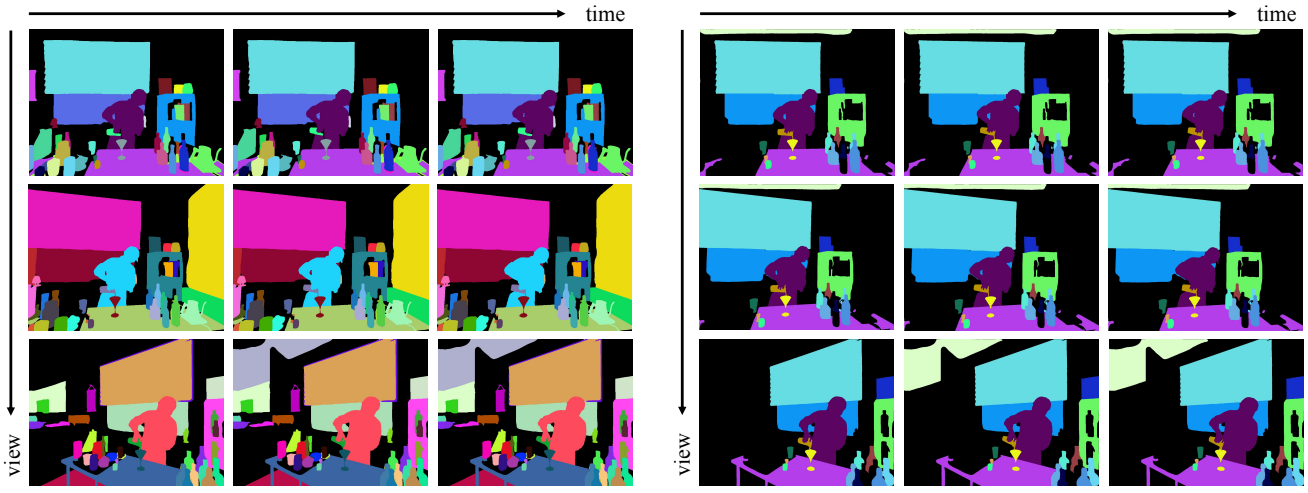
$$\mathcal{S}_{1:N}^{\text{ordered}} = \left[\mathcal{S}_1^{\text{ordered}}, \mathcal{S}_2^{\text{ordered}}, \ldots, \mathcal{S}_N^{\text{ordered}}\right]. \tag{S2}$$

This merged pseudo-monocular sequence ensures that adjacent frames vary smoothly across both time and viewpoint.

Video instance segmentation. Given a concatenated video $\mathcal{S}_{1:N}^{\text{ordered}}$, DEVA then produces per-frame instance masks with temporally propagated instance IDs.

Post-processing (visibility filtering). Due to occlusions and limited camera overlap, some instances may disappear entirely in certain views, leading to conflicting identities across cameras. To prevent incomplete or inconsistent masks, we retain only instances that remain visible in all views. Specifically, an instance k is considered valid if its mask has non-empty support in every view of the concatenated videos. Instances that do not meet this criterion are discarded.

Figure S1 compares the default DEVA results, where each video is segmented independently, with our multi-view



Default DEVA Our Multi-View Consistent Instance Segmentation

Figure S1. **Results of multi-view video segmentation on Neu3D [4] dataset.** (**Left**) DEVA [1] produces instance masks independently per camera view, leading to inconsistent instance identities across synchronized camera streams, (**Right**) our multi-view video segmentation results by merging multi-videos into a single pseudo-monocular sequence to produce multi-view consistent instance masks.

Table S1. **Quantitative comparison of our method with the state-of-the-art on open-vocabulary 4D querying using the HyperNeRF [7] dataset.** We report mAcc and mIoU metrics. The best, second best, and third best results are highlighted. * indicates failure of localizing the objects based on the text queries, as also demonstrated in Figure 5 of the main paper.

	americano				espresso			
	"glass cup"		*"mat"*		*"steel jug"*		*"coaster"*	
Method	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
4D LangSplat [5]	98.02	78.17	72.90	7.55*	96.33	35.61	86.69	0.43*
SA4D [2]	88.91	40.05	71.65	39.67	99.25	70.49	99.56	81.12
Ours	99.02	88.52	94.68	77.66	99.72	87.93	99.73	85.47

consistent instance segmentation. This strategy effectively converts the multi-view videos into a single coherent identity sequence, enabling cross-view consistent pseudo-labels for both supervision and evaluation.

B. Experimental Results

B.1. Additional Qualitative Results

To further demonstrate that our Consistent Instance Field provides coherent instance understanding across both space and time, we include additional qualitative video results comparing our method with recent state-of-the-art approaches SA4D [2], Trace3D [9], Dr.Splat [3], and VLGS [8] on the standard HyperNeRF [7] and Neu3D [4] benchmarks for both novel-view panoptic segmentation and open-vocabulary 4D querying tasks:

- Results on the monocular benchmark HyperNeRF [7] for novel-view panoptic segmentation: `panoptic_segmentation_hypernef.mp4`
- Results on the multi-view benchmark Neu3D [4] for novel-view panoptic segmentation: `panoptic_segmentation_neu3d.mp4`;
- Results on the monocular benchmark HyperNeRF [7] for open-vocabulary 4D querying: `open_vocabulary_4d_querying_hypernerf.mp4`

B.2. Additional Quantitative Results

To further assess the capability of the Consistent Instance Field in open-vocabulary 4D querying, we report additional quantitative results using standard metrics, mAcc and mIoU. We follow the experimental setting described in the main paper and use Grounded DINO [6] to obtain 2D masks for each text query as pseudo-ground-truth annotations.

As shown in Table S1, our method consistently achieves higher retrieval accuracy across various text prompts. On average, our method achieves 98.29 mAcc and 84.90 mIoU, surpassing the second-best approach (SA4D) by 8.45 and 27.07, respectively. In contrast, previous methods like 4D LangSplat [5] often struggle to localize the objects based on the text query, as demonstrated in Figure 5 of the main paper, thus yielding extremely low performance (*e.g.*, 7.55 mIoU on the "americano" scene with the "mat" query and 0.43 on the "espresso" scene with the "coaster" query). These results highlight the strength of our method in fine-grained and coherent instance modeling in complex dynamic scenes.

C. Discussions

C.1. Limitations

While the proposed Consistent Instance Field provides a principled formulation for identity modeling in dynamic scenes, several limitations remain. First, our formulation is instantiated via a deformable Gaussian representation [10, 11], which inherits its representational constraints. Scenes involving amorphous or continuously evolving materials (*e.g.*, smoke or liquids) lack stable structure and may not be faithfully represented through persistent Gaussian primitives. In such cases, identity assignments become less interpretable, as the Gaussians fail to maintain consistent spatial support or correspond to physically persistent entities. Future advances in dynamic scene representations may help extend our method to broader scene types. Second, although we construct pseudo-monocular sequences to synchronize multi-view pseudo-labels from video object tracking models [1], residual cross-view inconsistencies or missing annotations under severe occlusion can still bias identity estimation. Exploring more rigorous multi-view pseudo-label harmonization strategies represents a promising direction for enhancing robustness in such scenarios.

C.2. Societal Impact

Our approach provides structured, instance-consistent scene representations that extend beyond visual reconstruction to support simulation, prediction, and interaction within dynamic environments. These advances could improve the safety, efficiency, and interpretability of autonomous and interactive systems, provided that ethical and privacy standards are respected.

References

[1] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 1, 2

[2] Shengxiang Ji, Guanjun Wu, Jiemin Fang, Jiazhong Cen, Taoran Yi, Wenyu Liu, Qi Tian, and Xinggang Wang. Segment any 4d gaussians. *arXiv preprint arXiv:2407.04504*, 2024. 1, 2

[3] Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *CVPR*, 2025. 2

[4] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022. 1, 2

[5] Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. In *CVPR*, 2025. 1, 2

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2

[7] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In *SIGGRAPH Asia*, 2021. 2

[8] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*, 2024. 2

[9] Hongyu Shen, Junfeng Ni, Yixin Chen, Weishuo Li, Mingtao Pei, and Siyuan Huang. Trace3d: Consistent segmentation lifting via gaussian instance tracing. In *ICCV*, 2025. 2

[10] 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 *CVPR*, 2024. 2

[11] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024. 2