TALO: Pushing 3D Vision Foundation Models Towards Globally Consistent Online Reconstruction

Supplementary Material

A. Linear Transformation Assumptions

This section analyzes the hidden assumptions behind adopting a linear transformation between two submaps: different predictions differ only by a global scale in depth ($\mathrm{Sim}(3)$) and/or by intrinsics that can be corrected through a projective warp ($\mathrm{SL}(4)$). Specifically, consider two camera reconstructions $\{\tilde{\mathbf{u}}_1, \mathbf{K}_1, \mathbf{D}_1, \mathbf{T}_1\}$ and $\{\tilde{\mathbf{u}}_2, \mathbf{K}_2, \mathbf{D}_2, \mathbf{T}_2\}$ observing the same scene, where $\tilde{\mathbf{u}} = (u, v, 1)$ denotes homogeneous pixel coordinates, $\mathbf{K}$ the intrinsics, $\mathbf{D}$ the predicted depth, and $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}]$ the extrinsics. The 3D point $\mathbf{P}$ can be obtained by back-projection:

$$\mathbf{P} \;=\; [\mathbf{R} \mid \mathbf{t}]^{-1}\big(\mathbf{D}(u,v)\,\mathbf{K}^{-1}\tilde{\mathbf{u}}\big). \tag{1}$$

Now, if we wish to solve for the transformation $\mathbf{H}$ relating $\mathbf{T}_1$ and $\mathbf{T}_2$ through $\mathbf{P}_1$ and $\mathbf{P}_2$, three typical cases arise:

- Case 1: Consistent intrinsics and scaled depths. If $\mathbf{K}_1 = \mathbf{K}_2$ and $\mathbf{D}_2(u,v) = s\,\mathbf{D}_1(u,v)$ for a global scalar s, then the two reconstructions differ only by a uniform scale. This simplifying assumption is implicitly adopted by VGGT-Long [2]. There exists a similarity transformation $\mathbf{T} = [\,s\,\mathbf{R} \mid \mathbf{t}\,] \in \mathrm{Sim}(3)$ such that $\mathbf{P}_2 = s\,\mathbf{R}\,\mathbf{P}_1 + \mathbf{t}$, which perfectly aligns the two reconstructions.
- Case 2: Inconsistent intrinsics and scaled depths. When $\mathbf{K}_1 \neq \mathbf{K}_2$ and $\mathbf{D}_2(u,v) = s\,\mathbf{D}_1(u,v)$, the change in intrinsics induces a global projective distortion that a single Sim(3) cannot capture. This assumption is adopted by VGGT-SLAM [4]. Nevertheless, under the projective reconstruction theorem, there always exists a homogeneous linear mapping $\mathbf{H} \in \mathrm{SL}(4)$ that transforms one reconstruction into the other such that $\mathbf{P}_1 = \mathbf{H}\,\mathbf{P}_2$.
- Case 3: Nonlinear depth distortion. When the depth predictions deviate nonlinearly across spatial locations, i.e., $\mathbf{D}_2(u,v) = f(\mathbf{D}_1(u,v))$, where $f(\cdot)$ is non-linear and spatially varying, no global linear transformation can exactly align the two reconstructions, regardless of whether the internal calibration is consistent. Any linear alignment forces the optimizer to overfit one region at the expense of another, and bend the trajectory to compensate for local depth inconsistencies.

This analysis exposes the hidden assumptions underlying the two point-based submap alignment methods and explains their fundamental limitations. The more severely the corresponding assumption is violated, the poorer the alignment performance becomes. In scenarios with large submap size and few inter-submap alignments, the violation is often mild and aligns more closely with Case 1 and Case 2, allowing VGGT-Long [2] and VGGT-SLAM [4] to perform reasonably well. However, realistic outdoor multi-camera settings generally correspond to Case 3, where the assumptions in Case 1 and Case 2 rarely hold. As a result, enforcing a global rigid or projective transformation inevitably leaves residual inconsistencies and distorts the trajectory.

B. Additional Implementation Details

For models such as VGGT [6] and π^3 [7]which lack metric-scale, we estimate a global scale factor before control-point propagation to avoid using high-DOF TPS for global scale correction. For models like MapAnything [3]which maintain metric scale prediction, this step is omitted, and we also disable scale estimation in VGGT-Long's [2] $\mathrm{Sim}(3)$ alignment for fairness, whereas in VGGT-SLAM [4] ($\mathrm{SL}(4)$ formulation) the scale cannot be explicitly decoupled.

TALO makes **no assumptions** about the number of cameras, and therefore naturally supports **arbitrary** setups ranging from monocular and stereo to surround-view systems. To exploit the fixed inter-camera baselines in multi-camera configurations, we introduce a simple rig averaging strategy. Specifically, for each non-reference camera, we collect its predicted relative poses to the reference camera across time, and estimate a single time-invariant rig transform by averaging rotation and translation separately: rotations are averaged on $\mathrm{SO}(3)$ using a chordal ℓ_2 mean, implemented by summing rotation matrices and projecting the result back to $\mathrm{SO}(3)$ via SVD, while translations are averaged in Euclidean space. By aggregating these relative poses over time, rig averaging (**RA**) suppresses frame-wise prediction noise and enforces consistency with the underlying rigid camera setup. This simple yet effective strategy is applied fairly to both VGGT-SLAM and VGGT-Long.

C. Additional Qualitative Comparisons

We provide additional qualitative comparisons in Fig. 1, 2, 3, and 4. As shown, VGGT-Long [2] ($\mathrm{Sim}(3)$) exhibits noticeable trajectory drift in long-range sequences, together with local misalignments that manifest as multi-layer artifacts in the reconstructed geometry. VGGT-SLAM [4] ($\mathrm{SL}(4)$) is often disrupted by noise in the predicted point clouds, leading to severe geometric divergence or complete reconstruction failure. In contrast, TALO produces trajectories that closely match the ground truth and yields locally well-aligned geometry with clean, artifact-free details.

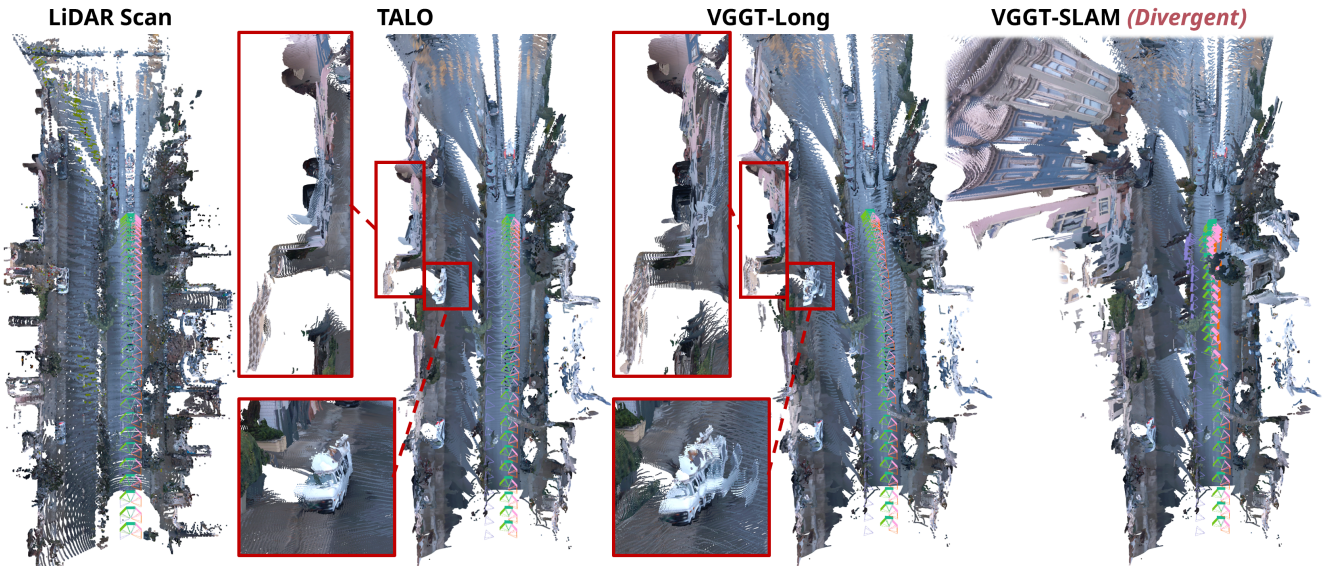

Figure 1. Qualitative comparison with VGGT [6] on **Waymo** [5] scene 6104545334635651714.

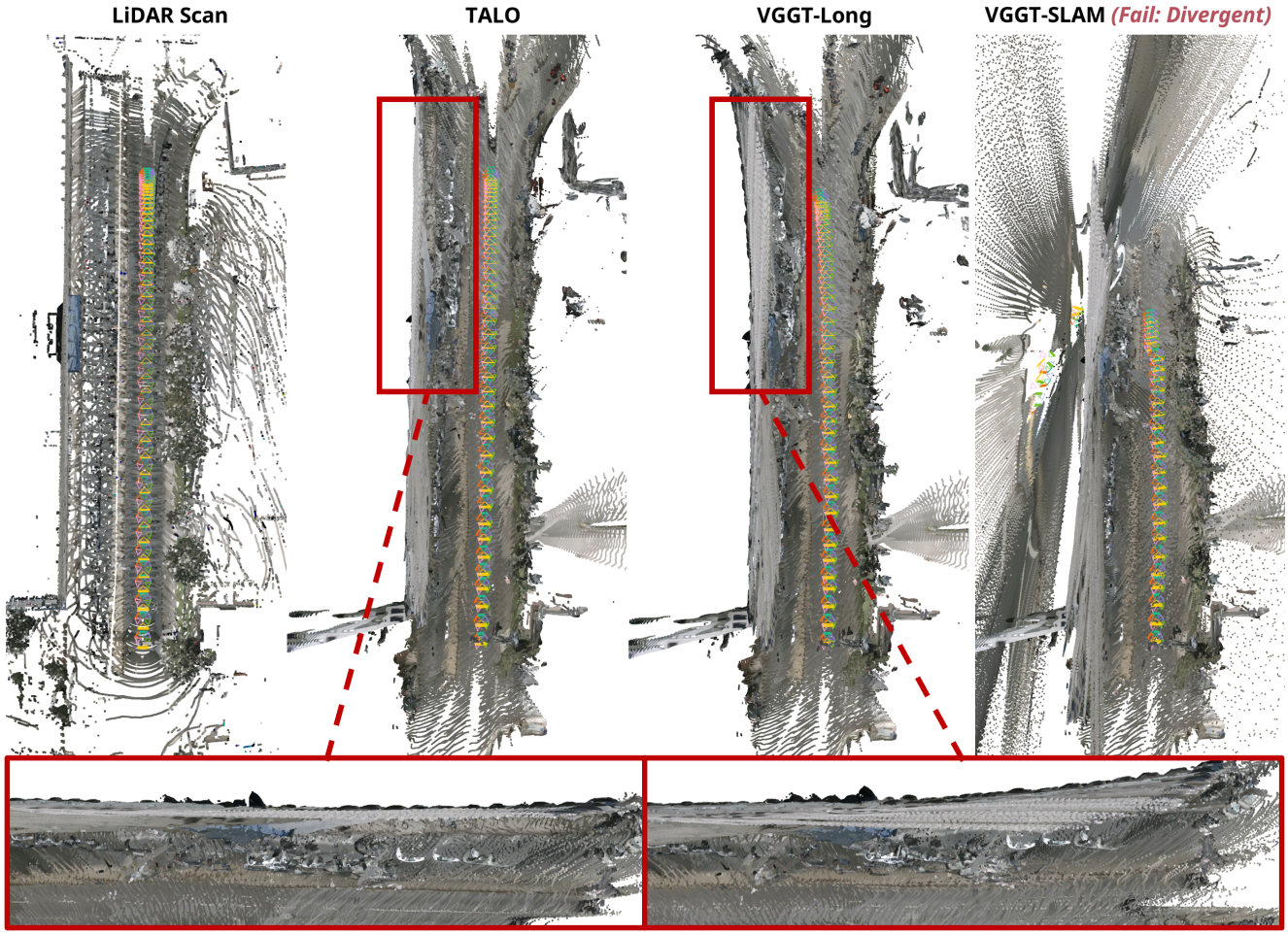

Figure 2. Qualitative comparison with VGGT [6] on **nuScenes** [1] scene-0094.

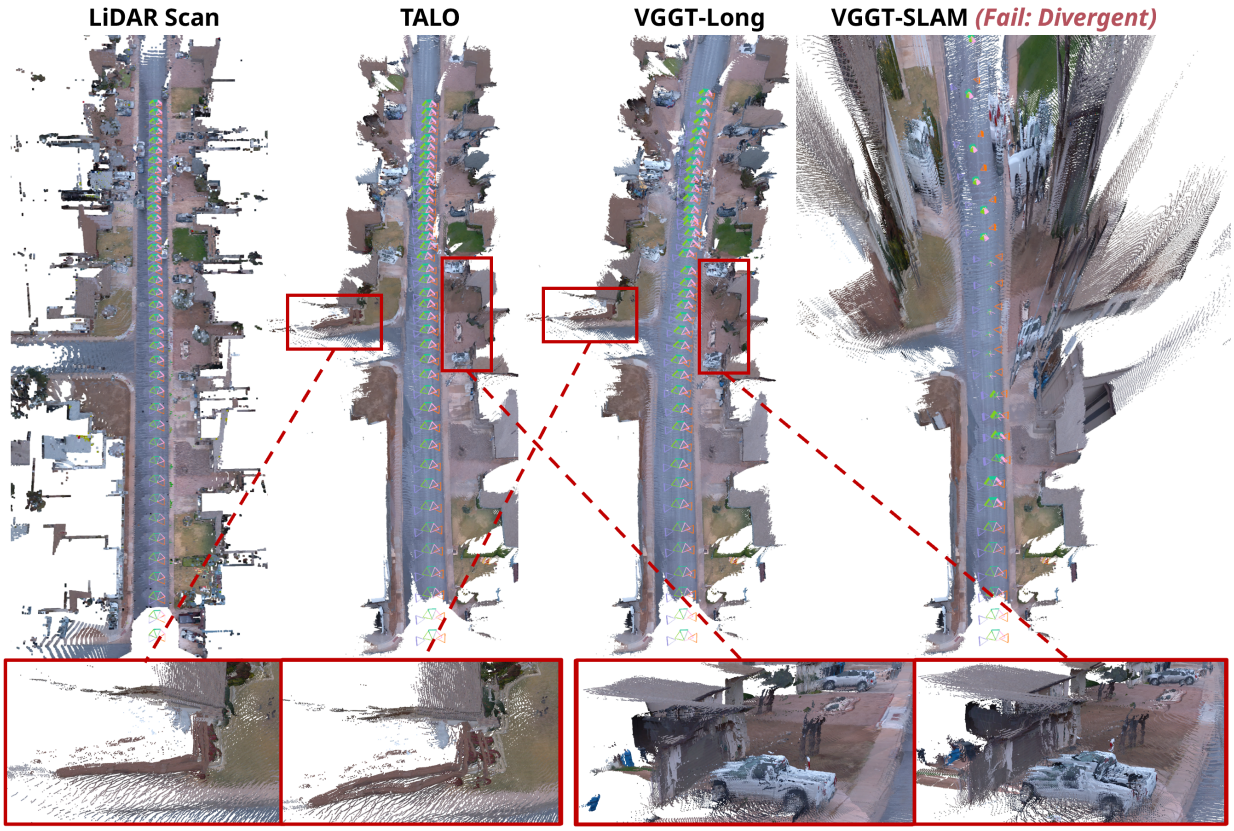

Figure 3. Qualitative comparison with π^3 [7] on **Waymo** [5] scene 3156155872654629090.

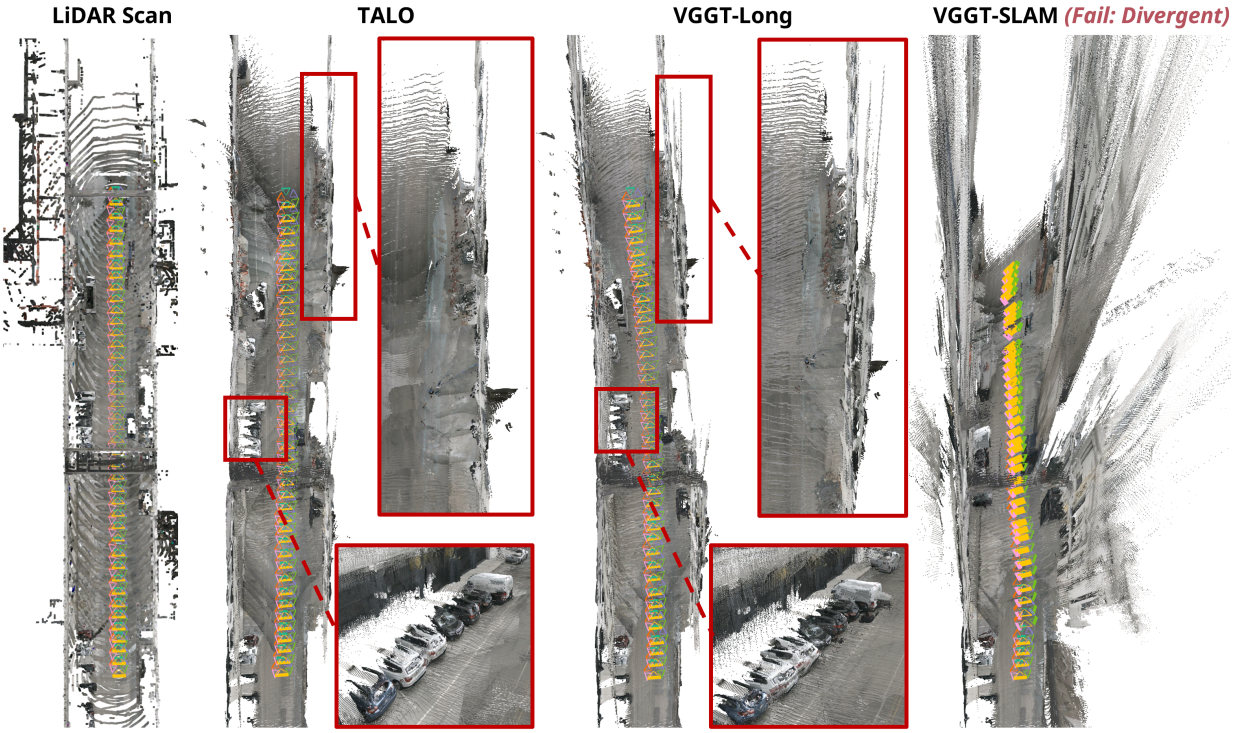

Figure 4. Qualitative comparison with π^3 [7] on **nuScenes** [1] scene-0092.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020. 2, 3

[2] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences, 2025. 1

[3] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction, 2025. 1

[4] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 1

[5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 3

[6] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2

[7] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 1, 3