LA-Pose: Latent Action Pretraining Meets Pose Estimation

Supplementary Material

The supplementary document provides additional qualitative visualizations and analysis:

- Qualitative comparison between LA-Pose and VGGT [1] under the low frame rate (1 fps) setting on Waymo. (§1)
- Additional qualitative results on the OpenDV–YouTube dataset. (§2)
- Analysis of failure modes across different motion regimes on the Waymo validation set. (§3)

1. Qualitative Results under Low Frame Rate

Figure 1 presents additional qualitative comparisons between LA-Pose and VGGT [1] under the low frame rate (1 fps) setting on the Waymo dataset. All visualizations follow the same protocol as in the main paper, where predicted camera trajectories are projected to the xz plane with camera frustums shown at frames 0, 5, 10, and 15.

At this extremely sparse temporal sampling, VGGT suffers from noticeable drift and unstable pose transitions, especially along long or turning trajectories. In contrast, LA-Pose produces smoother and more consistent camera trajectories, maintaining stable motion even with large temporal gaps between frames. These qualitative results highlight the robustness of our learned latent action representation when operating under low frame rate conditions.

2. Qualitative Results on OpenDV–YouTube

Figure 2 shows qualitative results of LA-Pose on the OpenDV–YouTube dataset [2]. The OpenDV–YouTube dataset [2] is a large-scale collection of unconstrained driving videos gathered from public YouTube channels. It forms the main component of OpenDV-2K, spanning over 1700 hours of front-view recordings captured across more than 40 countries and 240 cities, far exceeding the geographic coverage of our post-training datasets such as Waymo (San Francisco, Phoenix), nuScenes (Boston and Singapore), and Argoverse (six U.S. cities).This vast diversity covers a wide range of road types, weather conditions, lighting, and camera setups, making OpenDV–YouTube an extremely challenging for evaluating generalization.

OpenDV–YouTube consists of uncalibrated front-view driving videos collected from YouTube, recorded with diverse in-vehicle cameras of unknown intrinsic and extrinsic parameters. Since the dataset contains only raw RGB frames without ground-truth camera poses, we visualize our predicted camera trajectories to qualitatively assess pose consistency and realism.

We attribute this strong generalization capability to our pre-training stage, which learns a robust video representation from large-scale unlabeled data. The frozen backbone, pre-trained on massive driving video corpora, serves as a powerful feature extractor that captures high-level motion patterns and latent action structures. This foundation enables the model to transfer effectively to unseen conditions and datasets like OpenDV–YouTube, maintaining stable pose predictions even in unstructured, out-of-distribution environments.

3. Failure Mode Analysis

To further understand the limitations of our method, we analyze pose estimation performance across different trajectory curvatures and accelerations on the Waymo validation set. We categorize trajectories into bins based on curvature and acceleration to examine model performance under different motion regimes. Curvature is defined as $\kappa = d\psi/ds$, where ψ denotes the vehicle heading and s is the trajectory arc length. We divide curvature into three ranges: small (< 0.01 m^{-1}), medium (0.01–0.1 m^{-1}), and large (> 0.1 m^{-1}). Similarly, acceleration magnitude is divided into three ranges: < 0.3, 0.3–0.8, and > 0.8 m/s^2.

Table 1 reports AUC@5 (%) under these motion regimes. We observe that performance is lower on medium-curvature trajectories compared to straight or sharp-turn motions. Medium-curvature trajectories correspond to gradual steering behaviors where frame-to-frame geometric changes are subtle. In these situations, visual motion cues between consecutive frames are weak, making the motion representation learned through future-frame prediction less discriminative. By contrast, straight trajectories exhibit stable ego-motion patterns, while sharp turns introduce stronger geometric changes that are easier for the model to capture.

Table 1. AUC@5 (%) across different trajectory curvatures and accelerations on the Waymo validation set.

	Small	Medium	Large
Curvature	94.50	78.32	91.22
Acceleration	92.81	90.96	88.25

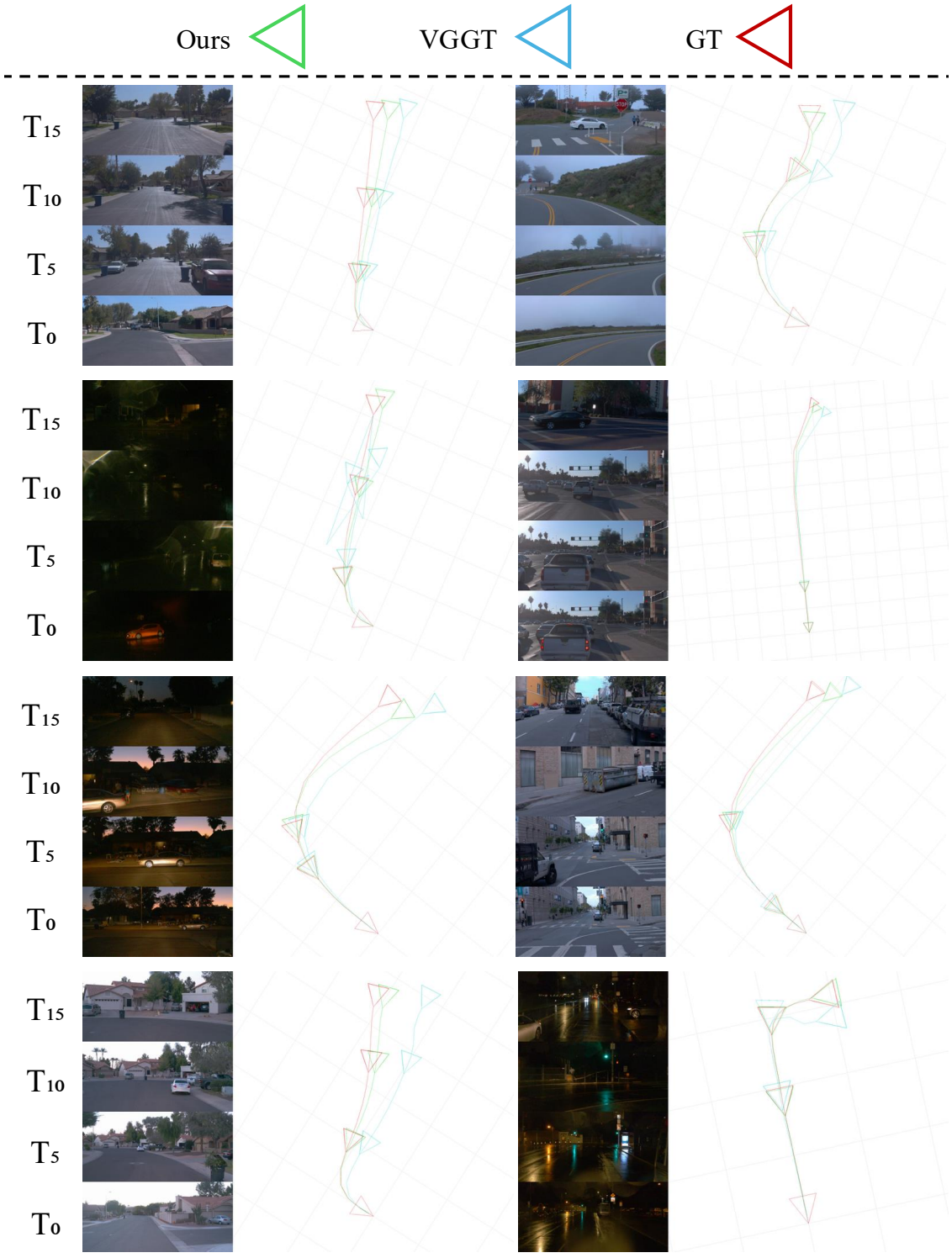

Figure 1. **Qualitative results under low frame rate (1 fps) on Waymo.** Each example shows camera poses projected onto the xz plane, with frustums drawn at frames 0, 5, 10, and 15. LA-Pose (green) maintains stable and temporally consistent motion across the sequence, whereas VGGT [1] (cyan) exhibits noticeable drift and discontinuities under sparse temporal sampling.

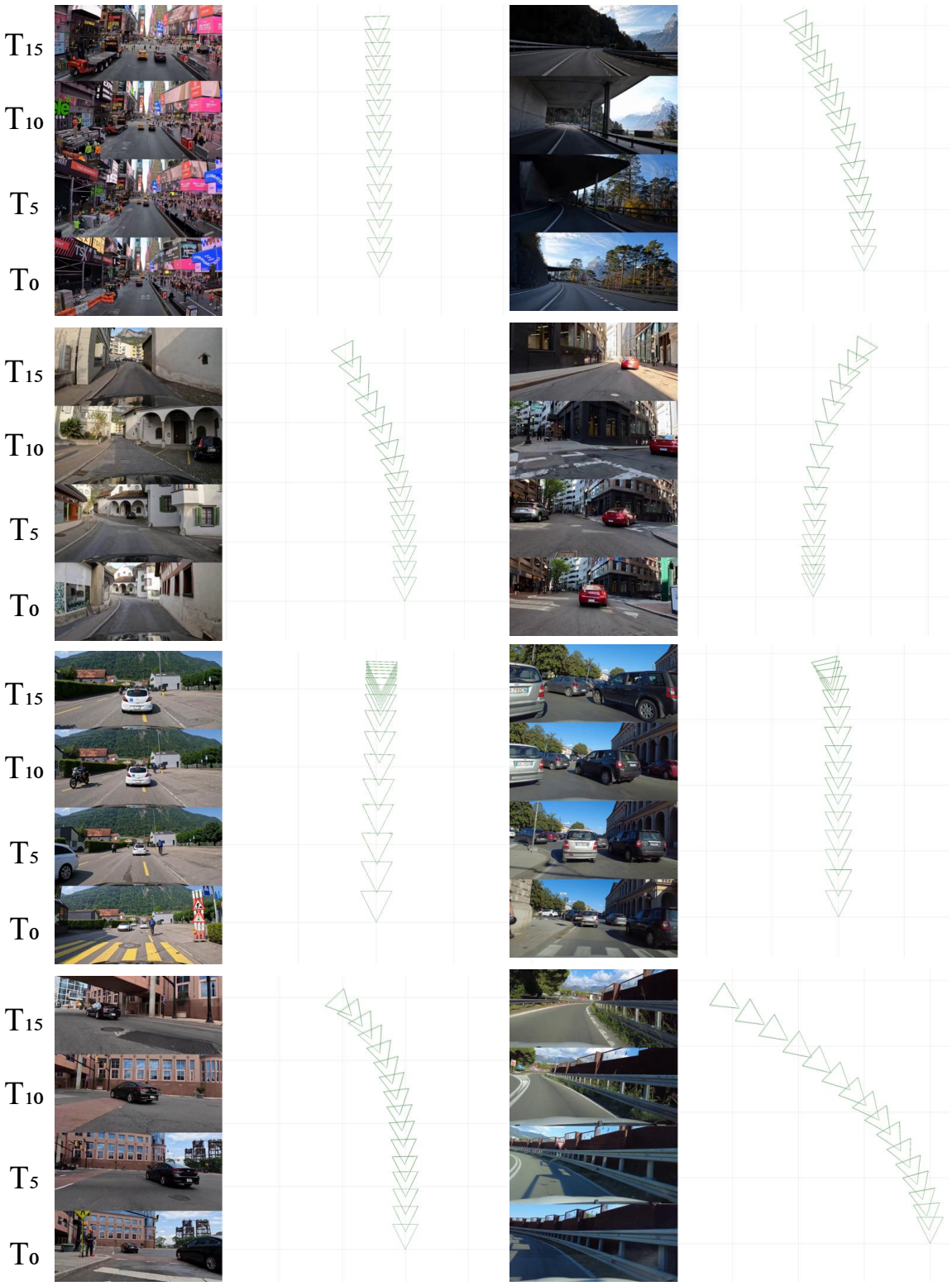

Figure 2. **Qualitative results on OpenDV–YouTube.** Each example shows scenes from diverse cities and viewpoints collected from online YouTube driving videos. LA-Pose produces stable and temporally consistent trajectories across a wide variety of conditions, including urban streets, highways, and curved mountain roads. The results qualitatively demonstrate strong generalization from our pre-trained backbone to uncalibrated, in-the-wild videos.

References

[1] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 2

[2] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Genad: Generalized predictive model for autonomous driving. *arXiv preprint arXiv:2403.09630*, 2024. 1