Supplementary Materials for

Co-Me: Confidence-Guided Token Merging for Visual Geometric Transformers

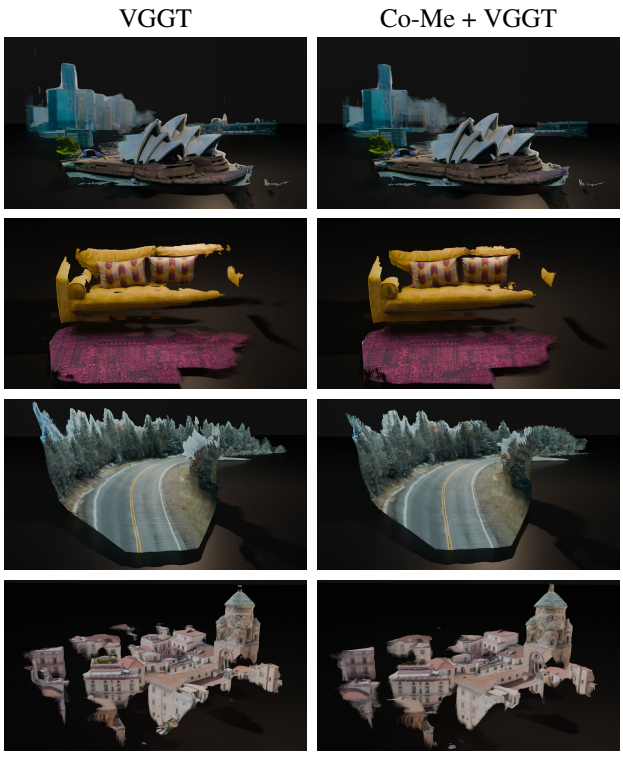

Figure 11. Qualitative comparison between VGGT (left) and Co-Me-accelerated VGGT (right). Best viewed digitally.

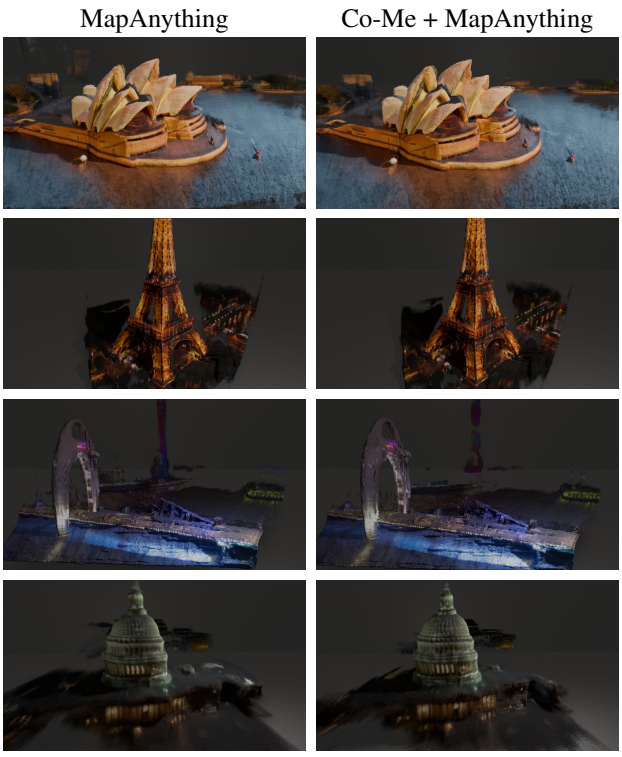

Figure 12. Qualitative comparison between MapAnything (left) and Co-Me-accelerated version (right). Best viewed digitally.

A. Qualitative Results

In this section we present qualitative results of Co-Me-accelerated VGGT and MapAnything. Specifically all results are created with the exact same configuration in Sec. 4 without finetuning or modification.

A.1. Success Cases

VGGT In Fig. 11, we show a qualitative comparison between VGGT (left) and Co-Me-accelerated VGGT (right) across eight representative scenes. Co-Me preserves the global scene structure and fine-grained geometry, including planar surfaces and prominent edges, despite operating with significantly fewer tokens. Minor differences appear primarily along the boundaries between high-confidence foreground regions and low-confidence background areas. These examples illustrate that confidence-guided merging maintains reconstruction fidelity with reduced computation.

MapAnything Figure 12 reports qualitative reconstructions from MapAnything and its Co-Me-accelerated variant across four diverse outdoor scenes. Despite aggressive token reduction, the accelerated model retains the characteristic large-scale structure that MapAnything recovers—such as façade geometry, smooth water surfaces, and distant skyline contours. Most observable differences are confined to peripheral regions where texture cues are weak or depth ambiguity is intrinsic to the input views. In these areas, Co-Me may slightly simplify fine-scale geometry, but the dominant scene layout and salient landmarks remain stable. These results show that token merging integrates cleanly with the MapAnything pipeline, preserving its strong global consistency while reducing inference cost.

A.2. Failure Modes

Figure 13 highlights scenarios where Co-Me introduces noticeable degradation. In both examples, the lost geometry corresponds to thin, high-frequency structures that occupy

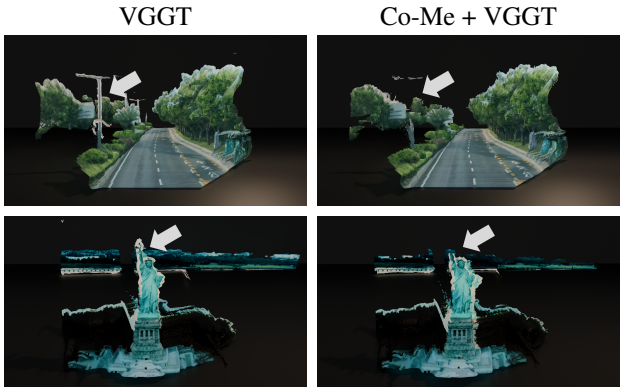

Figure 13. Failure cases of Co-Me-accelerated VGGT. Comparison between VGGT (left) and Co-Me-accelerated VGGT (right). Arrows indicate corrupted thin structures after token merging.

a small portion of the corresponding token. When these regions have low predicted confidence, merging discards their local resolution enough that the downstream decoder oversmooths the structure, causing incomplete reconstruction of the streetlight pole and the Statue of Liberty's raised arm. While these elements do not affect the global scene layout, they reveal a limitation of confidence-guided merging in handling small or elongated objects.

B. Additional Experiments

B.1. Confidence Distillation Layer Ablation

To investigate where the confidence predictor should be inserted within the ViT backbone, we trained the predictor on features extracted from different encoder layers of VGGT under identical training setups. Fig. 14 illustrates the ranking loss curves for predictors attached to layers 6, 9, 12, 15, 18, and 21 respectively. We observe that the predictor distilled from layer 15 achieves the lowest ranking loss across all layers. Earlier layers (e.g., 6, 9) provide insufficient semantic and geometric cues, leading to noisy confidence estimates, while later layers (e.g., 18, 21) have stronger geometric reasoning but reduced token diversity, which limits generalization and causes slower convergence. Therefore, we use the layer-15 configuration in all experiments, as it provides an optimal trade-off between confidence ranking accuracy and computational overhead.

B.2. Confidence Distillation Loss Ablation

In the Sec. 3.1, we replace the MSE objective with a ranking loss that supervises the relative ordering of token confidences. To validate the effectiveness of loss formulation, we conduct an ablation by retraining the model using MSE under identical settings. We then compare the resulting predictions by measuring the intersection-over-union (IoU) between the top-p merge masks derived from the distilled predictor and those obtained from the full VGGT model, with $p \in [0.2, 0.9]$. The IoU metric is defined as:

$$\text{IoU} = \frac{|\mathcal{M}_{\text{pred}} \cap \mathcal{M}_{\text{gt}}|}{|\mathcal{M}_{\text{pred}} \cup \mathcal{M}_{\text{gt}}|}, \tag{8}$$

where $\mathcal{M}_{\text{pred}}, \mathcal{M}_{\text{gt}}$ are the predicted and reference masks.

In Fig. 15, we can see that the ranking loss consistently outperforms MSE, demonstrating that supervising the relative ordering of confidences is more effective than regressing the confidence numerically for predicting merge masks.

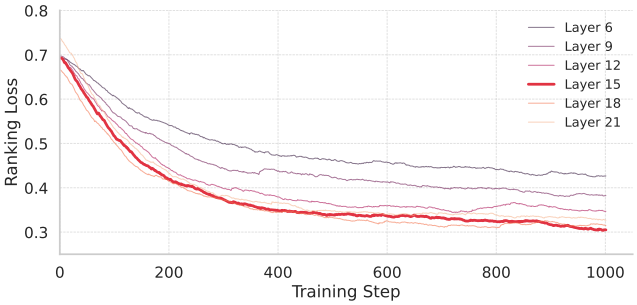

Figure 14. Distillation loss of confidence predictors distilled from various VGGT encoder layers. Layer 15 yields the lowest loss, indicating that mid-level encoder features contain the most information for confidence estimation. For readability, curves are smoothed with an exponential moving average with factor of 0.99.

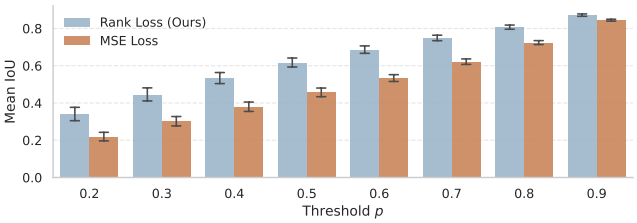

Figure 15. Confidence distillation with ranking loss achieves significantly higher IoU on DTU dataset than MSE loss. Error bar shows the 95% confidence interval.

B.3. Token Group Size Ablation

To evaluate the influence of token group size on the speed–accuracy trade-off, we tested Co-Me with group sizes of 2, 4, and 6 under identical merging ratios on DTU-MVS (32 frames). As shown in Fig. Fig. 16, smaller group sizes generally offer better accuracy retention for a given speedup, as they introduce finer-grained control and less information loss over which merged tokens. In contrast, larger groups provide stronger acceleration due to more aggressive token reduction, but incur slightly higher reconstruction error. Overall, group size 4 achieves the best balance between efficiency and accuracy and is therefore used for all experiments in Secs. 4 and 5.

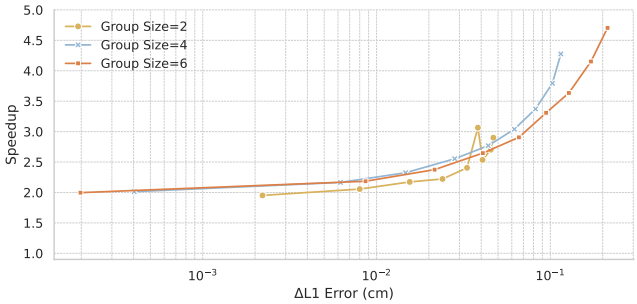

Figure 16. Speedup-accuracy trade-off of Co-Me-accelerated VGGT across various token group sizes on multi-view depth estimation (DTU-MVS, 32 frames). Smaller group sizes yield slightly better accuracy, while larger groups provide higher acceleration. Curves are plotted on a log-scaled error axis for clarity.

C. Edge Compute Deployment

In Fig. 9, we illustrate the real-world deployment setup and runtime profile used to evaluate edge performance. An NVIDIA Jetson Thor runs MapAnything and our Co-Me-accelerated variant while receiving stereo input from a Zed 2i camera. The system groups incoming frames into fixed segments of four images and accumulates the resulting reconstructions in a global world coordinate frame, effectively simulating a streaming visual-odometry pipeline.

The stacked runtime bars in Fig. 9 decompose per-segment latency into DINO, frame-level, and global attention components, linear projections, Co-Me overhead, and other operations. Applying Co-Me shrinks the attention-dominated portions while adding only a small confidence-prediction cost, yielding an overall 1.5× reduction in end-to-end runtime. On this platform, processing 4-image segments reaches 3.5 FPS, providing near real-time responsiveness under edge-compute constraints while preserving the stable 3D geometry observed in Sec. 5, H5.