# CogniMap3D: Cognitive 3D Mapping and Rapid Retrieval

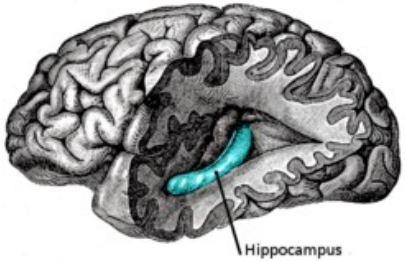
Feiran Wang[1], Junyi Wu[1], Dawen Cai[2], Yuan Hong[3], Yan Yan[1†]

[1]University of Illinois Chicago, [2]University of Michigan, [3]University of Connecticut

# Motivation

## How Humans Perceive Dynamic Scenes

Our attention naturally prioritizes moving objects while building persistent spatial representations of static environments. The hippocampus constructs internal "cognitive maps" in an egocentric reference frame. When revisiting familiar places, humans reliably recall static scenes even when dynamic elements have changed.




***Can we build 3D systems that mimic this cognitive process?***

# Key Challenges

## Challenge 1: Dynamic vs. Static Separation

How to accurately distinguish static and dynamic scene elements in monocular videos?

## Challenge 2: Persistent Memory Across Visits

How to construct and maintain persistent representations of static environments that can be efficiently recalled and updated when revisiting familiar scenes?

## Challenge 3: Geometrically Consistent Camera Poses

How to establish stable camera pose estimates that remain geometrically consistent despite the presence of dynamic objects?

# Our Contributions

## 1. Multi-stage Motion Cue Framework

Progressive 2D-3D motion analysis for accurate dynamic object identification.

## 2. Cognitive Mapping System

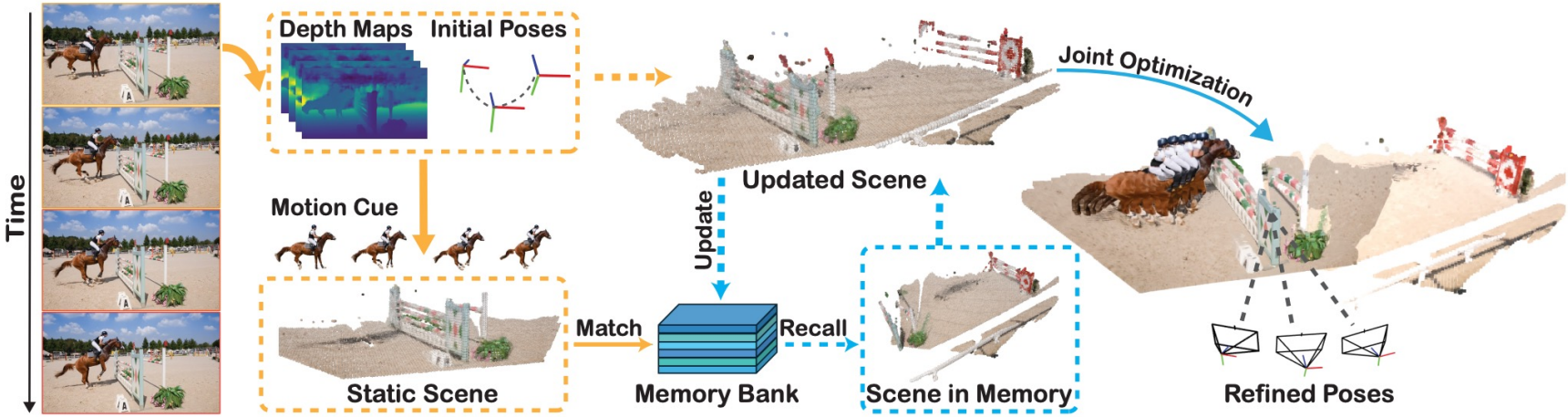
Store, recall, and update static scene memory across multiple visits.

## 3. Factor Graph Optimization

Joint camera pose refinement using constraints from static regions and updated memory.

# Pipeline Overview

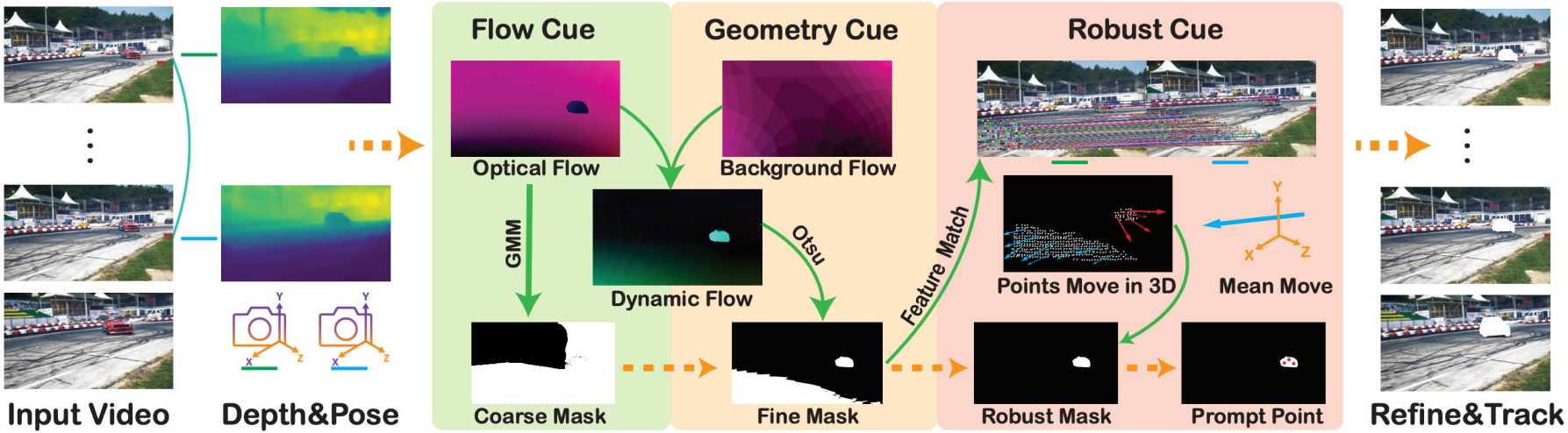
CogniMap3D maintains a cognitive mapping system that recalls, stores, and updates memories. Given an input video, it outputs camera poses and point clouds by isolating static scenes through motion cues, interacting with its memory bank, and optimizing across multiple visits.



# Multi-stage Motion Cue

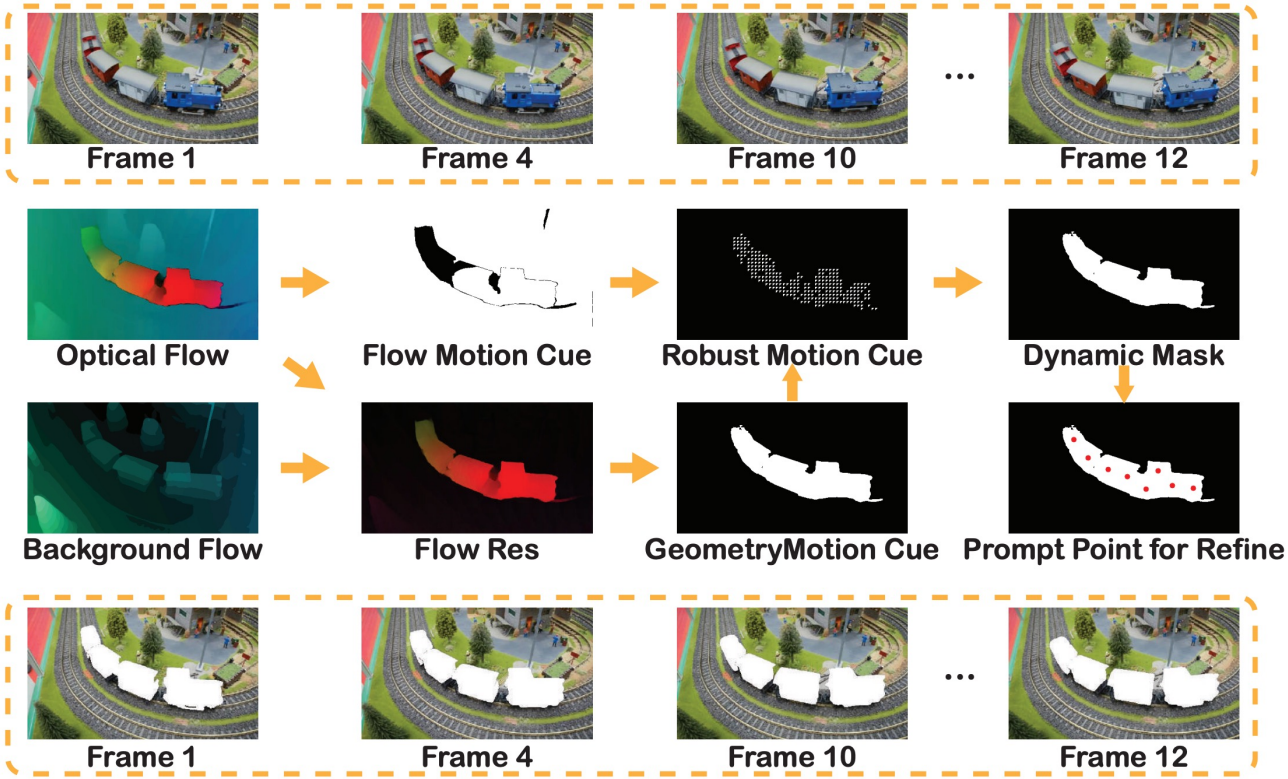
Given a pair of images, predict initial depth and camera pose via VFM to establish 3D prior. Three progressive motion cues through 2D-3D interaction:

**Flow Cue** (optical flow clustering) → **Geometry Cue** (depth-based 3D motion) → **Robust Cue** (3D keypoint refinement for precise dynamic masks).



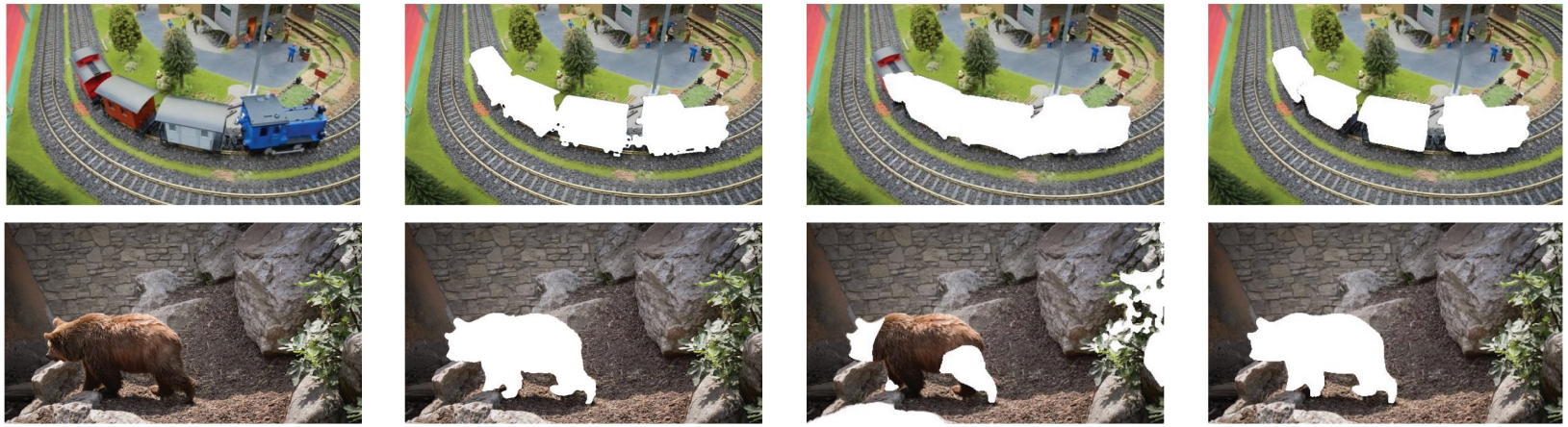
# Motion Cue: Example on DAVIS

Motion Cue Process of the Train Scene in DAVIS dataset.



# Dynamic Mask Comparison

Dynamic Mask Comparison. We visualize dynamic regions as white overlays on input images. Compared with MonST3R, our method achieves more complete and precise masks.



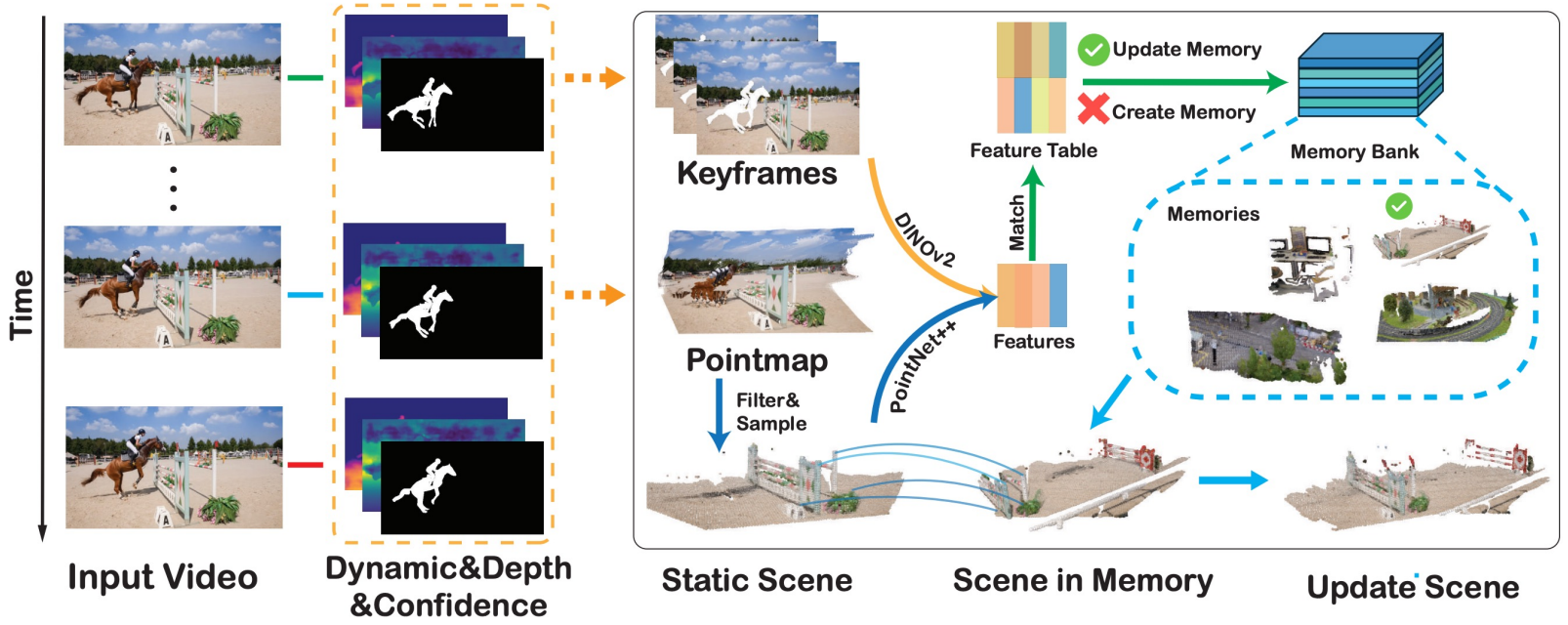
# Cognitive Mapping System

Estimate per-frame dynamic mask with depth, confidence, and camera pose priors. Filter static scenes using dynamic masks, then encode image and point cloud features.

Match against memory bank:

❎ No match → create new memory slot

✅ Match found → recall & update existing memory



# Camera Trajectory Optimization

We propose a factor graph optimization approach to refine camera trajectories, enhancing global geometric consistency through static scene constraints from both current observations and memory.

$$X^* = \text{argmin}_X \sum_{f \in F} \|C(X_f)\|^2_{\Sigma_f^{-1}}$$

## Step 1: Initial Landmark Selection

Extract candidate landmarks exclusively from **static regions** with high confidence. For recognized scenes, recalled memory points serve as additional landmarks with stronger geometric constraints.

## Step 2: Factor Graph Optimization

Jointly optimize camera poses and 3D landmarks with three complementary constraints:

- **Projection Factor:** 3D landmarks should project correctly onto their observed 2D positions
- **Prior Factor:** Anchor the first camera pose to prevent coordinate drift
- **Motion Factor:** Enforce smooth transitions between consecutive frames

$$C(X) = |f_{prior}|^2_{\Sigma^{-1}_{prior}} + \sum_{(i,j) \in \mathcal{O}} \rho(|f_{proj}|^2_{\Sigma^{-1}_{proj}}) + \sum_{i=1}^{N-1} |f_{motion}|^2_{\Sigma^{-1}_{motion}}$$

# Qualitative Results

Our method achieves cleaner reconstructions with better preservation of both static and dynamic elements.
The bottom rows demonstrate CogniMap3D's unique capability to store previous scenes in memory and recall them upon revisitation

# Quantitative Results

Table 1: **Video Depth Evaluation.** We report scale&shift-invariant depth accuracy and FPS. Methods requiring global alignment are marked "GA", while "Optim." and "FF" indicate optimization and online methods.

| | | | | **Sintel** Butler et al. (2012) | | **BONN** Palazzolo et al. (2019) | | **KITTI** Geiger et al. (2013) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | **Method** | **Optim.** | **FF** | Abs Rel ↓ | $\delta$<1.25 ↑ | Abs Rel ↓ | $\delta$<1.25 ↑ | Abs Rel ↓ | $\delta$ <1.25 ↑ | **FPS** |
| Depth Estimation model | Depth-Anything-V2 Yang et al. (2024) | | ✓ | 0.367 | 55.4 | 0.106 | 92.1 | 0.140 | 80.4 | 3.13 |
| | ChronoDepth Shao et al. (2024) | | ✓ | 0.687 | 48.6 | 0.100 | 91.1 | 0.167 | 75.9 | 1.89 |
| | DepthCrafter Hu et al. (2024) | | ✓ | **0.292** | **69.7** | **0.075** | **97.1** | **0.110** | **88.1** | 0.97 |
| Vision Foundation Model | DUSt3R-GA Wang et al. (2024) | ✓ | | 0.531 | 51.2 | 0.156 | 83.1 | 0.135 | 81.8 | 0.76 |
| | MASt3R-GA Leroy et al. (2024) | ✓ | | 0.327 | 59.4 | 0.167 | 78.5 | 0.137 | 83.6 | 0.31 |
| | MonST3R-GA Zhang et al. (2024) | ✓ | | 0.333 | 59.0 | 0.066 | 96.4 | 0.157 | 73.8 | 0.35 |
| | Spann3R Wang & Agapito (2024) | | ✓ | 0.508 | 50.8 | 0.157 | 82.1 | 0.207 | 73.0 | 13.55 |
| | CUT3R Wang et al. (2025b) | | ✓ | 0.454 | 55.7 | 0.074 | 94.5 | 0.106 | 88.7 | 16.58 |
| | VGGT Wang et al. (2025a) | | ✓ | 0.299 | 62.4 | **0.054** | 97.1 | 0.072 | **96.4** | 21.5 |
| | **Ours** | | ✓ | **0.295** | **68.6** | 0.058 | **97.9** | **0.069** | 96.2 | 14.32 |

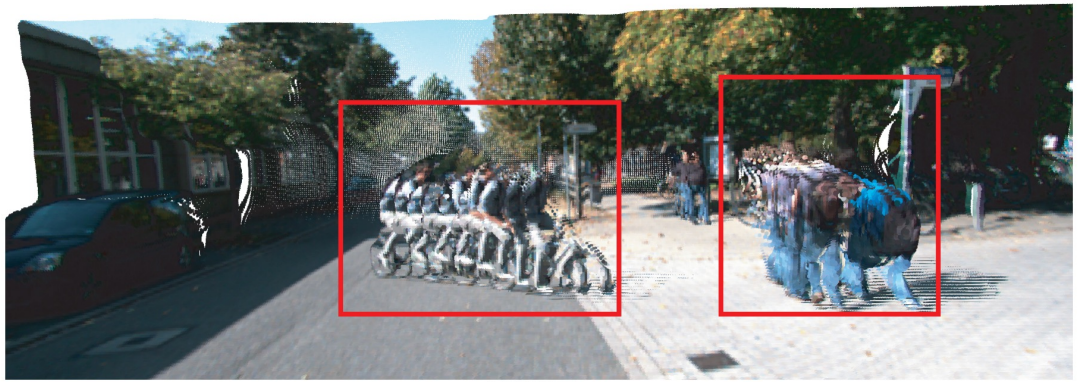
Table 2: **Quantitative Results of 3D reconstruction.** Evaluation on 7-Scenes dataset shows our approach without memory machinery still achieves comparable results to SOTA methods.

| | | | **7-Scenes Shotton et al. (2013)** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Optim.** | **FF** | Acc↓ Mean | Acc↓ Med. | Comp↓ Mean | Comp↓ Med. | NC↑ Mean | NC↑ Med. | **FPS** |
| DUSt3R-GA Wang et al. (2024) | ✓ | | 0.146 | 0.077 | 0.181 | 0.067 | 0.736 | 0.839 | 0.68 |
| MonST3R-GA Zhang et al. (2024) | ✓ | | 0.248 | 0.185 | 0.266 | 0.167 | 0.672 | 0.759 | 0.39 |
| CUT3R Wang et al. (2025b) | | ✓ | 0.126 | 0.047 | 0.154 | **0.031** | 0.727 | 0.834 | 17.0 |
| VGGT Wang et al. (2025a) | | ✓ | 0.088 | **0.040** | 0.092 | 0.040 | **0.784** | **0.888** | 21.5 |
| **Ours** | | ✓ | **0.086** | 0.041 | **0.089** | 0.039 | 0.751 | 0.863 | 14.3 |

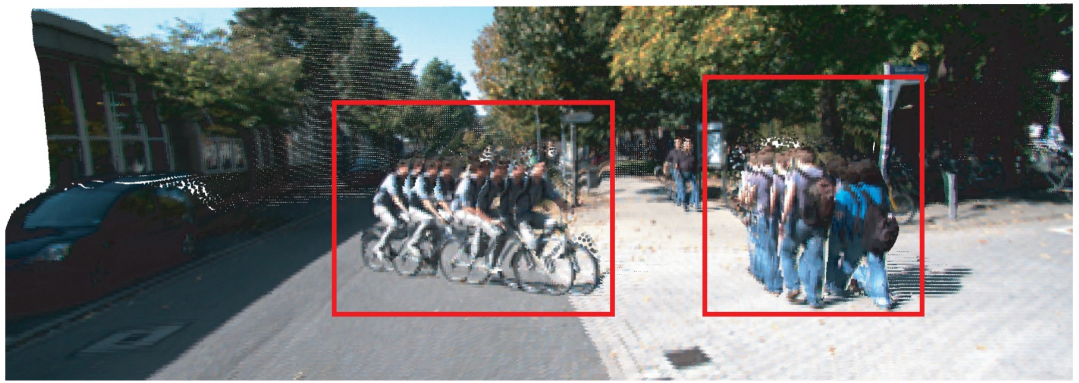
# Visual Comparison: VGGT vs. Ours

We compare reconstructed point clouds rendered over the input image for VGGT (top) and CogniMap3D (bottom).

The red boxes highlight regions with moving area (cyclist and pedestrians)



VGGT



CogniMap3D

# Camera Pose Estimation

Table 3: **Camera Pose Estimation Evaluation** on Sintel, TUM-dynamic, and ScanNet datasets. We group methods into (I) SLAM-based methods requiring intrinsics, (II) optimization-based VFM methods, and (III) feed-forward VFM methods.

| | | Sintel Butler et al. (2012) | | | TUM-dynamic Sturm et al. (2012) | | | ScanNet Dai et al. (2017) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | **Method** | ATE ↓ | RPE trans ↓ | RPE rot ↓ | ATE ↓ | RPE trans ↓ | RPE rot ↓ | ATE ↓ | RPE trans ↓ | RPE rot ↓ |
| I | DROID-SLAM Teed & Deng (2021) | 0.175 | 0.084 | 1.912 | - | - | - | - | - | - |
| | DPVO Teed et al. (2023) | 0.115 | 0.072 | 1.975 | - | - | - | - | - | - |
| | LEAP-VO Chen et al. (2024) | **0.089** | **0.066** | **1.250** | 0.068 | 0.008 | 1.686 | 0.070 | 0.018 | 0.535 |
| II | Robust-CVD Kopf et al. (2021) | 0.360 | **0.154** | 3.443 | 0.153 | 0.026 | 3.528 | 0.227 | 0.064 | 7.374 |
| | CasualSAM Zhang et al. (2022) | 0.141 | 0.035 | **0.615** | 0.071 | **0.010** | 1.712 | 0.158 | 0.034 | 1.618 |
| | DUSt3R-GA Wang et al. (2024) | 0.417 | 0.250 | 5.796 | 0.083 | 0.017 | 3.567 | 0.081 | 0.028 | 0.784 |
| | MASt3R-GA Duisterhof et al. (2024) | 0.185 | 0.060 | 1.496 | **0.038** | 0.012 | **0.448** | 0.078 | 0.020 | **0.475** |
| | MonST3R-GA Zhang et al. (2024) | **0.111** | 0.044 | 0.869 | 0.098 | 0.019 | 0.935 | **0.077** | **0.018** | 0.529 |
| III | DUSt3R Wang et al. (2024) | 0.290 | 0.132 | 7.869 | 0.140 | 0.106 | 3.286 | 0.246 | 0.108 | 8.210 |
| | Spann3R Wang & Agapito (2024) | 0.329 | 0.110 | 4.471 | 0.056 | 0.021 | 0.591 | 0.096 | 0.023 | 0.661 |
| | CUT3R Wang et al. (2025b) | 0.213 | **0.066** | 0.621 | 0.046 | 0.015 | 0.473 | 0.099 | 0.022 | 0.600 |
| | VGGT Wang et al. (2025a) | 0.189 | 0.069 | **0.529** | 0.028 | 0.020 | 0.350 | 0.023 | 0.015 | **0.326** |
| | **Ours** | **0.176** | 0.068 | 0.600 | **0.012** | **0.010** | **0.311** | **0.019** | **0.011** | 0.331 |

# Ablation Study

Table 4: **Memory Recall Analysis.**

| Method | Acc↓ | Comp↓ | NC↑ |
|---|---|---|---|
| MonST3R-GA | 0.248 | 0.266 | 0.672 |
| **Ours** | 0.086 | 0.089 | 0.751 |
| **Ours Update** | **0.082** | **0.085** | **0.789** |

We demonstrate this capability on the 7-Scenes dataset as shown in Table 4. For evaluation, we first initialize the memory bank with a single randomly selected frame from each scene. Leveraging memory recall, our method with updated memory outperforms both baseline methods and our model without memory.

Table 6: **Memory Size.**

| Number | Accuracy (%) |
|---|---|
| 1 | 100 |
| 50 | 96 |
| 100 | 97 |
| 200 | 97.5 |

Memory Matching. We evaluate CogniMap3D's memory matching on DAVIS (Perazzi et al., 2016) by dividing 50 scenes into thirds and performing 200 pairwise matches between segments.

# Conclusion

## CogniMap3D — A bioinspired framework for dynamic 3D scene reconstruction

✓ **Multi-stage Motion Cue:** Accurate dynamic/static separation via progressive 2D-3D motion analysis

✓ **Cognitive Mapping:** Persistent memory bank enabling scene recall & update across multiple visits

✓ **Factor Graph Optimization:** Geometrically consistent camera pose refinement

## Key Results

State-of-the-art performance on video depth estimation, camera pose reconstruction, and 3D mapping. Unique capability: continuous scene understanding across extended sequences and revisitations.

# Thank You!