# OmegaMotion: Mining Multi-Level Geometric Priors from 3D Foundation Models for Video Motion Segmentation

Jia-Hui Shen[1] Cheng Sun[2] Min-Hung Chen[2] Fu-En Yang[2] Yu-Lun Liu[1] Yen-Yu Lin[1]
[1]National Yang Ming Chiao Tung University [2]NVIDIA
jhshen.cs13@gmail.com chengsun@gapp.nthu.edu.tw vitec6@gmail.com
aoncpc@gmail.com yulunliu@cs.nycu.edu.tw lin@cs.nycu.edu.tw

## Abstract

*Video motion segmentation aims to identify moving objects in a video. It remains challenging because observed motion is jointly affected by object dynamics, camera movement, depth-dependent parallax, and scene geometry. Existing methods often rely on optical flow, point trajectories, mask propagation, or iterative test-time optimization, leading to error accumulation and high computational costs. To this end, we introduce **OmegaMotion**, a feed-forward framework that repurposes a frozen 3D foundation model for direct frame-wise motion mask prediction. OmegaMotion leverages the rich spatio-temporal geometric priors and camera-aware representations learned by VGGT-Ω to construct a multi-level encoding of scene structure and dynamics. By unifying these learned geometric priors with explicit dynamics cues, our model directly isolates independent object motion masks from camera-induced displacement in a single forward pass without requiring per-video, test-time optimization. Extensive experiments demonstrate that OmegaMotion achieves state-of-the-art performance across multiple motion segmentation benchmarks while delivering superior inference efficiency.*

## 1. Introduction

Video motion segmentation aims to predict category-agnostic binary masks that identify objects moving independently in a scene. It serves as a foundational building block for downstream tasks such as video understanding, robotic perception, autonomous navigation, and dynamic-scene reconstruction [7, 11, 20, 42, 48]. Despite its clear goal, segmenting independent object motion from raw visual input is intrinsically ill-posed, as observed 2D pixel displacement is governed by a complex mixture of object dynamics, camera motion, and depth-dependent parallax, and is further complicated by occlusion, non-rigid deformation, and transient movements [7, 11, 26]. Consequently, scene elements

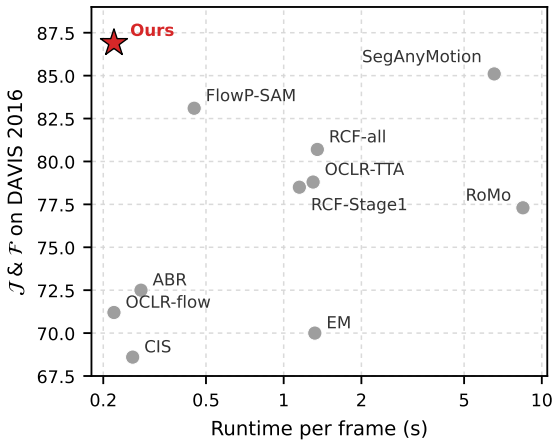



Figure 1. **Accuracy–efficiency trade-off on DAVIS 2016 (accuracy in $\mathcal{J}\&\mathcal{F}$ and per-frame inference time in seconds).** All runtime measurements are obtained from RGB inputs on a single NVIDIA RTX A6000 GPU. Methods toward the upper-left region demonstrate a more favorable trade-off, detailed in Section 4.2.

can exhibit large apparent motion under camera movement, whereas genuinely dynamic objects may exhibit only weak motion relative to the camera. Resolving these ambiguities therefore requires joint reasoning over 2D optical displacement, viewpoint dynamics, and 3D scene structure.

Existing motion segmentation paradigms can be broadly categorized according to their primary sources of motion reasoning. *2D motion-driven methods* estimate image-plane displacement using optical flow [23, 42–44], point trajectories [2, 11, 14, 27], or temporal correspondences [20, 47], and then group these estimates via clustering or object-centric decoding. Recent methods [11, 43, 44] further employ foundation models like SAM or SAM2 [15, 33] to restore mask completeness from sparse prompts. However, projected 2D displacement entangles camera-induced background motion with independent object motion, while flow and trajectories may become unreliable under occlusion, weak texture, rapid motion, or long-term drift.

Alternatively, *3D geometry-aware methods* incorporate epipolar constraints [7, 26], depth, camera pose, or scene structure [19, 32] to explain motion geometrically. While these explicit geometric cues resolve 2D motion ambiguities, their success depends heavily on accurate camera calibration and depth estimation. Furthermore, modern pipelines, such as OCLR-TTA [42], RoMo [7], and SegAnyMotion [11], rely on iterative pose-and-mask refinement and/or test-time optimization, incurring substantial computational latency during inference.

Concurrently, feed-forward 3D foundation models learn rich geometric priors from massive reconstruction data [38–40], which increasingly benefit dynamic-scene reasoning and motion segmentation [4, 9, 10, 45, 50]. However, existing methods often rely on geometric reconstruction outputs or iterative reconstruction, making them prone to estimation errors, imprecise motion boundaries, and substantial computational overhead. Moreover, predominantly static-scene supervision limits geometric reliability under non-rigid motion, severe occlusion, and rapid scene changes.

Rather than treating geometric reconstruction outputs as the final representation, we argue that the internal representations across different token types and network depths in 3D foundation models contain rich, complementary motion and structural cues. Specifically, multi-layer patch features capture hierarchical scene structure, camera and register tokens provide global view-conditioned context, confidence-aware depth describes depth-dependent geometry, and optical flow supplies explicit short-term displacement.

Driven by this insight, we present **OmegaMotion**, a feed-forward framework that mines multi-level priors from VGGT-Ω for video motion segmentation. We align and fuse latent 3D features, camera representations, confidence-aware dense depth, and explicit optical flow onto a unified patch grid, upon which a five-layer self-attention motion decoder and a mask head predict frame-wise motion masks in a single forward pass without per-video test-time optimization. During inference, each prediction can prompt a frozen SAM2 module [33] for spatial boundary refinement.

Our primary contributions are summarized as follows:

- We present **OmegaMotion**, a feed-forward framework that transfers geometric priors from a frozen 3D foundation model to facilitate video motion segmentation without test-time optimization.
- We design a structured aggregation and fusion scheme that consolidates multi-level geometric and motion priors into a unified representation, enabling a compact self-attention decoder to disentangle object motion from camera-induced displacement.
- Extensive evaluations across multiple benchmarks confirm that OmegaMotion achieves state-of-the-art accuracy while executing over 20× faster than representative iterative baselines, as showed in Figure 1.

## 2. Related Work

### 2.1. 3D Foundation Models

Recent feed-forward 3D foundation models learn generalizable geometric representations from large-scale multi-view data. DUSt3R [40] directly regresses pixel-aligned point maps from uncalibrated image pairs, replacing the conventional pipeline of separately estimating image correspondences, camera parameters, and depth. VGGT [38] extends this paradigm by jointly predicting camera parameters, depth maps, point maps, and point tracks within a unified transformer architecture based on alternating frame-wise and global attention. VGGT-Ω [39] further scales this architecture in both model capacity and training data, providing stronger multi-view geometric representations and camera-aware scene context.

Beyond explicit final predictions, the intermediate representations learned by these models encode rich structural relationships, viewpoint geometries, and cross-frame correspondences, making them highly effective for downstream visual reasoning tasks [3, 8, 17, 25]. Our work builds upon this foundation by exploiting frozen VGGT-Ω representations across network depths and representation types for feed-forward video motion segmentation.

### 2.2. Video Motion Segmentation

**2D motion-driven methods.** Conventional motion segmentation relies on the premise that pixels within an independently moving object exhibit coherent image-plane displacement. Early methods therefore group optical flow fields [1, 26, 29, 41] or long-term point trajectories based on motion orientation, temporal consistency, and trajectory affinity [2, 14, 27]. However, motion coherence alone frequently fails when objects deform, stop temporarily, or move similarly to the background. Learning-based methods mitigate these failures by supplementing displacement cues with appearance and object-level priors. Early learning models rank flow-derived spatio-temporal proposals [6], or jointly model appearance and motion [12, 35]. Subsequent approaches transition from hand-crafted grouping toward learned dynamic-static separation [20, 47, 49] and object-centric motion representations [42, 43].

More recently, foundation segmentation models and long-range point trackers have been incorporated to recover spatially complete object masks from noisy motion observations. FlowSAM [44] links optical flow fields with object proposals generated by SAM [15], while trajectory-based approaches leverage long-range point tracks as supervision, grouping cues, or segmentation prompts [11, 14].

Although these approaches improve object completeness and temporal consistency, their performance remains susceptible to optical flow degradation due to partial occlusion, weak texture, rapid motion, and tracking drift. In addition,

their multi-stage pipelines can be computationally expensive. More crucially, projected 2D displacement alone cannot resolve the intrinsic ambiguity between object motion, camera ego-motion, and depth-dependent parallax.

**3D geometry-aware methods.** Geometry-aware methods address projection ambiguities by explicitly accounting for scene depth, camera trajectory, and 3D structural constraints. Early frameworks explicitly decompose or jointly estimate these factors: SfM-Net [37] separates optical flow into depth, camera motion, and rigid object motion, while Ranjan et al. [32] coordinate depth, camera motion, optical flow, and motion segmentation through multiple interacting modules. While such geometric factorization provides a principled foundation for distinguishing dynamic objects from a moving background, it also couples segmentation accuracy to the reliability of depth, pose, flow, and correspondence estimates.

Modern methods like RoMo [7] mitigate this dependence by alternating epipolar reasoning with camera-pose and motion-mask refinement, while MegaSaM [19] jointly derive depth, flow, and cameras for structure-and-motion recovery in videos. Although these iterative refinement schemes can correct inconsistent estimates, they introduce substantial test-time computation and remain vulnerable to error accumulation across interdependent stages.

Pretrained 3D foundation models offer an alternative source of geometric knowledge. MonST3R [50] explores time-dependent point maps with flow-geometry consistency to distinguish static and dynamic regions, while DAS3R [45] incorporates predicted motion masks into dynamics-aware scene reconstruction. Moving beyond explicit reconstruction outputs, Easi3R [4] extracts camera-object motion cues from the cross-attention features in DUSt3R, while VGGT4D [10] shows that dynamic signals are distributed across multiple layers of VGGT and aggregates layer-wise attention for dynamic masking and 3D reconstruction. Moving from motion discovery toward direct segmentation, GeoMotion [9] integrates latent 4D geometry into a feed-forward motion-segmentation framework. While these works highlight the utility of pretrained geometric representations, the systematic extraction and fusion of complementary cues across varying network depths and token types remain largely underexplored.

Unlike prior cue-level fusion schemes, our OmegaMotion systematically mines motion-relevant representations across network depths and token types within a frozen VGGT-Ω backbone. Cross-depth patch features capture complementary object-level and global geometric structure, while camera and register tokens condition spatial features to isolate object motion from camera displacement. Together with confidence-aware depth and bidirectional flow, these priors enable direct frame-wise motion segmentation in a single forward pass without test-time optimization.

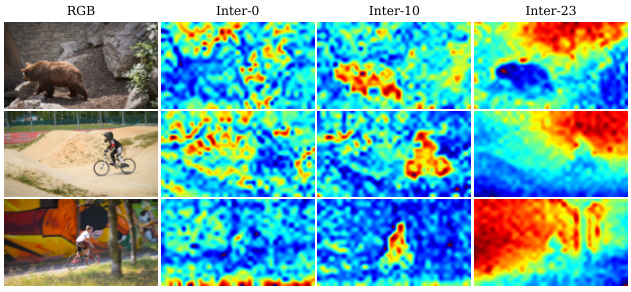



Figure 2. **Layer-wise visualization of frozen VGGT-Ω patch features.** The shallow layer (Inter-0) captures localized, small-scale salient regions; the intermediate layer (Inter-10) forms coherent regions aligned with object geometry and contours; and the deep layer (Inter-23) exhibits broad scene-level organization.

## 3. Method

Given a video sequence $\mathcal{I} = \{I^t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{N}$ containing $N$ frames, our objective is to predict a category-agnostic binary motion mask for each frame. We introduce **OmegaMotion**, a feed-forward framework that combines explicit motion evidence with multi-level geometric priors from a frozen 3D foundation model. While adaptable to any backbone that provides intermediate spatial features and geometric predictions, we instantiate our framework using VGGT-Ω [39] for its strong multi-view geometric representations. As illustrated in Figure 3, these representations in OmegaMotion are aligned and fused onto a common patch grid to form a unified geometry-aware feature sequence, which is directly decoded into frame-wise motion masks. During inference, each prediction prompts a frozen SAM2 [33] model once for boundary refinement, without requiring per-video test-time optimization.

### 3.1. Design Observations

**Observation #1: Hierarchical Feature Evolution in 3D Backbones.** Although VGGT-Ω is trained primarily for multi-view reconstruction, its internal representations exhibit distinct spatial organization across network depths. As shown in Figure 2, early-, intermediate-, and final-layer features emphasize complementary structural properties: shallow layers highlight localized saliency, mid-level layers align with object boundaries and contours, and deep layers capture broad, scene-level geometry. This reveals that relying solely on the final-layer output discards critical structural details, motivating a feature aggregation strategy across multiple backbone layers.

**Observation #2: Complementarity of Motion and Geometric Uncertainty.** Optical flow explicitly captures local displacement, but such displacement conflates object motion, camera motion, or depth-dependent parallax. As shown in Figure 4, camera motion induces flow magnitude in static background regions, whereas scene depth helps ex-

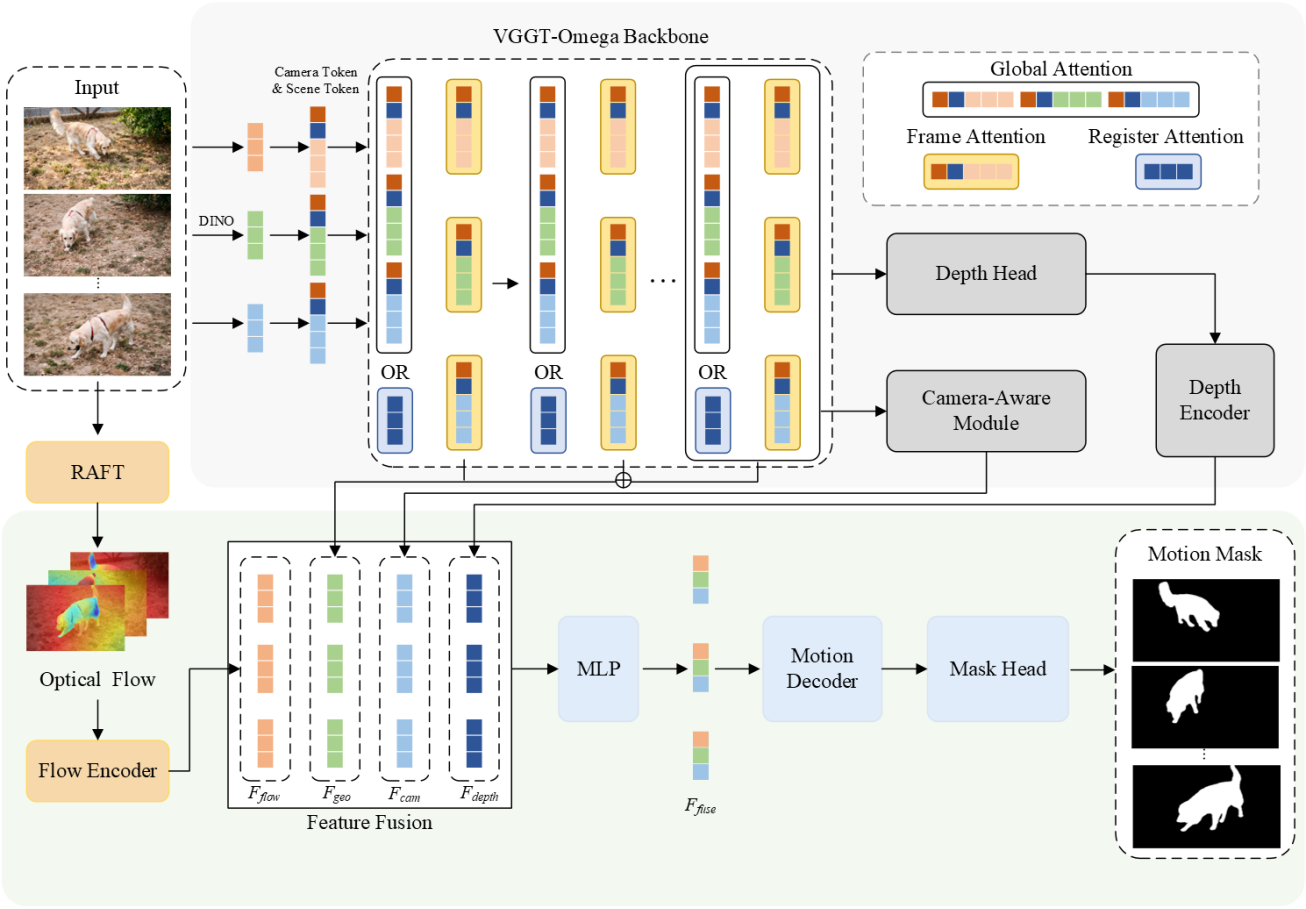



Figure 3. **Architecture of OmegaMotion**. A frozen VGGT-Ω backbone provides multi-layer latent representation, camera-aware context, and dense depth with confidence, while a frozen RAFT model estimates bidirectional optical-flow magnitude. The four complementary cues are aligned on a common patch grid and integrated into a unified geometry-aware representation, which is processed by a Transformer-based motion decoder and a mask head to predict frame-wise motion masks.

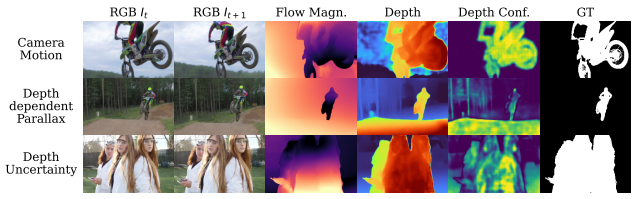



Figure 4. **Qualitative visualization of complementary motion and geometry cues.** Given two consecutive frames, optical flow magnitude conflates object and camera motion; Dense depth accounts for depth-dependent parallax, while depth confidence indicates the spatial reliability of geometric estimates. Integrating these representations facilitates object-background separation.

plain spatially varying parallax. Meanwhile, the predicted depth confidence varies across objects, boundaries, and occluded regions, indicating that depth estimates should not be treated as uniformly reliable. These observations motivate jointly modeling explicit flow, hierarchical scene structure, global camera context, and confidence-aware depth.

### 3.2. Multi-Prior Feed-Forward Architecture

Given a sequence of $N$ frames, DINOv2 [28] first extracts per-frame patch tokens, which are fed into the frozen Visual Geometry Backbone of VGGT-Ω [39] to derive multi-layer geometry features $\mathbf{F}_{\text{geo}}$, camera-aware features $\mathbf{F}_{\text{cam}}$, and confidence-aware depth features $\mathbf{F}_{\text{depth}}$. Concurrently, bidirectional optical flow fields estimated by RAFT [34] are processed by a flow encoder to produce $\mathbf{F}_{\text{flow}}$. These four patch-aligned feature streams are fused into $\mathbf{F}_{\text{fuse}}$ and decoded into frame-wise motion masks.

**Multi-layer geometry features.** Motivated by Observation #1, we aggregate sparse representations from early, intermediate, and final backbone stages to retain geometric information that evolves across network depth. We select six representative layers: Inter-0, Inter-3, Inter-10, Inter-15, Frame-23, and Inter-23, balancing cross-depth coverage with representation redundancy and computation.

Let $\mathbf{X}^{\ell}$ denote the aligned token sequence from layer $\ell$, we concatenate the selected representations along the fea-

ture dimension and apply a learnable projection $\phi_{geo}$:

$$(\mathbf{S}_{\text{geo}}, \mathbf{F}_{\text{geo}}) = \phi_{\text{geo}} \left( \underset{\ell \in \mathcal{L}}{\text{Concat}}(\mathbf{X}^{\ell}) \right) \tag{1}$$

where $\mathbf{S}_{\text{geo}}$ and $\mathbf{F}_{\text{geo}}$ denote the projected special and patch tokens, respectively. The latter serves as the multi-layer geometric feature for subsequent fusion.

**Camera-aware features.** To account for camera-induced displacement, we construct camera-conditioned features from the final two representations, each containing one camera token and 16 register tokens. Special and patch tokens are projected into $\mathbf{S}_c$ and $\mathbf{P}_c$, respectively, and fused through multi-head cross-attention [36]:

$$\mathbf{F}_{\text{cam}} = \text{LN}\left(\mathbf{P}_c + \text{MHA}\left(\mathbf{P}_c, \mathbf{S}_c, \mathbf{S}_c\right)\right). \tag{2}$$

This operation conditions each spatial token on global camera and scene context without explicitly computing camera-induced motion.

**Confidence-aware depth features.** VGGT-Ω [39] predicts dense depth maps and their confidence. The two maps are concatenated and embedded by a pointwise MLP to obtain $\mathbf{F}_{\text{depth}}$, providing confidence-aware geometric context by jointly modeling scene geometry and its spatially varying reliability.

**Bidirectional optical-flow features.** Geometric representations provide scene-level structure but do not explicitly measure short-term pixel displacement. We therefore use RAFT [34] to estimate bidirectional optical flow $\mathbf{F}_{\text{flow}}$.

**Multi-prior fusion and motion decoding.** We concatenate four patch-aligned streams that provide complementary geometry, camera context, depth, and motion cues:

$$\mathbf{F}_{\text{fuse}} = \phi_{\text{fuse}}\left([\mathbf{F}_{\text{geo}}; \mathbf{F}_{\text{cam}}; \mathbf{F}_{\text{flow}}; \mathbf{F}_{\text{depth}}]\right). \tag{3}$$

The fused representation is processed by a transformer-based decoder and a mask head to predict frame-wise motion masks.

### 3.3. Training Objective

For each frame $t$, the mask decoder predicts a soft motion mask $\hat{\mathbf{M}}_t \in [0,1]^{H \times W}$, where each value represents the probability that the corresponding pixel belongs to a moving object, and the prediction is supervised by a binary ground-truth mask $\mathbf{M}_t^{\text{gt}} \in \{0,1\}^{H \times W}$, in which foreground moving objects and the static background are labeled as 1 and 0, respectively. For a sequence of $N$ frames, the training objective is

$$\mathcal{L} = \frac{1}{N} \sum_{t=1}^{N} \left[ \lambda_1 \mathcal{L}_{\text{bce}}\left(\hat{\mathbf{M}}_t, \mathbf{M}_t^{\text{gt}}\right) + \lambda_2 \mathcal{L}_{\text{dice}}\left(\hat{\mathbf{M}}_t, \mathbf{M}_t^{\text{gt}}\right) \right]. \tag{4}$$

Binary cross-entropy loss provides pixel-wise supervision, while Dice loss [24] promotes spatial overlap and mitigates foreground-background imbalance.

## 4. Experiments

### 4.1. Experimental Settings

**Training datasets.** We train on Dynamic Replica [13], HOI4D [21], MOSE [5], and YTVOS18-m [43, 46], which jointly cover synthetic scenes, egocentric interactions, crowded multi-object motion, and diverse object categories. During training, the four datasets are sampled with probabilities 0.15, 0.35, 0.25, and 0.25, respectively.

**Benchmarks and metrics.** We evaluate our model using four popular moving object video segmentation benchmarks: DAVIS2016 [30], DAVIS2017 [31], FBMS-59 [27], and SegTrackV2 [18]. Following prior works [11, 43], we evaluate region similarity ($\mathcal{J}$), contour similarity ($\mathcal{F}$), and their average ($\mathcal{J}\&\mathcal{F}$). For comparison with reconstruction-based methods, we additionally report the sequence-level mean IoU ($\mathcal{J}_{\text{M}}$) and the percentage of frames with IoU above 0.5 ($\mathcal{J}_{\text{R}}$), following Easi3R [4].

**Implementation details.** We keep VGGT-Ω [39] and RAFT [34] frozen. The model takes 12-frame clips at a resolution of $512 \times 512$ and is trained for 40 epochs on a single NVIDIA RTX A6000 GPU. We use AdamW [22] with an initial learning rate of $5 \times 10^{-5}$ and set $\lambda_1 = \lambda_2 = 1$. To improve temporal diversity, frames are randomly resampled from each training video at every epoch.

### 4.2. Comparison with Prior Methods

#### 4.2.1. Quantitative Comparison

**Motion segmentation methods.** We compare our method with representative motion segmentation approaches under two inference paradigms. Iterative methods include OCLR-TTA [42], RoMo [7], and SegAnyMotion [11], which perform test-time adaptation, optimization, or iterative mask propagation. Non-iterative baselines include CIS [48], EM [23], RCF [20], OCLR-flow [42], ABR [43], and FlowP-SAM [44]. We report region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ on DAVIS2016 [30], DAVIS2017 [31], FBMS-59 [27], and SegTrackV2 [18].

As shown in Table 1, our method achieves the best $\mathcal{J}\&\mathcal{F}$ scores on three of the four benchmarks. On DAVIS2016 [30], it achieves a $\mathcal{J}\&\mathcal{F}$ score of 86.9, outperforming the second-best SegAnyMotion by 1.8 points. In particular, it achieves a $\mathcal{J}$ score of 87.3, exceeding the strongest competing result by 3.1 points, while its $\mathcal{F}$ score remains competitive with SegAnyMotion [11].

On DAVIS2017 [31], OmegaMotion achieves a $\mathcal{J}\&\mathcal{F}$ score of 81.2, ranking second and trailing SegAnyMotion [11] by only 1.7 points. Notably, this performance is obtained through feed-forward prediction without iterative test-time optimization. Its narrow gap in $\mathcal{J}$ score (81.8 versus 82.0) demonstrates its competitiveness in region localization on this more challenging multi-object benchmark.

Table 1. **Quantitative comparison with motion segmentation methods on four benchmarks.** The best result in each column is shown in **bold**, and the second best is underlined.

| Method | Iter. Opt. | Trainable Params. | DAVIS2016 $\mathcal{J}\&\mathcal{F}\uparrow$ | DAVIS2016 $\mathcal{J}\uparrow$ | DAVIS2016 $\mathcal{F}\uparrow$ | DAVIS2017 $\mathcal{J}\&\mathcal{F}\uparrow$ | DAVIS2017 $\mathcal{J}\uparrow$ | DAVIS2017 $\mathcal{F}\uparrow$ | FBMS-59 $\mathcal{J}\&\mathcal{F}\uparrow$ | FBMS-59 $\mathcal{J}\uparrow$ | FBMS-59 $\mathcal{F}\uparrow$ | SegTrackV2 $\mathcal{J}\&\mathcal{F}\uparrow$ | SegTrackV2 $\mathcal{J}\uparrow$ | SegTrackV2 $\mathcal{F}\uparrow$ | Avg. $\mathcal{J}\&\mathcal{F}\uparrow$ | Run Time per frame |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OCLR-TTA [42] | Yes | 42.33M | 78.8 | 80.7 | 76.9 | 74.5 | 76.0 | 73.0 | 69.1 | 69.9 | 68.3 | 73.4 | 72.3 | 74.5 | 74.0 | 1.3s |
| RoMo [7] | Yes | 74.37M | 75.7 | 77.3 | 74.1 | 71.1 | 72.1 | 70.2 | 74.4 | 75.5 | 73.3 | 68.8 | 67.7 | 69.8 | 72.5 | 8.45s |
| SegAnyMotion [11] | Yes | **3.44M** | 85.1 | 81.7 | **88.5** | **82.9** | **82.0** | **84.8** | 74.5 | 71.9 | 77.2 | 64.0 | 61.6 | 66.4 | 76.6 | 6.54s |
| CIS [48] | No | 18.92M | 68.6 | 70.4 | 66.8 | 65.8 | 67.6 | 64.0 | 61.9 | 63.6 | 60.2 | 63.5 | 62.0 | 65.1 | 65.0 | 0.26s |
| EM [23] | No | 497.67M | 70.0 | 69.4 | 70.6 | 66.2 | 65.2 | 67.2 | 57.0 | 57.9 | 56.0 | 54.8 | 55.5 | 54.1 | 62.0 | 1.32s |
| RCF-Stage1 [20] | No | 39.48M | 78.5 | 80.2 | 76.8 | 75.2 | 76.7 | 73.8 | 70.5 | 69.9 | 71.0 | 76.1 | 76.7 | 75.5 | 75.1 | 1.15s |
| OCLR-flow [42] | No | 20.66M | 71.2 | 72.0 | 70.4 | 68.9 | 69.9 | 67.8 | 65.2 | 65.5 | 64.9 | 66.8 | 67.6 | 66.0 | 68.0 | **0.22s** |
| RCF-all [20] | No | 39.48M | 80.7 | 82.1 | 79.2 | 77.2 | 78.6 | 76.7 | 73.8 | 72.4 | 75.2 | 78.2 | **78.6** | 77.8 | 77.5 | 1.35s |
| ABR [43] | No | 19.55M | 72.5 | 71.8 | 73.2 | 74.8 | 74.6 | 75.1 | 80.7 | **81.9** | 79.6 | 77.2 | 76.6 | 77.8 | 76.3 | 0.28s |
| FlowP-SAM [44] | No | 5.77M | 83.1 | 84.2 | 82.0 | 73.7 | 71.2 | 76.3 | 72.0 | 71.7 | 72.3 | 78.4 | 77.7 | 79.1 | 76.8 | 0.45s |
| **Ours** | No | 273.72M | **86.9** | **87.3** | 86.5 | 81.2 | 81.8 | 80.7 | **82.9** | 81.5 | **84.3** | **79.4** | 75.5 | **83.3** | **82.6** | **0.22s** |




Figure 5. **Qualitative comparisons on multiple benchmarks.** Competing methods include FlowP-SAM [44], ABR[43], SegAnyMotion[11], and ours (OmegaMotion).

Our method performs particularly well on FBMS-59 [27], achieving the best $\mathcal{J}\&\mathcal{F}$ score of 82.9, which surpasses ABR by 2.2 points. Although ABR obtains a slightly higher $\mathcal{J}$ (81.9 versus 81.5), OmegaMotion improves $\mathcal{F}$ score by 4.7 points, resulting in a stronger overall balance between region and contour accuracy. On SegTrackV2 [18], our method also ranks first with a $\mathcal{J}\&\mathcal{F}$ score of 79.4 and achieves the highest $\mathcal{F}$ score of 83.3, exceeding the second-best result by 4.2 points. Its comparatively lower $\mathcal{J}$ score of 75.5, however, suggests that complete region coverage remains more challenging than boundary localization in some low-resolution sequences.

No competing baseline maintains comparable performance across all datasets: SegAnyMotion [11] is strongest on DAVIS2017 [31] but degrades substantially on SegTrackV2 [18], whereas ABR [43] performs well on FBMS-59 [27] but is less competitive on DAVIS. In contrast, OmegaMotion consistently ranks first or second, demonstrating stronger cross-dataset generalization across diverse motion patterns, camera motion, and video resolutions.

**Reconstruction-based methods.** We further compare our method with representative 3D/4D reconstruction approaches that derive dynamic masks from geometric estimation, including DUSt3R [40], MonST3R [50], DAS3R [45], Easi3R [4], and VGGT4D [10]. Following the evaluation protocol of Easi3R [4], we report mean IoU $\mathcal{J}_{\mathrm{M}}$ and IoU recall $\mathcal{J}_{\mathrm{R}}$ before and after SAM2 refinement.

As shown in Table 2, it substantially outperforms all reconstruction-based methods before SAM2 refinement. On DAVIS2016 [30], DAVIS2017 [31], and DAVIS-all, it improves the best competing $\mathcal{J}_{\mathrm{M}}$ scores by 15.9, 19.3, and 19.3 points, respectively, with corresponding gains of 19.9, 24.4, and 24.5 points in $\mathcal{J}_{\mathrm{R}}$. These consistent margins

Table 2. **Comparisons among reconstruction-based methods on DAVIS benchmarks.** The best result in each column is shown in **bold**, and the second best is underlined.

| Method | Optical Flow | DAVIS2016 | | | | DAVIS2017 | | | | DAVIS-all | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | w/o SAM2 | | w/ SAM2 | | w/o SAM2 | | w/ SAM2 | | w/o SAM2 | | w/ SAM2 | |
| | | $\mathcal{J}_M\uparrow$ | $\mathcal{J}_R\uparrow$ | $\mathcal{J}_M\uparrow$ | $\mathcal{J}_R\uparrow$ | $\mathcal{J}_M\uparrow$ | $\mathcal{J}_R\uparrow$ | $\mathcal{J}_M\uparrow$ | $\mathcal{J}_R\uparrow$ | $\mathcal{J}_M\uparrow$ | $\mathcal{J}_R\uparrow$ | $\mathcal{J}_M\uparrow$ | $\mathcal{J}_R\uparrow$ |
| DUSt3R [40] | ✓ | 42.1 | 45.7 | 58.5 | 63.4 | 35.2 | 35.3 | 48.7 | 50.2 | 35.9 | 34.0 | 47.6 | 48.7 |
| MonST3R [50] | ✓ | 40.9 | 42.2 | 64.3 | 73.3 | 38.6 | 38.2 | 56.4 | 59.6 | 36.7 | 34.3 | 51.9 | 54.1 |
| DAS3R [45] | ✗ | 41.6 | 39.0 | 54.2 | 55.8 | 43.5 | 42.1 | 57.4 | 61.3 | 40.4 | 39.3 | 53.9 | 54.8 |
| Easi3R$_{\text{DUSt3R}}$ [4] | ✗ | 53.1 | 60.4 | 67.9 | 71.4 | 49.0 | 56.4 | 60.1 | 65.3 | 44.5 | 49.6 | 54.7 | 60.6 |
| Easi3R$_{\text{MonST3R}}$ [4] | ✗ | 57.7 | 71.6 | 70.7 | 79.9 | 56.5 | 68.6 | 67.9 | 76.1 | 53.0 | 63.4 | 63.1 | 72.6 |
| VGGT4D [10] | ✗ | 60.3 | 74.5 | – | – | 53.8 | 61.2 | – | – | 49.2 | 53.2 | – | – |
| **Ours** | ✓ | **76.2** | **94.4** | **87.3** | **97.0** | **75.8** | **93.0** | **81.8** | **90.3** | **72.3** | **87.9** | **77.8** | **86.8** |

show that directly learning motion-oriented representations is more effective for dynamic-region segmentation than deriving masks indirectly from reconstruction consistency.

SAM2 further improves the region-level segmentation quality of OmegaMotion predictions. After refinement, OmegaMotion achieves $\mathcal{J}_M$ scores of 87.3, 81.8, and 77.8 on DAVIS2016 [30], DAVIS2017 [31], and DAVIS-all, respectively, outperforming the best reconstruction-based baseline on each benchmark by 16.6, 13.9, and 14.7 points. Importantly, the substantial margins already observed before refinement confirm that the gains do not originate solely from SAM2; instead, SAM2 further enhances masks that are already considerably more accurate than those produced by reconstruction-oriented approaches.

#### 4.2.2. Qualitative comparison

Figure 5 presents qualitative comparisons on representative DAVIS sequences. Compared with FlowP-SAM [44], SegAnyMotion [11], and ABR [43], our method generally produces more complete object masks while suppressing background responses caused by camera motion and parallax. The improvements are particularly visible around articulated structures and object boundaries, consistent with the quantitative gains in boundary accuracy measured by $\mathcal{F}$. These examples illustrate how explicit displacement and multi-level geometric context complement each other when local motion cues alone are ambiguous.

#### 4.2.3. Accuracy-efficiency trade-off

Figure 1 further compares segmentation accuracy and inference speed. OmegaMotion processes videos at 0.22 seconds per frame while achieving a $\mathcal{J}\&\mathcal{F}$ score of 86.9 on DAVIS2016 [30]. It is over 20× faster than both SegAnyMotion [11] and RoMo [7], while obtaining higher accuracy on this benchmark. These results show that the feed-forward architecture avoids the substantial overhead of iterative optimization without sacrificing segmentation quality.

### 4.3. Ablation Study

#### 4.3.1. Contribution of Fusion component

We conduct component ablations on DAVIS2017 [31] to evaluate the contributions of multi-layer geometry, camera-aware context, optical flow, and confidence-aware depth. As shown in Table 3, replacing the final-stage baseline representation with the selected multi-layer geometric features improves the $\mathcal{J}\&\mathcal{F}$ score from 71.2 to 76.2, yielding a substantial gain of 5.0 points. This result validates the benefit of aggregating geometric representations across network depths rather than relying solely on the final stage.

Table 3. **Component ablation on DAVIS2017.** The baseline uses the final two features, while Geo denotes the selected multi-layer VGGT-Ω features. Cam, Flow, and Depth denote camera-aware, optical-flow, and confidence-aware depth cues, respectively. All settings share the same training and inference protocol.

| Configuration | Geo | Cam | Flow | Depth | $\mathcal{J}\&\mathcal{F}\uparrow$ |
|---|---|---|---|---|---|
| Baseline | – | – | – | – | 71.2 |
| Geo | ✓ | – | – | – | 76.2 |
| Geo + Cam | ✓ | ✓ | – | – | 78.4 |
| Geo + Flow | ✓ | – | ✓ | – | 80.0 |
| Geo + Depth | ✓ | – | – | ✓ | 76.8 |
| Geo + Cam + Depth | ✓ | ✓ | – | ✓ | 79.5 |
| Geo + Flow + Depth | ✓ | – | ✓ | ✓ | 80.2 |
| Geo + Cam + Flow | ✓ | ✓ | ✓ | – | 80.4 |
| Ours | ✓ | ✓ | ✓ | ✓ | **81.2** |

Table 4. **Effect of training-data diversity.** Datasets are added cumulatively from top to bottom. #D denotes the number of training datasets.

| Training datasets | #D | DAVIS2016 | | DAVIS2017 | |
|---|---|---|---|---|---|
| | | $\mathcal{J}\uparrow$ | $\mathcal{F}\uparrow$ | $\mathcal{J}\uparrow$ | $\mathcal{F}\uparrow$ |
| Dynamic Replica [13] | 1 | 72.3 | 69.8 | 68.6 | 66.4 |
| + HOI4D [21] | 2 | 86.5 | 85.4 | 78.0 | 77.5 |
| + MOSE [5] | 3 | 86.6 | 85.8 | 78.1 | 77.6 |
| + YTVOS18-m [46] | 4 | **87.3** | **86.5** | **81.8** | **80.7** |

Starting from the multi-layer Geo configuration, adding Cam, Flow, and Depth individually improves performance by 2.2, 3.8, and 0.6 points, respectively. Flow provides the largest individual gain by supplying explicit local displacement, while the improvement from Cam is consistent with the importance of global camera-conditioned context for interpreting image motion. Although Depth yields a relatively modest standalone gain, it remains complementary to the other cues: removing Cam, Flow, or Depth from the complete model decreases performance by 1.0, 1.7, and 0.8 points, respectively. Integrating all four components achieves the best result of 81.2, demonstrating that hierarchical geometry, camera context, depth, and optical flow provide complementary information for motion segmentation.

#### 4.3.2. Contribution of Data Scale

We analyze how the composition and diversity of the training data affect the model performance by cumulatively adding HOI4D [21], MOSE [5], and YTVOS18-m [46] to the Dynamic Replica baseline. As shown in Table 4, incor-

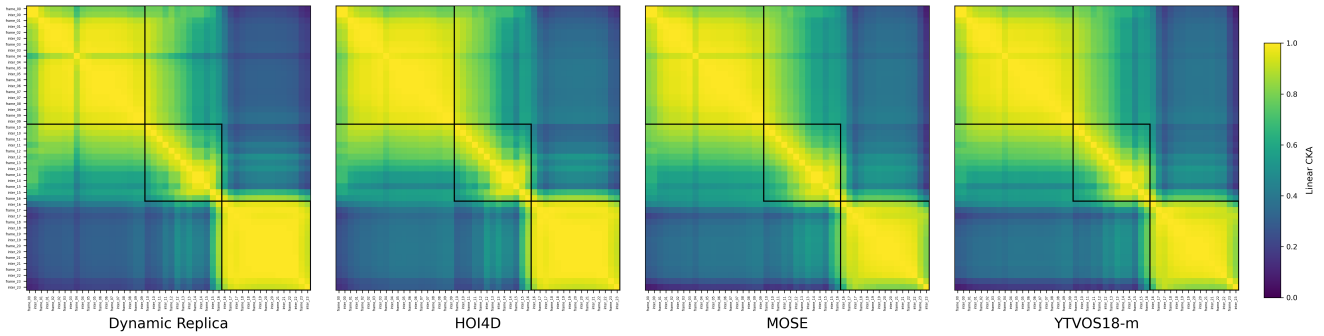



Figure 6. **Layer-wise analysis of frozen VGGT-Ω representations.** Pairwise linear CKA across DynamicReplica, HOI4D, MOSE, and YTVIS shows consistent block-wise similarity patterns, revealing distinct early-, middle-, and late-stage representation regimes.

porating HOI4D [21] produces the greatest improvement, increasing the $\mathcal{J}$ and $\mathcal{F}$ scores by 14.2 and 15.6 points, respectively, on DAVIS2016 [30], and by 9.4/11.1 points, respectively, on DAVIS2017 [31]. Since Dynamic Replica contains synthetic scenes whereas HOI4D [21] provides real-world interactions, this substantial gain suggests that diversity in real-world appearance and motion is important for bridging the synthetic-to-real domain gap.

Adding MOSE [5] produces modest but consistent improvements across both benchmarks. The subsequent inclusion of YTVOS18-m [46] provides a more pronounced gain on DAVIS2017 [31], improving $\mathcal{J}$ and $\mathcal{F}$ by 3.7 and 3.1 points, respectively, while further improving both metrics on DAVIS2016 [30] by 0.7 points. Training on all four datasets yields the best performance, with $\mathcal{J}$ and $\mathcal{F}$ scores of 87.3 and 86.5, respectively, on DAVIS2016 [30], and 81.8 and 80.7, respectively, on DAVIS2017 [31]. These results show that OmegaMotion benefits from combining synthetic scenes with diverse real-world motion and object appearances, leading to stronger generalization across benchmarks.

#### 4.3.3. Multi-layer Feature Analysis

We analyze the layer-wise representations of the frozen VGGT-Ω backbone using pairwise linear centered kernel alignment (CKA) [16]. To examine whether the observed representation structure is dataset-specific, we independently compute CKA on DynamicReplica [13], HOI4D [21], MOSE [5], and YTVOS18-m [46] datasets.

As shown in Figure 6, all four datasets exhibit a consistent structure of stage similarity. Representations within the early and late stages are highly correlated, while the intermediate layers form a transition regime between them. In contrast, similarity decreases substantially across different stages, indicating that distant network depths encode less redundant representations. The consistency of this pattern across datasets suggests that layer-wise organization is an intrinsic property of the frozen backbone rather than a dataset-specific phenomenon.

These observations motivate sparse feature aggregation across network depth. Densely sampling neighboring layers would introduce substantial representational redundancy, whereas selecting features from different CKA regimes provides more complementary information under a limited computational budget. Accordingly, we select Inter-0, Inter-3, Inter-10, Inter-15, Frame-23, and Inter-23 to span early, intermediate, and late representations while retaining the final frame-wise and inter-frame outputs. This selection is guided by the observed layer-wise representation structure rather than exhaustively optimized.

## 5. Conclusion

We presented **OmegaMotion**, a feed-forward framework that repurposes a frozen 3D foundation model for category-agnostic video motion segmentation. By unifying multi-layer VGGT-Ω representations, camera and register tokens, confidence-aware depth, and bidirectional optical flow, OmegaMotion effectively disentangles independent object dynamics from camera ego-motion and depth-dependent parallax without iterative test-time optimization. Extensive evaluations across four standard benchmarks demonstrate OmegaMotion achieves state-of-the-art segmentation accuracy while operating 20× faster than iterative methods. Furthermore, OmegaMotion substantially outperforms reconstruction-based approaches before and after SAM2 refinement, showing that its gains originate primarily from the learned motion-oriented representation. Component ablations confirm the complementary contributions of the four feature streams, while layer-wise CKA analyses support sparse feature aggregation across network depths.

**Limitations.** Despite strong feed-forward performance, OmegaMotion currently relies on off-the-shelf optical flow estimation and an external SAM2 module for boundary refinement. Future work will focus on efficient motion cues, adaptive layer selection, and end-to-end boundary refinement to further improve robustness and reduce reliance on external modules.

## Acknowledgments

This work was supported in part by the National Science and Technology Council (NSTC) under grants 114-2221-E-A49-038-MY3, 115-2221-E-A49-119-MY3, and 115-2634-F-A49-011, and by the NVIDIA Taiwan AI Research & Development Center (TRDC).

## References

[1] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. *arXiv preprint arXiv:1604.00136*, 2016. 2

[2] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, pages 282–295, 2010. 1, 2

[3] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1050–1060, 2025. 2

[4] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2, 3, 5, 6, 7

[5] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H. S. Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20224–20234, 2023. 5, 7, 8

[6] Katerina Fragkiadaki, Pablo Arbeláez, Panna Felsen, and Jitendra Malik. Learning to segment moving objects in videos. In *CVPR*, pages 4083–4090, 2015. 2

[7] Lily Goli, Sara Sabour, Mark Matthews, Marcus A. Brubaker, Dmitry Lagun, Alec Jacobson, David J. Fleet, Saurabh Saxena, and Andrea Tagliasacchi. RoMo: Robust motion segmentation improves structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6155–6164, 2025. 1, 2, 3, 5, 6, 7

[8] Rhodri Guerrier, Adam W. Harley, and Dima Damen. PointSt3R: Point tracking through 3d ground correspondence. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7430–7439, 2026. 2

[9] Xiankang He, Peile Lin, Ying Cui, Dongyan Guo, Chunhua Shen, and Xiaoqin Zhang. GeoMotion: Rethinking motion segmentation via latent 4D geometry. In *CVPR*, 2026. 2, 3

[10] Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. VGGT4D: Mining motion cues in visual geometry transformers for 4D scene reconstruction. *arXiv preprint arXiv:2511.19971*, 2025. 2, 3, 6, 7

[11] Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3406–3416, 2025. 1, 2, 5, 6, 7

[12] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, pages 2117–2126, 2017. 2

[13] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5, 7, 8

[14] Laurynas Karazija, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Learning segmentation from point trajectories. In *NeurIPS*, 2024. 1, 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1, 2

[16] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019. 8

[17] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, pages 71–91, 2024. 2

[18] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 5, 6

[19] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *CVPR*, 2025. 2, 3

[20] Long Lian, Zhirong Wu, and Stella X. Yu. Bootstrapping objectness from videos by relaxed common fate and visual grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14582–14591, 2023. 1, 2, 5, 6

[21] Yunze Liu, Yizhou Liu, Chen Jiang, Kun Lyu, Weikai Wan, Hao Shen, Bo Liang, Zhenyu Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human–object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5, 7, 8

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. 5

[23] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. EM-driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4462–4473, 2023. 1, 5, 6

[24] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571, 2016. 5

[25] Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *CVPR*, pages 16695–16705, 2025. 2

[26] Manjunath Narayana, Allen Hanson, and Erik Learned-Miller. Coherent motion segmentation in moving camera videos using optical flow orientations. In *ICCV*, pages 1577–1584, 2013. 1, 2

[27] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6): 1187–1200, 2014. 1, 2, 5, 6

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[29] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013. 2

[30] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5, 6, 7, 8

[31] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5, 6, 7, 8

[32] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, pages 12240–12249, 2019. 2, 3

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025. 1, 2, 3

[34] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 4, 5

[35] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, pages 531–539, 2017. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 5

[37] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfM-Net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 3

[38] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[39] Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-$\omega$. In *CVPR*, 2026. 2, 3, 4, 5

[40] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6, 7

[41] Andreas Wedel, Annemarie Meißner, Clemens Rabe, Uwe Franke, and Daniel Cremers. Detection and segmentation of independently moving objects from dense scene flow. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 14–27. Springer, 2009. 2

[42] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 5, 6

[43] Junyu Xie, Weidi Xie, and Andrew Zisserman. Appearance-based refinement for object-centric motion segmentation. In *European Conference on Computer Vision*, 2024. 1, 2, 5, 6, 7

[44] Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is SAM (and flow). In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2024. 1, 2, 5, 6, 7

[45] Kai Xu, Tze Ho Elden Tse, Jizong Peng, and Angela Yao. DAS3R: Dynamics-aware gaussian splatting for static scene reconstruction. *arXiv preprint arXiv:2412.19584*, 2024. 2, 3, 6, 7

[46] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5, 7, 8

[47] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, pages 7177–7188, 2021. 1, 2

[48] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019. 1, 5, 6

[49] Yanchao Yang, Brian Lai, and Stefano Soatto. DyStaB: Unsupervised object segmentation via dynamic-static bootstrapping. In *CVPR*, pages 2826–2836, 2021. 2

[50] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, 2025. 2, 3, 6, 7