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

Supplementary Material

1. Architecture and Implementation Details

1.1. Detailed Architecture

OmegaMotion takes a clip of T RGB frames resized to 512×512 as input. With a patch stride of 16, each frame is represented by a 32×32 patch grid containing $N = 1024$ patch tokens. In addition to the patch tokens, each VGGT-Ω representation contains one camera token and 16 register tokens, resulting in 17 special tokens per frame. The frozen VGGT-Ω backbone provides multi-level geometric features, camera-aware context, dense depth, and depth confidence, while a frozen RAFT model provides bidirectional optical-flow magnitude. All feature streams are aligned to the same 32×32 patch grid before fusion.

Multi-layer geometric features. We extract six representations distributed across the VGGT-Ω backbone: Inter-0, Inter-3, Inter-10, Inter-15, Frame-23, and Inter-23. Each representation has a channel dimension of 1024 and contains both the 17 special tokens and the N patch tokens. The six representations are concatenated along the channel dimension, producing a 6144-dimensional feature for each token. A linear projection followed by ReLU maps the concatenated representation to 2048 dimensions, producing the projected special tokens $\mathbf{S}_{\text{geo}} \in \mathbb{R}^{17\times 2048}$ and patch features $\mathbf{F}_{\text{geo}} \in \mathbb{R}^{N\times 2048}$.

Camera-aware features. The camera branch uses the final frame-wise and inter-frame representations, namely Frame-23 and Inter-23. Their corresponding tokens are concatenated to obtain a 2048-dimensional representation. The 17 special tokens and N patch tokens are separately projected to 512 dimensions. We then apply eight-head special-to-patch attention, in which the patch tokens serve as queries and the camera and register tokens serve as keys and values. A residual connection and layer normalization produce the camera-aware patch features $\mathbf{F}_{\text{cam}} \in \mathbb{R}^{N\times 512}$.

Confidence-aware depth features. The dense depth and confidence maps predicted by VGGT-Ω are first transformed using $\log(1+x)$ and bilinearly resized to the 32×32 patch grid. The resulting depth and confidence values are concatenated at each patch location and embedded using a two-layer pointwise MLP with GELU activation:

$$2 \to 32 \to 32.$$

This produces $\mathbf{F}_{\text{depth}} \in \mathbb{R}^{N\times 32}$.

Optical-flow features. RAFT estimates optical flow in both temporal directions, from which a scalar bidirectional flow-magnitude map is constructed for every frame. The magnitude map is processed by three 3×3 convolutional layers with channel dimensions

$$1 \to 32 \to 64 \to 128,$$

each followed by ReLU. The encoded feature map is bilinearly resized to the patch grid and flattened into $\mathbf{F}_{\text{flow}} \in \mathbb{R}^{N\times 128}$.

Feature fusion and motion decoding. The four patch-aligned feature streams are concatenated to form a 2720-dimensional representation:

$$2048 + 512 + 128 + 32 = 2720.$$

A linear fusion layer projects this representation to 2048 dimensions. The 17 projected geometric special tokens are then prepended to the fused patch tokens. The resulting sequence is processed by five Transformer blocks, each using 16 attention heads and an MLP expansion ratio of 4. Finally, a linear mask head maps each decoded patch token to a scalar motion logit. The patch-level logits are reshaped to 32×32 and bilinearly upsampled to the input resolution.

Table 1 summarizes the detailed feature configuration.

Algorithm 1 summarizes the feature extraction, alignment, and fusion procedure. Batch and temporal dimensions are omitted for clarity.

1.2. Training and Inference Details

Training. We train OmegaMotion using 12-frame clips resized to 512×512. To increase temporal diversity, the

Table 1. Detailed architecture of OmegaMotion. Output dimensions denote the channel dimension of each patch token.

Feature	Source	Encoding	Dim.
$\mathbf{F}_{\text{geo}}$	Six VGGT-Ω layers	Concatenation and projection	2048
$\mathbf{F}_{\text{cam}}$	Frame-23 and Inter-23	17-token cross-attention	512
$\mathbf{F}_{\text{flow}}$	Bidirectional RAFT magnitude	Three-layer CNN	128
$\mathbf{F}_{\text{depth}}$	Depth and confidence	Log transform and pointwise MLP	32
$\mathbf{F}_{\text{fuse}}$	Four aligned feature streams	$2720 \rightarrow 2048$ projection	2048
Motion decoder	Special and fused patch tokens	Five Transformer blocks	2048

Algorithm 1 Multi-prior feature fusion in OmegaMotion

Require: RGB clip $\mathbf{I} \in \mathbb{R}^{T\times 3\times 512\times 512}$
Ensure: Motion logits $\mathbf{M} \in \mathbb{R}^{T\times 512\times 512}$
Extract VGGT-Ω representations, depth $\mathbf{D}$, and confidence $\mathbf{C}$
Select $\mathcal{L} = \{$Inter-0, Inter-3, Inter-10, Inter-15, Frame-23, Inter-23$\}$
Extract $\{\mathbf{X}_\ell\}_{\ell\in\mathcal{L}}$
$(\mathbf{S}_{\text{geo}}, \mathbf{F}_{\text{geo}}) \leftarrow \phi_{\text{geo}}(\text{Concat}_{\ell\in\mathcal{L}}\, \mathbf{X}_\ell)$
Concatenate the Frame-23 and Inter-23 representations
Separate and project the special and patch tokens as $\mathbf{S}_c$ and $\mathbf{P}_c$
$\mathbf{F}_{\text{cam}} \leftarrow \text{LN}(\mathbf{P}_c + \text{MHA}(\mathbf{P}_c, \mathbf{S}_c, \mathbf{S}_c))$
$\widetilde{\mathbf{D}} \leftarrow \text{Resize}(\log(1+\mathbf{D}), 32, 32)$
$\widetilde{\mathbf{C}} \leftarrow \text{Resize}(\log(1+\mathbf{C}), 32, 32)$
$\mathbf{F}_{\text{depth}} \leftarrow \phi_{\text{depth}}([\widetilde{\mathbf{D}}; \widetilde{\mathbf{C}}])$
Estimate bidirectional flow magnitude $\mathbf{O}$ using frozen RAFT
$\mathbf{F}_{\text{flow}} \leftarrow \text{Resize}(\phi_{\text{flow}}(\mathbf{O}), 32, 32)$
$\mathbf{F}_{\text{fuse}} \leftarrow \phi_{\text{fuse}}\big([\mathbf{F}_{\text{geo}}; \mathbf{F}_{\text{cam}}; \mathbf{F}_{\text{flow}}; \mathbf{F}_{\text{depth}}]\big)$
$\mathbf{H} \leftarrow [\mathbf{S}_{\text{geo}}; \mathbf{F}_{\text{fuse}}]$
for $k = 1, \ldots, 5$ **do**
    $\mathbf{H} \leftarrow \text{TransformerBlock}_k(\mathbf{H})$
end for
Apply the mask head to the decoded patch tokens
Reshape and bilinearly upsample the logits to 512×512
return $\mathbf{M}$

frames forming each clip are randomly resampled from the corresponding training video at every epoch. The training set combines Dynamic Replica, HOI4D, MOSE, and YTVOS18-m, which are sampled with probabilities 0.15, 0.35, 0.25, and 0.25, respectively.

The VGGT-Ω backbone and RAFT optical-flow estimator remain frozen throughout training. We optimize only the geometric projection, camera-attention branch, flow and depth encoders, feature-fusion layer, motion decoder, and mask head. Training is performed for 40 epochs on a single NVIDIA RTX A6000 GPU using AdamW with an initial learning rate of 5×10^{-5} and a cosine-annealing learning-rate schedule. We use bfloat16 mixed-precision training.

The model is supervised using equally weighted binary cross-entropy and Dice losses. Binary cross-entropy provides pixel-wise supervision, whereas the Dice term encourages foreground overlap and alleviates foreground–background imbalance.

Chunked inference. At inference time, each video is resized to 512×512 and processed in fixed-length chunks. Our implementation uses 16-frame inference chunks. Full chunks are processed without overlap. If the final chunk contains fewer than 16 frames, we construct a 16-frame window ending at the last video frame and retain only the predictions corresponding to the previously unprocessed frames. This strategy provides full temporal context for the end of a sequence without duplicating predictions. All network parameters remain fixed, and no per-video optimization or test-time adaptation is performed.

SAM2 refinement. We use the frozen SAM2.1 Hiera-Large model for final spatial refinement. Raw motion probabilities are thresholded at 0.8 to construct binary seed masks. Refinement is performed using two complementary propagation passes. In the first pass, predictions from odd-indexed frames are used as prompts to refine even-indexed frames; in the second pass, even-indexed predictions are used to refine odd-indexed frames. Consequently, each raw prediction is used once as a seed, while every frame receives a propagated refinement. If a pass contains no valid seed mask, the corresponding raw predictions are retained. SAM2 is used only for this final refinement stage and does not update OmegaMotion or participate in motion-feature extraction.

2. Analysis of Multi-Level Geometric Features

We provide additional quantitative analyses of the frozen VGGT-Ω representations from two perspectives. First, we evaluate how linearly separable moving foreground and static background patches are at different network depths. Second, we measure the representational similarity among the six layers selected by OmegaMotion. The former examines whether motion-related information can be decoded from individual layers, whereas the latter characterizes the extent to which the selected representations occupy similar or distinct feature regimes.

2.1. Foreground–Background Separability

Evaluation protocol. We evaluate foreground–background separability using 40 video sequences drawn equally from the four training datasets: Dynamic Replica, HOI4D, MOSE, and YTVOS18-m, with 10 sequences from each dataset. For every sequence, we extract patch-token representations from all 24 frame-wise and 24 inter-frame blocks of the frozen VGGT-Ω backbone. We use at most 12 frames per sequence, resize each frame to 512×512, and obtain a 32×32 patch grid containing 1024 patch tokens per frame.

Ground-truth motion masks are converted into patch-level labels by area interpolation. A patch is labeled as foreground if at least 50% of its pixels belong to a moving object; otherwise, it is labeled as background. We perform

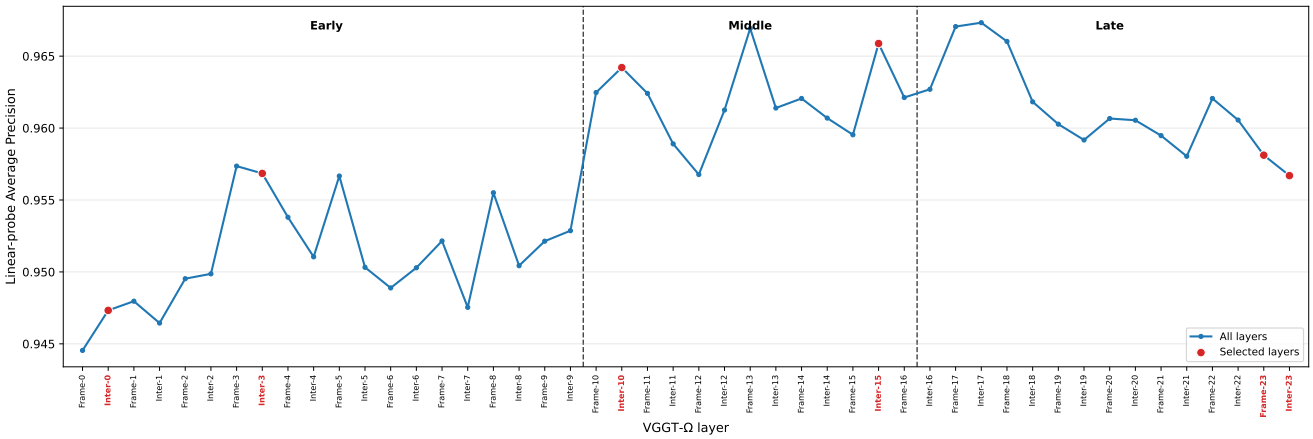


Figure 1. Layer-wise foreground–background separability of frozen VGGT-Ω patch representations. We report the Average Precision of a linear probe evaluated on held-out video sequences. The dashed lines indicate the early, middle, and late representation regimes identified by the full-layer CKA analysis. Red markers denote the six layers selected by OmegaMotion. Although motion-related information is decodable throughout the backbone, its separability varies across network depth and is generally stronger in the middle and late layers.

a dataset-stratified split at the video level, assigning seven sequences from each dataset to the training set and three to the test set. This produces 28 training sequences and 12 test sequences while ensuring that patches from the same video never occur in both subsets.

For every layer, we sample up to 50,000 foreground and 50,000 background patches from the training sequences and up to 20,000 patches per class from the test sequences. Features are standardized using statistics computed only from the training split. We then train a single linear classification layer for 30 epochs using class-weighted binary cross-entropy. The linear probe is optimized with AdamW using a learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size of 1024. We report Average Precision (AP) on the held-out sequences because it evaluates foreground ranking without requiring a layer-specific decision threshold.

Results. As shown in Fig. 1, all frozen VGGT-Ω representations exhibit strong foreground–background decodability, with linear-probe AP values ranging approximately from 0.94 to 0.97. Nevertheless, the separability is not uniform across depth. The early layers exhibit comparatively lower and more variable AP, whereas a noticeable increase occurs when entering the middle representation regime. The middle and late layers generally retain stronger separability, although their performance continues to fluctuate rather than increasing monotonically with depth.

These results indicate that motion-related foreground–background information is distributed throughout the backbone and changes as the representations evolve. Importantly, the six layers used by OmegaMotion are not selected simply as the six AP maxima. Instead, they sample different depth regimes while retaining strong linear decodability. The separability analysis therefore supports the availability of motion-relevant information at the selected depths, while the representational relationships among these layers are examined next through CKA.

2.2. CKA among the Selected Layers

Evaluation protocol. We use linear centered kernel alignment (CKA) to quantify similarity among the patch-token representations of the six selected layers: Inter-0, Inter-3, Inter-10, Inter-15, Frame-23, and Inter-23. The analysis uses the same 40 sequences, comprising 10 sequences from each of Dynamic Replica, HOI4D, MOSE, and YTVOS18-m. From every sequence, we uniformly sample 2,500 positions from the flattened frame–patch grid, resulting in 100,000 samples per layer. Exactly the same frame–patch positions are used for all six layers, ensuring that each pairwise comparison is sample-aligned.

CKA sampling is label-agnostic: foreground and background patches are included according to their occurrence in the sampled frame–patch positions and are not analyzed separately. This differs from the separability experiment, where ground-truth motion masks are explicitly used to construct foreground and background labels.

For centered feature matrices $\mathbf{X} \in \mathbb{R}^{n \times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$, linear CKA is computed as

$$\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\left\|\mathbf{X}^\top \mathbf{Y}\right\|_F^2}{\left\|\mathbf{X}^\top \mathbf{X}\right\|_F \left\|\mathbf{Y}^\top \mathbf{Y}\right\|_F}.$$

Feature centering is performed independently for every layer across the 100,000 sampled positions.

Results. Figure 2 reveals a clear organization across network depth. Inter-0, Inter-3, and Inter-10 form a relatively

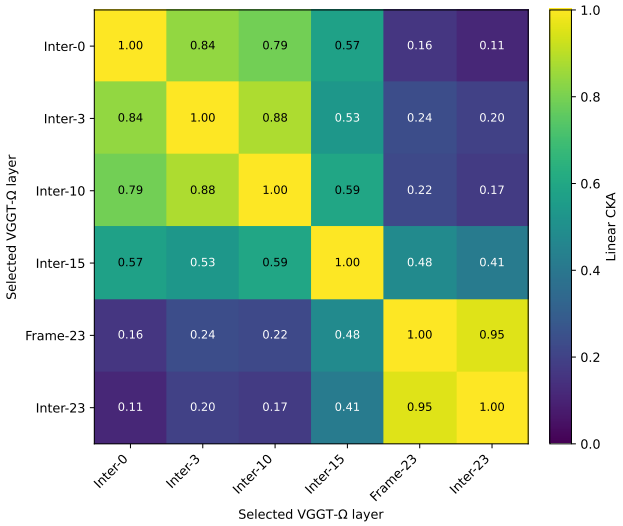


Figure 2. Linear CKA among the six VGGT-Ω layers selected by OmegaMotion. The selected representations exhibit high similarity within the early group and between the two final outputs, but considerably lower similarity across early, transitional, and late network depths. CKA characterizes representational similarity and does not by itself establish task-level complementarity.

similar early-to-middle group, with pairwise CKA values between 0.79 and 0.88. In contrast, their similarities to the final Frame-23 and Inter-23 representations are substantially lower, ranging from 0.11 to 0.24. Inter-15 occupies an intermediate position: its similarity to the earlier layers ranges from 0.53 to 0.59, while its similarity to the two final representations ranges from 0.41 to 0.48. This pattern suggests a gradual transition between the early and late representation regimes rather than a single abrupt change.

Frame-23 and Inter-23 exhibit a high CKA value of 0.95, indicating that their patch-token representations are closely related. This is expected because they are adjacent final frame-wise and inter-frame outputs. Nevertheless, they preserve the two endpoint representation types used by OmegaMotion. Moreover, the present CKA analysis is restricted to patch tokens and does not characterize the camera and register tokens separately exploited by the camera-aware branch.

Overall, the varying pairwise similarities indicate that the selected layers span distinct representation regimes rather than repeatedly sampling only neighboring features. This observation should not be interpreted as CKA alone proving that the layers are complementary for motion segmentation. Instead, CKA provides representational evidence that the selected features are not uniformly redundant, while the improvement of the multi-layer geometric configuration over the final-stage baseline in the component ablation provides task-level evidence for the effectiveness of their aggregation.

3. Qualitative Analysis of Component Ablation

We complement the quantitative component ablation in the main paper with qualitative comparisons of the progressive configurations in Fig. 3. The baseline uses the final-stage VGGT-Ω representations, while +*Geo* replaces them with the selected multi-layer geometric features. Camera-aware context, bidirectional optical flow, and confidence-aware depth are then incorporated cumulatively. Each configuration is trained independently under the same protocol, and all predictions are processed using the same frozen SAM2 model and refinement settings. Differences across columns therefore reflect changes in the motion predictions provided to a fixed refinement procedure.

The *breakdance* and *loading* examples most clearly illustrate the effect of multi-layer geometry. Predictions based only on the final-stage representations contain fragmented foreground regions and noisy responses around the moving objects. Replacing these features with representations aggregated across network depth produces more coherent silhouettes and recovers a larger portion of the moving extent. This observation is consistent with the quantitative improvement of the multi-layer Geo configuration over the final-stage baseline.

The subsequent components provide progressively finer corrections. Camera-aware features condition spatial tokens on global view-dependent context, helping the model interpret displacement in the presence of camera motion. Bidirectional optical flow supplies explicit short-term motion evidence and improves the continuity of moving regions, particularly around articulated or extended object parts. Confidence-aware depth further introduces geometric and reliability information, correcting localized omissions and suppressing residual spatial inconsistencies. In cases such as *car-shadow*, where the foreground is already largely localized, the changes are comparatively subtle but further improve its alignment with the ground truth.

For all four visualized frames, the frame-level $\mathcal{J}\&\mathcal{F}$ score increases monotonically as multi-layer geometry, camera-aware context, optical flow, and confidence-aware depth are progressively incorporated. Since all configurations use identical SAM2 refinement, this consistent improvement indicates that the additional components provide increasingly informative motion predictions before the fixed refinement stage. These qualitative results therefore offer frame-level evidence that complements the aggregate component ablation reported in the main paper.

4. Qualitative Comparison without SAM2

To isolate the quality of the motion representations from the effect of foundation-model-based mask refinement, we

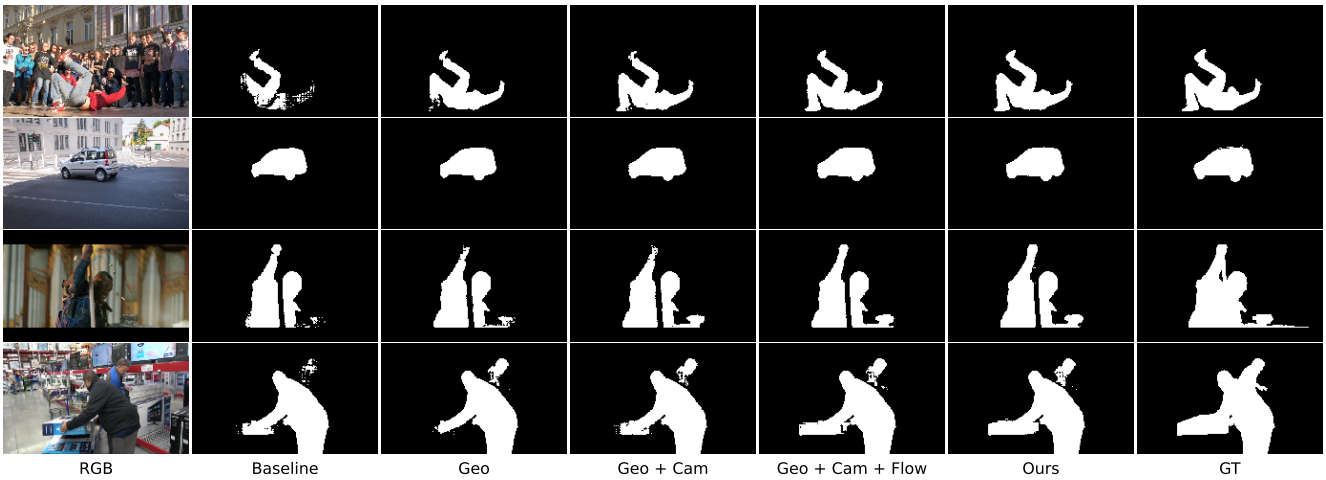


Figure 3. Qualitative comparison of the cumulative component ablation on representative DAVIS2017 frames. All configurations use the same SAM2 refinement protocol. The baseline relies on the final-stage VGGT-Ω representations; *+Geo* replaces them with the selected multi-layer geometric features, followed by the progressive addition of camera-aware context, bidirectional optical flow, and confidence-aware depth. The predictions become progressively more complete and spatially coherent, consistent with the monotonic improvement in frame-level $\mathcal{J}\&\mathcal{F}$ across the shown examples.

compare OmegaMotion with Easi3R and VGGT4D before applying SAM2. Unlike the final predictions reported in the main motion-segmentation comparison, none of the masks shown in Fig. 4 are refined by SAM2. This comparison therefore directly evaluates the dynamic regions identified by each method.

As shown in Fig. 4, Easi3R often identifies only a small subset of the moving region. For example, its predictions may cover only part of a person, vehicle, or animal, while other object parts remain undetected. Its masks also contain scattered responses in several scenes, indicating that dynamics inferred indirectly from reconstruction consistency do not always translate into spatially complete object masks.

VGGT4D recovers larger dynamic regions in some cases, but its predictions remain sensitive to background structure and may be fragmented or overextended. This is particularly visible in scenes containing large camera-induced displacement, complex backgrounds, or multiple independently moving objects. Although the inferred dynamic regions often overlap with the target objects, their spatial support does not consistently correspond to complete object silhouettes.

OmegaMotion produces substantially more coherent predictions across the shown examples. It recovers complete human silhouettes in the multi-person and egocentric scenes, captures the spatial extent of vehicles rather than isolated moving parts, and better preserves animals and small objects under challenging appearance and scale variations. In the scene containing a person and multiple dogs, OmegaMotion identifies the independently moving regions more completely, whereas the reconstruction-based methods recover only partial foreground objects. Similarly, in the distant-motorcycle example, OmegaMotion localizes the small moving subject while suppressing the camera-induced background response.

These differences arise from the distinct objectives of the methods. Easi3R and VGGT4D infer dynamic regions primarily from representations or consistency constraints associated with geometric reconstruction. OmegaMotion instead learns task-specific aggregation and decoding modules that combine multi-level geometric structure, camera-aware context, confidence-aware depth, and explicit optical flow for direct motion-mask prediction. Consequently, its unrefined masks already exhibit substantially stronger region completeness and foreground localization.

Some fine contours and thin structures remain imperfect before refinement, as expected from predictions decoded on a patch grid. Nevertheless, the overall object locations and extents are already considerably closer to the ground truth than those of the reconstruction-based alternatives. This qualitative comparison is consistent with the quantitative results reported in the main paper and confirms that SAM2 acts primarily as a final spatial refinement module rather than being the main source of OmegaMotion's performance advantage.

5. Qualitative Analysis of SAM2 Refinement

Figure 5 qualitatively isolates the effect of the final SAM2 refinement stage. We compare the raw binary masks predicted by OmegaMotion with the corresponding masks after refinement by the frozen SAM2.1 Hiera-Large model. Both



Figure 4. Qualitative comparison with reconstruction-based methods before SAM2 refinement. Easi3R and VGGT4D frequently produce sparse, fragmented, or spatially incomplete dynamic masks and may respond to static background regions. In contrast, OmegaMotion recovers more complete and coherent moving-object regions across diverse scenes, including multiple people, articulated objects, vehicles, animals, and small distant objects. These results show that the improvement of OmegaMotion originates primarily from its learned motion-oriented representation rather than from SAM2 refinement.

variants use the same trained OmegaMotion parameters; the only difference is whether SAM2 refinement is applied. No per-video optimization, test-time adaptation, or parameter update is performed.

Quality of the raw predictions. The predictions without SAM2 already capture the identity and approximate extent of the moving foreground across diverse rigid and non-rigid objects, including vehicles, animals, and articulated humans. This observation is important because SAM2 receives the OmegaMotion predictions as mask prompts rather than independently discovering which regions are moving. The raw masks must therefore provide sufficiently accurate motion localization before spatial refinement can

be effective. As shown in Figure 5, OmegaMotion identifies the correct moving instance even under substantial appearance variation, articulation, viewpoint change, and complex background structure.

Effect of spatial refinement. SAM2 primarily improves the spatial quality of these predictions. For rigid objects, such as the vehicle in the first row, refinement produces a cleaner silhouette and better alignment with the visible object boundary. For articulated animals and humans, SAM2 recovers finer structures that are partially merged or omitted in the raw masks, including legs, arms, and narrow extremities. It also removes small isolated artifacts and corrects locally irregular contours. These changes bring the refined masks closer to the ground truth while generally preserving the foreground identity and localization established by OmegaMotion.

The comparison also clarifies the respective roles of the two components. OmegaMotion performs the semantic decision of separating independently moving regions from the static scene, whereas SAM2 acts as a frozen, category-agnostic boundary refiner. Because SAM2 is prompted by the predicted masks, it cannot reliably compensate for a completely missing foreground instance or an incorrectly localized motion seed. The strong raw predictions in Figure 5 therefore demonstrate that the motion-separation capability originates from the learned multi-prior representation, while SAM2 provides a final improvement in mask completeness and boundary precision.

6. More Qualitative Results

Figure 6 presents additional qualitative results on three motion-segmentation benchmarks: DAVIS 2016, FBMS59, and SegTrackv2. We select two representative sequences from each benchmark and visualize predictions at multiple time steps. For each sequence, the first row shows the input RGB frames and the second row overlays the predicted moving regions in red.

DAVIS 2016. The first two sequences contain highly articulated human and animal motion. In the breakdance sequence, the foreground undergoes rapid pose changes, large limb displacement, and self-occlusion. OmegaMotion follows the changing body configuration and preserves the spatial extent of the dancer across the sequence. In the dog sequence, the object exhibits both non-rigid motion and noticeable scale variation as it moves relative to the camera. The predicted masks remain attached to the dog and adapt to its changing appearance and image-plane size.

FBMS59. The middle two sequences illustrate motion under cluttered indoor and outdoor backgrounds. The rabbit undergoes partial occlusion by foreground bedding and changes its visible shape while moving across the scene. OmegaMotion consistently separates the visible moving region from the surrounding static content. The approaching dog produces a particularly large scale change, ranging from a small distant object to one occupying a substantial portion of the frame. The predictions remain temporally stable throughout this transition and continue to identify the same independently moving instance.

SegTrackv2. The final two sequences demonstrate performance on small objects and challenging appearance conditions. In the parachute sequence, the moving target occupies only a small portion of the frame and is observed against a low-contrast, highly textured background. OmegaMotion maintains the target across successive viewpoints without responding strongly to the surrounding water pattern. In the bird sequence, the foreground contains thin articulated structures and changes its configuration substantially during motion. The predicted masks follow these variations while suppressing most of the visually similar background vegetation.

Overall, these examples demonstrate that OmegaMotion generalizes across different object categories and scene types. Its predictions remain coherent under articulation, scale changes, camera motion, motion blur, partial occlusion, and small foreground size. This cross-benchmark consistency supports the use of combined geometric, camera-aware, depth, and optical-flow priors for category-agnostic video motion segmentation.

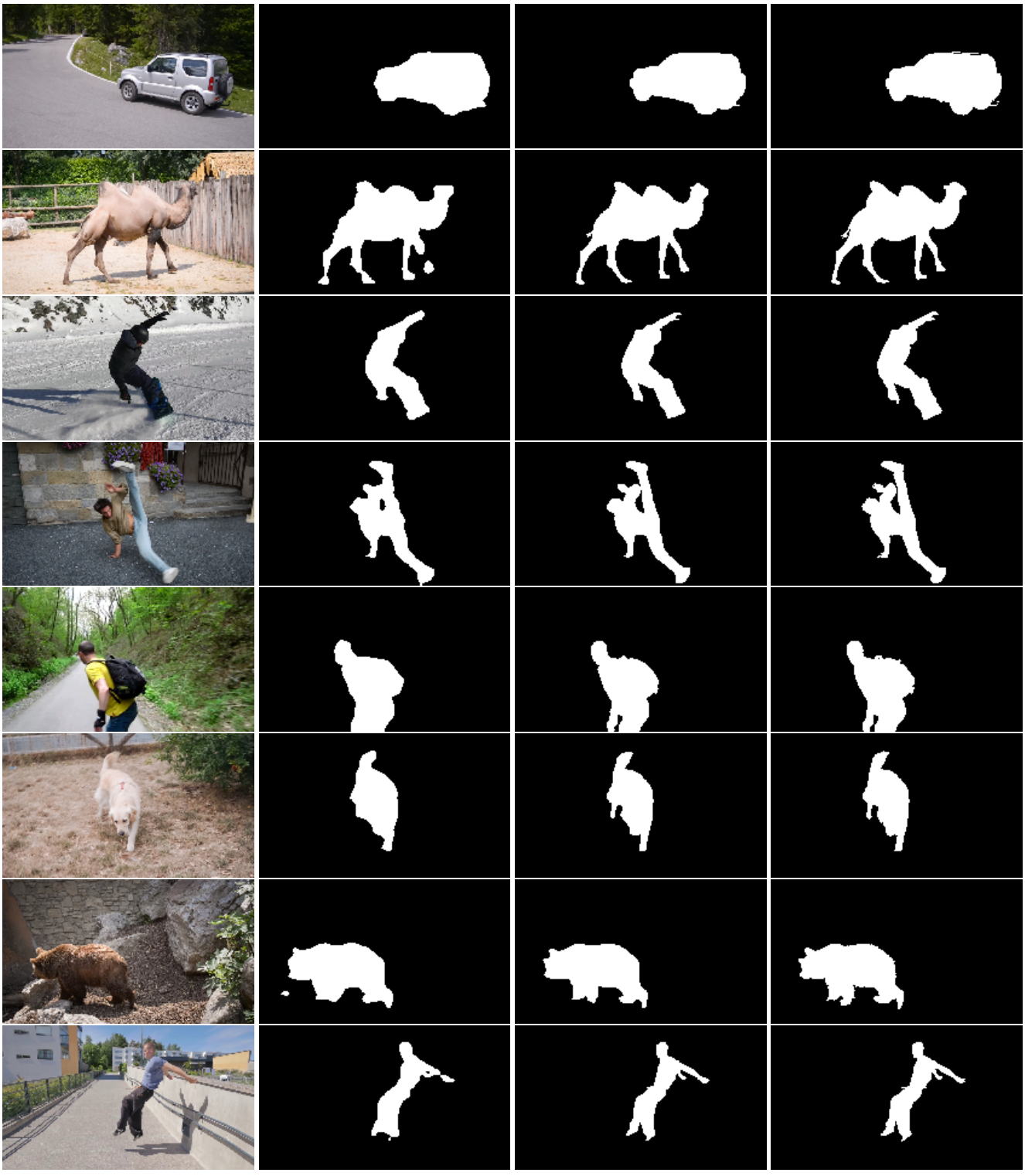


Figure 5. **Qualitative analysis of SAM2 refinement.** From left to right: input RGB frame, the raw OmegaMotion prediction without SAM2, the refined prediction with SAM2, and the ground-truth motion mask. White denotes the predicted or annotated moving foreground. The raw predictions already localize the moving objects and recover their principal spatial extent. SAM2 further improves object completeness, suppresses local mask artifacts, and aligns the predictions more closely with fine object boundaries, particularly around limbs, legs, and other thin structures.

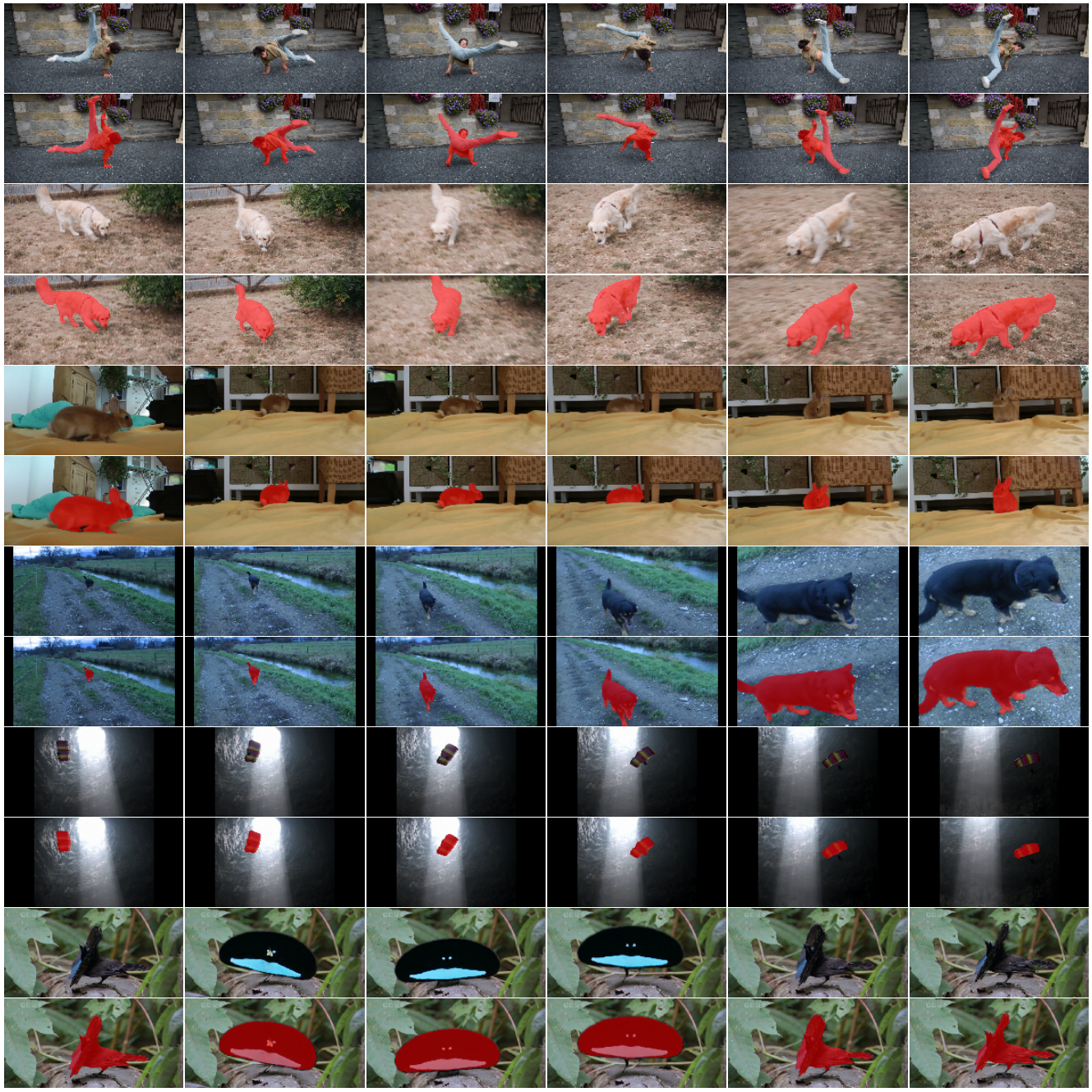


Figure 6. **Additional qualitative results on three benchmarks.** From top to bottom, the three groups show two sequences each from DAVIS 2016, FBMS59, and SegTrackv2. For every sequence, the upper row contains RGB frames sampled across the video, while the lower row shows the final OmegaMotion predictions overlaid in red. OmegaMotion consistently tracks the independently moving foreground across substantial articulation, scale variation, viewpoint changes, motion blur, and complex backgrounds.