

ProMo3D: Probing Motion Cues in Frozen 3D Foundation Models

Can a lightweight probe recover motion information from frozen 3D tokens?

Xiaoang Zhang, Mert Kiray, Yordanka Velikova, Manoj Biswanath, Benjamin Busam

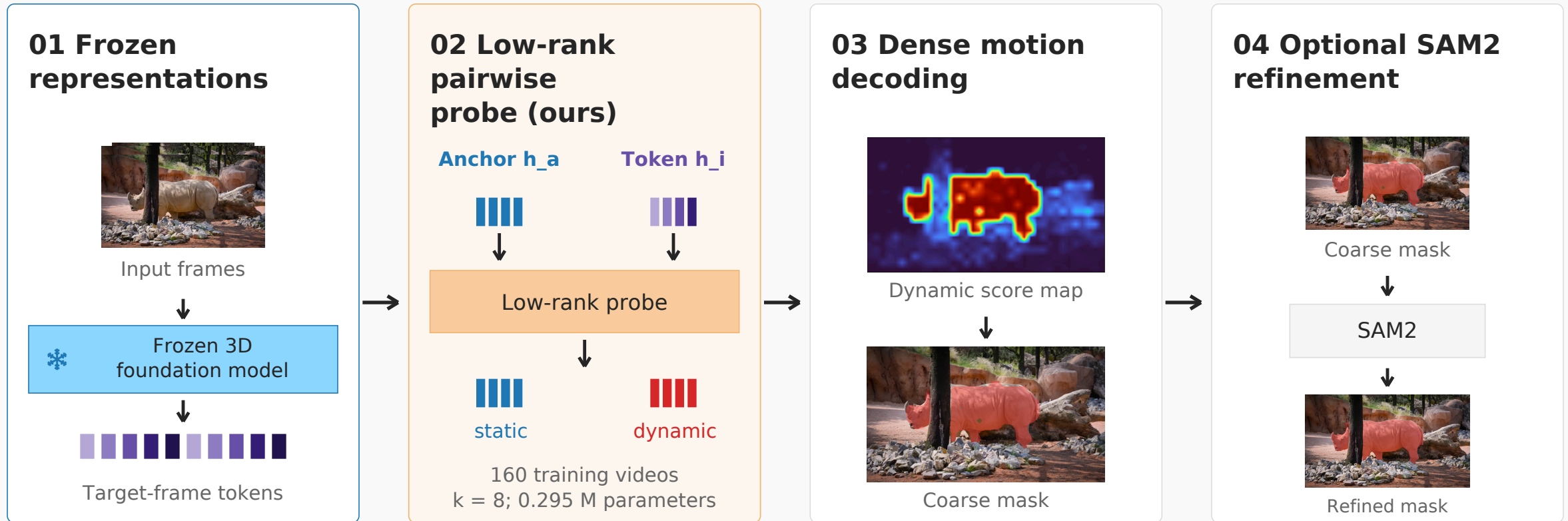
Technical University of Munich (TUM) | MCML | Obsphera

Input / coarse PMD mask



From frozen tokens to motion masks

The probe learns whether two frozen tokens share the same motion state.



$$f(\mathbf{h}_i, \mathbf{h}_j) = \sigma(\mathbf{h}_i^\top W \mathbf{h}_j + b)$$

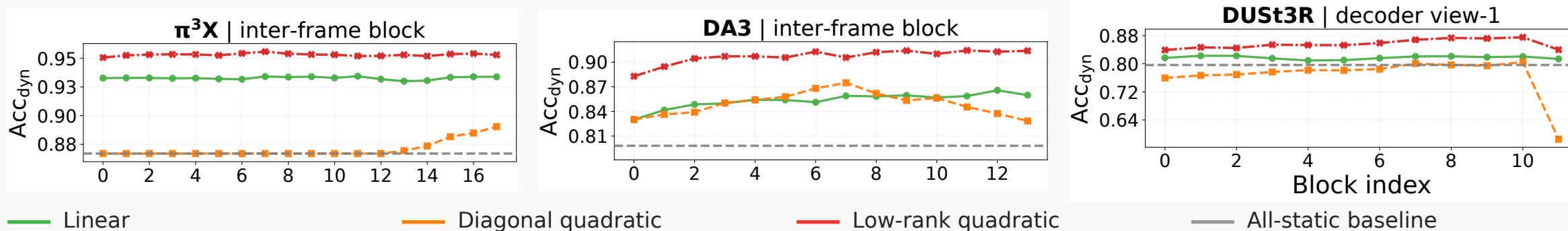
$$W = \frac{U^\top V + V^\top U}{2\sqrt{d}}, \quad U, V \in \mathbb{R}^{k \times d}$$

$$\hat{p}_i^{\text{dyn}} = 1 - f(\mathbf{h}_a, \mathbf{h}_i)$$

Pi3X: anchor token 0 lies in static white padding.
Average blocks, bilinearly upsample, then threshold.

Motion is decodable across geometric layers

Rank-8 low-rank quadratic probes compared with linear and diagonal probes

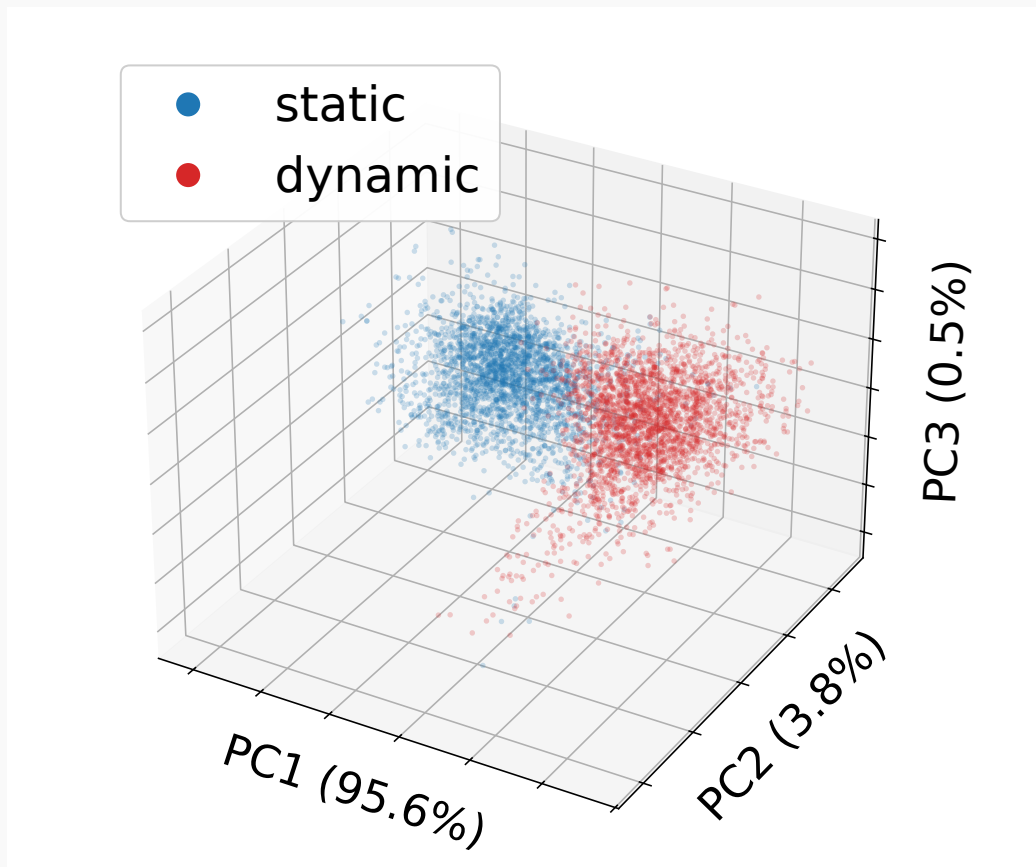


Motion is decodable across many blocks. PMD therefore averages a fixed group, without per-sequence best-layer selection.

Source: camera-ready Figure 2. DAVIS-all diagnostic; original vector plot crops.

A compact probe-induced representation

PCA visualizes the learned subspace; the rank sweep tests how much probe capacity is needed.



Evaluation dataset: DAVIS-all

Probe-induced 3D PCA | Pi3X, $k = 8$

Low factor ranks retain validation accuracy

k	Acc_dyn (%)	Params
4	96.10	0.147 M
8	96.12	0.295 M
16	96.10	0.590 M

Best-block Acc_dyn on YouTube-VOS validation (80 videos). Counts correct static and dynamic token labels.

Zero-shot benchmark transfer

J-Mean / J-Recall (%) | PMD uses no target-benchmark fine-tuning

Without SAM/SAM2	DAVIS2017	SegTrackv2	FBMS-59
VGGT4D	56.5 / 65.6	--	--
Easi3R (DUST3R)	46.9 / 50.5	--	--
GeoMotion	75.4 / 87.2	--	--
π^3X-PMD (ours)	73.7 / 91.0	72.9 / 94.0	73.5 / 88.8

Coarse PMD improves over training-free baselines without any SAM2 refinement.

With SAM/SAM2	DAVIS2017	SegTrackv2	FBMS-59
GeoMotion + SAM2	81.1 / 92.6	77.3 / --	72.5 / --
Flow-SAM*	71.6 / --	84.2 / --	88.7 / --
SegAnyMo	--	76.3 / --	78.3 / --
π^3X-PMD + SAM2 (ours)	81.9 / 92.3	86.4 / 99.3	83.2 / 89.3

J-M / J-R (%). Non-PMD scores quoted from camera-ready Table 2; supervision and priors differ.
Flow-SAM and SegAnyMo include native SAM/SAM2. * Flow-SAM uses DAVIS fine-tuning.

Coarse-mask threshold robustness

Pi3X-PMD J-Mean (%) without SAM2 | Each rule is fixed across all three benchmarks

Rule	DAVIS-all	SegTrackv2	FBMS-59	Mean	Std.
Fixed 0.30	64.5	69.4	66.6	66.8	1.98
Fixed 0.40	68.1	73.3	71.1	70.8	2.15
Fixed 0.50	69.9	73.3	73.9	72.4	1.75
Fixed 0.60	69.7	69.4	74.2	71.1	2.20
Fixed 0.70	66.5	61.6	71.2	66.4	3.93
Otsu	69.1	72.2	70.8	70.7	1.28
Otsu + 0.1	71.2	74.5	73.2	73.0	1.34
Otsu + 0.2 (default)	70.7	72.9	73.5	72.4	1.17
Otsu + 0.3	67.5	66.2	70.8	68.2	1.93

Fixed 0.40-0.60: std. 0.69

Otsu + 0.0-0.2: std. 0.97

Callouts: across-rule std. of the three-benchmark mean J-Mean. Table Std.: across benchmarks.

The default is not the best tested rule: Otsu + 0.1 reaches 73.0 vs. 72.4. Source: camera-ready threshold table.

Supervision, capacity, and performance

Compare motion supervision, trainable capacity, and benchmark performance.

Method	Motion-mask supervision	Params (M)	DAVIS2017 J-M	SegTrackv2 J-M	FBMS-59 J-M
SegAnyMo	Kubric, Dynamic Replica, HOI4D	3.44	--	76.3	78.3
Flow-SAM	Synthetic pretraining; DAVIS fine-tuning	9.83	71.6	84.2	88.7
GeoMotion +SAM2	HOI4D, Dynamic Replica, GOT-Motion YTVOS18-m, OmniWorld-motion	80.3	81.1	77.3	72.5
PMD (ours) π^3 X+SAM2	160 YouTube-VOS videos	0.295	81.9	86.4	83.2

~272x fewer trained parameters

160 source videos

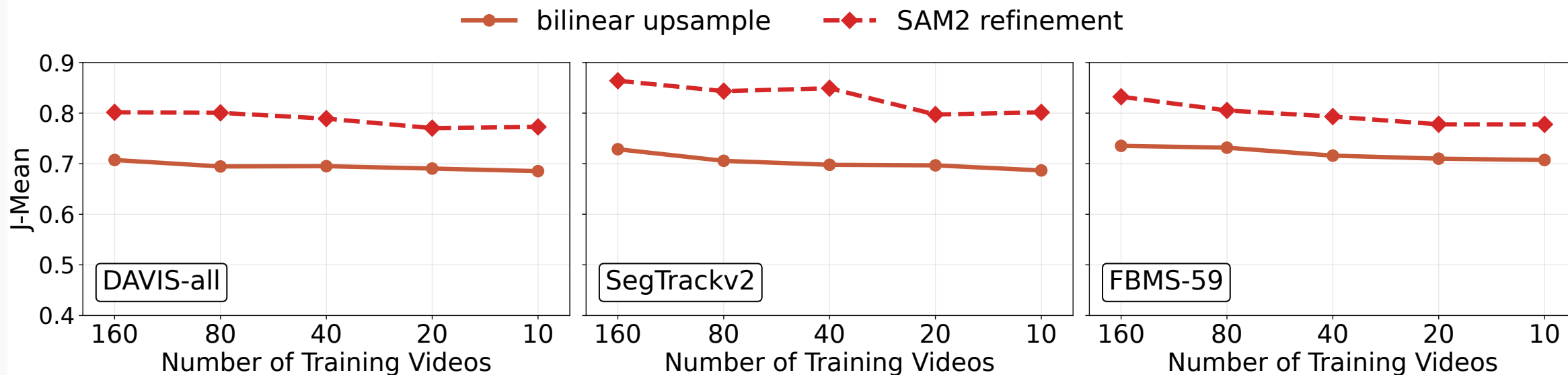
than GeoMotion's motion branch

no target-benchmark fine-tuning

J-M (%); scores include SAM/SAM2. Quoted baselines use different training sets and priors.
Trainable parameters exclude frozen backbones/priors. Flow-SAM uses DAVIS fine-tuning.

Limited supervision and small readout overhead

Reducing the source data tests data efficiency; profiling measures the readout's added cost.



160 to 10 training videos
Gradual degradation in coarse-mask
quality

0.295 M trained parameters
+4.2 ms/frame over Pi3X

Figure 5(b), three target benchmarks. Matched profile: 480.6 ms backbone / 484.8 ms coarse PMD; not a fastest-pipeline claim.

Frozen 3D tokens support motion decoding

A low-rank supervised probe recovers transferable motion information with limited source data.



Rhino | coarse PMD



Koala | coarse PMD



Cheetah | coarse PMD

The backbone stays frozen; only the low-rank probe is trained on source videos.



Full-sequence videos
on the project page

Backup: probe and anchor decoding

Training learns same-state compatibility; inference converts it to a dynamic score.

Train on token pairs

Use two spatial tokens from the same frame. The pair label is 1 when their static/dynamic states match, and 0 otherwise.

$$z_{ij} = \mathbf{1}[y_i = y_j]$$

$$f(\mathbf{h}_i, \mathbf{h}_j) = \sigma(\mathbf{h}_i^\top W \mathbf{h}_j + b)$$

$$W = \frac{U^\top V + V^\top U}{2\sqrt{d}}, \quad U, V \in \mathbb{R}^{k \times d}$$

Only U , V and b are learned with binary cross entropy; the 3D backbone stays frozen.

$$P = 18(2kd + 1) = 294,930 \quad (k = 8, d = 1024)$$

Pi3X: 18 blocks. k is the factor rank; the symmetric matrix W has rank at most $2k$.

Decode a dense motion mask

At inference, $\mathbf{h}_a$ is a fixed static reference: token 0 in Pi3X's white padding. Compatibility gives the static score; its complement gives the dynamic score.

$$\hat{p}_i^{\text{dyn}} = 1 - f(\mathbf{h}_a, \mathbf{h}_i)$$

Average blocks, bilinearly upsample, and threshold. SAM2 can optionally refine the resulting coarse mask.

Backup: metrics and selectivity

What each metric measures

$$J_{s,t} = \frac{|M_{s,t} \cap G_{s,t}|}{|M_{s,t} \cup G_{s,t}|}$$

J is mask IoU: overlap divided by the union of prediction and GT.

$$J\text{-Mean} = \frac{1}{S} \sum_s \frac{1}{T_s} \sum_t J_{s,t}$$

J-Mean averages frame IoU within each sequence, then averages sequences.

$$J\text{-Recall} = \frac{1}{S} \sum_s \frac{1}{T_s} \sum_t \mathbf{1}[J_{s,t} > 0.5]$$

J-Recall is the fraction of frames with $J > 0.5$, averaged across sequences.

$$\text{Acc}_{\text{dyn}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Acc_dyn is binary token accuracy, including static and dynamic tokens.

Token cutoffs: prediction 0.5; GT dynamic-pixel ratio 0.35.
Masks: valid GT pixels only; FBMS-59: annotated frames only.

Matched shuffled-label control

Pi3X, rank 8, YouTube-VOS validation. Label shuffling within frame groups preserves counts; features and probe capacity stay fixed.

Pairwise control	Pair acc. (%)	Balanced acc. (%)
Correct labels	95.72	91.67
Shuffled labels	84.48	50.34
Always-same baseline	84.47	50.00

Correct labels retain high balanced accuracy; shuffled labels fall to chance. This tests whether pairwise decodability depends on the correct token-label correspondence.

Backup: six frozen backbones

Backbone / stage	DAVIS2017	SegTrackv2	FBMS-59
π^3 X / coarse	73.7 / 91.0	72.9 / 94.0	73.5 / 88.8
π^3 X / SAM2	81.9 / 92.3	86.4 / 99.3	83.2 / 89.3
π^3 / coarse	72.7 / 89.9	72.9 / 92.9	72.6 / 83.4
π^3 / SAM2	80.2 / 91.1	84.5 / 97.5	81.1 / 85.4
VGGT / coarse	72.6 / 89.9	70.8 / 84.7	73.5 / 85.9
VGGT / SAM2	80.8 / 92.3	83.6 / 96.4	82.9 / 90.8
DA3 / coarse	69.6 / 85.8	66.3 / 80.3	69.8 / 81.4
DA3 / SAM2	77.7 / 88.2	75.7 / 89.8	82.3 / 88.6
MapAnything / coarse	69.2 / 86.0	62.3 / 75.7	69.1 / 79.9
MapAnything / SAM2	77.3 / 86.6	75.5 / 88.2	75.8 / 78.5
DUST3R / coarse	62.2 / 70.0	58.6 / 77.3	64.0 / 74.5
DUST3R / SAM2	66.5 / 70.6	66.0 / 84.0	74.0 / 78.3

J-M / J-R (%). Same 160-video source training; no target-benchmark fine-tuning.