# 3AM: 3EGMENT ANYTHING WITH ADAPTIVE GEOMETRIC CONSISTENCY IN VIDEOS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Consistent video object tracking must preserve identity under large viewpoint changes, temporary absence, and nearby distractors. Memory-based trackers such as SAM2 rely on appearance matching and fail in these cases, whereas 3D-based methods preserve identity across views but need camera poses, depth, or a reconstructed point cloud, run offline, and are not promptable. We present 3AM, a promptable online tracker that achieves geometric consistency from RGB alone by coupling memory-based tracking with the online multi-view reconstruction model MUSt3R. We identify the MUSt3R layers whose features preserve instance identity across views and fuse them into the tracker's image features by feature-wise modulation whose parameters are decoded from *routing tokens*. Geometry alone is insufficient, since a location cue holds only while the object stays in place; the routing tokens therefore traverse the tracker's memory attention and decide per frame how much the fusion should rely on geometry, so that one model handles static and moving objects without a hand-designed fallback. On wide-baseline benchmarks from ScanNet++ and Replica, 3AM reaches 90.6% IoU and 65.7% tracking recall on our selected subset of ScanNet++ (+12.3 and +28.2 points over the best prior tracker), matches or exceeds SAM2 on standard VOS and VOT benchmarks, and transfers zero-shot to ego-exo correspondence on Ego-Exo4D.

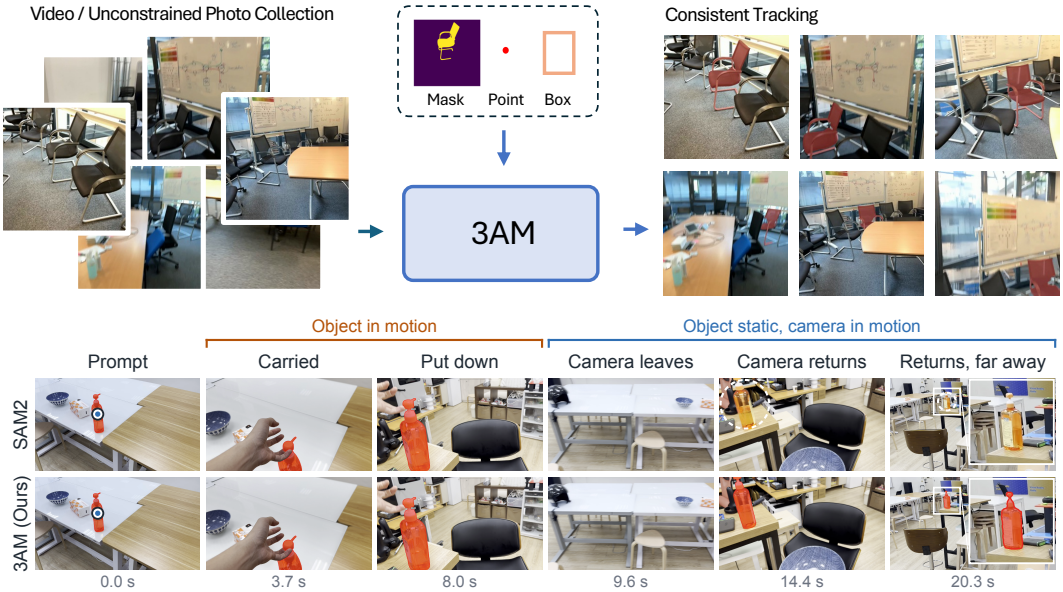


Figure 1: **Overview of 3AM and tracking through object motion and reappearance.** Upper: given a video or unordered photo collection and a mask, point, or box prompt on one frame, 3AM tracks the object across all views, online and from RGB only, adapting between geometry and appearance. Lower: in a phone video prompted with one click, the bottle is carried, put down elsewhere, and revisited twice; both trackers follow it while it moves, but only 3AM re-identifies it on both returns, whereas SAM2 (Ravi et al., 2024) never recovers it (dashed circle, magnified in the inset).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 2: **Two existing paradigms and ours.** (a) Memory-based trackers such as SAM2 (Ravi et al., 2024) lose the object under a large viewpoint change. (b) 3D segmentation methods preserve identity in scene coordinates but need camera poses and 3D mask merging, and propagate reconstruction errors. (c) 3AM preserves identity across drastic viewpoint changes from RGB alone.

## 1 INTRODUCTION

Video object segmentation (VOS) propagates a prompt on one frame to masks of the same object in all other frames, and for embodied agents the central requirement is consistency. As a robot or wearable camera moves through a room, the object leaves the view and reappears from another direction, at another distance, or next to similar objects (Figure 1, lower); the tracker must return the same object, and no mask while the object is absent. Memory-based models such as SAM2 (Ravi et al., 2024) match each frame against stored appearance, which fails under large viewpoint changes because no stored view matches (Figure 2a); on ScanNet++ (Yeshwanth et al., 2023), SAM2 and its memory-selection extensions (Ding et al., 2024; Videnovic et al., 2025) degrade sharply, as memory management cannot recover an object whose appearance has changed.

Appearance features lack the object's location. 3D-based methods preserve identity by construction, by segmenting point clouds (Schult et al., 2022; Kolodiazhnyi et al., 2024) or by lifting and merging 2D masks (Yang et al., 2023; Nguyen et al., 2024; Zhao et al., 2025), but need poses, depth, and a reconstruction whose cost grows with sequence length, and are neither promptable nor online (Figure 2b). Trackers thus offer the interface without location, and 3D-based methods the reverse.

We bridge this gap by coupling a promptable memory-based tracker with an online reconstruction model that provides geometry at every frame (Figure 2c). MUSt3R (Cabon et al., 2025) predicts point maps in a common frame from RGB, online and without poses; we identify its layers that preserve instance identity across views and fuse them into SAM2's image features by lightweight feature-wise modulation, so that memory attention, mask decoder, and memory encoder see geometry-aware features. Training uses 2D masks only, and inference needs no poses or depth.

Integrating geometry is necessary but not sufficient. Geometry is reliable only while the object stays in place, but objects are picked up, carried, and put down, so a tracker that always trusts geometry fails on a carried object and one that always trusts appearance fails on reappearance. We make this trade-off adaptive with *routing tokens* that traverse the tracker's memory attention, read the memory bank and the current frame, and decide how much the next frame's fusion relies on each geometric source; the geometric path opens for static objects and closes for moving ones. In the lower part of Figure 1, one model follows the carried bottle and re-identifies it once put down.

Our contributions are threefold. **(i) A view-consistency gap in video object tracking.** Memory-based trackers lose identity under large viewpoint changes and reappearance, and 3D-based methods need poses, depth, and offline reconstruction; our ScanNet++ and Replica protocol isolates wide-baseline reappearance. **(ii) Geometry-consistent tracking from RGB alone.** A layer-wise analysis of MUSt3R identifies the layers that preserve instance identity across views, which lightweight

feature-wise modulation fuses into a promptable memory-based tracker. **(iii) Adaptive routing between geometry and appearance.** Routing tokens in SAM2's memory attention set every frame's fusion parameters, so one model tracks static and moving objects without a hand-designed fallback. 3AM raises IoU on our ScanNet++ Selected Subset from 78.3 (SAM2Long) to 90.6 and Tracking Recall from 37.5 (DAM4SAM) to 65.7, matches or exceeds SAM2 on dynamic benchmarks, and transfers zero-shot to ego-exo correspondence on Ego-Exo4D (Grauman et al., 2024).

## 2 Related Work

**Video Object Segmentation.** VOS propagates a mask from one frame through a video (Perazzi et al., 2016; Caelles et al., 2017; Voigtlaender et al., 2019a; Maninis et al., 2018; Yang et al., 2019; Oh et al., 2019; Bhat et al., 2020). Memory-based methods match past frames by spatio-temporal attention (Yang et al., 2021; Cheng & Schwing, 2022; Cheng et al., 2021; Liu et al., 2022; Park et al., 2022; Xu et al., 2022; Yang et al., 2024b; Yang & Yang, 2022; Li et al., 2024a; Cheng et al., 2024), with later work on memory size (Zhou et al., 2024a; Liu et al., 2025a), unified transformers (Li et al., 2024b; Zhang et al., 2025b), anchor queries (Zhou et al., 2024b), and decoupled segmentation and propagation (Cheng et al., 2023). SAM2 (Ravi et al., 2024) adds streaming memory to SAM (Kirillov et al., 2023); follow-ups improve memory selection (Ding et al., 2024; Videnovic et al., 2025; Yang et al., 2024a; 2025b), efficiency (Zhou et al., 2025), and language grounding (Cuttano et al., 2025; Yuan et al., 2025), and SAM 3 (Carion et al., 2026) adds concept prompts. Recent variants separate short- and long-term memory (Kim & Lim, 2026; Alansari et al., 2025), gate memory writes by presence (Chen, 2026), unify mask, box, and point granularities (Zhang et al., 2025c), add motion and semantic cues (Zhu et al., 2026b), or build object concepts with a vision-language model (Zhang et al., 2026b). Long-term benchmarks require absence reporting and re-detection (Valmadre et al., 2018; Tang et al., 2023), object permanence is learned for multi-object tracking (Tokmakov et al., 2021), appearance memories re-identify indoor re-entries (Diaz-Pereda et al., 2026), and Whareformer (Chalk et al., 2026) associates observations of moved objects with tracks through appearance and 3D-location memories, given masks from a separate segmenter and depth aligned to a reconstruction. All rely on appearance matching, which benchmarks with large viewpoint changes and reappearance (Ding et al., 2025; Hong et al., 2023; Zhang et al., 2026b) show to be fragile; 3AM instead makes a promptable tracker's memory geometry-aware.

**Consistency through 3D Segmentation.** Methods that preserve identity across views operate in scene coordinates. Proposal-based methods segment point clouds (Schult et al., 2022; Jiang et al., 2020; Vu et al., 2022; Sun et al., 2023; Kolodiazhnyi et al., 2024; Takmaz et al., 2023), and lifting methods merge per-view masks in 3D by linear assignment, contrastive fusion, or view-consensus clustering (Siddiqui et al., 2023; Bhalgat et al., 2023; Yan et al., 2024; Yang et al., 2023; Nguyen et al., 2024; Yin et al., 2024; Jung et al., 2025; Zhao et al., 2025), sometimes through feature fields or Gaussians (Kerr et al., 2023; Qin et al., 2024; Ye et al., 2024; Shen et al., 2024), as do object-centric maps for embodied agents (Gu et al., 2024); SAM3D (Yang et al., 2023) is unrelated to the single-image SAM 3D models (Chen et al., 2026c; Yang et al., 2026). Most need posed RGB-D and offline processing; online variants work on posed RGB-D (Xu et al., 2025c; Wang et al., 2025a) or monocular RGB with reconstruction priors (Du et al., 2026; Gelencsér-Horváth et al., 2026) but output scene-level instances, not a prompted object. Merging is needed because per-view masks are inconsistent (Siddiqui et al., 2023; Bhalgat et al., 2023; Peng et al., 2023; Jia et al., 2024); 3AM instead makes the 2D tracks consistent, so their projection into 3D needs little merging (Section 4.6).

**Feed-Forward 3D Reconstruction and 3D-Aware Segmentation.** DUSt3R-style models regress point maps from RGB (Wang et al., 2024; Leroy et al., 2024) and have become multi-view and online. Spann3R (Wang & Agapito, 2025) adds a spatial memory, MUSt3R (Cabon et al., 2025) and CUT3R (Wang et al., 2025c) keep a persistent memory or state, STream3R (Lan et al., 2026) and StreamVGGT (Zhuo et al., 2026) use causal attention, Point3R (Wu et al., 2025), LingBot-Map (Chen et al., 2026a), and Anchor3R (Tao et al., 2026) bound the memory of long streams, MonST3R (Zhang et al., 2024a) and Easi3R (Chen et al., 2025a) handle dynamic scenes, and SLAM systems build on these priors (Murai et al., 2025; Liu et al., 2025b; Maggio et al., 2025), whereas VGGT (Wang et al., 2025b), $\pi^3$ (Wang et al., 2026), Depth Anything 3 (Lin et al., 2026), Fast3R (Yang et al., 2025a), and MapAnything (Keetha et al., 2026) process frame sets offline. For segmentation, PanSt3R (Zust et al., 2025) predicts panoptic labels from MUSt3R features offline, SegMASt3R (Jayanti et al., 2025) matches segments between two images, ODIN (Jain et al., 2024)

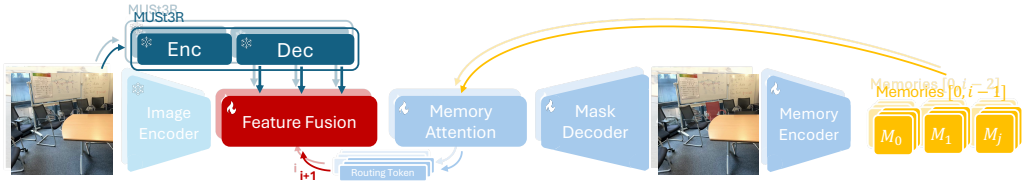


Figure 3: **3AM pipeline.** Frozen SAM2 and MUSt3R encode each frame, MUSt3R with its own memory of past frames. Feature Fusion (Section 3.2) merges multi-layer MUSt3R features into the SAM2 feature used by memory attention, the mask decoder, and the memory encoder, so the memories $M_0, \ldots, M_j$ are geometry-aware; routing tokens pass through memory attention, and their output $r^{i+1}$ sets the fusion of frame $i{+}1$.

fuses 2D and 3D backbones, IGGT (Li et al., 2026) clusters learned 3D-consistent instance embeddings, and SAM-V (Gong et al., 2026) and MV-SAM (Jeong et al., 2026) inject VGGT features or point maps into SAM to segment one object in a fixed set of static views without temporal memory. None is both promptable and online. Probing studies measure the 3D awareness of 2D foundation features (El Banani et al., 2024; Man et al., 2024; Chen et al., 2025b), find epipolar geometry in the middle layers of reconstruction models (Bratulić et al., 2025), and show that VGGT layers specialize (Ziliotto et al., 2026); Section 3.1 measures instance identity across MUSt3R layers.

**Tracking through 4D Reconstruction.** Point trackers follow sparse points in 2D (Doersch et al., 2023; Karaev et al., 2024), online with a memory (Aydemir et al., 2025; 2026), or in 3D (Xiao et al., 2024; Ngo et al., 2025), and long-range tracks with semantic features can prompt SAM2 to segment moving objects (Huang et al., 2025). Among feed-forward 4D models, TAPIP3D (Zhang et al., 2025a) and SpatialTrackerV2 (Xiao et al., 2025) track points in a camera-stabilized 3D space, TrackingWorld (Lu et al., 2025) optimizes world-centric trajectories per video, D4RT (Zhang et al., 2026a), 4RC (Luo et al., 2026), and Track4World (Lu et al., 2026) encode a video once and answer motion queries for any pixel and time, and Point4D (Jeon et al., 2026) extends such queries to multi-hundred-frame videos by chunking. They process whole videos or long chunks offline and output trajectories rather than masks, so identity after a long absence still rests on a mask model; we pair their tracks with SAM2 as baselines. SAM4D (Xu et al., 2025a) extends SAM2's memory to camera and LiDAR streams and GOT-Edit (Chen et al., 2026b) injects VGGT features into a box tracker, but the former needs LiDAR and the latter outputs boxes, and IGGT4D (Zou et al., 2026) streams instance-grounded geometry but clusters all instances of a scene rather than following a prompt and needs 4D instance annotations. Closest to 3AM, G$^2$TAM (Zhu et al., 2026a) builds on $\pi^3$ and uses aligned geometric representations as an implicit memory for promptable tracking, but processes image sets offline; 3AM keeps an online tracker's appearance memory and learns per frame how much to rely on geometry, so that one model also tracks moving objects.

**Cross-View Object Correspondence.** Ego-Exo4D (Grauman et al., 2024) defines object correspondence between synchronized ego and exo cameras, with co-segmentation and VOS baselines (XSegTx and XView-XMem (Cheng & Schwing, 2022)) and PSALM (Zhang et al., 2024b) as a zero-shot reference; later methods add language cues (Fu et al., 2025), DINO-based mask matching (Mur-Labadia et al., 2025), long-term memory (Hu et al., 2025), cycle-consistent mask prediction (Yan et al., 2026), and multi-prompt experts on SAM2 (Pan et al., 2026). All train on Ego-Exo4D, whereas we evaluate 3AM zero-shot on ego-exo pairs as extreme viewpoint changes.

## 3 METHOD

Given a video and a mask, point, or box prompt on one frame, the tracker outputs the object's mask in every frame where it is visible and an empty mask otherwise, as the object leaves the view, returns from another viewpoint, or moves. In 3AM (Figure 3), each frame $i$ is encoded by SAM2's frozen image encoder (Ravi et al., 2024), giving an appearance feature $F_{\text{2D}}$, and by frozen MUSt3R (Cabon et al., 2025), an online reconstruction model with its own memory of past frames, giving geometric features $F_{\text{3D}}$ that encode where the content lies. Feature Fusion (Section 3.2) merges them into $F_{\text{fused}}$, which replaces $F_{\text{2D}}$ in memory attention, the mask decoder, and the memory encoder, and routing tokens computed in memory attention at frame $i$ produce routing vectors $r^{i+1}$ that set the fusion of frame $i{+}1$. Inference requires RGB only. We keep SAM2 as the segmentation backbone instead of segmenting the predicted point cloud, which would inherit reconstruction errors, run offline, and forgo abundant 2D video data; geometry serves only as an association cue.

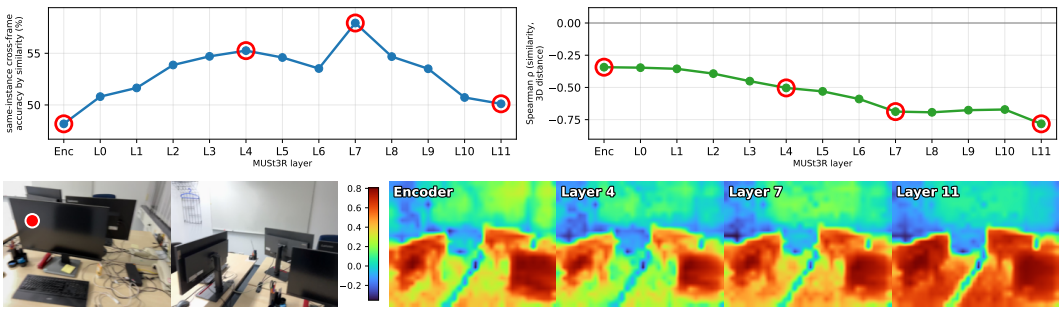


Figure 4: **MUSt3R features move from semantics to geometry with depth.** Top: per-layer statistics on ScanNet++ frame pairs (selected layers circled); same-instance accuracy peaks at middle layers, and the Spearman correlation with 3D distance grows with depth. Bottom: similarity of a red query on a monitor to another view; the encoder fires on every monitor, while geometry-bound deeper layers stop at the desk edge.

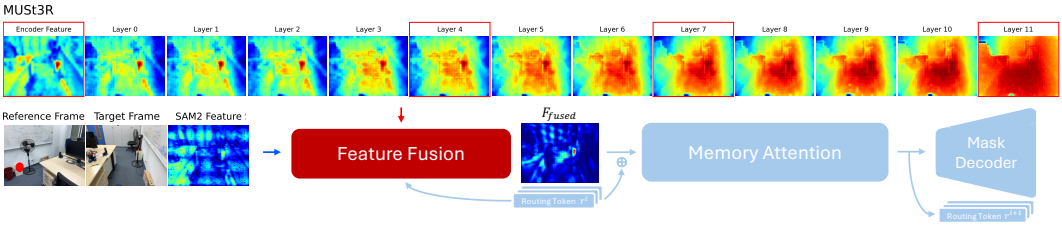


Figure 5: **Feature Fusion and routing.** Top: similarity of a reference-frame query point to the target frame for the MUSt3R encoder and decoder layers 0 to 11 (selected layers boxed). Bottom: the selected layers and the explicit geometry are projected, scaled and shifted per channel, gated, and added to the SAM2 feature with parameters decoded from the routing vector $r^i$; $F_{\text{fused}}$ and the routing tokens are concatenated along the token dimension for memory attention, whose token output $r^{i+1}$ parameterizes the fusion of the next frame.

### 3.1 What Do MUSt3R Features Encode?

MUSt3R predicts point maps in a common coordinate frame from an RGB stream, so its features encode cross-view correspondence, but which layers encode it in a form useful for object identity is unknown. On pairs of ScanNet++ frames, we measure for the encoder output and the twelve decoder blocks (i) same-instance accuracy, the fraction of query patches whose most similar patch in the other frame belongs to the same instance, and (ii) the Spearman correlation between the cosine similarity and the 3D distance of two patches (Figure 4). The curves show one gradual transition. Early layers are semantic, so a query on a monitor activates every monitor; intermediate decoder blocks attain the highest instance accuracy; deeper blocks correlate more with 3D distance and follow geometry rather than appearance, stopping at the desk edge. As no single layer serves both roles, we take the encoder output and decoder blocks 4, 7, and 11 and let the model weight them.

### 3.2 Feature Fusion with Routing Tokens

Let $F_{2\text{D}} \in \mathbb{R}^{C\times H\times W}$ be the SAM2 feature of frame $i$ at stride 16 (frame index omitted). The five geometric sources $F_{3\text{D}}^{(k)}$, $k = 1, \ldots, 5$, are the four selected MUSt3R layers and the frame's MUSt3R point map with its camera rays in Plücker coordinates, encoded by sinusoidal bands and a patch embedding. A small convolutional head ($1\times1$ convolution, LayerNorm, GELU, $3\times3$ convolution) and bilinear upsampling project each source to $C \times H \times W$, giving $\hat{F}_{3\text{D}}^{(k)}$, and the fusion is a gated, feature-wise modulated sum (Figure 5),

$$F_{\text{fused}} = \gamma_0 \odot F_{2\text{D}} + \beta_0 \; + \; \sum_{k=1}^{5} \alpha_k \left(\gamma_k \odot \hat{F}_{3\text{D}}^{(k)} + \beta_k\right), \qquad (\gamma_k, \beta_k, \alpha_k) = \text{MLP}_k(r_k^i), \qquad (1)$$

where $\gamma_k, \beta_k \in \mathbb{R}^C$ scale and shift each channel, $\alpha_k \in \mathbb{R}$ gates the whole source, $\odot$ is channel-wise multiplication, and the SAM2 feature ($k = 0$) is not gated. A two-layer $\text{MLP}_k$ decodes $(\gamma_k, \beta_k, \alpha_k)$ from the routing vector $r_k^i \in \mathbb{R}^{256}$ that the routing tokens produced at the previous frame. The routing vectors are the state that makes the fusion adaptive; each source enters under one global gate, which is cheap and can be switched off, and with all gates closed the model reduces to SAM2.

**Routing Tokens.** SAM2's memory attention has four layers, each with self-attention over the $HW$ frame tokens, cross-attention to the memory bank (spatial memories and object pointers of past frames), and a feed-forward layer. We concatenate $N = 12$ routing tokens, two per term $k = 0, \ldots, 5$ of Eq. (1), to the frame tokens along the token dimension and pass all tokens through these layers with shared weights; the routing tokens carry no position and receive no rotary encoding, and at every layer they see the fused frame, the memory bank, and the spatial tokens. At frame $i$, each token is a learned embedding plus a linear projection of $r_k^i$; after the last layer, the two tokens of term $k$ are concatenated and projected to $r_k^{i+1}$, which sets $(\gamma_k, \beta_k, \alpha_k)$ of frame $i{+}1$. A learned initial state $r_k^0$ serves the prompted frame, and $r_k^1 = r_k^0$ plus the projected object pointer of the prompt conditions routing on the tracked object. When the current 3D features agree with the object's stored memories, the geometric gates stay open; when the object has moved relative to the scene, memories and geometry disagree, and the gates close so that the model reverts to appearance matching, per frame and per source. Since the memory encoder stores fused features, the memory bank keeps the geometric cue whenever it was used. The routing output projection and the object-pointer projection are zero-initialized, so training starts as plain SAM2 with gates near zero and opens the geometric path where geometry reduces the loss.

### 3.3 Training

We fine-tune SAM2.1-L on eight-frame clips with the SAM2 losses (focal and dice on the mask, $\ell_1$ on the predicted IoU, cross-entropy on object presence), training Feature Fusion, routing, memory attention, the memory encoder, and the mask decoder with the image encoder and MUSt3R frozen.

**Data.** Training mixes two static datasets with long camera trajectories, ScanNet++ (Yeshwanth et al., 2023) with masks projected from 3D labels and the synthetic ASE (Avetisyan et al., 2024) with clean masks, and three dynamic video datasets, MOSE (Ding et al., 2023), MOSEv2 (Ding et al., 2025), and SA-V (Ravi et al., 2024). Dynamic data are essential, as with static data alone the routing tokens never see geometry mislead and trust it unconditionally (Section 4.5). Static clips use *anchor-then-contiguous* sampling, which pairs a prompt frame from anywhere in the sequence with a contiguous seven-frame block elsewhere, and dynamic clips use consecutive frames.

## 4 Experiments

### 4.1 Setup

**Implementation.** We build on SAM2.1-L ($1024^2$ input) and MUSt3R-512 (512 pixels on the long side); the fused feature has $C = 256$ channels at $64 \times 64$, routing uses $N = 12$ tokens and 256-dimensional vectors, and 3AM keeps SAM2's memory selection throughout. Training uses AdamW with per-module learning rates between $10^{-6}$ and $10^{-5}$, cosine decay, and bfloat16 on one RTX 4090, with MUSt3R on a second GPU (details in Appendices A to C).

**Benchmarks and Metrics.** Conventional VOS and VOT benchmarks such as DAVIS (Pont-Tuset et al., 2017), YouTube-VOS (Xu et al., 2018), LVOS (Hong et al., 2023), and LaSOT (Fan et al., 2019) mostly keep the object in view and the camera near its initial pose. For large viewpoint changes we use ScanNet++ (Yeshwanth et al., 2023) and Replica (Straub et al., 2019), whose trajectories cover whole rooms; 3D instance labels are projected into every frame with ground-truth poses, and the frame with the largest visible mask is the prompt, tracked forward and backward; our ScanNet++ training masks come from the same projection. Besides the Whole Set of ScanNet++ objects, we filter a *Selected Subset* of 286 objects whose tracks split into many disjoint visible segments, *i.e.*, frequent reappearance under large viewpoint changes. **IoU** averages over all frames, including those where the object is absent (the tracking quality of VOTS, Kristan et al., 2023), and thus penalizes false positives; **Tracking Recall** (Lukežič et al., 2021) is the mean IoU over frames where the object is present; **Accuracy** is the mean IoU over present frames where the prediction overlaps the ground truth (the VOTS accuracy), *i.e.*, mask quality given correct localization. Dynamic objects are evaluated on DAVIS17, LaSOT, $\text{LaSOT}_{\text{ext}}$, KITTI-MOTS (Voigtlaender et al., 2019b), and EPIC-KITCHENS (Damen et al., 2018; Darkhalil et al., 2022), objects that are moved and left in place on Aria Digital Twin (ADT) (Pan et al., 2023), cross-view correspondence on Ego-Exo4D (Grauman et al., 2024) with the protocol of $\text{V}^2$-SAM (Pan et al., 2026), and 3D-projected tracks on class-agnostic ScanNet200.

Table 1: **Tracking under large viewpoint changes and on dynamic benchmarks** (%) for SAM2 (Ravi et al., 2024), SAM2Long (Ding et al., 2024), and DAM4SAM (Videnovic et al., 2025); KITTI-MOTS and EPIC-KITCHENS use mask prompts; ADT: Aria Digital Twin (Pan et al., 2023). Best in red.

| | ScanNet++ | | | | | | Replica | | | DAVIS17 | LaSOT | $\text{LaSOT}_{\text{ext}}$ | KITTI-MOTS | | | EPIC-KITCHENS | | | ADT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Whole Set | | | Selected Subset | | | | | | | | | | | | | | | | | |
| Method | IoU | Recall | Acc. | IoU | Recall | Acc. | IoU | Recall | Acc. | $\mathcal{J}\&\mathcal{F}$ | AUC | AUC | IoU | Recall | Acc. | IoU | Recall | Acc. | IoU | Recall | Acc. |
| SAM2 | 81.2 | 32.5 | 69.3 | 74.3 | 28.2 | 66.6 | 80.0 | 56.0 | 69.0 | 89.0 | 70.0 | 56.9 | 81.9 | 83.8 | 84.0 | 78.5 | 83.1 | 86.8 | 81.3 | 68.1 | 71.9 |
| SAM2Long | 81.3 | 41.7 | 69.6 | 78.3 | 36.1 | 62.4 | 76.9 | 52.0 | 62.7 | 88.3 | 73.9 | – | – | – | – | – | – | – | 81.2 | 68.3 | 72.9 |
| DAM4SAM | 83.1 | 41.9 | 67.8 | 76.6 | 37.5 | 63.8 | 79.3 | 58.5 | 67.8 | 86.6 | 75.1 | 60.9 | 81.9 | 83.8 | 84.7 | 77.7 | 83.0 | 86.4 | 81.0 | 69.1 | 73.1 |
| 3AM (Ours) | 89.5 | 59.3 | 71.6 | 90.6 | 65.7 | 72.4 | 83.4 | 67.8 | 70.9 | 90.4 | 75.4 | 60.1 | 82.6 | 83.9 | 84.0 | 78.8 | 83.1 | 86.7 | 85.3 | 76.9 | 75.5 |

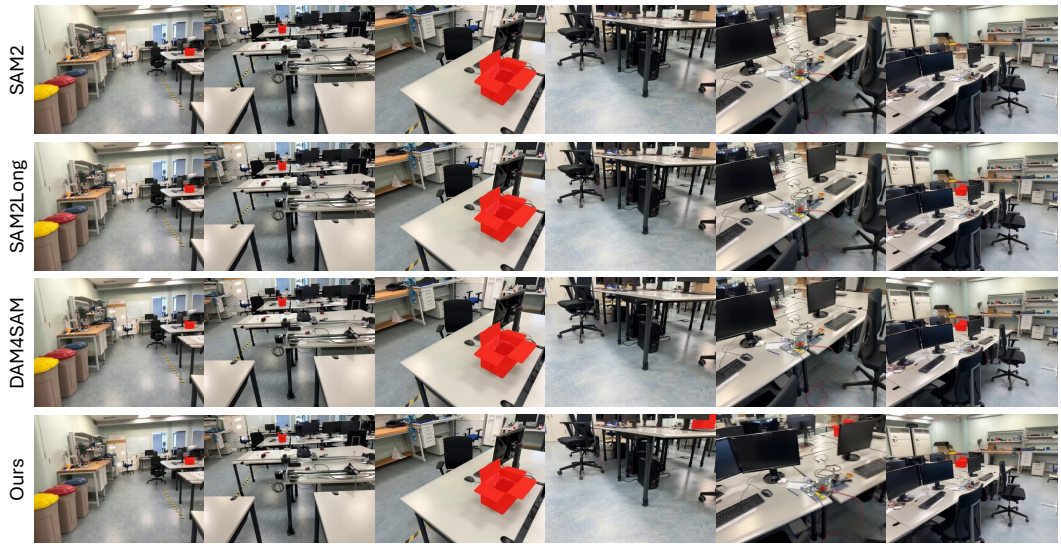


Figure 6: **Qualitative comparison.** The leftmost frame is prompted. SAM2-based trackers drift to distractors or return nothing after the object leaves the view; 3AM recovers the same instance from new viewpoints.

### 4.2 Tracking under Large Viewpoint Changes

Table 1 compares 3AM with SAM2 (Ravi et al., 2024) and its memory-selection extensions SAM2Long (Ding et al., 2024) and DAM4SAM (Videnovic et al., 2025), Table 2 adds SAM2 fine-tuned on our data (SAM2$^\dagger$), which separates the data from the architecture, and other tracker families, and Figure 6 shows examples. SAM2 recovers few objects after they leave the view (28.2 Recall on the Selected Subset), SAM 3 (Carion et al., 2026) and memory selection stay below 38, and SAM2$^\dagger$ reaches only 40.9, since memory attention without a geometric cue cannot separate the target from look-alikes under a new viewpoint. Point-based baselines track points sampled in the prompt mask with CoTracker3 (Karaev et al., 2025), OmniMotion (Wang et al., 2023), or the 4D model 4RC (Luo et al., 2026) and prompt SAM2 with them; point trackers cannot re-identify the object after a long absence, OmniMotion needs per-video optimization, and 4RC reaches 41.3 Recall but runs offline and yields coarse masks (51.7 Accuracy). The 3D-aware baselines are consistent by construction but are not online promptable trackers, since PanSt3R (Zust et al., 2025) predicts panoptic segments offline, SegMASt3R (Jayanti et al., 2025) matches segments between two views, EmbodiedSAM (Xu et al., 2025c) segments 3D instances online from posed RGB-D, and the training-free MUSt3R-Thr+SAM2, which prompts SAM2 with the thresholded MUSt3R feature similarity to the prompt mask, is not identity-aware. Reprojecting the prompt mask through MUSt3R's point maps (MUSt3R-Reproj+SAM2) reaches 40.7 Recall but yields coarse masks (50.4 Accuracy). IGGT4D supports online inference but relies on instance-feature clustering without native prompt-conditioned segmentation. We cannot compare with G$^2$TAM (Zhu et al., 2026a), whose code has not been released. 3AM is best on all datasets and metrics, with 90.6 IoU and 65.7 Recall on the Selected Subset, +12.3 IoU over SAM2Long and +28.2 Recall over DAM4SAM. 3AM also improves on the Whole Set, and the advantage holds for point and box prompts and for unordered photos, which are first reordered by feature retrieval.

Table 2: **Comparison with other tracker families** on the ScanNet++ Selected Subset (%). Prompt.: promptable; Online: streaming; RGB: no poses, depth, or per-video optimization; †: SAM2 fine-tuned on our data. Best in red.

| Method | Prompt. | Online | RGB | IoU | Recall | Acc. |
|---|---|---|---|---|---|---|
| *Memory-based VOS* | | | | | | |
| SAM2 (Ravi et al., 2024) | ✓ | ✓ | ✓ | 74.3 | 28.2 | 66.6 |
| SAM2† | ✓ | ✓ | ✓ | 77.5 | 40.9 | 67.0 |
| DAM4SAM (Videnovic et al., 2025) | ✓ | ✓ | ✓ | 76.6 | 37.5 | 63.8 |
| SAM 3 (Carion et al., 2026) | ✓ | ✓ | ✓ | 76.8 | 31.6 | 60.7 |
| *Point and 4D tracking, masks decoded by SAM2* | | | | | | |
| CoTracker3 (Karaev et al., 2025)+SAM2 | ✓ | ✓ | ✓ | 73.6 | 26.7 | 66.7 |
| OmniMotion (Wang et al., 2023)+SAM2 | ✓ | ✗ | ✗ | 58.0 | 9.3 | 56.4 |
| 4RC (Luo et al., 2026)+SAM2 | ✓ | ✗ | ✓ | 80.6 | 41.3 | 51.7 |
| *3D-aware segmentation and geometric matching* | | | | | | |
| PanSt3R (Zust et al., 2025) | ✗ | ✗ | ✓ | 52.1 | 7.1 | 9.2 |
| SegMASt3R (Jayanti et al., 2025) | ✗ | ✗ | ✓ | 54.1 | 35.1 | 40.4 |
| EmbodiedSAM (Xu et al., 2025c) | ✗ | ✓ | ✗ | 70.7 | 16.9 | 28.1 |
| MUSt3R-Thr+SAM2 (training-free) | ✓ | ✓ | ✓ | 65.1 | 30.0 | 31.0 |
| MUSt3R-Reproj+SAM2 (training-free) | ✓ | ✓ | ✓ | 81.5 | 40.7 | 50.4 |
| IGGT4D (Zou et al., 2026) | ✗ | ✓ | ✓ | 78.1 | 52.6 | 57.9 |
| 3AM (Ours) | ✓ | ✓ | ✓ | 90.6 | 65.7 | 72.4 |

Table 3: **Routing shares** (%) of the appearance terms (SAM2 feature, MUSt3R encoder) and geometric terms (decoder blocks 4, 7, 11, explicit geometry) on static ScanNet++ and dynamic DAVIS17.

| Terms | Static | Dynamic |
|---|---|---|
| Appearance | 49.3 | 82.2 |
| Geometric | 50.7 | 17.8 |

Table 4: **Efficiency** on an RTX 4090 ($1024^2$ input, MUSt3R at 512), single-object tracking.

| Method | FPS | Mem. (GB) |
|---|---|---|
| SAM2 (Ravi et al., 2024) | 14.4 | 3.6 |
| SAM2Long (Ding et al., 2024) | 7.1 | 6.8 |
| DAM4SAM (Videnovic et al., 2025) | 12.3 | 3.5 |
| 3AM (Ours) | 10.2 | 4.3 |

Table 5: **Ego-exo object correspondence on Ego-Exo4D** (IoU, %). Tr.: trained on Ego-Exo4D; SAM2 and 3AM are zero-shot. Baselines from Pan et al. (2026); best in red.

| Method | Tr. | Ego→Exo | Exo→Ego | Mean |
|---|---|---|---|---|
| PSALM (Zhang et al., 2024b) (zero-shot) | ✗ | 7.4 | 2.1 | 4.8 |
| CMX (Zhang et al., 2023) | ✓ | 6.8 | 12.0 | 9.4 |
| XSegTx (Grauman et al., 2024) | ✓ | 18.9 | 27.1 | 23.0 |
| XView-XMem (Grauman et al., 2024) | ✓ | 19.3 | 16.6 | 18.0 |
| XMem + XSegTx (Grauman et al., 2024) | ✓ | 34.9 | 25.0 | 30.0 |
| ObjectRelator (Fu et al., 2025) | ✓ | 35.3 | 40.3 | 37.8 |
| O-MaMa (Mur-Labadia et al., 2025) | ✓ | 42.6 | 44.1 | 43.4 |
| $V^2$-SAM (Pan et al., 2026) (single expert) | ✓ | 44.5 | 47.3 | 45.9 |
| $V^2$-SAM (Pan et al., 2026) (multi-expert) | ✓ | 46.3 | 49.6 | 48.0 |
| SAM2 (zero-shot) | ✗ | 14.0 | 20.6 | 17.3 |
| 3AM (Ours, zero-shot) | ✗ | 49.1 | 56.0 | 52.6 |

Table 6: **Projected 3AM tracks as class-agnostic 3D instances on ScanNet200**, a consistency check rather than a 3D segmentation method; all methods use ground-truth depth and poses. Best in red.

| Method | AP | $AP_{50}$ | $AP_{25}$ |
|---|---|---|---|
| *Offline* | | | |
| SAMPro3D (Xu et al., 2025b) | 18.0 | 32.8 | 56.1 |
| Open3DIS (Nguyen et al., 2024) | 34.6 | 43.1 | 48.5 |
| SAI3D (Yin et al., 2024) | 28.2 | 47.2 | 67.9 |
| SAM2Object (Zhao et al., 2025) | 34.0 | 52.7 | 70.3 |
| *Online* | | | |
| SAM3D (Yang et al., 2023) | 20.2 | 35.7 | 55.5 |
| EmbodiedSAM (Xu et al., 2025c) | 42.2 | 63.7 | 79.6 |
| 3AM (Ours) | 47.3 | 59.7 | 75.3 |

### 4.3 Dynamic Objects and Routing

The geometric prior must not hurt moving objects. On benchmarks dominated by object motion and stable cameras, including egocentric hand-object manipulation in EPIC-KITCHENS (Table 1, right), 3AM matches or exceeds SAM2 everywhere (90.4 versus 89.0 on DAVIS17, 75.4 versus 70.0 on LaSOT) and is on par with the distractor-aware DAM4SAM, as the routing tokens close the geometric gates when the current geometry disagrees with the memories. The lower part of Figure 1 shows the case that neither regime alone solves, a bottle carried across a room, put down, and revisited after the camera has left; both trackers follow it in the hand, but only 3AM re-identifies it on every return, even when small and distant. On ADT, where many objects are moved and then left in place, 3AM reaches 76.9 Recall, 7.8 points above the best SAM2-based tracker (Table 1).

**What Do the Routing Tokens Learn?** On static ScanNet++, geometric terms receive half of the routing-coefficient magnitude (50.7%, Table 3), whereas on dynamic DAVIS17 they receive 17.8%; decoder blocks 4 and 11 carry most of the static share, their signed gains flip or vanish on dynamic data, and the first 100 frames after the prompt give the same shares. Training the fusion without routing, with $(\gamma_k, \beta_k, \alpha_k)$ learned as constants shared by all frames, not set by hand (Table 7c), keeps most of the wide-baseline gain (59.6 versus 65.7 Recall) but loses 3.5 points on DAVIS17.

### 4.4 Cross-View Correspondence on Ego-Exo4D

An egocentric and an exocentric view of one object form an extreme viewpoint change. Following $V^2$-SAM (Pan et al., 2026), the target-view mask is predicted from the query-view image and mask in both directions; 3AM treats the pair as a two-frame video prompted on the query view, with MUSt3R initialized on both views. Without Ego-Exo4D training data, unlike all other methods in Table 5, 3AM reaches 52.6 mean IoU, 4.6 points above the best trained method and 35.3 above zero-shot SAM2, with the larger margin for Exo→Ego (56.0 versus 49.6).

| | Variant | IoU | Recall | Acc. | DAVIS17 |
|---|---|---|---|---|---|
| | 3AM (full) | 90.6 | 65.7 | 72.4 | 90.4 |
| (a) | SAM2 features only (SAM2$^\dagger$) | 77.5 | 40.9 | 67.0 | 87.3 |
| (b) | MUSt3R features only | 84.6 | 52.7 | 61.3 | 84.9 |
| (c) | No routing, trained fixed fusion | 89.1 | 59.6 | 68.5 | 86.9 |
| (d) | Cross-attention merger, +103M parameters | 91.2 | 66.8 | 72.5 | 90.5 |
| (e) | Static data only (ScanNet++, ASE) | 92.3 | 68.5 | 73.6 | 80.2 |
| (f) | Dynamic data only (MOSE, MOSEv2, SA-V) | 76.6 | 45.7 | 64.3 | 92.8 |
| (g) | Consecutive-frame sampling | 89.6 | 63.8 | 70.1 | 90.5 |

Table 7: **Ablations** on the ScanNet++ Selected Subset (%) and DAVIS17 ($\mathcal{J}\&\mathcal{F}$): feature streams (a, b), routing (c), fusion design (d), training data (e, f), and clip sampling for static scenes (g).

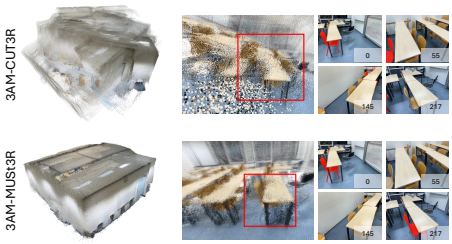


Figure 7: **CUT3R vs. MUSt3R geometry.** CUT3R (top) places the same table at inconsistent 3D locations; MUSt3R (bottom) keeps objects aligned, which its features inherit.

### 4.5 Ablation Study

**Feature Streams and Fusion Design.** Both streams are required (Table 7a, b). SAM2 features alone reach 40.9 Recall, lacking a geometric anchor under new viewpoints, and MUSt3R features alone localize the object (52.7 Recall) but give coarser masks (61.3 versus 72.4 Accuracy) without SAM2's appearance detail. A hierarchical cross-attention merger (d) gains at most 1.1 points while adding 103M parameters and attention over all positions per source; the lightweight fusion suffices because the two feature maps are spatially aligned, and its scalar gates are what routing switches.

**Training Data and Sampling.** Static data only (e) trusts geometry everywhere, the best variant on ScanNet++ (68.5 Recall) but 10.2 points worse on DAVIS17, and dynamic data only (f) never opens the geometric path, the best on DAVIS17 (92.8) but at 45.7 Recall on ScanNet++, so routing needs both domains to learn when geometry is reliable. Consecutive frames (g) give fewer reappearances; anchor-then-contiguous sampling adds 1.9 Recall and 2.3 Accuracy at no cost on DAVIS17.

**3D Backbone.** As 3D feature sources, VGGT and $\pi^3$ attend over all frames jointly and cannot run online on long sequences, and the online CUT3R drifts (Figure 7) and gives features too coarse for re-identification, reaching 27.5 Recall against 65.7 with MUSt3R. MUSt3R keeps objects aligned, and its frozen features are the strongest correspondence source among online models (Appendix D).

**Efficiency.** 3AM fine-tunes 22.455M parameters and runs at 10.2 FPS with 4.3 GB of peak memory, faster and lighter than SAM2Long (Table 4); MUSt3R dominates the extra time, and the cost stays bounded on long videos as SAM2's memory bank is fixed and MUSt3R's memory, which otherwise grows with the sequence, discards old entries at inference beyond a fixed budget.

### 4.6 Lifting Consistent Tracks to 3D

As a final check that the identity preserved in 2D is consistent in 3D, we propagate SAM2 automatic-mask proposals from keyframes with 3AM, associate tracks by temporal overlap, and project them into 3D with ground-truth depth and poses instead of MUSt3R's geometry, since the annotations are defined on the ground-truth 3D scans. Lifting methods must merge per-view masks in 3D because the masks are inconsistent, whether from 2D proposals (Jung et al., 2025; Zhao et al., 2025) or 3D detectors (Boudjoghra et al., 2025; Nguyen et al., 2024), whereas consistent tracks need little merging. Under the class-agnostic ScanNet200 protocol (Table 6), the projected tracks reach 47.3 AP, on par with or above offline (Xu et al., 2025b; Nguyen et al., 2024; Yin et al., 2024; Zhao et al., 2025) and online (Yang et al., 2023; Xu et al., 2025c) methods that merge in 3D.

## 5 Conclusion

We presented 3AM, a promptable tracker that preserves object identity through reappearance and large viewpoint changes while running online on RGB alone, without poses, depth, or offline reconstruction. Coupling SAM2 with the online reconstruction model MUSt3R supplies the location cue that appearance matching lacks, and because this cue holds only while the object stays in place, routing tokens in SAM2's memory attention decide per frame how much the fusion relies on geometry. One model thus re-identifies static objects from new viewpoints, follows objects that move, transfers zero-shot to ego-exo correspondence, and produces tracks that stay consistent in 3D.

### AI USE STATEMENT

In this work, we used generative AI tools for editing and polishing the writing of this paper and for assistance with implementing and debugging experimental code. We have not used generative AI tools for generating research ideas, designing experiments, or producing or interpreting results. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses publicly available datasets under their respective licenses (ScanNet++, Replica, ASE, MOSE, MOSEv2, SA-V, DAVIS, LaSOT, KITTI-MOTS, EPIC-KITCHENS, Ego-Exo4D, and ScanNet200). Ego-Exo4D and EPIC-KITCHENS contain recordings of people; we use only the released, consented data and do not attempt to identify individuals. The method is a general-purpose object tracker and shares the dual-use considerations of prior VOS work.

### REPRODUCIBILITY STATEMENT

Architecture and training details are given in Section 3 and Appendix A, including the selected MUSt3R layers, the fusion and routing dimensions, optimizer settings, and the clip sampling policy. The construction of the ScanNet++ selected subset, the evaluation protocols for all benchmarks, and the class-agnostic 3D instance segmentation pipeline are described in Appendix B and Appendix E. Code, the trained model, and evaluation scripts will be released.

## REFERENCES

Mohamad Alansari, Muzammal Naseer, Hasan Al Marzouqi, Naoufel Werghi, and Sajid Javed. Rethinking memory design in SAM-based visual object tracking. *arXiv preprint arXiv:2512.22624*, 2025.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, pp. 247–263. Springer, 2024.

Görkay Aydemir, Xiongyi Cai, Weidi Xie, and Fatma Guney. Track-On: Transformer-based online point tracking with memory. In *International Conference on Learning Representations*, 2025.

Görkay Aydemir, Weidi Xie, and Fatma Güney. Track-on2: Enhancing online point tracking with memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yash Bhalgat, Iro Laina, João F. Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3D object instance segmentation by slow-fast contrastive fusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9092–9116, 2023.

Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *European Conference on Computer Vision*, pp. 777–794. Springer, 2020.

Mohamed El Amine Boudjoghra, Angela Dai, Jean Lahoud, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Open-YOLO 3d: Towards fast and accurate open-vocabulary 3d instance segmentation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3d reconstruction models. *arXiv preprint arXiv:2512.11508*, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 1050–1060, 2025.

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 221–230, 2017.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Jacob Chalk, Saptarshi Sinha, Dima Damen, Yannis Kalantidis, and Diane Larlus. Whareformer: Learning to track what is where in long egocentric videos. In *European Conference on Computer Vision*, pp. 473–491. Springer, 2026.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a.

Shih-Fang Chen, Jun-Cheng Chen, I-Hong Jhuo, and Yen-Yu Lin. GOT-Edit: Geometry-aware generic object tracking via online model editing. In *International Conference on Learning Representations*, 2026b.

Xiang Chen. StreamDAM: Presence-aware memory for real-time streaming video object segmentation. *arXiv preprint arXiv:2608.03912*, 2026.

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*, pp. 9158–9168, 2025a.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026c.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6348–6361. IEEE, 2025b.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in neural information processing systems*, 34:11781–11794, 2021.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326. IEEE, 2023.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3151–3161, 2024.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3395–3405, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13745–13758, 2022.

Pablo Diaz-Pereda, Alejandro Rodriguez-Ramos, David Perez-Saura, and Pascual Campoy. Remind: Re-identification with memory for indoor navigation. *arXiv preprint arXiv:2607.09267*, 2026.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023.

Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. Mosev2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10027–10038. IEEE, 2023.

Zhipeng Du, Duolikun Danier, Jan Eric Lenssen, and Hakan Bilen. MoonSeg3R: Monocular online zero-shot segment anything in 3D with reconstructive foundation priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7304–7314, 2026.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806. IEEE, 2024.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5374–5383, 2019.

Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. ObjectRelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6530–6540, 2025.

Anna Gelencsér-Horváth, Gergely Dinya, Dorka Boglárka Erős, Péter Halász, Islam Muhammad Muqsit, and Kristóf Karacs. Scenevggt: Vggt-based online 3d semantic slam for indoor scene understanding and navigation. *arXiv preprint arXiv:2602.15899*, 2026.

Jiangshan Gong, Yuqun Wu, Qiqian Fu, Yao Xiao, Chuhang Zou, Shenlong Wang, and Derek Hoiem. SAM-V: Geometry-aware segment anything for multi-view instance segmentation. *arXiv preprint arXiv:2609.25490*, 2026.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2024.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13480–13492, 2023.

Yijun Hu, Bing Fan, Xin Gu, Haiqing Ren, Dongfang Liu, Heng Fan, and Libo Zhang. Robust ego-exo correspondence with long-term memory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 24047–24071, 2025.

Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3406–3416, June 2025.

Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3564–3574, 2024.

Rohit Jayanti, Swayam Agrawal, Vansh Garg, Siddharth Tourani, Muhammad Haris Khan, Sourav Garg, and Madhava Krishna. SegMASt3r: Geometry grounded segment matching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DI2AAFnLrc.

Minsik Jeon, Jay Karhade, Deva Ramanan, and Shubham Tulsiani. Point4D: Long-range 4D motion reconstruction. *arXiv preprint arXiv:2609.09145*, 2026.

Yoonwoo Jeong, Cheng Sun, Yu-Chiang Frank Wang, Minsu Cho, and Jaesung Choe. Mv-sam: Multi-view promptable segmentation using pointmap guidance. *arXiv preprint arXiv:2601.17866*, 2026.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pp. 289–310. Springer, 2024.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pp. 4867–4876, 2020.

Sanghun Jung, Jingjing Zheng, Ke Zhang, Nan Qiao, Albert YC Chen, Lu Xia, Chi Liu, Yuyin Sun, Xiao Zeng, Hsiang-Wei Huang, et al. Details matter for indoor open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9637, 2025.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. MapAnything: Universal feed-forward metric 3D reconstruction. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19729–19739, 2023.

JeongRae Kim and Changwon Lim. SAM2Dual: Training-free, dual memory for long-term video object segmentation. *arXiv preprint arXiv:2608.18640*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20943–20953, 2024.

Matej Kristan, Jiří Matas, Martin Danelljan, Michael Felsberg, Hyung Jin Chang, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Zhongqun Zhang, Khanh-Tung Tran, et al. The first visual object tracking segmentation VOTS2023 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1796–1818, October 2023.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *International Conference on Learning Representations*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, and Ziwei Liu. IGGT: Instance-grounded geometry transformer for semantic 3D reconstruction. In *International Conference on Learning Representations*, 2026.

Minghan Li, Shuai Li, Xindong Zhang, and Lei Zhang. Univs: Unified and universal video segmentation with prompts as queries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3227–3238, 2024a.

Wanyun Li, Pinxue Guo, Xinyu Zhou, Lingyi Hong, Yangji He, Xiangyu Zheng, Wei Zhang, and Wenqiang Zhang. Onevos: unifying video object segmentation with all-in-one transformer framework. In *European Conference on Computer Vision*, pp. 20–40. Springer, 2024b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Qin Liu, Jianfeng Wang, Zhengyuan Yang, Linjie Li, Kevin Lin, Marc Niethammer, and Lijuan Wang. Livos: Light video object segmentation with gated linear matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8668–8678, 2025a.

Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *European Conference on Computer Vision*, pp. 468–486. Springer, 2022.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16651–16662. IEEE, 2025b.

Jiahao Lu, Weitao Xiong, Jiacheng Deng, Peng Li, Tianyu Huang, Zhiyang Dou, Cheng Lin, Sai-Kit Yeung, and Yuan Liu. TrackingWorld: World-centric monocular 3D tracking of almost all pixels. In *Advances in Neural Information Processing Systems*, volume 38, pp. 141339–141359, 2025.

Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4World: Feedforward world-centric dense 3D tracking of all pixels. In *European Conference on Computer Vision*, pp. 184–204. Springer, 2026.

Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Performance evaluation methodology for long-term single-object tracking. *IEEE Transactions on Cybernetics*, 51 (12):6305–6318, 2021.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D reconstruction via conditional querying anytime and anywhere. In *International Conference on Machine Learning*, 2026.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129839–129867, 2025.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024.

K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018.

Lorenzo Mur-Labadia, Maria Santos-Villafranca, Jesus Bermudez-Cameo, Alejandro Perez-Yus, Ruben Martinez-Cantin, and Jose J. Guerrero. O-MaMa: Learning object mask matching between egocentric and exocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6892–6903, 2025.

Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16695–16705, 2025.

Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. DELTA: Dense efficient long-range 3D tracking for any video. In *International Conference on Learning Representations*, 2025.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4018–4028, 2024.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9226–9235, 2019.

Jiancheng Pan, Runze Wang, Tianwen Qian, Mohammad Mahdi, Yanwei Fu, Xiangyang Xue, Xiaomeng Huang, Luc Van Gool, Danda Pani Paudel, and Yuqian Fu. $V^2$-SAM: Marrying SAM2 with multi-prompt experts for cross-view object correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16910–16919, 2026.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria Digital Twin: A new benchmark dataset for egocentric 3D machine perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20076–20086. IEEE, 2023.

Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1352–1361, 2022.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

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*, pp. 724–732, 2016.

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:1704.00675*, 2017.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

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. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. *ICML*, 2023.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. In *European Conference on Computer Vision*, pp. 456–472. Springer, 2024.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9043–9052. IEEE, 2023.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2393–2401, 2023.

Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 36:75716–75739, 2023.

Peilin Tao, Chong Cheng, Yuansen Du, Caiwei Song, Zhengqing Chen, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, Hainan Cui, and Shuhan Shen. Anchor3R: Streaming 3D reconstruction with transient anchors for long-horizon visual mapping. *arXiv preprint arXiv:2606.05035*, 2026.

Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10840–10849. IEEE, 2021.

Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *European Conference on Computer Vision*, pp. 692–707. Springer, 2018.

Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24255–24264, 2025.

Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9481–9490, 2019a.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7934–7943, 2019b.

Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2708–2717, 2022.

Hanshi Wang, Zijian Cai, Jin Gao, Yiwei Zhang, Weiming Hu, Ke Wang, and Zhipeng Zhang. Online segment any 3D thing as instance tracking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49076–49095, 2025a.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19795–19806, 2023.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20406–20417. IEEE, 2024.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: Advancing 3D point tracking with explicit camera motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6726–6737, 2025.

Jianyun Xu, Song Wang, Ziqian Ni, Chunyong Hu, Sheng Yang, Jianke Zhu, and Qiang Li. Sam4d: Segment anything in camera and lidar streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28535–28545. IEEE, 2025a.

Mutian Xu, Xingyilang Yin, Lingteng Qiu, Yang Liu, Xin Tong, and Xiaoguang Han. Sampro3d: Locating sam prompts in 3d for zero-shot instance segmentation. In *2025 International Conference on 3D Vision (3DV)*, pp. 1222–1232. IEEE, 2025b.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2946–2954, 2022.

Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam: Online segment any 3d thing in real time. In *International Conference on Learning Representations*, volume 2025, pp. 38020–38035, 2025c.

Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28274–28284. IEEE, 2024.

Shannan Yan, Leqi Zheng, Keyu Lv, Jingchen Ni, Hongyang Wei, Jiajun Zhang, Guangting Wang, Jing Lyu, Chun Yuan, and Fengyun Rao. Learning cross-view object correspondence via cycle-consistent mask prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6653–6663, 2026.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024a.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025a.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5188–5197, 2019.

Qiushi Yang, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mosam: Motion-guided segment anything model with spatial-temporal memory selection. *arXiv preprint arXiv:2505.00739*, 2025b.

Xitong Yang, Devansh Kukreja, Don Pinkus, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jia-Wei Liu, Nicolás Ugrinovic, Anushka Sagar, et al. Sam 3d body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7209–7219, 2026.

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.

Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021.

Zongxin Yang, Jiaxu Miao, Yunchao Wei, Wenguan Wang, Xiaohan Wang, and Yi Yang. Scalable video object segmentation with identification mechanism. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6247–6262, 2024b.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European conference on computer vision*, pp. 162–179. Springer, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3292–3302, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Bowei Zhang, Lei Ke, Adam W. Harley, and Katerina Fragkiadaki. TAPIP3D: Tracking any point in persistent 3D geometry. In *Advances in Neural Information Processing Systems*, volume 38, pp. 135284–135303, 2025a.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, 2026a.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023.

Jiaming Zhang, Yutao Cui, Gangshan Wu, and Limin Wang. Jointformer: A unified framework with joint modeling for video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Jiaming Zhang, Cheng Liang, Yichun Yang, Chenkai Zeng, Yutao Cui, Xinwen Zhang, Xin Zhou, Kai Ma, Gangshan Wu, and Limin Wang. SAM 2++: Tracking anything at any granularity. *arXiv preprint arXiv:2510.18822*, 2025c.

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. *arXiv preprint arXiv:2410.03825*, 2024a.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.

Zhixiong Zhang, Shuangrui Ding, Xiaoyi Dong, Songxin He, Jianfan Lin, Junsong Tang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Advancing complex video object segmentation via progressive concept construction. In *International Conference on Learning Representations*, 2026b.

Jihuai Zhao, Junbao Zhuo, Jiansheng Chen, and Huimin Ma. Sam2object: Consolidating view consistency via sam2 for zero-shot 3d instance segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19325–19334, 2025.

Chong Zhou, Chenchen Zhu, Yunyang Xiong, Saksham Suri, Fanyi Xiao, Lemeng Wu, Raghuraman Krishnamoorthi, Bo Dai, Chen Change Loy, Vikas Chandra, et al. Edgetam: On-device track anything model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13832–13842, 2025.

Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18602–18611, 2024a.

Yikang Zhou, Tao Zhang, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Improving video segmentation via dynamic anchor queries. In *European Conference on Computer Vision*, pp. 446–463. Springer, 2024b.

Chenming Zhu, Peizhou Cao, Jingli Lin, Wenbo Hu, Yunlong Ran, Jiangmiao Pang, Tai Wang, and Xihui Liu. $G^2$TAM: Geometry grounded track anything model. In *International Conference on Machine Learning*, 2026a.

Deyi Zhu, Yuji Wang, Yong Liu, Yansong Tang, Bingyao Yu, Jiwen Lu, and Jie Zhou. Segment anything with motion, geometry, and semantic adaptation for complex nonlinear visual object tracking. *arXiv preprint arXiv:2605.22538*, 2026b.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, 2026.

Filippo Ziliotto, Luciano Serafini, Lamberto Ballan, and Tommaso Campari. What vggt knows about overlap: Probing geometric foundation models for co-visibility. *arXiv preprint arXiv:2607.09503*, 2026.

Zhengyu Zou, Hao Li, Kuixuan Jiao, Liu Liu, Tingyang Xiao, Xiaolin Zhou, Fangzhou Hong, Zhizhong Su, Dingwen Zhang, and Ziwei Liu. Iggt4d: Streaming 4d instance-grounded geometry transformer. *arXiv preprint arXiv:2607.19228*, 2026.

Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. Panst3r: Multi-view consistent panoptic segmentation. *arXiv preprint arXiv:2506.21348*, 2025.

# A Implementation Details

## A.1 SAM2 Backbone

All experiments use SAM2.1-L. A Hiera (Ryali et al., 2023) image encoder produces multi-scale features; the stride-16 and stride-32 outputs are fused by an FPN (Lin et al., 2017) into the frame embedding ($256\times64\times64$ for a $1024^2$ input), while the stride-4 and stride-8 features are injected only into the mask decoder to recover fine boundaries. Memory attention has four layers with sinusoidal absolute positional embeddings and 2D RoPE on the spatial tokens; object-pointer tokens do not use RoPE. The prompt encoder follows SAM, and the mask decoder produces the object pointer stored in the memory bank, an object-presence score, and several mask candidates, from which the one with the highest predicted IoU is selected when the prompt is ambiguous. The memory encoder projects features to 64 channels; the 256-dimensional object pointer is split into four 64-dimensional tokens for cross-attention. The memory bank holds the prompted frame and up to six recent frames, plus up to 16 object pointers. For multi-object videos, the image encoder is shared and each object keeps its own memory bank.

## A.2 Feature Fusion and Routing

**MUSt3R Features.** We use MUSt3R-512 with its own view-coverage based memory selection; since this memory otherwise grows with the sequence, at inference we discard old entries once it exceeds a fixed budget. From each frame we take the encoder output (1024 channels) and the outputs of decoder blocks 4, 7, and 11 (768 channels each), all at stride 16. Block 11 is the last decoder block, whose normalized output feeds the point-map head; we also read the head output (point map, confidence, and the camera rays derived from the estimated focal length) for the explicit geometry source.

**Projection Heads.** Each of the four MUSt3R features passes through a $1\times1$ convolution to 384 channels, a channel-wise LayerNorm, GELU, and a $3\times3$ convolution to 256 channels, followed by bilinear upsampling to $64 \times 64$. The explicit geometry source concatenates the point map, encoded with a sinusoidal band encoding of six frequencies (39 channels), and the Plücker coordinates of the camera rays (6 channels), and applies a patch embedding (kernel and stride 16) to 256 channels, followed by the same upsampling.

**Modulation MLPs.** For each geometric source $k = 1, \dots, 5$, a two-layer $\mathrm{MLP}_k$ ($256 \to 512 \to 513$, GELU) maps $r_k^i$ to the per-channel scale and shift $\gamma_k, \beta_k \in \mathbb{R}^{256}$ and the scalar gate $\alpha_k$. $\mathrm{MLP}_0$ ($256 \to 512 \to 512$) produces $\gamma_0, \beta_0$ for the SAM2 feature; its output layer is zero-initialized so that this term starts as the identity.

**Routing Tokens.** There are six terms $k = 0, \dots, 5$ (the SAM2 feature, four MUSt3R layers, and the explicit geometry source) with two routing tokens each, $N = 12$ in total. Each token is a learned embedding plus a linear projection ($256 \to 256$) of $r_k^i$. Routing tokens are concatenated to the $HW = 4096$ spatial tokens and pass through the four memory-attention layers with the shared self-attention, cross-attention, and feed-forward weights; in cross-attention the rotary encoding is applied to the spatial queries only. After the last layer, the two tokens of term $k$ are concatenated and projected ($512 \to 256$) to $r_k^{i+1}$, used by frame $i{+}1$. This output projection, the initial states $r_k^0$, and the projection of the prompt object pointer ($256 \to 6 \times 256$) are zero-initialized, so training starts from $r_k^i = r_k^0$ with gates near zero. The prompted frame is fused with $r_k^0$ and its fused feature is stored in memory; every later frame is fused with the routing vectors produced by the previous frame, and the memory encoder stores the fused feature of that frame together with its predicted mask.

## A.3 Training Details

**Losses and Optimizer.** We follow the SAM2 loss design with focal loss (weight 20) and dice loss (weight 1) on the mask, an $\ell_1$ loss on the predicted IoU (weight 1), and a cross-entropy loss on object presence (weight 1). Trainable parameters (22.455M in total) are the Feature Fusion module, the routing tokens and projections, memory attention, the memory encoder, and the mask decoder; the image encoder, its high-resolution projection convolutions, and MUSt3R are frozen. We use AdamW with learning rates of $5\times10^{-6}$ (fusion and decoder), $10^{-5}$ (memory attention), and $10^{-6}$

(memory encoder), cosine decay to one tenth of the initial rate, gradient clipping at norm 1, bfloat16 autocast, and one clip of eight frames per iteration. Training runs on one RTX 4090; MUSt3R inference for the clip runs in worker processes on a second GPU.

**Prompts.** On ScanNet++ the prompt is always a mask, because the projected 2D masks are too noisy for reliable point sampling and the model should learn from the mask directly. On ASE the prompt is a point without iterative corrections. On MOSE, MOSEv2, and SA-V we keep SAM2's default protocol, with point, box, or mask prompts on the first frame and correction clicks on up to two frames.

**Clip Sampling for Static Scenes.** A training clip must contain a reappearance under a large viewpoint change, but SAM2's memory holds few frames, and frames sampled at random from a large object may show disjoint parts of it, contradicting the goal of locating the object by its 3D position. For ASE and ScanNet++ we therefore use anchor-then-contiguous sampling, in which the prompt frame is drawn from anywhere in the sequence and the remaining seven frames form a contiguous block elsewhere, before or after the anchor. The anchor supplies the viewpoint jump, and the block retains temporally coherent motion and matches the inference regime. Fully random frames would expose the model to more viewpoint variation, but the resulting model attends at inference to non-conditioning memories that are wrong, which degrades the results. The sequence is the list of keyframes for which masks and cached MUSt3R features exist (ScanNet++ frames at stride 10; ASE keyframes selected by co-visibility). The prompted frame is drawn uniformly among frames in which the target is visible with sufficient area, and the remaining seven frames form a block of consecutive keyframes drawn uniformly among valid block starts elsewhere in the sequence, before or after the anchor. The block is kept in forward temporal order and is not filtered for target visibility, so the target may be absent in part or all of the block, which supervises absence prediction. For MOSE, MOSEv2, and SA-V the eight frames are consecutive.

**MUSt3R Features During Training.** For the static datasets we precompute MUSt3R features by streaming the full keyframe sequence of every scene once and caching the selected layers, so that training sees the same long-term features as inference; contiguous frames also keep these features consistent, whereas frames far apart can be reconstructed poorly. Because at inference MUSt3R only sees the frames it is given, 20% of the ScanNet++ clips and 10% of the ASE clips are instead processed online over the eight sampled frames. Dynamic datasets are always processed online.

### A.4 DATASETS

Table 8: **Training data.**

| | ScanNet++ | ASE | MOSE | MOSEv2 | SA-V |
|---|---|---|---|---|---|
| Real | ✓ | ✗ | ✓ | ✓ | ✓ |
| Dynamic | ✗ | ✗ | ✓ | ✓ | ✓ |
| Clip sampling | Anchor + contiguous | Anchor + contiguous | Consecutive | Consecutive | Consecutive |
| Prompt | Mask | Point | Point / box / mask | Point / box / mask | Point / box / mask |
| MUSt3R features | Cached (20% online) | Cached (10% online) | Online | Online | Online |

**ScanNet++ (Yeshwanth et al., 2023)** is a high-quality indoor RGB-D dataset with accurate poses, dense trajectories, and 3D instance annotations. Its 2D masks are projections of the 3D labels, so they carry reprojection noise from depth and pose errors, and the pose annotation omits stretches of frames that the SfM pipeline discards. It nevertheless provides the realistic geometry and viewpoint diversity needed to learn view-consistent identification.

**Aria Synthetic Environments (ASE) (Avetisyan et al., 2024)** contains procedurally generated indoor scenes rendered with simulated Aria sensors, with clean 2D and 3D annotations. We use scenes whose number of views lies in a practical range, since too few views give little viewpoint variation and too many exceed the memory of MUSt3R.

**MOSE (Ding et al., 2023), MOSEv2 (Ding et al., 2025), and SA-V (Ravi et al., 2024)** are video segmentation datasets with heavy occlusion, disappearance and reappearance, small objects, and, in SA-V, a very large number of videos. They keep the appearance-based VOS ability of the model and, more importantly, supply moving objects, which the routing tokens need in order to learn when geometry is unreliable. They also prevent the model from hallucinating tracks from geometry alone.

# B Evaluation Details

## B.1 ScanNet++ and Replica

For each scene we project the 3D instance labels into every frame with the ground-truth poses and depth, select the frame with the largest visible mask of an instance as the prompt frame, and track forward and backward. Objects that are absent in the first frame are therefore still evaluated from a valid prompt. We construct the Selected Subset of ScanNet++ ourselves: a visible segment is a maximal run of consecutive frames in which the instance is visible, and an instance is selected when more than eight of its visible segments last at least 12 frames. Table 9 summarizes both groups. Selected instances disappear more often, stay absent longer, and span most of the sequence, which stresses re-identification after long absences.

Table 9: **ScanNet++ Selected Subset statistics**, median (25th–75th percentile) per instance. The fraction of empty frames, in which the object is absent, is 0.6961 on the Selected Subset and 0.7871 on the Whole Set.

| Set | # disjoint visible segments with >12 consecutive frames | # disjoint visible segments | # absent frames | Frame coverage from first to last appearance |
|---|---|---|---|---|
| Selected ($N$=286) | 9 (8–10) | 20 (14–26) | 552 (384–696) | 0.948 (0.908–0.981) |
| Non-selected ($N$=3800) | 3 (2–5) | 8 (5–13) | 367 (227–545) | 0.787 (0.565–0.933) |

**Overlap with MUSt3R Training Data.** MUSt3R was trained on, among other datasets, a ScanNet++ split. MUSt3R is used as a frozen feature extractor only, none of our validation scenes appear in MUSt3R's training split or in our own training data, and our evaluation asks for online promptable tracking over thousand-frame videos rather than the short multi-view tuples MUSt3R was trained on.

**Multiple Objects.** As in SAM2, the image encoder, MUSt3R, and the Feature Fusion inputs are shared across objects in the same video; each object keeps its own memory bank and its own routing state, so the per-object cost of 3AM is the cost of memory attention and the mask decoder only.

## B.2 Viewpoint Analysis

For the breakdown in Figure 8, both trackers are initialized with the same maximum-area ground-truth mask and track all RGB frames in both directions; the prompt frame is excluded from the scores. Camera translation is the metric distance between the COLMAP camera centres of the current and the prompt frame, and rotation is the geodesic angle between their orientations; both describe camera-pose changes rather than object-centric viewing angles. Recall in a bin is the mean IoU over GT-positive frames in that bin, averaged within an object and then over objects; the number of objects with frames in a bin ranges from 285 (0–15°) to 199 (120–180°) and from 285 (0–0.25 m) to 128 ($\geq$ 4 m). Confidence intervals come from 2,000 scene-cluster bootstrap resamples over the 33 scenes. Table 10 lists the values plotted in Figure 8. 3AM keeps 65% of its near-view Recall beyond 120° of rotation and 51% beyond 4 m of translation, versus 16% and 8% for SAM2; its gain over SAM2 grows from 15 to 40 points over rotation and from 7 to 42 points over translation, and within objects the far-minus-near gain is 18.6 points for rotation ($\geq$ 60° versus $<$ 30°) and 31.6 points for translation ($\geq$ 1 m versus $<$ 0.25 m), with confidence intervals excluding zero.

## B.3 Baseline Protocols

**CoTracker3+SAM2 and OmniMotion+SAM2.** Query points are sampled inside the prompt mask and tracked through the video; in every frame, the tracked points predicted as visible are given to SAM2 as point prompts, and the resulting mask is scored. OmniMotion is optimized per video before tracking. **PanSt3R.** The sequence is processed offline; the predicted panoptic instance with the largest overlap with the prompt mask is used as the track. **SegMASt3R.** For every frame, segments are matched to the prompt frame by two-view matching, and the segment matched to the prompt mask is the prediction. **EmbodiedSAM.** Online 3D instance segmentation is run with the ground-truth depth and poses; the 3D instance overlapping the prompt is projected back into every frame. **MUSt3R-Thr+SAM2.** The cosine similarity between the MUSt3R features of the current frame and the mean feature of the prompt mask is thresholded and passed to SAM2 as a mask prompt. **MUSt3R-Reproj+SAM2.** The prompt mask is reprojected into every frame through MUSt3R's point maps with a visibility check, and the reprojected mask prompts SAM2. **SAM 3.** SAM 3 is prompted with the same mask on the prompt frame as SAM2. **Unstructured collections.** When

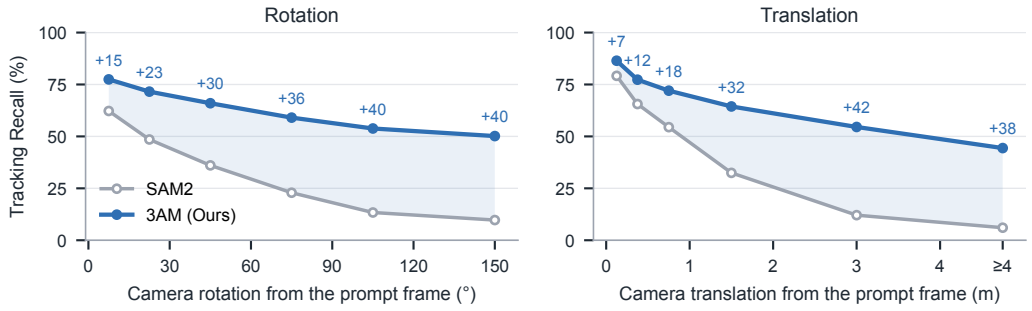


Figure 8: **Tracking Recall versus camera-pose change from the prompt frame** on the 286 objects of the ScanNet++ Selected Subset. Frames are binned by the rotation (left) or translation (right) of the camera relative to the prompt frame; each bin averages frames within an object and then objects. Numbers above the curve give the gain of 3AM over SAM2 in points; all gains have scene-bootstrap 95% confidence intervals excluding zero (Appendix B.2).

Table 10: **Tracking Recall by camera-pose change** on the ScanNet++ Selected Subset (%), with the gain of 3AM over SAM2 in points and its scene-bootstrap 95% confidence interval.

| Bin | Objects | SAM2 | 3AM | Gain [95% CI] |
|---|---|---|---|---|
| *Camera rotation from the prompt frame* | | | | |
| 0–15° | 286 | 62.2 | 77.5 | +15.2 [12.6, 18.1] |
| 15–30° | 285 | 48.5 | 71.6 | +23.0 [20.2, 26.1] |
| 30–60° | 286 | 36.1 | 66.0 | +29.9 [27.1, 32.3] |
| 60–90° | 284 | 22.9 | 59.0 | +36.2 [31.8, 39.6] |
| 90–120° | 265 | 13.4 | 53.8 | +40.5 [35.3, 44.6] |
| 120–180° | 199 | 9.7 | 50.2 | +40.4 [32.9, 46.4] |
| *Camera translation from the prompt frame* | | | | |
| 0–0.25 m | 285 | 79.2 | 86.4 | +7.3 [6.3, 8.2] |
| 0.25–0.5 m | 278 | 65.6 | 77.3 | +11.7 [9.5, 13.7] |
| 0.5–1 m | 282 | 54.5 | 72.0 | +17.6 [14.1, 20.7] |
| 1–2 m | 284 | 32.4 | 64.4 | +32.0 [26.3, 36.7] |
| 2–4 m | 278 | 12.1 | 54.5 | +42.5 [37.7, 46.2] |
| $\geq$ 4 m | 128 | 6.0 | 44.4 | +38.3 [29.3, 46.9] |
| *Within-object far-minus-near gain* | | | | |
| Rotation, $\geq 60°$ vs. $< 30°$ | 285 | near +20.1, far +38.7 | | +18.6 [14.0, 23.0] |
| Translation, $\geq$ 1 m vs. $<$ 0.25 m | 284 | near +7.3, far +38.9 | | +31.6 [26.6, 35.6] |

the input is an unordered photo collection, frames are reordered by SAM2 feature retrieval before tracking.

### B.4 Ego-Exo4D

We follow the correspondence benchmark of Ego-Exo4D (Grauman et al., 2024) as used by $V^2$-SAM (Pan et al., 2026). Each test sample consists of a query-view image with the ground-truth mask of one object and the synchronized target-view image. 3AM takes the query view as the prompted frame with the mask prompt and predicts the mask in the target view as the second frame; MUSt3R is initialized with the two images. Ego frames are undistorted to a pinhole model with the nominal Aria calibration.

### B.5 Dynamic Benchmarks

DAVIS17, LaSOT, and $\text{LaSOT}_{\text{ext}}$ are evaluated with their official protocols and the prompts used by DAM4SAM (Videnovic et al., 2025). For KITTI-MOTS (Voigtlaender et al., 2019b) and EPIC-KITCHENS (Darkhalil et al., 2022), each annotated object is prompted with its mask on the first annotated frame and tracked through the sequence; we report IoU, Tracking Recall, and Accuracy as defined in Section 4.1.

# C Additional Results

**Routing Analysis.** For every GT-positive, non-prompt frame we measure the magnitude of each routing coefficient, namely the channel-wise RMS of $\gamma_0$ for the SAM2 term and of $\alpha_k\gamma_k$ for each geometric term $k$, where $\gamma$ includes its identity offset. Magnitudes are normalized to shares within a frame, averaged over frames within an object, and then over objects. These are shares of coefficient magnitudes; they exclude the shifts $\beta_k$ and the feature amplitudes, and they do not measure the contribution of each term to the fused feature. Table 11 lists the per-term shares behind Table 3. Decoder blocks 4 and 11 carry the largest geometric shares on static data and almost none on dynamic data, and their signed gains ($\alpha_k\gamma_k$, mean $\pm$ SD over objects) change accordingly, block 4 from $0.48 \pm 0.03$ to $-0.04 \pm 0.04$ and block 11 from $0.47 \pm 0.03$ to $0.03 \pm 0.02$, whereas the encoder output keeps a similar gain on both ($0.39 \pm 0.05$ and $0.42 \pm 0.10$). Decoder block 7 and the explicit geometry receive similar shares on static and dynamic data (8.7 and 10.8%, 0.3 and 0.5%); even so, we found that removing these two sources lowers performance, so both are kept.

Table 11: **Per-term routing shares** (%) on the ScanNet++ Selected Subset (static, 286 objects) and DAVIS17 (dynamic, 61 objects), with scene-bootstrap 95% confidence intervals for the two groups.

| Term | Static | Dynamic |
|---|---|---|
| SAM2 feature ($\gamma_0$) | 32.1 | 54.3 |
| MUSt3R encoder output | 17.2 | 28.0 |
| MUSt3R decoder block 4 | 20.9 | 4.2 |
| MUSt3R decoder block 7 | 8.7 | 10.8 |
| MUSt3R decoder block 11 | 20.8 | 2.4 |
| Explicit geometry (point map and rays) | 0.3 | 0.5 |
| Appearance terms (SAM2, encoder) | 49.3 | 82.2 |
| Geometric terms (blocks 4, 7, 11, explicit) | 50.7 [50.4, 51.0] | 17.8 [16.8, 18.9] |

**Memory Selection.** 3AM uses SAM2's original memory selection. Table 12 combines 3AM with the memory-selection policies of DAM4SAM and SAM2Long on the ScanNet++ Selected Subset. Both policies change the results by less than one point in either direction, far less than the gain of 3AM over the same policies applied to SAM2 (Table 1), which shows that the improvement comes from the features rather than from memory management. Memory selection tailored to geometry-aware memories is an interesting direction for future work.

Table 12: **Memory-selection policies on top of 3AM**, ScanNet++ Selected Subset (%).

| Method | IoU | Recall | Acc. |
|---|---|---|---|
| 3AM (SAM2 memory selection) | 90.6 | 65.7 | 72.4 |
| 3AM + DAM4SAM (Videnovic et al., 2025) policy | 90.4 | 66.4 | 72.0 |
| 3AM + SAM2Long (Ding et al., 2024) policy | 90.8 | 65.1 | 73.6 |

**Prompt Type and Unordered Input.** Table 13 evaluates mask, point, box, and point-plus-box prompts, and unordered photo collections, on the Selected Subset. 3AM keeps its advantage for every prompt type, and the drop on unordered collections is small, because MUSt3R's memory does not require temporal continuity.

Table 13: **Prompt types and unordered input**, ScanNet++ Selected Subset (IoU / Tracking Recall / Accuracy, %). For unordered photo collections the frames are shuffled and reordered by SAM2 feature retrieval before tracking.

| Prompt / input | SAM2 | DAM4SAM | 3AM |
|---|---|---|---|
| Mask | 74.3 / 28.2 / 66.6 | 76.6 / 37.5 / 63.8 | 90.6 / 65.7 / 72.4 |
| Mask, unordered collection | – | 71.7 / 32.1 / 58.6 | 90.1 / 65.2 / 72.1 |
| Point | – | 75.1 / 35.1 / 61.4 | 88.7 / 64.2 / 71.4 |
| Box | – | 74.3 / 34.7 / 60.7 | 87.6 / 63.3 / 70.9 |
| Point + box | – | 74.6 / 34.4 / 61.1 | 88.0 / 62.6 / 71.0 |

**Repeated Trajectories.** To test long-range stability, each Selected Subset sequence is repeated one, two, or four times so that the object reappears again and again from the same viewpoints. Table 14 shows that 3AM stays stable across cycles and DAM4SAM does not improve with repetition, whereas SAM2 gains from the repeated views (39.0 Recall at two cycles) but remains far below 3AM.

Table 14: **Repeated trajectories**, ScanNet++ Selected Subset (IoU / Tracking Recall / Accuracy, %).

| Cycles | SAM2 | DAM4SAM | 3AM |
|---|---|---|---|
| 1 | 74.3 / 28.2 / 66.6 | 76.6 / 37.5 / 63.8 | 90.6 / 65.7 / 72.4 |
| 2 | 84.5 / 39.0 / 80.2 | 75.5 / 35.7 / 61.5 | 91.8 / 67.9 / 74.0 |
| 4 | 83.2 / 36.6 / 79.5 | 73.9 / 34.4 / 61.2 | 91.4 / 67.1 / 74.5 |

# D MUSt3R Compared with Other Reconstruction Backbones as Feature Sources

Table 15 compares reconstruction backbones as the feature source of 3AM (Section 4.5); VGGT and $\pi^3$ cannot process the evaluation sequences online, and CUT3R features give low Recall.

Table 15: **3D reconstruction backbones** as the feature source of 3AM. Frame Num.: frames processable in one sequence; Tracking Recall on the ScanNet++ Selected Subset. Best in red.

| Backbone | Online | Frame Num. | Recall |
|---|---|---|---|
| VGGT (Wang et al., 2025b) | ✗ | ∼200 | – |
| $\pi^3$ (Wang et al., 2026) | ✗ | ∼300 | – |
| CUT3R (Wang et al., 2025c) | ✓ | >10,000 | 27.5 |
| MUSt3R (Cabon et al., 2025) | ✓ | >10,000 | 65.7 |

Section 3.1 analyzes MUSt3R across layers. Here we ask whether another reconstruction backbone would be a better feature source, using the frozen features of each model without fine-tuning, which is how 3AM consumes them. We compare MUSt3R-512 with LingBot-Map (Chen et al., 2026a), a recent streaming model with a compact geometric context, and with the offline VGGT-1B (Wang et al., 2025b), on ASE (6 scenes, 16 keyframes each), ScanNet++ (8 scenes, 16 frames within a 64-frame window), and MOSE (8 dynamic videos, 12 frames each). Online models stream over the same capped window; VGGT sees the 16 evaluation frames jointly. Ground-truth correspondences are obtained by unprojecting the depth of one frame into the other with a depth-consistency check. For every model we report its best layer. Table 16 summarizes the results. MUSt3R gives the best dense correspondence (PCK and retrieval accuracy) on both static datasets and the most peaked similarity maps (lowest entropy) on ScanNet++, better than VGGT although VGGT sees all evaluation frames jointly and cannot be run online over a long stream; on ASE, VGGT's similarity maps are more peaked (0.751 versus 0.826 entropy) while MUSt3R still matches more points correctly. LingBot-Map's similarities are globally saturated, with cosine values above 0.9 almost everywhere and near-zero margins between the true match and the hardest negative. On the dynamic MOSE videos, MUSt3R keeps a clear contrast between an object and its background over time (cosine 0.88 within the object versus 0.34 to the background), whereas LingBot-Map hardly separates them; this contrast is what memory attention needs to track a specific object. VGGT is also best at instance-level discrimination on ASE and at metric geometry (Spearman correlation with 3D distance), which matches the layer-wise picture in Section 3.1, where correspondence and metric geometry are carried by different representations, and MUSt3R's middle layers give the correspondence 3AM relies on.

Table 16: **Frozen features of reconstruction backbones as correspondence sources.** PCK@5 is the fraction of dense matches within 5% of the image diagonal, Ret. the retrieval accuracy of the true match, Ent. the normalized entropy of the similarity map (lower is more peaked), Inst. the cross-frame instance nearest-neighbour accuracy, $\rho$ the magnitude of the Spearman correlation between feature similarity and 3D distance, Prop. the MOSE mask-propagation mIoU by feature nearest neighbours, and Contrast the cosine gap between same-object and object-to-background similarities over time. Best values in red.

| | | ASE | | | | | ScanNet++ | | | | | MOSE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Backbone | Online | PCK@5 | Ret. | Ent.↓ | Inst. | $\rho$ | PCK@5 | Ret. | Ent.↓ | Inst. | $\rho$ | Prop. | Contrast |
| MUSt3R (Cabon et al., 2025) | ✓ | 45.6 | 46.5 | 0.826 | 50.6 | 0.648 | 72.4 | 75.3 | 0.807 | 73.7 | 0.777 | 54.5 | 0.54 |
| LingBot-Map (Chen et al., 2026a) | ✓ | 43.0 | 39.3 | 0.971 | 58.3 | 0.568 | 54.8 | 55.8 | 0.935 | 69.0 | 0.605 | 50.1 | 0.02 |
| VGGT-1B (Wang et al., 2025b) | ✗ | 44.3 | 40.7 | 0.751 | 64.8 | 0.728 | 55.4 | 56.1 | 0.839 | 74.6 | 0.838 | 56.8 | 0.33 |

# E Class-Agnostic 3D Instance Segmentation

Given keyframes sampled at a fixed stride, we first generate 2D instance proposals with SAM2's automatic mask generator. For each keyframe, its proposals are propagated forward with 3AM to all later frames; no backward tracking is performed. The propagated masks are lifted into 3D by

back-projecting their pixels with the ground-truth depth and poses. Reprojection errors concentrate near object boundaries, so we compute a reprojection score from the agreement between projected and observed depth and erode the 2D masks slightly before projection. Per-frame 3D fragments are then merged into instances with two scores, a 3D overlap score in the point-cloud domain and a 2D temporal-overlap score along the propagated tracks. Because 3AM keeps the identity of an object across the video, an object seen earlier is recognized again when it reappears, which provides reliable temporal evidence for merging. Duplicate assignments are resolved by majority voting at the superpoint level, so that each superpoint is assigned to the instance in which it is observed most often.

## F LIMITATIONS

An object that is moved while it is outside the field of view, for example behind the camera, cannot be recovered immediately. Its geometric anchor points to the old location, no visual evidence links the old and the new location, and neither appearance nor geometry can re-establish the identity until the object is observed again.

## G ADDITIONAL QUALITATIVE RESULTS

Figures 9 to 16 show more video object tracking results on ScanNet++, and Figure 17 shows class-agnostic 3D instance segmentation results.

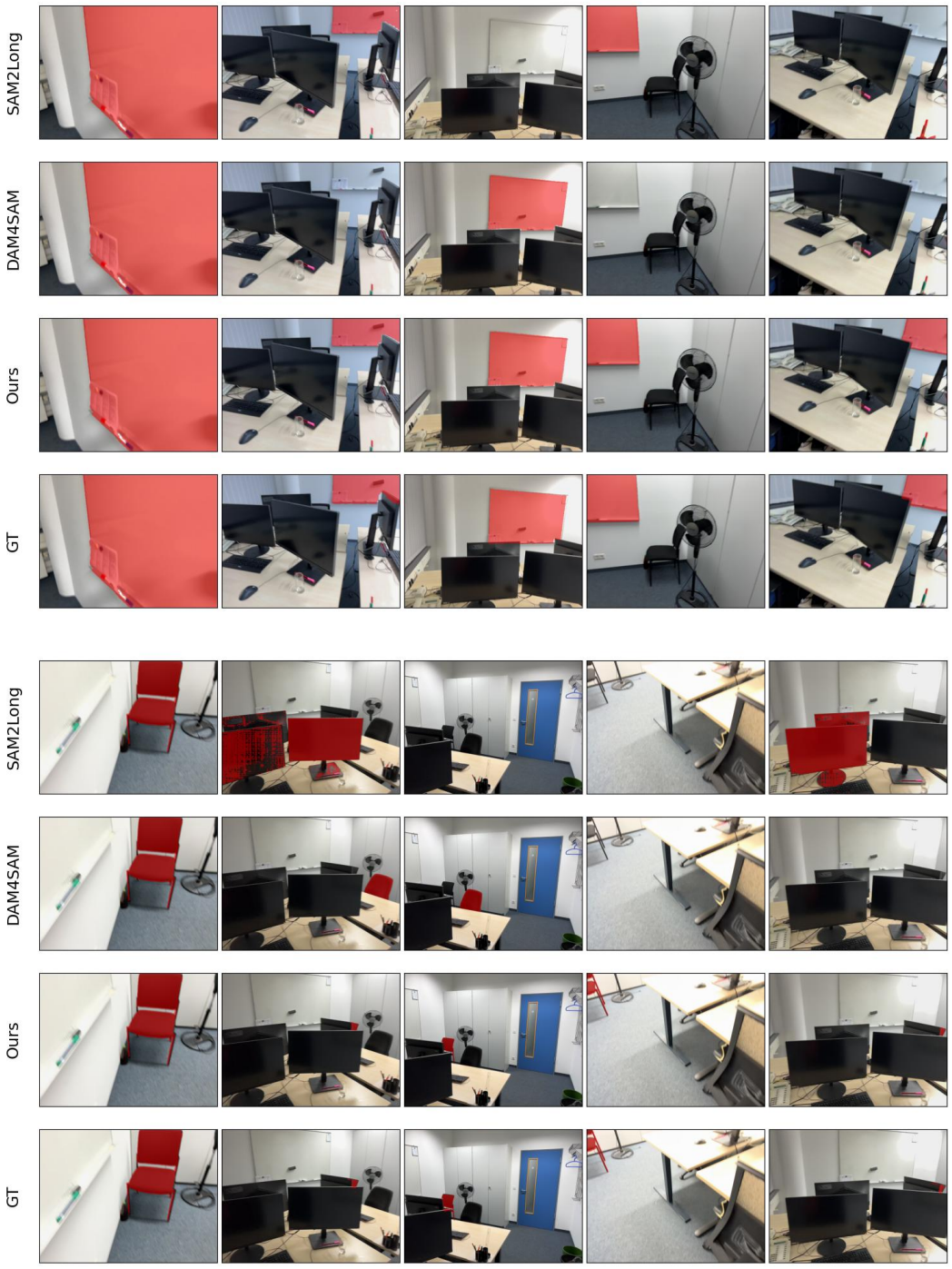


Figure 9: **Qualitative comparison of VOS methods** (scene 1ada7a0617).

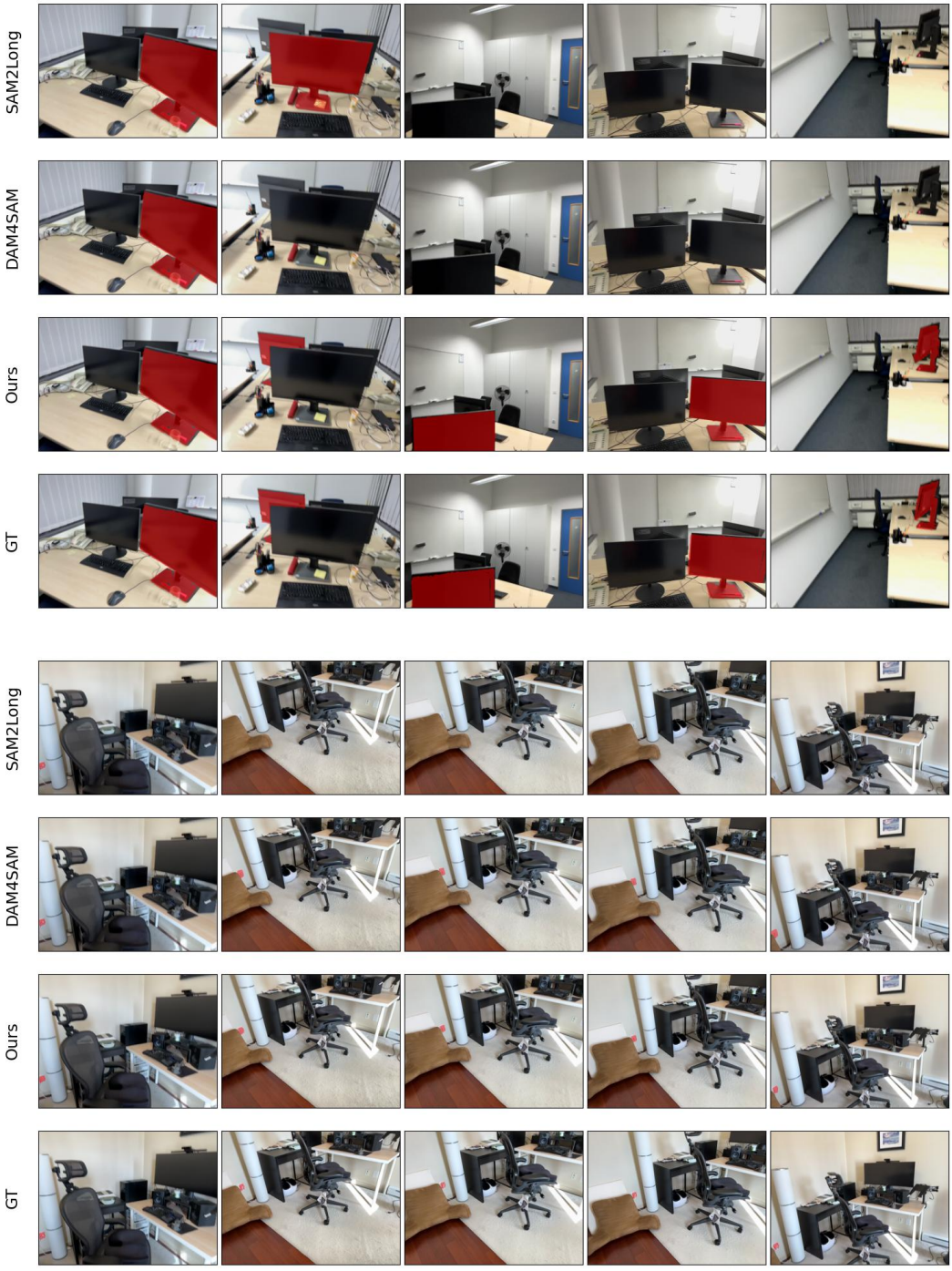


Figure 10: **Qualitative comparison of VOS methods** (scenes 1ada7a0617 and a24f64f7fb).

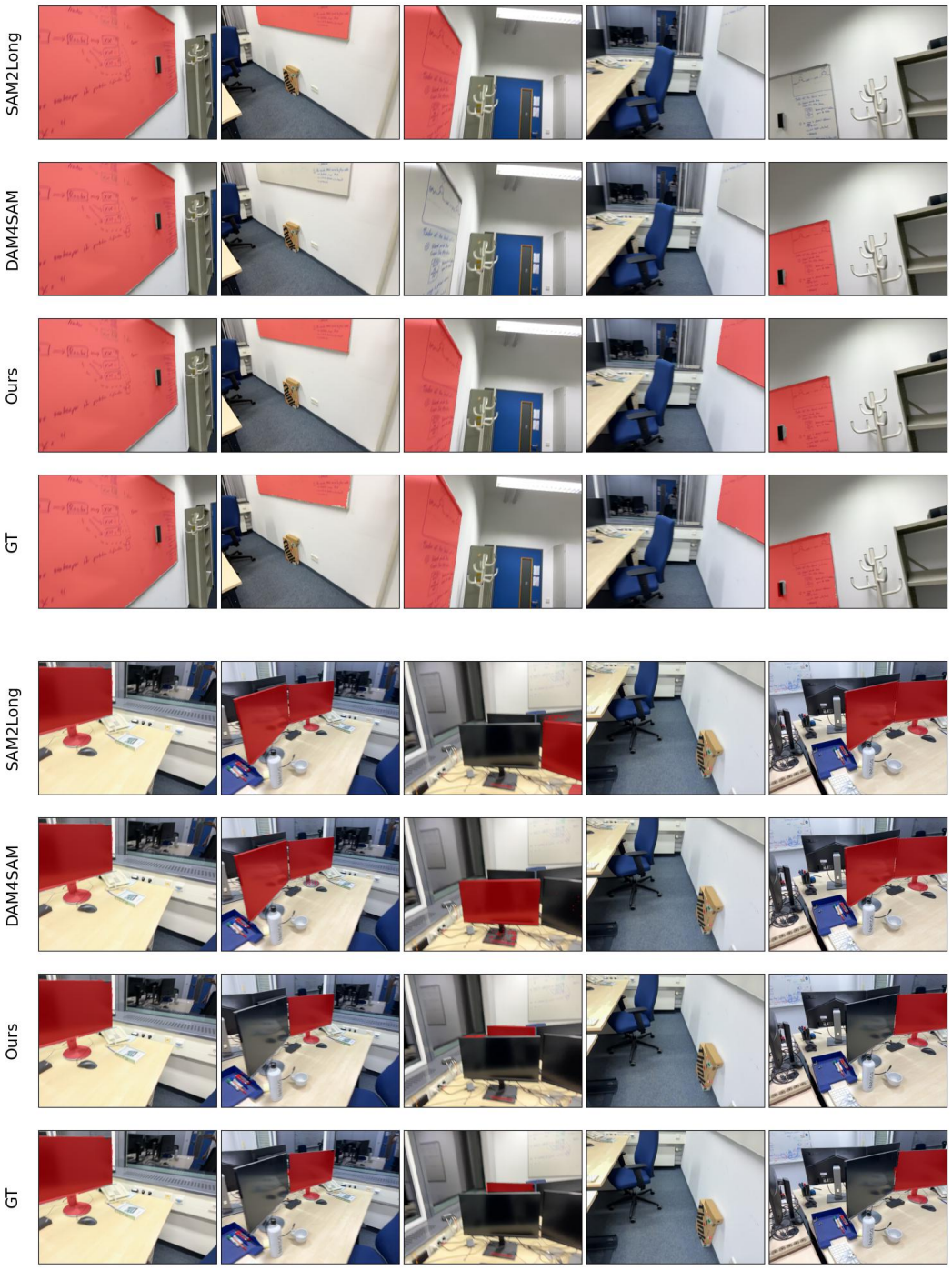


Figure 11: **Qualitative comparison of VOS methods** (scene 7b6477cb95).

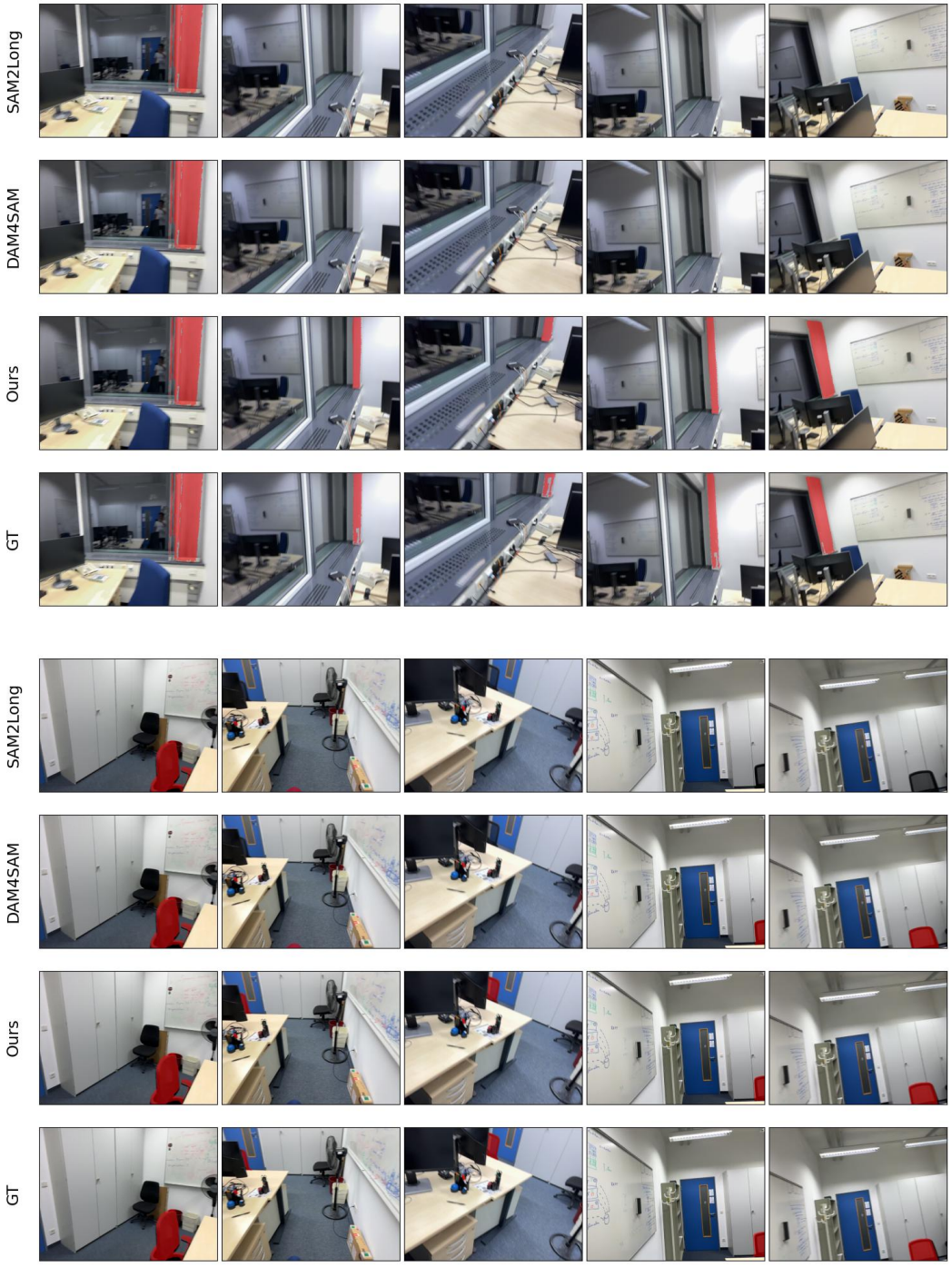


Figure 12: **Qualitative comparison of VOS methods** (scene 7b6477cb95).

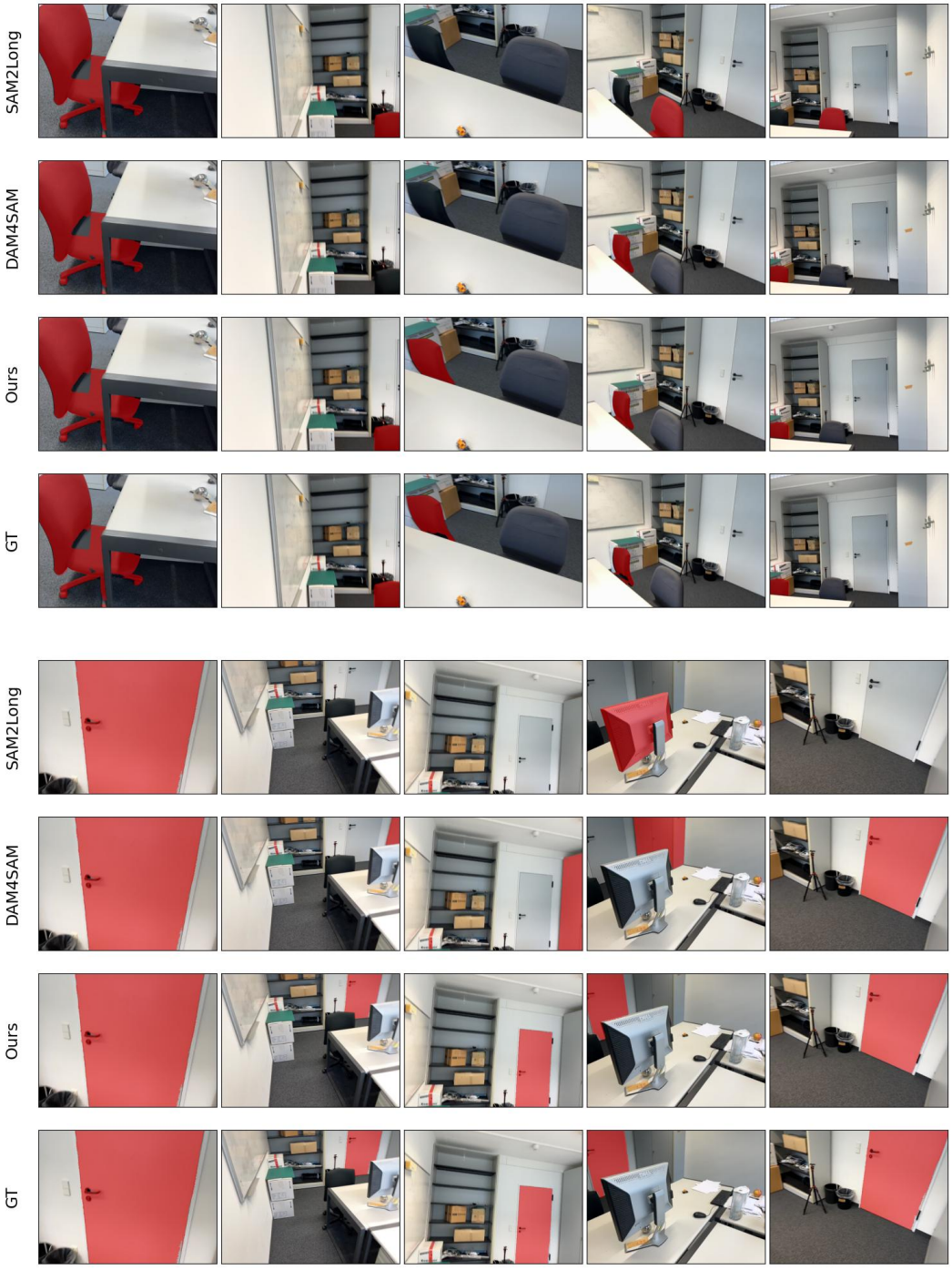


Figure 13: **Qualitative comparison of VOS methods** (scene acd95847c5).

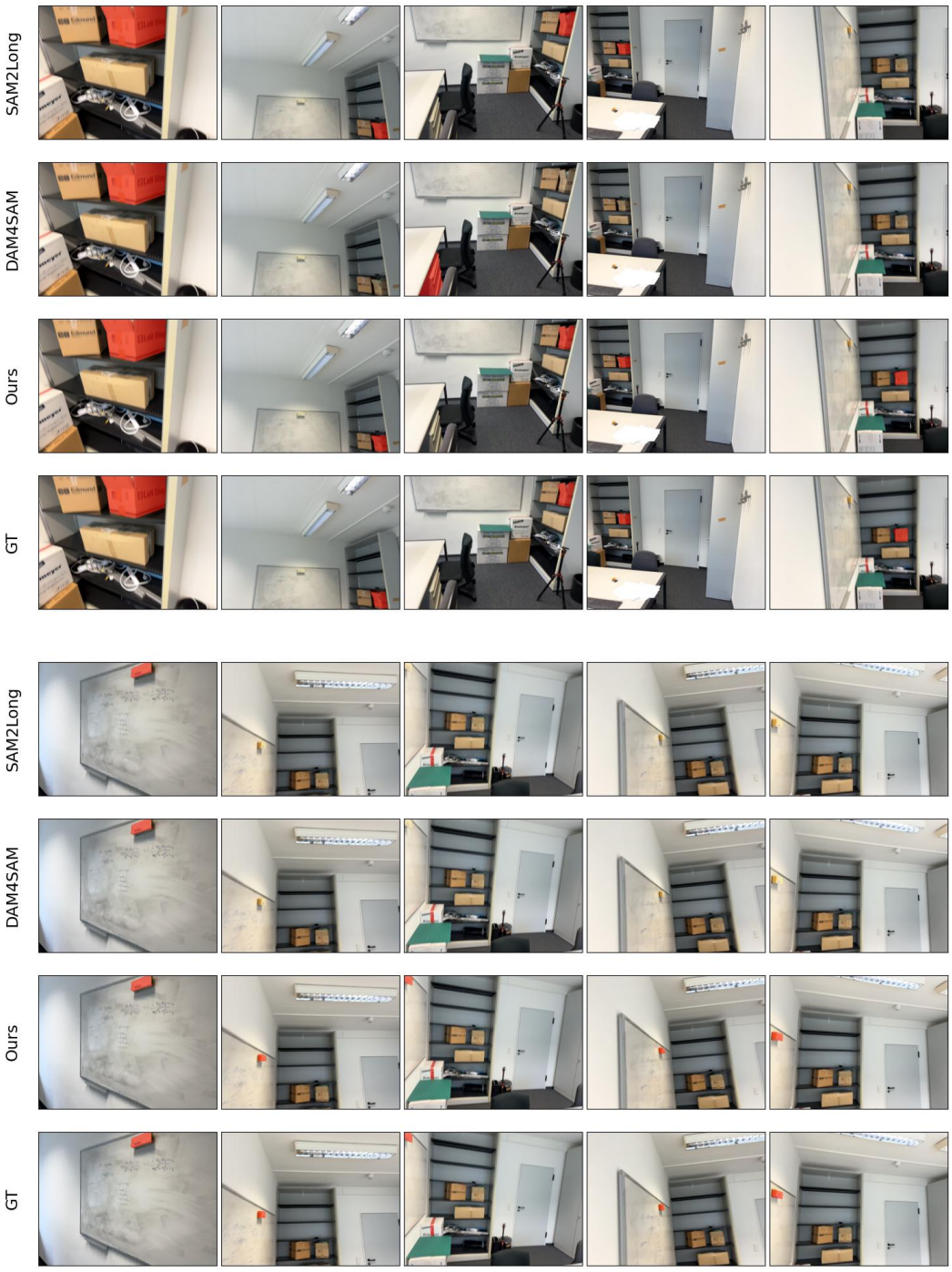


Figure 14: **Qualitative comparison of VOS methods** (scene acd95847c5).




Figure 15: **Qualitative comparison of VOS methods** (scenes acd95847c5 and c50d2d1d42).

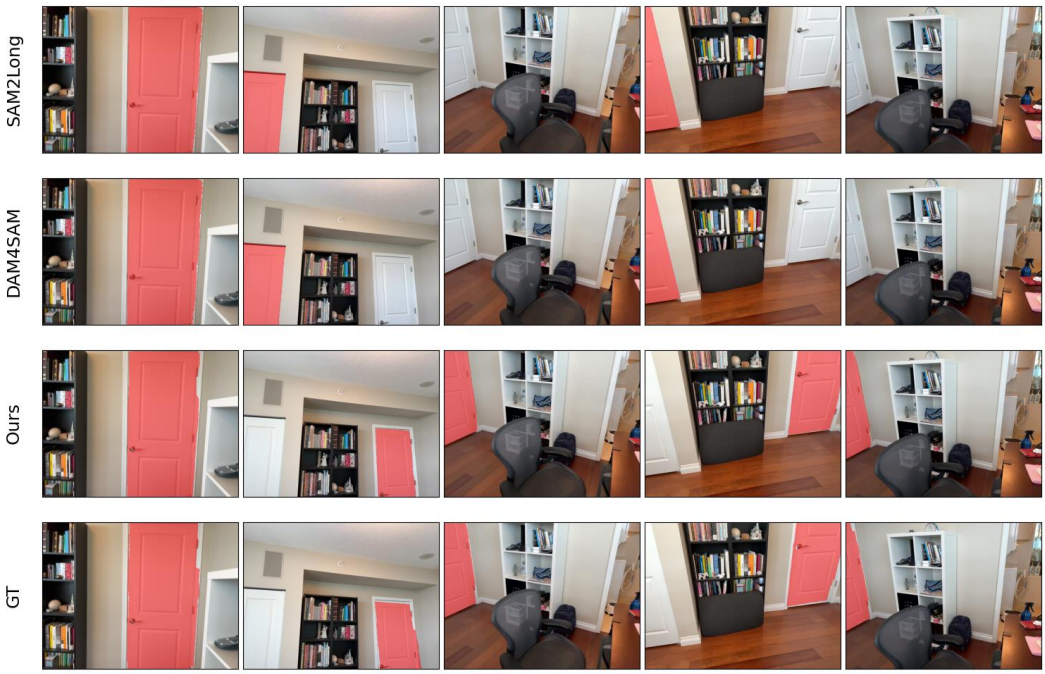


Figure 16: **Qualitative comparison of VOS methods** (scene cc5237fd77).

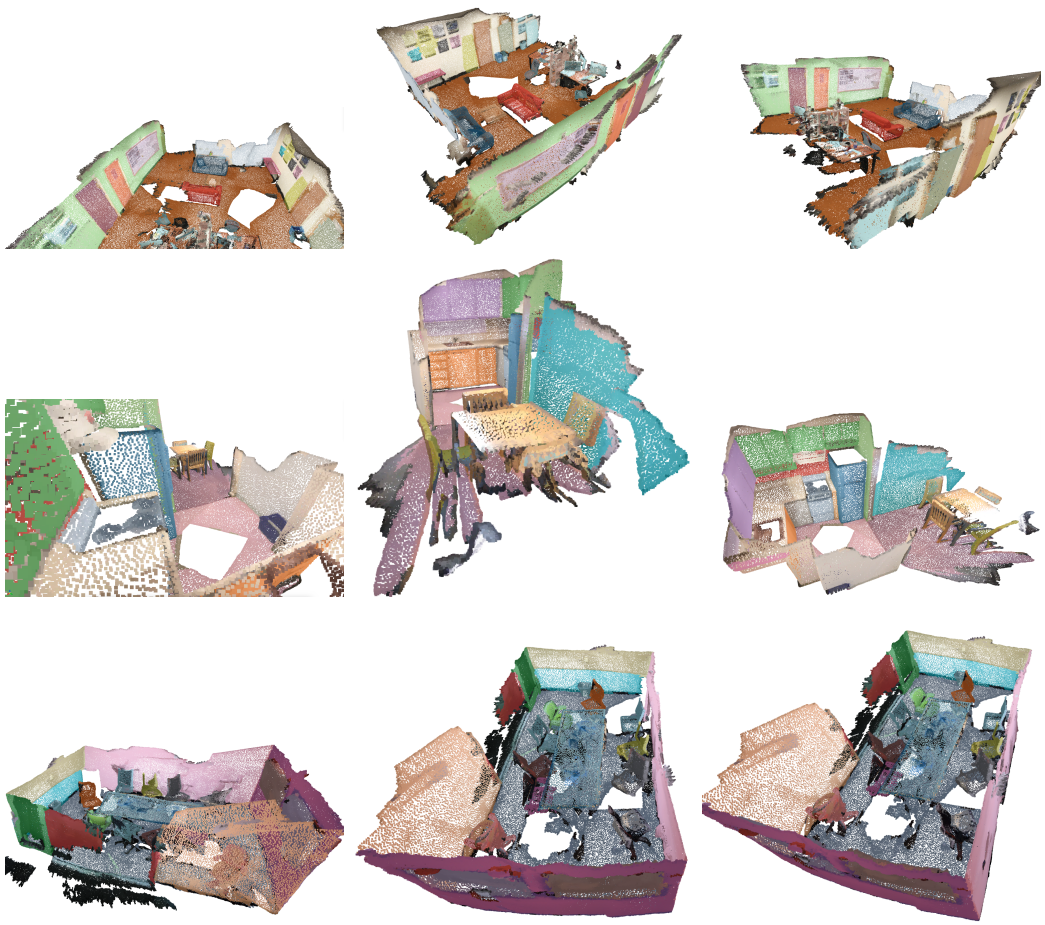

Figure 17: **Class-agnostic 3D instance segmentation** obtained by projecting 3AM tracks into 3D.