TORUS: SCALING OPEN-VOCABULARY TRACKING WITH HETEROGENEOUS DATA AND STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary multi-object tracking (OVMOT) is short of training data: images with large vocabularies are plentiful, but annotated tracking videos are scarce and temporally sparse. Existing methods either simulate motion from images—which poorly represents real-world deformation and occlusion—or label video by hand, which fails to scale. We introduce TORUS, an online tracker that bypasses this trade-off by unifying both data sources, jointly learning from detection images, disjoint-vocabulary videos, and synthetic pseudo-videos—without using any TAO annotations. TORUS freezes an open-vocabulary detector, adding a selective state-space (Mamba) memory to propagate tracks across frames. To learn jointly from this data mixture, we introduce two key designs: i) a class-agnostic confidence head that decouples track lifecycle management from classification, keeping tracks alive when predicted classes flip; and ii) a spatial-semantic federated loss that masks objectness supervision to avoid background penalization where labels are incomplete. With a ResNet-50 backbone, TORUS achieves 42.5 Base and 42.1 Novel TETA on TAO OVMOT, outperforming the state of the art by +2.9 and +6.8 points. A single set of untuned weights generalizes zero-shot across DanceTrack (57.7 HOTA), YouTube-VIS 2021 (42.9 mAP), and VisDrone-MOT (42.3 TETA)—operating entirely online at 14.7 FPS under 1,203 prompts.

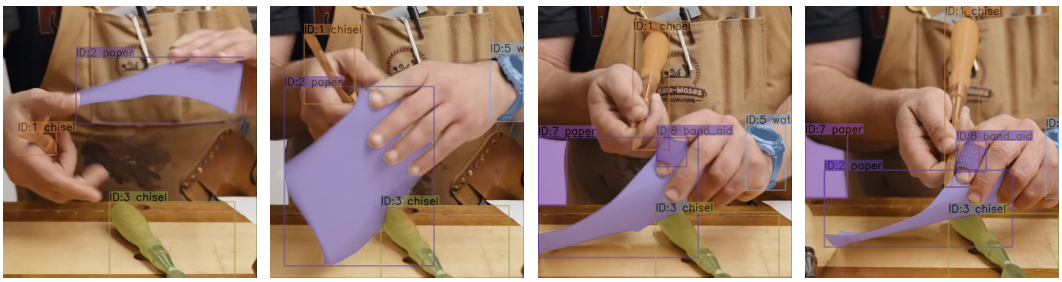

Figure 1: **Open-vocabulary tracking with TORUS.** Our framework enables the detection, segmentation, and stable tracking of long-tail categories (e.g., *chisel*, *band-aid*, *watch*) and generalizes zero-shot to unseen classes such as *paper*. By fusing static image priors with temporal kinematics via Mamba-based state-space models and a novel federated loss, TORUS maintains robust identities across diverse, unconstrained vocabularies.

1 INTRODUCTION

Open-Vocabulary Multi-Object Tracking (OVMOT) requires detecting, segmenting, and tracking arbitrary object categories defined by natural language queries (Li et al., 2023).

While open-vocabulary object detection has matured significantly, temporal association of arbitrary categories remains an open challenge. Modern detectors learn from large-vocabulary image datasets such as LVIS (Gupta et al., 2019) to successfully localize diverse, open-set vocabularies (Zhou

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.

et al., 2022; Wu et al., 2024; Segù et al., 2025a). However, tracking arbitrary categories requires learning open-set dynamic motion representations, which is hindered by the state of current open-vocabulary video resources: existing datasets either feature extensive vocabularies with highly sparse annotations (e.g., TAO; Dave et al., 2020), display severe category-wise data scarcity (e.g., LV-VIS; Wang et al., 2023), or restrict dense trajectories to a single class (e.g., DanceTrack; Sun et al., 2022).

The field has worked around this shortage by: i) manufacturing motion from still images (Li et al., 2023; 2025)—though synthesized sequences rarely capture complex 3D rotations, non-rigid deformations, or occlusions—and ii) maximizing the utility of TAO, the singular data source containing both open-vocabulary and video labels; specifically, COVTrack (Qian et al., 2025b) densifies TAO's annotations, while VOVTrack (Qian et al., 2025a) leverages its unlabeled frames to learn association. This dependency binds both approaches to the target benchmark, restricting their generalization to expanding taxonomies. Moreover, prior works rely on an offline majority vote over completed tracks to stabilize class predictions, a non-causal step that prevents their use in online settings.

In this paper, we argue that this data shortage is an artifact of requiring a singular training source to annotate spatial semantics and temporal association jointly, forcing a compromise on annotation density. These two objectives are fundamentally separable: temporal association can be learned class-agnostically via spatiotemporal continuity on video datasets, while semantics can be acquired motion-agnostically from static images. While each domain has abundant data in isolation, their conjunction is scarce. The central unanswered challenge is whether and how a unified model can jointly acquire both representations without cross-modal interference.

We answer this challenge with **TORUS** (**T**racking **O**pen vocabularies with hete**R**ogeneo**U**s data and **S**tate-space models), a unified online framework (Figure 1) trained on heterogeneous data—static images, disjoint-vocabulary videos, and synthetic pseudo-videos—without any TAO annotations. Architecturally, TORUS extends the frozen MOBIUS (Segù et al., 2025a) detector with joint decoding of detection and track queries, propagating temporal states via a selective state-space memory (Gu & Dao, 2023). To prevent cross-modal interference, we: i) decouple track lifecycle management from classification via a class-agnostic confidence head, preventing semantic misidentifications from terminating valid tracks; and ii) introduce a spatial-semantic federated loss that masks objectness supervision for unannotated categories (e.g., chairs in DanceTrack) to avoid penalizing them as background, extending federated loss principles (Zhou et al., 2021; 2022) to localization. Our motivating ablation confirms the power of this decoupling: while trained on static images alone, TORUS matches prior methods at 6.5 Novel ClsA on TAO; adding videos with virtually none of those classes surges performance to 19.3. A single model sets the state of the art on open-vocabulary and standard TAO benchmarks, generalizing zero-shot to DanceTrack, VisDrone, and YouTube-VIS entirely online.

Our contributions are summarized as follows:

- **A decoupled training paradigm for open-vocabulary tracking** that eliminates the need for open-vocabulary video annotations, demonstrating that dynamic motion representations learned from restricted-vocabulary videos generalize zero-shot to unseen semantic categories.
- **A spatial-semantic federated masking formulation** that generalizes federated supervision from classification logits to localization-level objectness targets. This, combined with our decoupled confidence head, allows seamless joint training on datasets with disjoint or incomplete vocabularies.
- **A unified state-space open-vocabulary tracking framework** that strips redundancy and specificity to a minimum. Evaluated with a single set of weights per backbone—and without benchmark-specific tuning—our model generalizes robustly across four diverse benchmarks and achieves the strongest public results on TAO OVMOT.

2 Related work

Open-vocabulary detection and partial labels. Detectors that align region features with text embeddings recognize classes named at test time (Radford et al., 2021; Zhou et al., 2022); universal models such as GLEE (Wu et al., 2024), UNINEXT (Yan et al., 2023) and MOBIUS (Segù et al., 2025a) unify detection and segmentation across datasets. Their data sources such as LVIS (Gupta et al., 2019) are not exhaustively labeled. Federated losses (Zhou et al., 2022) address this by evaluating the classification loss only on positives and a random sample of negative classes. They

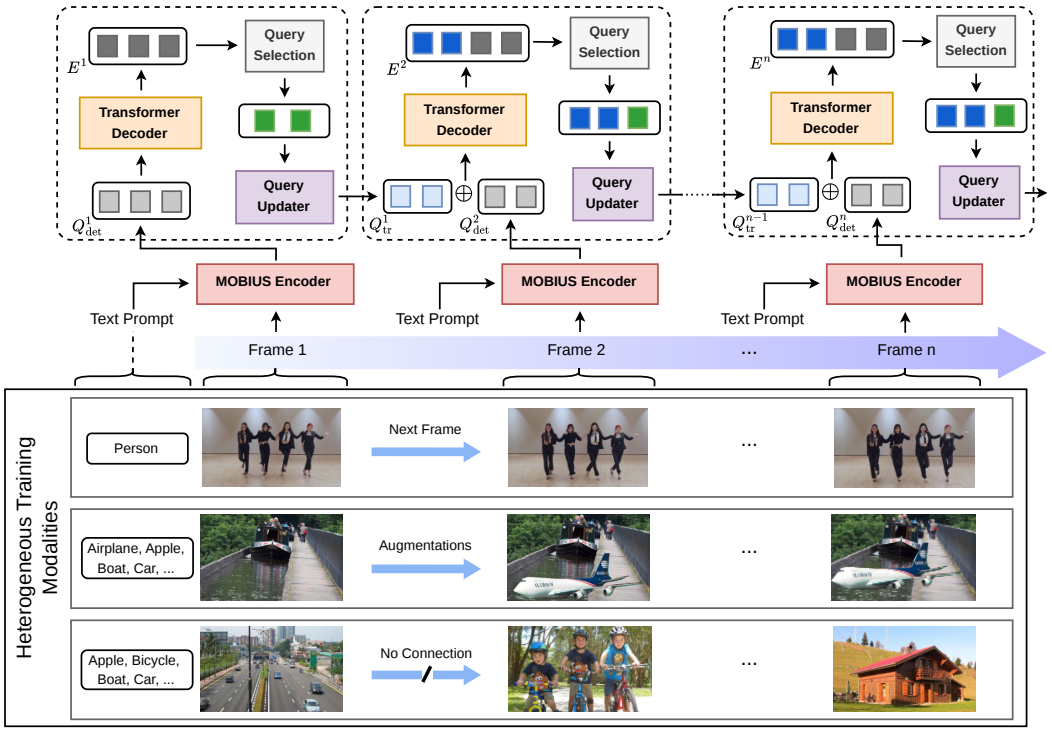

Figure 2: **Overview of the TORUS** temporal architecture and heterogeneous data training. **(Bottom)** To achieve generalist perception, the model is jointly trained on a mixture of three data modalities: Real Videos to learn real-world kinematics, synthetic Pseudo-Clips to expose the temporal memory to massive open-vocabulary taxonomies, and Independent Images to preserve spatial detection priors. **(Top)** The unrolled auto-regressive tracking process across n frames. At each frame t, the frozen MOBIUS encoder extracts semantic detection queries (Q^t_{det}). For frames $t \geq 2$, these are jointly decoded with continuous kinematic track queries (Q^{t-1}_{tr}) propagated by the Query Updater. For an in-depth schematic of the operations within a single time step, refer to Figure 4 in the Appendix.

leave the objectness target alone, so an unlabeled object is still pushed toward background. Our mask is deterministic, reads the per-image metadata directly, and acts on the tracking-confidence target.

End-to-end trackers and memories. Tracking-by-detection links detections with a motion model and an assignment step (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2022), with thresholds tuned per dataset. Query-based trackers (Zeng et al., 2022; Meinhardt et al., 2022) carry one query per track through a DETR-style decoder (Carion et al., 2020) and learn association end to end; later work adds long-term memory (Gao & Wang, 2023) or keeps a state-space hidden state per track (Gu & Dao, 2023; Segù et al., 2025b; Xiao et al., 2024; Huang et al., 2025). Video instance segmentation (VIS) extends the same idea to masks (Cheng et al., 2021); MinVIS (Huang et al., 2022) tracks by bipartite matching of per-frame instance queries, and OV2Seg (Wang et al., 2023) and OVFormer (Fang et al., 2024) align mask embeddings with text. To achieve real-time efficiency, SegFS (Barsellotti et al., 2026) decouples semantics from tracking by evaluating the foundation model only on sparse keyframes. In contrast, TORUS uses the query-propagation framework with a per-track Mamba memory and drives its box and mask heads with the same track queries.

Open-vocabulary tracking. OVTrack (Li et al., 2023) defined the task on TAO (Dave et al., 2020) and distilled vision-language knowledge into a two-stage tracker; SLAck (Li et al., 2024b) fuses semantic, location and appearance cues; MASA (Li et al., 2024a) learns association from SAM (Kirillov et al., 2023) masks. The first end-to-end OVMOT transformer OVTR (Li et al., 2025) and SAM2-OV (Chen et al., 2026) train on augmented images, while COVTrack (Qian et al., 2025b) and VOVTrack (Qian et al., 2025a) rely on relabeled or unannotated TAO videos. All of them train on images alone or on TAO video, labeled or not. TORUS trains on neither TAO nor any new annotation, and it takes its motion supervision from datasets whose vocabularies are small or disjoint from the target taxonomy.

3 TORUS

Figure 2 illustrates the auto-regressive tracking process across consecutive frames. A frozen open-vocabulary detector produces detection queries for frame t. A decoder refines them together with the track queries carried over from frame $t-1$. A query selection step starts new and ends lost tracks. Finally, a query updater with a Mamba memory writes the next frame's track queries.

Notation. At frame t, the prompted vocabulary $\mathcal{C}$ is encoded into text embeddings Z_{text}. The detector outputs K detection queries Q^t_{det}. Track queries from the previous frame are Q^{t-1}_{tr}, one per active track. The decoder maps the joint set $Q^t = [Q^{t-1}_{\text{tr}}; Q^t_{\text{det}}]$ to embeddings $E^t = [E^t_{\text{tr}}; E^t_{\text{det}}]$. Four heads read each embedding $e \in E^t$: a box $\hat{b}$, a mask $\hat{m}$, a confidence logit s, and class logits $z \in \mathbb{R}^{|\mathcal{C}|}$. We write σ for the sigmoid.

3.1 A FROZEN DETECTOR AND A STATE-SPACE TRACK MEMORY

Frozen detector. We keep the image backbone, text encoder and bottleneck encoder of MOBIUS (Segù et al., 2025a) frozen. The bottleneck fuses multi-scale image features with CLIP (Radford et al., 2021) text embeddings and selects the $K = 300$ tokens with the highest objectness as Q^t_{det}. Freezing keeps the vocabulary MOBIUS learned from five detection datasets and leaves only the decoder, heads and query updater to train (16.1 million of 71.3 million parameters with ResNet-50).

Joint decoding. The transformer decoder, initialized from MOBIUS, processes Q^t with self-attention across all queries and cross-attention to the frame's features, as in MOTR (Zeng et al., 2022). Track queries and detection queries therefore see each other, which lets a detection query yield to the track that already explains its object. Query pruning during inference is only applied to detect queries.

Query selection. A detection embedding whose fused score $\hat{s}$ (Eq. 3, Sec. 3.3) exceeds $\tau_{\text{new}} = 0.25$ starts a new track, with hidden state zero. A track whose fused score stays below $\tau_{\text{keep}} = 0.2$ for 20 consecutive frames is deleted. Nothing else governs the life of a track, and the thresholds are shared across every benchmark in this paper if not stated otherwise.

Query updater. Appearance alone fails under occlusion, motion blur and among look-alike targets, so each track keeps a hidden state h_t that summarizes its history. For a track with decoder embedding $e^t \in E^t_{\text{tr}}$ and positional embedding p^t of its predicted box, $N = 4$ stacked Mamba blocks (Gu & Dao, 2023) update the state from thresholded observations:

$$\Delta q^t,\ h_t = \text{Mamba}\big(\tilde{x}^t,\ h_{t-1}\big)\,, \quad \text{where} \quad \tilde{x}^t = \mathbb{I}(\hat{s}^t > \theta_{\text{update}})\,(e^t + p^t), \tag{1}$$

with $\theta_{\text{update}} = 0.4$. Each block applies $h_t = \bar{A} h_{t-1} + \bar{B}\,\tilde{x}^t$ and $\Delta q^t = C\,h_t$. Crucially, when an object is occluded ($\tilde{x}^t = 0$), the transition reduces to $h_t = \bar{A} h_{t-1}$, allowing Mamba to inertia-extrapolate the track trajectory without assimilating background noise. The next track query is a residual update of the previous one,

$$Q^t_{\text{tr}} = \text{Norm}\big(Q^{t-1}_{\text{tr}} + \text{FFN}(\Delta q^t)\big)\,. \tag{2}$$

Each track is its own sequence. Unlike Samba (Segù et al., 2025b), we do not synchronize hidden states across tracks; tracks interact only through decoder self-attention. The state carries a track through frames in which its appearance is unusable, and it is what the video data of Section 3.4 train.

3.2 A CONFIDENCE HEAD SEPARATE FROM THE CLASSIFIER

In a single-head tracker, the class score doubles as the track score, so a frame in which a rare object is read as a similar but incorrect class lowers its score and can end the track. TETer (Li et al., 2022), MeMOT (Cai et al., 2022) and OVTrack (Li et al., 2023) separate the two for association. In TORUS, track identity is maintained recurrently without semantic re-matching, while birth, memory update, and deletion are governed by a fused score of confidence and class.

Assignment. Following MOTR, a track query keeps the ground-truth identity it was born with, with target IoU zero in frames where the object is absent, and detection queries are matched to the remaining ground truth by the Hungarian algorithm (Kuhn, 1955) on a cost of class, L_1 box and GIoU terms with weights $(4.0, 5.0, 2.0)$.

Class loss. $\mathcal{L}_{\text{cls}}$ is a sigmoid focal loss (Lin et al., 2017) over the prompted vocabulary, applied to matched queries only ($\alpha = 0.5$, $\gamma = 2.0$). Unmatched queries receive no class loss. This is safe because the confidence head, not the classifier, suppresses background, and it means that an unlabeled object can never teach the classifier that its class is absent.

Box, mask and total loss. Following MOBIUS, the box sub-loss is $\mathcal{L}_{\text{box}} = 5\,\mathcal{L}_{L_1} + 2\,\mathcal{L}_{\text{giou}}$ and the mask sub-loss is $\mathcal{L}_{\text{mask}} = 5\,\mathcal{L}_{\text{ce}} + 5\,\mathcal{L}_{\text{dice}}$, with mask terms disabled for box-only datasets. The total loss is $\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$ with weights $\lambda_{\text{cls}} = 4$ and $\lambda_{\text{conf}} = 4$, summed over the frames of a video sequence.

3.3 A SPATIAL-SEMANTIC FEDERATED MASK FOR INCOMPLETE LABELS

The confidence head must see unmatched queries, or it cannot learn to reject background. On a non-exhaustively labeled image, however, an unmatched query with high confidence is either a false positive or a correct detection of an unlabeled object, and the loss cannot tell which. Pushing all of them to zero teaches the model that unlabeled classes are background. Federated losses (Zhou et al., 2022) avoid this for the classifier by sampling negative classes; nothing protects the class-agnostic target. We use the metadata that large-vocabulary datasets already ship to decide, per query, whether the label can be trusted in a deterministic way.

Fused score. To rank queries we combine confidence with the strongest class,

$$\hat{s} = \sigma(s) \cdot \max_{c \in \mathcal{C}} \sigma(z_c), \tag{3}$$

with the gradient of the class term stopped, so that $\mathcal{L}_{\text{conf}}$ trains s alone.

Mask. Each training image comes with an exhaustively annotated or confirmed absent category set $\mathcal{C}_{\text{exh}}$ and a set of ignore regions $\mathcal{I}$ (crowd boxes and dataset ignore labels). For an unmatched query q with box b_q and predicted class $c_q^* = \arg\max_c z_c$,

$$M_q = \Big(1 - \mathbb{1}\big[\,\text{IoP}(b_q, \mathcal{I}) > 0.5\big]\Big) \cdot \Big(1 - \mathbb{1}\big[\hat{s}_q > 0.5 \,\wedge\, c_q^* \notin \mathcal{C}_{\text{exh}}\big]\Big), \tag{4}$$

where $\text{IoP}(b, \mathcal{I}) = |b \cap \mathcal{I}|/|b|$ is the fraction of the box that lies in ignore regions. The first factor drops queries that lie in unlabeled areas. The second drops confident predictions of classes the image does not annotate exhaustively. On COCO, Objects365 and the closed-set videos, $\mathcal{C}_{\text{exh}}$ is the dataset vocabulary, so on DanceTrack a confident *chair* costs nothing and a spurious *person* is penalized. On LVIS, $\mathcal{C}_{\text{exh}}$ is the per-image list. An image labeled exhaustively for *dog* but not *leash* lets a confident *leash* query pass. A masked query receives no loss at all instead of turning into a positive.

Confidence loss. We use the Varifocal loss (Zhang et al., 2021), in the calibrated form of MOBIUS, with the IoU to the matched ground truth as target:

$$\mathcal{L}_{\text{conf}} = \sum_{q \in \text{matched}} \text{VFL}(\hat{s}_q, \text{IoU}_q) + \sum_{q \notin \text{matched}} M_q\, \text{VFL}(\hat{s}_q, 0)\,. \tag{5}$$

The mask has a price: a confident false positive of a class outside $\mathcal{C}_{\text{exh}}$ is never corrected.

3.4 TRAINING ON HETEROGENEOUS DATA

TORUS trains on three kinds of sample drawn from ten datasets; Table 12 in the appendix gives the sampling weights. The images are COCO (Lin et al., 2014), LVIS, Objects365 (Shao et al., 2019), OpenImages (Kuznetsova et al., 2020) and BDD100K (Yu et al., 2020) detection, the sets MOBIUS was pretrained on. The videos are LV-VIS (Wang et al., 2023), OVIS (Qi et al., 2021), BDD100K-MOTS, DanceTrack (Sun et al., 2022) and BFT (Zheng et al., 2024). Every sample is a clip of $T = 10$ frames processed autoregressively, as at test time.

Independent images (80% of image samples). T unrelated images form a clip. The track pool is reset after every frame and populated with dummy track queries initialized to zeros (zero embeddings and reference points, with a randomized count of 0–9), while tracking losses are weighted zero. These samples preserve detection quality and keep the full vocabulary in the decoder.

Video clips. We take T frames from a video with a random stride and at most 0.75 s between neighbors, so that the memory sees slow and fast motion, occlusion and deformation. These datasets label few classes and almost never a rare one; what they teach is class-agnostic.

Table 1: **TORUS leads every published ResNet-50 method on both TAO OVMOT splits and uses no TAO annotation.** Data: I = images, V = videos other than TAO, T = TAO videos, T^u = unlabeled TAO videos. †Offline majority vote over completed tracks. *Swin-T backbone of comparable size; code unpublished, so voting unknown. ‡Novel classes are LVIS-rare classes, which the MOBIUS backbone saw in pretraining, so Novel scores measure retention, not zero-shot discovery.

Method	Data	Base				Novel‡				FPS
		TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA	
OVTrack†	I	35.5	49.3	36.9	20.2	27.8	48.8	33.6	1.5	–
MASA†	I	36.9	55.1	36.4	19.3	30.0	54.2	34.6	1.0	–
SLAck*(†)	I+T	37.2	55.0	37.6	19.1	31.1	54.3	37.8	1.3	–
OVTR†	I	36.6	52.2	37.6	20.1	31.4	54.4	34.5	5.4	–
VOVTrack†	I+T^u	39.6	58.9	40.9	19.1	35.3	58.5	40.9	6.5	–
COVTrack†	I+T	39.6	57.3	42.0	19.6	34.3	58.2	41.3	3.5	–
TORUS (ours)	I+V	**42.5**	**60.4**	**43.1**	**24.1**	**42.1**	**63.3**	**43.8**	**19.3**	14.7
TORUS-MNv4	I+V	40.3	57.3	42.0	21.6	38.4	58.5	42.6	14.0	17.9
TORUS-FViT-3	I+V	44.8	60.0	44.6	29.9	43.7	64.9	43.2	22.9	12.3

Pseudo-videos (20% of image samples). An image becomes a clip through a moving crop window (crop ratio 0.7, speeds of 10–50 px/frame), the dynamic insertion and translation of segmented objects sampled from other images, and, with probability $p = 0.2$, a "sandwich" that splices frames of another clip to imitate a camera cut. Pseudo-videos show the memory the full vocabulary in motion. They cannot show real deformation, which is why they complement real video rather than replace it (Table 3).

Track lifecycle noise. To stop the decoder from leaning on propagated queries, we drop each active track with probability 0.2 after every frame, so that has to be rediscovered; 30% of newborn tracks are protected from dropping for the rest of the clip, so that the memory still learns long dependencies.

4 Experiments

4.1 Setup

Backbones. We use ResNet-50 (R50) (He et al., 2016) for comparison with prior work, MobileNetV4-Medium (MNv4) (Qin et al., 2024) for efficiency and FasterViT-3 (FViT-3) (Hatamizadeh et al., 2024) for capacity. Unless stated otherwise, results use R50 and one set of weights per backbone, trained once on the mixture of Section 3.4.

Benchmarks and metrics. TAO OVMOT (Li et al., 2023) evaluates the TAO validation set with all 1,203 LVIS classes as prompts, split into Base and Novel classes; the Novel classes are the LVIS-rare classes. TETA (Li et al., 2022) is the mean of localization (LocA), association (AssocA) and classification (ClsA) accuracy. Standard TAO uses TETA under the closed-set protocol of TETer. DanceTrack (Sun et al., 2022) uses HOTA (Luiten et al., 2021), which combines detection (DetA) and association (AssA) accuracy, together with MOTA (Bernardin & Stiefelhagen, 2008) and IDF1 (Ristani et al., 2016). YouTube-VIS 2021 (YTVIS-21) (Yang et al., 2019) uses mask mAP, split into base (mAP_b) and novel (mAP_n) classes as in OV2Seg. VisDrone-MOT (Zhu et al., 2022) uses TETA over eight classes. Appendix A.4 gives the protocol.

Inference protocol. All TORUS results are produced online. The class logits of each track are smoothed by an exponential moving average (EMA), $\bar{z}_t = \beta \bar{z}_{t-1} + (1 - \beta)\, z_t$ with $\beta = 0.9$, and no vote over completed tracks is applied. Baselines run from their published code with their default TAO configurations. Speed is measured on one RTX 4090 with all 1,203 LVIS prompts on a crowded scene with more than 200 concurrent tracks, the worst case for the decoder. In tables, bold marks the best value per column within a block, larger-backbone TORUS rows are never bolded against R50 baselines, and all numbers are single runs, so differences below about 0.5 TETA are inconclusive.

Table 2: **One set of ResNet-50 weights competes with specialists on three benchmarks it was never tuned for.** (a) *DanceTrack*: SambaMOTR and the first TORUS row train on DanceTrack only; the other TORUS rows use the weights of Table 1, whose mixture contains DanceTrack (3% of samples); OVTR and COVTrack run zero-shot with their TAO configurations; the gray row swaps the Mamba memory for a Transformer memory of matched capacity. (b) *VisDrone-MOT*, zero-shot for every method, prompted with the eight target classes or with all 1,203 LVIS classes filtered afterwards. (c) *YouTube-VIS 2021*. *TORUS saw the novel classes through LVIS and trains on video masks from OVIS, LV-VIS and BDD100K-MOTS, so only mAP_b compares with OV2Seg and OVFormer. All methods use a majority vote since track-mAP requires a single class per track.

(a) DanceTrack

Model	DanceTrack in training	HOTA	AssA	DetA	MOTA	IDF1
Overfitting Scenario						
SambaMOTR	✓	67.2	57.5	78.8	88.1	70.5
TORUS (ours)	✓	66.1	56.2	78.1	88.3	70.1
Open-Vocabulary Scenario						
OVTR	✗	30.0	15.5	58.5	65.9	28.8
COVTrack	✗	8.9	3.4	23.9	−13.1	6.9
TORUS (ours)	✓ (in mixture)	**57.7**	**43.3**	**77.4**	**87.9**	**59.7**
TORUS, Transformer memory	✓ (in mixture)	53.7	38.5	75.2	84.8	53.9

(b) VisDrone-MOT

Method	TETA	LocA	AssocA	ClsA
Closed Set				
OVTR	23.2	18.1	22.1	29.5
COVTrack	41.1	44.5	46.0	32.7
TORUS (ours)	**48.2**	**52.0**	**53.6**	**39.1**
Open Vocabulary				
OVTR	33.8	35.7	31.7	33.9
COVTrack	35.5	33.3	40.0	33.3
TORUS (ours)	**42.3**	**41.5**	**49.7**	**35.8**

(c) YouTube-VIS 2021

Method	YTVIS train	Backbone	mAP	mAP_b	mAP_n
Mask2Former	✓	R50	46.4	–	–
MinVIS	✓	R50	44.2	–	–
OV2Seg	✗	R50	27.2	30.1	11.1
OVFormer	✗	R50	29.8	32.3	15.7
TORUS (ours)	✗	R50	42.9*	**44.3**	35.2*
Mask2Former	✓	Swin-L	52.6	–	–
MinVIS	✓	Swin-L	55.3	–	–
OV2Seg	✗	Swin-B	33.9	36.7	18.2
OVFormer	✗	Swin-B	37.6	41.0	18.3
TORUS (ours)	✗	FViT-3	50.7*	**53.2**	36.7*

4.2 Open-vocabulary tracking on TAO

Table 1 compares TORUS with published OVMOT methods at ResNet-50 scale. TORUS reaches 42.5 Base and 42.1 Novel TETA, 2.9 and 6.8 points above VOVTrack and COVTrack, the best prior methods, both of which train on TAO videos. TORUS uses no TAO annotation. The largest gain is Novel ClsA: 19.3 against 6.5 for the best baseline, and every baseline sits inside the 1.0–6.5 range that our image-only ablation reproduces (Table 3). The Novel classes are LVIS-rare classes, which the frozen MOBIUS encoder saw during pretraining, so the Novel scores measure how much of that knowledge survives tracking training rather than zero-shot discovery; Section 4.5 shows that without the recipe the same encoder keeps a third of it. The two other backbones follow the same pattern (Table 1). Under the standard closed-set TAO protocol, TORUS-R50 reaches 42.8 TETA against 36.9 for COVTrack. Associating the detections of the same frozen MOBIUS encoder with MinVIS gives 36.2, so 6.6 points of the gain come from the learned tracker and not from the detector.

4.3 One model on three more benchmarks

Table 2 runs the same R50 weights on three real-world benchmarks that differ from TAO in motion, viewpoint, and output.

DanceTrack has one class, fast non-linear motion, and near-identical dancers. Trained on DanceTrack alone, TORUS is within 1.1 HOTA of SambaMOTR, a tracker built for this data, so the memory itself is competitive. With the single set of weights of Table 1, in which DanceTrack is 3% of the training mixture, TORUS reaches 57.7 HOTA. OVTR and COVTrack, which cannot add such data without

Table 3: **Real video teaches Novel-class stability; pseudo-videos add the last 2.8 ClsA.** Detection on COCO and LVIS (box AP; AP^r = LVIS-rare classes) and tracking on TAO OVMOT for models trained on subsets of the mixture, all with ResNet-50. Images alone match or beat the full recipe on Base classes and fail on Novel classes (6.5 vs. 19.3 ClsA); the failure is in classification, not association. "Images only" includes pseudo-videos; "LVIS+COCO only" is trained without them and was not evaluated on detection (–); "Full recipe" adds pseudo-videos to images and videos.

Training data	COCO	LVIS		TAO Base			TAO Novel		
	AP	AP	AP^r	TETA	AssocA	ClsA	TETA	AssocA	ClsA
Videos only	49.3	41.0	34.0	39.8	41.5	18.6	39.4	41.3	17.0
Images only	**53.2**	**48.6**	35.6	43.0	43.7	23.8	37.1	41.7	6.5
LVIS+COCO only	52.8	42.4	32.2	**43.3**	**45.6**	23.2	37.8	**44.6**	6.7
Images + videos	53.1	48.4	37.9	42.0	42.3	23.2	40.9	43.2	16.5
Full recipe (TORUS)	52.7	48.1	**39.3**	42.5	43.1	**24.1**	**42.1**	43.8	**19.3**

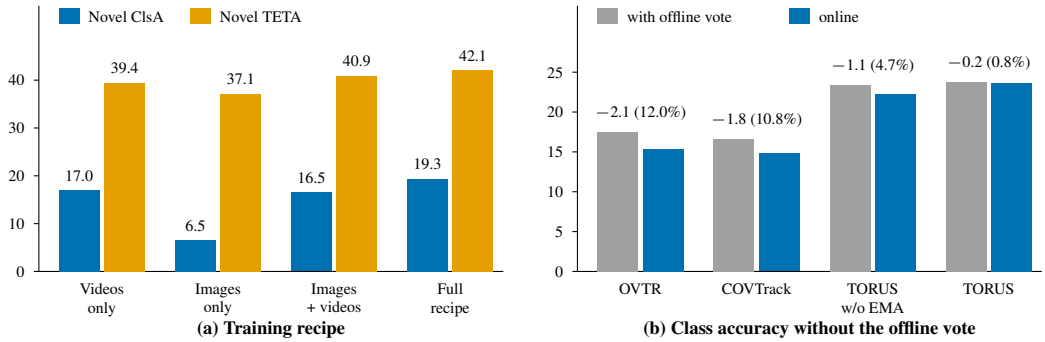

Figure 3: **Real video makes Novel-class tracking stable, and the stability holds online.** (a) Novel ClsA and Novel TETA on TAO OVMOT for the four training recipes of Table 3: images alone give 6.5 ClsA, adding videos 16.5, adding pseudo-videos 19.3. (b) ClsA on TAO OVMOT (Base and Novel combined) with and without the offline majority vote (Table 7). TORUS loses 0.2 points; OVTR and COVTrack lose 2.1 and 1.8.

narrowing their vocabulary, reach 30.0 and 8.9 when run with their TAO configurations. COVTrack's negative MOTA points to thresholds calibrated for 1 FPS video as much as to a modeling failure. Replacing the Mamba memory by a Transformer of matched capacity in the same model (gray row) costs 4.8 AssA and 4.0 HOTA, while detection accuracy barely moves since the encoder is shared.

VisDrone-MOT moves the camera to a drone and shrinks the objects; no method trains on it. Prompted with the eight target classes, TORUS reaches 48.2 TETA against 41.1 for COVTrack; prompted with all 1,203 LVIS classes and filtered afterwards (see Appendix A.4), 42.3 against 35.5. OVTR scores lower with eight prompts than with 1,203 (23.2 vs. 33.8), which we attribute to score thresholds calibrated for the long prompt list; we did not re-tune any method.

YouTube-VIS 2021 asks for masks. TORUS-R50 reaches 42.9 mAP, within 1.3 of MinVIS trained on YTVIS-21, and 44.3 base-class mAP against 32.3 for OVFormer, the best open-vocabulary VIS method; with FasterViT-3 it reaches 50.7 and 53.2. TORUS trains on video masks from OVIS, LV-VIS and BDD100K-MOTS but not on YTVIS-21, and its novel classes appear in LVIS, so only mAP_b compares directly with OV2Seg and OVFormer. Appendix A.7 shows further sequences with closed-set and 1,203-class prompts.

4.4 Online Semantic Stability

OVMOT methods report class accuracy after an offline majority vote over each completed track, which a live system cannot apply. Figure 3b removes the vote (full numbers in Appendix A.2). OVTR loses 2.1 ClsA (12.0%) and COVTrack 1.8 (10.8%). TORUS loses 0.2 (0.8%). Two parts explain

Table 4: **The federated mask protects rare classes. The separate confidence head protects association.** One mechanism removed at a time, ResNet-50. Detection on COCO and LVIS (box AP), tracking on TAO OVMOT evaluated at at 6 FPS, so the tracking columns compare with each other but not with Table 1. MOBIUS is the frozen detector before tracking training. Bold: best per column.

Model	Sep. head	Fed. mask	COCO	LVIS		TAO Base				TAO Novel			
			AP	AP	AP^r	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
MOBIUS	✗	✗	**54.4**	**50.4**	34.7	–	–	–	–	–	–	–	–
w/o sep. head	✗	✓	–	–	–	40.0	55.7	39.2	25.0	36.4	57.4	36.6	15.3
w/o fed. mask	✓	✗	52.9	48.2	36.7	40.9	**57.0**	40.7	25.0	41.1	**61.4**	**44.6**	17.4
TORUS	✓	✓	52.7	48.1	**39.3**	**41.5**	56.8	**42.0**	**25.8**	**41.2**	59.9	44.0	**19.6**

the difference. The Mamba memory alone, with the EMA switched off, limits the loss to 1.1 points (4.7%), less than half of the baselines'; the online EMA over class logits removes most of the rest.

4.5 Ablations

All ablations use ResNet-50. Detection is measured on COCO and LVIS, tracking on TAO OVMOT.

Training recipe. Table 3 evaluates subsets of the mixture. Videos alone keep LVIS-rare AP (34.0 against 34.7 for the frozen detector) and reach 17.0 Novel ClsA, but lose common detection (41.0 LVIS AP against 48.6) and Base tracking. Images alone, with pseudo-videos, are the mirror image: the best COCO and LVIS AP and 43.0 Base TETA, but 6.5 Novel ClsA. The failure is semantic, not associative: Novel AssocA drops by 2.1 points while Novel ClsA drops by 12.8. A crop window moving over a still image leaves the object's appearance fixed, so a memory trained this way never learns to hold a class through real deformation; at test time the model falls back to per-frame predictions, which flicker for rare classes. Adding the videos restores Novel ClsA to 16.5 despite lacking these classes, and adding pseudo-videos on top gives 19.3 and the best LVIS-rare AP (39.3), because they show the memory the full vocabulary in motion. Against images alone, the full recipe pays 0.5 Base TETA for 5.0 Novel TETA and 12.8 Novel ClsA. Training on LVIS and COCO images only gives the best Base association (45.6), confirming static supervision suffices for common classes.

Components. Table 4 removes one mechanism at a time. Adding tracking costs 1.7 COCO AP and 2.3 LVIS AP and gains 4.6 LVIS-rare AP. Removing the federated mask costs 2.6 LVIS-rare AP and 2.2 Novel ClsA; TETA remains stable, and Novel LocA and AssocA rise by 1.5 and 0.6, the price of never correcting a confident false positive on a non-exhaustive class. Removing the separate confidence head, so that the class score governs the track, costs 7.4 Novel AssocA and 2.5 Novel LocA: rare objects whose class flickers lose their identity and, through the lifecycle thresholds, recall.

5 Conclusion

TORUS shows that open-vocabulary tracking need not wait for open-vocabulary video labels. Vocabulary comes from images, motion from any video, and two conditions keep them from interfering: a tracking confidence that does not depend on class, and an objectness loss that is silent where labels are incomplete. One set of weights then reaches 42.5 Base and 42.1 Novel TETA on TAO OVMOT, 57.7 HOTA on DanceTrack and 42.3 TETA on unseen aerial video, all online. Mining motion supervision from unlabeled video is the natural next step. The code and trained models will be made public.

Limitations. The frozen encoder sets a recall ceiling: an object it never proposes cannot be recovered by the memory. The federated mask needs per-image category metadata, which uncurated video lacks. Pseudo-videos imitate camera motion, not deformation, so a rare class learns its appearance changes only from the few real videos that contain it. Novel-class scores measure retention rather than zero-shot discovery, because the encoder saw the LVIS-rare classes. Results are single runs, VisDrone uses our own protocol, and two thresholds and an EMA coefficient remain hand-set.

AI USE STATEMENT

In this work, we used generative AI tools for literature review, code implementation, and the initial drafting of specific sections. We have manually reviewed all AI-assisted work: AI-generated literature summaries were manually verified against the original sources, code implementations were thoroughly tested for correctness by the authors, and all AI-drafted text was extensively revised and rewritten to ensure scientific accuracy. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

While we do not foresee immediate ethical risks directly arising from this specific methodology, we acknowledge the broader implications of Open-Vocabulary Multi-Object Tracking (OVMOT). Because our framework is capable of tracking arbitrary, user-defined categories in unconstrained environments—including people and vehicles—it carries potential dual-use risks related to unauthorized mass surveillance and privacy infringement. We urge researchers and practitioners to ensure that such models are deployed responsibly, in compliance with privacy regulations, and exclusively for applications that contribute positively to society.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we have provided comprehensive details throughout the manuscript and supplementary materials. The complete model architecture, including the Mamba memory module and functional head disentanglement, is detailed in Section 3 and Appendix A.6. Our heterogeneous data mixture, optimization hyperparameters, and exact dataset sampling ratios are explicitly defined in Section 3.4, Table 12, and Appendix A.8. Furthermore, the complete source code, evaluation scripts, and pre-trained model weights will be released publicly to facilitate future research.

REFERENCES

Luca Barsellotti, Martin Sundermeyer, Mattia Segù, Nikita Araslanov, Muhammad Ferjad Naeem, Marcella Cornia, Yongqin Xian, and Maxim Berman. Segmenting, fast and slow: Real-time open-vocabulary video instance segmentation with dual-path processing. *CoRR*, abs/2607.00124, 2026. doi: 10.48550/ARXIV.2607.00124. URL https://doi.org/10.48550/arXiv.2607.00124.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. doi: 10.1155/2008/246309.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016. doi: 10.1109/ICIP.2016.7533003.

Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8080–8090. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00792. URL https://doi.org/10.1109/CVPR52688.2022.00792.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://doi.org/10.1007/978-3-030-58452-8_13.

Yangkai Chen, Qiangqiang Wu, Guangyao Li, Junlong Gao, Guanglin Niu, and Hanzi Wang. SAM2-OV: A novel detection-only tuning paradigm for open-vocabulary multi-object tracking. In Sven

Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 3083–3091. AAAI Press, 2026. doi: 10.1609/AAAI.V40I4.37301. URL https://doi.org/10.1609/aaai.v40i4.37301.

Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. *CoRR*, abs/2112.10764, 2021. URL https://arxiv.org/abs/2112.10764.

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pp. 436–454. Springer, 2020. doi: 10.1007/978-3-030-58558-7_26. URL https://doi.org/10.1007/978-3-030-58558-7_26.

Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXX*, Lecture Notes in Computer Science, pp. 225–241. Springer, 2024. doi: 10.1007/978-3-031-72897-6_13. URL https://doi.org/10.1007/978-3-031-72897-6_13.

Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9867–9876, 2023. doi: 10.1109/ICCV51070.2023.00908.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752, 2023. doi: 10.48550/ARXIV.2312.00752. URL https://doi.org/10.48550/arXiv.2312.00752.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5351–5359. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00550. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html.

Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, José M. Álvarez, Jan Kautz, and Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=kB4yBiNmXX.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ca9567d8ef6b2ea2da0d7eed57b933ee-Abstract-Conference.html.

Hsiang-Wei Huang, Cheng-Yen Yang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Mambamot: State-space model as motion predictor for multi-object tracking. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10890199. URL https://doi.org/10.1109/ICASSP49660.2025.10890199.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3992–4003. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981, 2020. doi: 10.1007/S11263-020-01316-Z. URL https://doi.org/10.1007/s11263-020-01316-z.

Jinyang Li, En Yu, Sijia Chen, and Wenbing Tao. OVTR: end-to-end open-vocabulary multiple object tracking with transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=GDS5eN65QY.

Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E. Huang, and Fisher Yu. Tracking every thing in the wild. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, volume 13682 of *Lecture Notes in Computer Science*, pp. 498–515. Springer, 2022. doi: 10.1007/978-3-031-20047-2_29. URL https://doi.org/10.1007/978-3-031-20047-2_29.

Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 5567–5577. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00539. URL https://doi.org/10.1109/CVPR52729.2023.00539.

Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segù, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18963–18973. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.01794. URL https://doi.org/10.1109/CVPR52733.2024.01794.

Siyuan Li, Lei Ke, Yung-Hsu Yang, Luigi Piccinelli, Mattia Segù, Martin Danelljan, and Luc Van Gool. Slack: Semantic, location, and appearance aware open-vocabulary tracking. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXVII*, volume 15085 of *Lecture Notes in Computer Science*, pp. 1–18. Springer, 2024b. doi: 10.1007/978-3-031-73383-3_1. URL https://doi.org/10.1007/978-3-031-73383-3_1.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip H. S. Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. doi: 10.1007/s11263-020-01375-2.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8834–8844. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00864. URL https://doi.org/10.1109/CVPR52688.2022.00864.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge J. Belongie, Alan L. Yuille, Philip H. S. Torr, and Song Bai. Occluded video instance segmentation: Dataset and ICCV 2021 challenge. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/7f39f8317fbdb1988ef4c628eba02591-Abstract-round2.html.

Zekun Qian, Ruize Han, Junhui Hou, Linqi Song, and Wei Feng. VOVTrack: Exploring the potentiality in raw videos for open-vocabulary multi-object tracking. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7472–7482, 2025a. doi: 10.1109/ICCV51701.2025.00701.

Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. COVTrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10054–10063, 2025b. doi: 10.1109/ICCV51701.2025.00937.

Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby R. Banbury, Chengxi Ye, Berkin Akin, Vaibhav Aggarwal, Tenghui Zhu, Daniele Moro, and Andrew G. Howard. Mobilenetv4: Universal models for the mobile ecosystem. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XL*, Lecture Notes in Computer Science, pp. 78–96. Springer, 2024. doi: 10.1007/978-3-031-73661-2_5. URL https://doi.org/10.1007/978-3-031-73661-2_5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshops (ECCVW)*, volume 9914 of *Lecture Notes in Computer Science*, pp. 17–35, 2016. doi: 10.1007/978-3-319-48881-3_2.

Mattia Segù, Marta Tintore Gazulla, Yongqin Xian, Luc Van Gool, and Federico Tombari. MOBIUS: Big-to-mobile universal instance segmentation via multi-modal bottleneck fusion and calibrated decoder pruning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20726–20736, 2025a. doi: 10.1109/ICCV51701.2025.01927.

Mattia Segù, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Luc Van Gool, and Bernt Schiele. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=OeBY9XqiTz.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8429–8438. IEEE, 2019. doi: 10.1109/ICCV.2019.00852. URL https://doi.org/10.1109/ICCV.2019.00852.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 20961–20970. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02032. URL https://doi.org/10.1109/CVPR52688.2022.02032.

Haochen Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Cilin Yan, Weidi Xie, Shuai Wang, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4034–4043. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00375. URL https://doi.org/10.1109/ICCV51070.2023.00375.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, 2017. doi: 10.1109/ICIP.2017.8296962.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 3783–3795. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00363. URL https://doi.org/10.1109/CVPR52733.2024.00363.

Changcheng Xiao, Qiong Cao, Zhigang Luo, and Long Lan. Mambatrack: A simple baseline for multiple object tracking with state space model. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 4082–4091. ACM, 2024. doi: 10.1145/3664647.3680944. URL https://doi.org/10.1145/3664647.3680944.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15325–15336. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01471. URL https://doi.org/10.1109/CVPR52729.2023.01471.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 5187–5196. IEEE, 2019. doi: 10.1109/ICCV.2019.00529. URL https://doi.org/10.1109/ICCV.2019.00529.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2633–2642. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00271. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yu_BDD100K_A_Diverse_Driving_Dataset_for_Heterogeneous_Multitask_Learning_CVPR_2020_paper.html.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: end-to-end multiple-object tracking with transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVII*, volume 13687 of *Lecture Notes in Computer Science*, pp. 659–675. Springer, 2022. doi: 10.1007/978-3-031-19812-0_38. URL https://doi.org/10.1007/978-3-031-19812-0_38.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object detector. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 8510–8519. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00841. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VarifocalNet_An_IoU-Aware_Dense_Object_Detector_CVPR_2021_paper.html.

Table 5: **Comparison of state-of-the-art models on the standard TAO benchmark.** Unless otherwise specified, all methods utilize a ResNet-50 backbone. *SLAck uses a Swin-Tiny backbone, which possesses a comparable parameter count. Models highlighted in gray represent offline tracking methods.

Method	TETA	LocA	AssocA	ClsA
TETer	33.3	51.6	35.0	13.2
GLEE-Lite+MinVIS	36.1	50.5	37.2	20.5
MOBIUS-R50+MinVIS	36.2	48.4	38.0	22.2
UNINEXT (R50)	31.9	43.3	35.5	17.1
OVTrack	34.7	49.3	36.7	18.1
MASA-R50	34.1	52.1	35.7	15.0
SLAck-T*	35.5	52.2	38.9	15.6
OVTR	35.3	52.1	36.3	17.6
COVTrack	36.9	55.1	41.1	14.6
TORUS (ours)	**42.8**	**60.6**	**43.0**	**24.8**
TORUS MNv4-CM	39.7	56.9	41.4	20.6
TORUS FasterViT-3	45.6	60.5	45.4	30.8

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, volume 13682 of *Lecture Notes in Computer Science*, pp. 1–21. Springer, 2022. doi: 10.1007/978-3-031-20047-2_1. URL https://doi.org/10.1007/978-3-031-20047-2_1.

Guangze Zheng, Shijie Lin, Haobo Zuo, Changhong Fu, and Jia Pan. Nettrack: Tracking highly dynamic objects with a net. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 19145–19155. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01811. URL https://doi.org/10.1109/CVPR52733.2024.01811.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *CoRR*, abs/2103.07461, 2021. URL https://arxiv.org/abs/2103.07461.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pp. 350–368. Springer, 2022. doi: 10.1007/978-3-031-20077-9_21. URL https://doi.org/10.1007/978-3-031-20077-9_21.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022. doi: 10.1109/TPAMI.2021.3119563.

A Appendix

A.1 Standard TAO benchmark evaluation

While our primary evaluation focuses on the open-vocabulary capabilities of TORUS (Section 4.2), we also evaluate our framework on the standard, closed-set TAO benchmark for completeness and to provide direct comparability with earlier tracking literature.

As shown in Table 5, our ResNet-50 model achieves a state-of-the-art TETA of 42.8, significantly outperforming recent baselines. Furthermore, our FasterViT-3 variant pushes tracking capabilities to

45.6 TETA, demonstrating that our architecture scales efficiently with foundational model capacity even in standard tracking scenarios.

A.2 Extended ablation study

In this section, we provide detailed quantitative analyses supplementing the core ablations presented in the main text. Specifically, we evaluate the architectural choice of our temporal memory and the online semantic stability of our framework.

Table 6: **Temporal memory ablation on DanceTrack.** While detection accuracy (DetA) remains stable due to the shared spatial backbone, replacing our continuous SSM memory with a standard Transformer severely degrades Association Accuracy (AssA). This demonstrates that state-space models are superior at maintaining track identities through complex kinematics and heavy occlusions.

Memory Model	HOTA	AssA	DetA	MOTA	IDF1
Transformer-based	53.7	38.5	75.2	84.8	53.9
SSM-based (Default)	**57.7**	**43.3**	**77.4**	**87.9**	**59.7**

Temporal memory: Mamba vs. Transformer To validate our choice of Selective State Space Models for temporal propagation, we ablate the memory architecture by replacing the Mamba memory model with standard Transformer-based memory (utilizing self-attention followed by two linear layers). We evaluate this substitution on the DanceTrack dataset to explicitly stress-test the memory module against heavy occlusions, non-rigid deformations, and visually identical targets. As shown in Table 6, basic detection accuracy remains relatively stable between the two models due to the shared MOBIUS spatial backbone, whereas temporal tracking degrades significantly. The Transformer-based memory struggles to maintain identities through complex kinematics, resulting in a severe 4.8-point drop in Association Accuracy (AssA). This confirms that the continuous, sequence-dependent hidden state of Mamba is fundamentally superior to standard self-attention for modeling complex, non-linear real-world trajectories.

Detailed analysis of online classification stability. Expanding on the main text, Table 7 provides the comprehensive metrics for our online stability analysis. A persistent limitation in current OVMOT literature is the reliance on non-causal offline heuristics—specifically, majority-voting across a completed tracklet—to artificially boost Classification Accuracy (ClsA).

To evaluate true online robustness, we strip this offline correction from the baselines. Without it, OVTR and COVTrack suffer severe ClsA degradation (dropping by 12.0% and 10.8%, respectively), exposing their reliance on frame-independent, jittery semantic predictions.

In contrast, TORUS maintains stable semantics natively. We can isolate this stability into two causal components:

- **Continuous memory (no EMA):** Because our Mamba module propagates a continuous temporal hidden state, TORUS exhibits inherently lower semantic jitter. Even when we completely disable our online Exponential Moving Average (EMA) smoothing, TORUS experiences only a 4.7% drop in ClsA—less than half the degradation of the baselines.
- **Online EMA smoothing:** By applying a lightweight, strictly online EMA over the predicted class logits, we suppress the remaining frame-level variance. This reduces the ClsA drop to a negligible 0.8%.

Ultimately, this demonstrates that the combination of continuous hidden states and online temporal smoothing successfully eliminates the need for artificial offline post-processing, making TORUS highly suitable for real-time deployment.

A.3 Image dataset ablation: LVIS and COCO

To further isolate the impact of our training data composition on tracking performance, we conduct an additional ablation by fine-tuning the temporal memory exclusively on a strictly smaller set of static image datasets. Specifically, we use LVIS to make the size of our training data directly

Table 7: **Online classification stability on TAO OVMOT.** Results combine Base and Novel categories. Traditional baselines (OVTR, COVTrack) rely on an offline majority vote across completed trajectories to maintain high Classification Accuracy (ClsA), suffering noticeable drops without it. In contrast, our MOBIUS+Mamba framework leverages continuous hidden states and an online Exponential Moving Average (EMA) to maintain near-identical ClsA under strict online constraints.

Model	With Majority Vote				Without Majority Vote				Δ ClsA
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA	
OVTR	35.3	52.1	36.3	17.5	34.6	52.2	36.3	15.4	-12.0%
COVTrack	38.1	56.3	41.5	16.6	37.5	56.3	41.5	14.8	-10.8%
TORUS w/o EMA	42.5	60.7	43.2	23.4	42.5	60.7	43.2	22.3	-4.7%
TORUS (ours)	42.5	60.7	43.2	**23.8**	42.5	60.7	43.2	**23.6**	**-0.8%**

comparable to OVTR, which also utilizes LVIS for training. In these experiments, we omit all real video datasets, relying entirely on synthetic pseudo-video augmentations to simulate motion. We retain the pretrained MOBIUS backbone, which inherently possesses broad spatial priors from its massive-scale pre-training.

Because LVIS annotations are non-exhaustive, training solely on LVIS requires removing our spatial-semantic federated mask, leaving the model without a reliable negative training signal. To evaluate static training while preserving the federated mask, we add an experiment combining LVIS with COCO. COCO provides exhaustive labels for the same underlying images (albeit for fewer classes), allowing the federated mask to safely penalize false positives.

As shown in Table 8, training the tracking framework solely on static images yields highly competitive performance on Base categories. The LVIS+COCO configuration achieves a Base TETA of 43.2, which marginally outperforms the full heterogeneous TORUS baseline (42.5 Base TETA). This indicates that dense, static spatial supervision is highly effective at stabilizing common classes where abundant data is available.

Table 8: **Image dataset ablation on the TAO OVMOT benchmark.** Comparison of models trained exclusively on static images (LVIS only, LVIS+COCO, and all images) against the full heterogeneous TORUS recipe (ResNet-50). While static image training yields competitive Base performance, it causes severe degradation in Classification Accuracy (ClsA) on Novel/Rare categories due to the lack of authentic temporal dynamics.

Training Data	Base				Novel / Rare			
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
LVIS Only	42.8	60.0	44.7	23.5	36.4	59.6	42.0	7.7
LVIS + COCO	**43.2**	60.9	**45.4**	23.4	36.8	60.7	43.0	6.6
All Images	43.0	**61.6**	43.7	23.8	37.1	63.0	41.7	6.5
TORUS (Full Recipe)	42.5	60.4	43.1	**24.1**	**42.1**	**63.3**	**43.8**	**19.3**

However, as already discussed in the data ablation in section 4.5, all image-only configurations experience a severe degradation in Classification Accuracy (ClsA) and overall TETA compared to the full TORUS recipe on the Novel (Rare) split. While the foundation model provides strong spatial priors, exposing the continuous temporal memory exclusively to synthetic rigid augmentations fails to teach it how to handle authentic deformations for rare categories. The inclusion of our full heterogeneous training mixture—specifically bridging real motion with pseudo-video augmentations—remains strictly necessary to achieve temporal semantic stability across long-tail vocabularies.

A.4 VisDrone-MOT evaluation protocol and class mapping

To evaluate the zero-shot domain transfer capabilities of the models on the VisDrone-MOT benchmark, it is necessary to align the expansive open-vocabulary predictions (which span the 1,203 LVIS categories) with the specific target classes of the VisDrone dataset.

VisDrone class consolidation The original VisDrone-MOT dataset annotates the following target categories: *pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus* and *motor*. To ensure a fair and consistent evaluation against foundation models trained on standard taxonomies, we map and consolidate these original annotations into 8 distinct target classes: *person, bicycle, car, van, truck, tricycle, bus,* and *motorcycle*.

Open-vocabulary filtering process During the "Open-Vocabulary" evaluation setting (as reported in the main text), models are completely unconstrained and prompted with the full 1,203-class LVIS vocabulary during inference. To evaluate these unconstrained outputs against the VisDrone ground truth, we apply a deterministic post-inference filtering script to the raw output files.

The script parses the predicted trajectories and maps the predicted LVIS categories to the 8 consolidated VisDrone target classes according to the predefined relationships detailed in Table 9. If a model predicts a class that maps to a valid VisDrone target, the class ID is converted, and the tracklet is retained for evaluation. Conversely, any detection or track whose highest-confidence predicted class does not exist in this mapping is considered irrelevant to the benchmark context and is subsequently dropped from the final evaluation.

Table 9: Explicit mapping from the unconstrained LVIS vocabulary to the consolidated VisDrone-MOT target classes. This mapping is used for the post-inference filtering script during Open-Vocabulary evaluation.

Consolidated VisDrone Class	**Mapped LVIS Categories (LVIS ID)**
Person	person
Bicycle	bicycle
Car	car_(automobile), motor_vehicle, cab_(taxi), convertible
Van	minivan
Truck	truck, pickup_truck, trailer_truck
Tricycle	tricycle
Bus	bus_(vehicle)
Motorcycle	motorcycle, motor_scooter

A.5 Tracking training efficiency

To contextualize the training requirements of our temporal memory, we evaluate the parameter scale and sample efficiency of TORUS compared to recent baselines. As detailed in Table 10, our ResNet-50 variant operates at a comparable parameter scale to both OVTR and COVTrack.

By leveraging the continuous state-space memory to adapt a frozen spatial backbone, TORUS requires only 120k tracking-specific training samples. In contrast, OVTR requires approximately 1.5M samples. COVTrack utilizes the C-TAO dataset, which consists of over 490k annotated images, resulting in roughly 4.9M samples processed over its training schedule. While we acknowledge that both OVTR and COVTrack jointly optimize for spatial detection alongside tracking, our comparison highlights the efficiency of adapting an existing spatial foundation model into a temporal tracker—a strategy that significantly reduces the computational overhead required for training. Despite this drastic reduction in tracking-specific supervision, TORUS achieves superior Association Accuracy (AssocA) across benchmarks and maintains real-time inference speeds of 14.7 FPS on a NVIDIA RTX 4090 GPU, demonstrating an efficient translation of static spatial priors into stable trajectories.

Table 10: **Tracking training efficiency and parameter scale.** Comparison of model capacity and the total number of tracking-specific samples processed during temporal training. TORUS achieves robust temporal association while requiring significantly fewer tracking samples than current baselines.

Method	Backbone	Parameters	Tracking Samples
OVTR	ResNet-50	58.8M	$\sim$1.5M
COVTrack	ResNet-50	74.4M	$\sim$4.9M
TORUS (ours)	ResNet-50	71.3M	120k

A.6 Detailed Model Architecture

Figure 4 Shows the full TORUS architecture of a single time step.

TORUS builds upon the open-vocabulary multi-modal foundation model MOBIUS. The multi-scale visual-linguistic feature extractor remains frozen to preserve broad open-vocabulary priors, while the transformer decoder and the sequence query updater are fine-tuned for temporal tracking. Table 11 summarizes the key architectural hyperparameters.

Foundation Backbone and Bottleneck (Frozen)

- **Visual Backbone:** Standard ResNet-50 pre-trained on ImageNet. Multi-scale feature maps from stages $\{C_2, C_3, C_4, C_5\}$ with strides $\{4, 8, 16, 32\}$ are extracted and frozen.
- **Visual-Linguistic Bottleneck (VLBottleneck):** A 9-layer multi-scale deformable transformer encoder fusing multi-scale visual features with textual representations from a frozen CLIP text encoder ($d_{lang} = 512$). It outputs 4 multi-scale feature levels $\{P_2, P_3, P_4, P_5\}$ with a uniform hidden dimension $d_{model} = 256$.

Disentangled Transformer Decoder

The transformer decoder consists of $L_{dec} = 9$ stacked deformable attention layers following the MaskDINO selective decoder formulation.

Query Initialization and Composition. At frame t, the decoder processes the joint query set $Q^t = [Q_{tr}^{t-1}, Q_{det}^t]$:

- **Detection Queries (Q_{det}^t):** Following the two-stage MaskDINO paradigm, dense proposals are extracted from encoder memory features.
- **Track Queries (Q_{tr}^{t-1}):** Up to $N_{tr} \leq 500$ active track instances propagated from the previous frame.
- **Joint Attention:** In every decoder layer, all queries interact via standard multi-head self-attention ($n_{heads} = 8$, $d_{model} = 256$, dropout $= 0.0$), enabling mutual inhibition between overlapping track and detection queries.
- **Multi-Scale Cross-Attention:** Each query samples key points across all 4 feature levels using Multi-Scale Deformable Attention (`MSDeformAttn`) with $n_{points} = 4$ sampling points per head per level.

Functional Head Disentanglement. We functionally separate class-agnostic spatial localization from semantic categorization using independent prediction heads:

1. **Class-Agnostic Confidence Head (s):** A 2-layer MLP ($\mathbb{R}^{256} \to \mathbb{R}^{256} \xrightarrow{\text{ReLU}} \mathbb{R}^1$) predicting a scalar objectness logit representing target existence independent of category.
2. **Semantic Classification Head (z):** A linear projection $W_{cat} \in \mathbb{R}^{256\times512}$ mapping decoder output embeddings to the multi-modal joint space, followed by temperature-scaled cosine similarity against prompted CLIP text embeddings $T \in \mathbb{R}^{C\times512}$ with a learnable logit scale τ. Supervised exclusively on matched foreground targets with Binary Focal Loss ($\alpha = 0.5, \gamma = 2.0$).
3. **Bounding Box Head:** A 3-layer MLP ($\mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^4$) shared across all decoder layers, iteratively predicting box offsets $\Delta[c_x, c_y, w, h]$ in normalized inverse-sigmoid space.
4. **Mask Head:** A 3-layer MLP ($\mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^{256}$) combined via tensor contraction (Einsum) with high-resolution $1/4$-scale mask features from the pixel decoder.

Selective State Space Query Updater (Mamba)

The query updater models inter-frame temporal kinematics and updates track query representations without relying on frame-by-frame visual appearance matching.

Input Embedding and Spatial Conditioning. For each active track instance $i \in \{1, \ldots, N_{tr}\}$ at frame t, the input combines:

- The refined decoder output embedding $e_i^t \in \mathbb{R}^{256}$.
- The updated spatial reference box $b_i^t \in [0,1]^4$, mapped via sinusoidal positional encoding into $\mathbb{R}^{512}$ and transformed by a 2-layer positional MLP ($\mathbb{R}^{512} \to \mathbb{R}^{256} \xrightarrow{\text{ReLU}} \mathbb{R}^{256}$) into positional embedding $p_i^t \in \mathbb{R}^{256}$.

The combined observation $e_i^t + p_i^t$ serves as the input to Mamba as defined in the main text.

Low-Confidence Observation Masking. During severe occlusion or motion blur, noisy decoder outputs can degrade temporal memory. If the fused confidence score falls below the threshold ($\hat{s}_i^t \leq \theta_{update}$, with $\theta_{update} = 0.4$), the visual observation is masked to zero: $e_i^t + p_i^t \leftarrow \mathbf{0}$. The recurrence continues through the state space model, allowing Mamba's internal continuous state h_{t-1} to extrapolate the tracklet using learned kinematic motion priors alone.

Mamba SSM Sequence Block. The sequence backbone consists of $N_{ssm} = 4$ stacked Residual Mamba Blocks:

- **Input / Output Dimension (d_{model}):** 256.
- **State Dimension (d_{state}):** 16 (SSM latent memory capacity per channel).
- **Expansion Factor (E):** 2 (internal projection dimension $= 2 \times 256 = 512$).
- **Causal 1D Convolution (d_{conv}):** Kernel width 4 with bias, capturing local inter-frame velocity and acceleration.
- **Normalization:** LayerNorm applied at the block input and inside the gated branch prior to the SSM discretizer.

Recurrent Query Refinement. The output temporal adjustment $\Delta q_i^t \in \mathbb{R}^{256}$ is processed by a lightweight Feed-Forward Network ($d_{ffn} = 128$) and added to $Q_{tr,i}^{t-1}$ before final LayerNorm. Gradients are backpropagated through time (BPTT) across all $T = 10$ training frames without detachment, enabling end-to-end learning of track survival and motion dynamics.

A.7 Further qualitative results

In this section, we provide further qualitative visualizations of TORUS's tracking performance to supplement our quantitative benchmarks. Figure 5 demonstrates our model's capability in a closed-set scenario, showcasing stable temporal propagation and precise localization. Additionally, Figure 6 results illustrate TORUS's robustness in unconstrained open-vocabulary environments, where the continuous state-space memory successfully preserves physical object identities despite severe semantic crowding.

A.8 Detailed dataset sampling ratios

To ensure complete reproducibility and provide transparency regarding our heterogeneous training mixture, we detail the exact dataset sampling ratios used during the joint training phase.

As discussed in Section 3.4, balancing expansive spatial discovery with robust temporal dynamics requires careful calibration. Table 12 outlines the specific sampling weights assigned to each dataset. Massive static image datasets are heavily weighted to preserve the foundation model's large-vocabulary capacity and spatial priors. Conversely, video datasets are sampled proportionally to impart complex kinematic tracking priors without unlearning the semantic representations.

Table 11: **Architecture and training hyperparameters of TORUS (ResNet-50 variant)**.

Hyperparameter	Value
Backbone (Frozen)	
Architecture	ResNet-50
Feature Strides	$\{4, 8, 16, 32\}$
Text Projection Dimension	512
Bottleneck (Frozen)	
Architecture	VLBottleneck (Deformable Transf. Encoder)
Encoder Layers	9
Transformer Decoder	
Decoder Layers (L_{dec})	9
Hidden Dimension (d_{model})	256
Feed-Forward Dimension (d_{ffn})	2048
Multi-Head Attention Heads (n_{heads})	8
Deformable Points per Level (n_{points})	4
Detection Queries (N_{det})	300
Max Track Queries (N_{tr}^{max})	500
Query Merge Layer	0 (Full cross-query interaction)
Dropout	0.0
Prediction Heads	
Confidence Head (s)	MLP: $\mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^{1}$
Classification Head (z)	Linear: $\mathbb{R}^{256} \to \mathbb{R}^{512}$
Box Head	MLP: $\mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^{4}$
Mask Head	MLP: $\mathbb{R}^{256} \to \mathbb{R}^{256} \to \mathbb{R}^{256}$
Query Updater (Mamba)	
Sequence Model	Mamba State-Space Model
Number of Layers (N_{ssm})	4
Hidden State Dimension (d_{state})	16
Expansion Factor (E)	2 (Inner dimension: 512)
1D Convolution Kernel (d_{conv})	4
Positional Encoder	Sinusoidal + MLP($512 \to 256 \to 256$)
FFN Dimension	128
Observation Update Threshold (θ_{update})	0.4
Optimization & Data	
Training Sequence Length (T)	10 frames
Base LR / Weight Decay	1×10^{-4} / 0.05
Effective Batch Size	16 sequences (Acc. Steps: 4)
Iterations / Schedule	30,000 (decays at 10k, 25k)

Table 12: Heterogeneous training data composition and corresponding sampling weights.

Dataset	Modality	Sampling Weight
OpenImages	Static Image	10
Objects365 v2	Static Image	10
COCO 2017	Static Image	10
LVIS v1	Static Image	10
BDD100K (Detection)	Static Image	5
LVVIS	Video	8
OVIS	Video	5
BDD100K (MOTS)	Video	4
DanceTrack	Video	2
BFT	Video	1

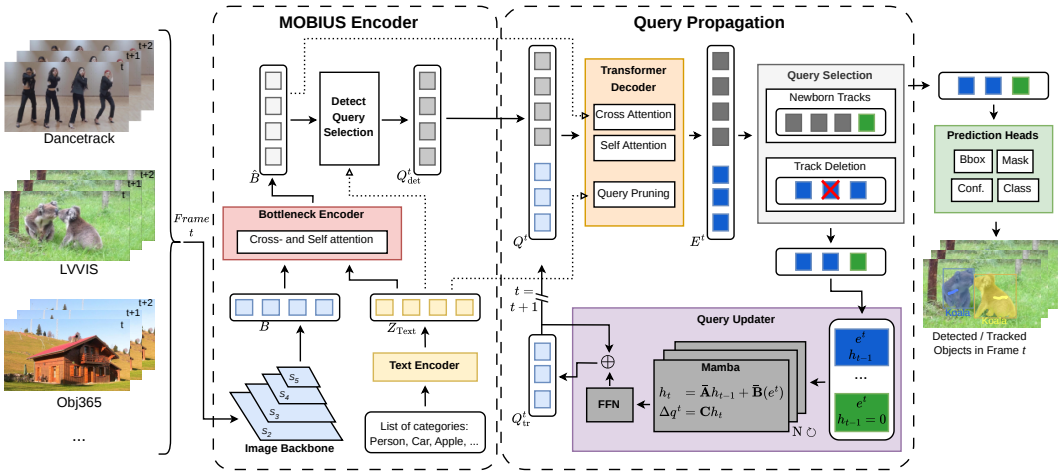

Figure 4: **Detailed architecture of the TORUS framework for a single time step t.** The model builds upon the frozen **Default MOBIUS** foundation model, which leverages an Image Backbone, a Text Encoder, and a Bottleneck Encoder to extract language-guided features, ultimately outputting semantic detection queries (Q^t_{det}). To propagate trajectories, the **Transformer Decoder** jointly processes these newborn detection queries alongside the active temporal track queries (Q^t_{tr}). The **Query Selection** module then manages the track lifecycle, initializing highly confident detections as newborn tracks and purging lost targets via track deletion. Finally, the **Query Updater** models the non-linear kinematics of the surviving tracks by passing each embedding through N Selective State Space (Mamba) blocks and an FFN, applying a residual connection from the previous query embedding to generate the updated track queries for the subsequent frame.



Figure 5: **Closed set qualitative results of TORUS.** TORUS is only prompted with present classes.



Figure 6: **Qualitative results demonstrating open vocabulary tracking in crowded scenes.** TORUS is prompted with all 1,203 LVIS classes.