HYPERDAM: HYPERSPECTRAL DISTRACTOR-AWARE MEMORY WITH AMODAL EXPANSION FOR SAM 3 TRACKING

Ryoga Yuzawa, Tasuku Takagi

Independent Researcher, Japan

ABSTRACT

Hyperspectral video provides material cues that can disambiguate targets with similar false-color appearance, yet foundation-model trackers update memory primarily from spatial and appearance evidence. We present HyperDAM, a DAM4SAM3-based hyperspectral tracker with three principal contributions. First, HOTC2026-Modal adds human-verified frame-wise modal masks and mask-tight boxes to all 481 organizer-provided HOTC 2026 videos. Second, a frame-zero-calibrated HSI gate rejects spectrally inconsistent updates to the distractor-resolving memory (DRM) without altering the current prediction. Third, a causal spatiotemporal expander adds outward-only amodal corrections from frozen SAM features. Static-scene recovery and empty-mask RTS smoothing address target switches and full occlusion. Model selection prioritizes cross-domain robustness over leaderboard-specific optimization. The final system ranked second in HOTC 2026, achieving 68.0093% AUC and 87.7703% DP@20 in the organizer's private evaluation.

***Index Terms*—** hyperspectral object tracking, modal segmentation dataset, spectral identity, amodal bounding box

1. INTRODUCTION

Single-object tracking estimates a target trajectory from an initial bounding box. Hyperspectral video augments conventional appearance with densely sampled spectral responses, providing evidence about material identity when color or texture alone is ambiguous. In camouflaged scenes, a target and its background may have nearly identical visual appearances while retaining distinct spectral signatures, making spectral evidence essential when appearance is unreliable [1]. Existing hyperspectral benchmarks and trackers demonstrate the value of spectral–spatial modeling [2, 3, 4, 5], while limited training data and the spectral-band gap continue to favor adaptation of strong RGB trackers [6].

Promptable segmentation models provide a complementary path. The Segment Anything family supports transferable segmentation from visual prompts [7, 8], and DAM4SAM improves tracking robustness by managing distractors in a dedicated memory [9]. Two gaps remain in the HOTC setting. First, an incorrect high-confidence mask can still enter memory and contaminate future predictions. Second, a segmentation mask describes only the visible part of an occluded target, whereas the required tracking box may extend beyond the visible support.

Figure 1 previews how the proposed output trajectory handles the latter gap. The learned expander restores latent extent while the target is partially visible, and RTS supplies a box only when the SAM mask is absent.

We address these gaps with HyperDAM, a system that separates three responsibilities: spectral evidence protects target identity, a learned head estimates latent object extent, and the underlying frozen tracker owns the base trajectory. Our contributions are:

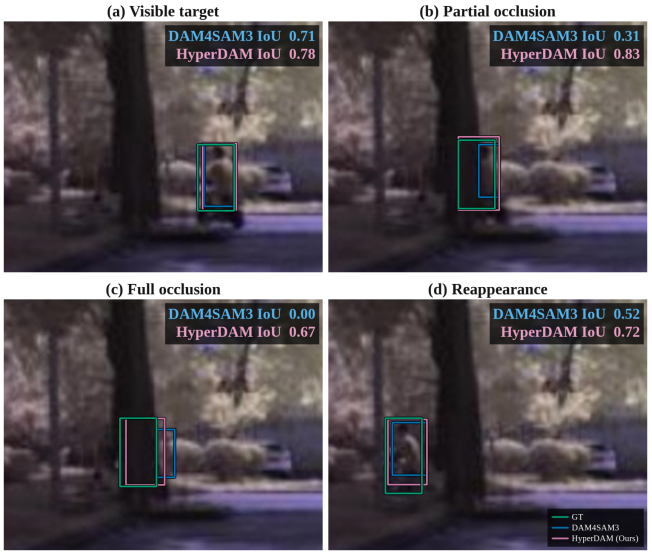


Fig. 1. HyperDAM through an occlusion on `nir-rider6`. From left to right and top to bottom, the target is visible, partially occluded, fully occluded, and visible again. The amodal expander corrects partial visibility, while empty-mask RTS bridges full occlusion. Green, blue, and pink denote ground truth, DAM4SAM3, and HyperDAM, respectively.

- **A reviewed modal-annotation extension.** We augment the organizer-provided HOTC 2026 competition dataset with binary visible/modal masks and mask-tight boxes for the 406 training/update videos and the organizer's 75-video public-leaderboard set (denoted Public-LB75).
- **A hyperspectral memory gate.** A sequence-local model calibrated from the frame-zero target evaluates material consistency at DAM4SAM's native distractor-resolving memory (DRM) update boundary. It can reject a requested update but cannot request one or alter the current-frame output.
- **An amodal expander for DAM4SAM3 boxes.** A causal six-frame 3-D head trained from paired modal masks and challenge boxes predicts nonnegative per-side expansion. The expansion is transferred as a residual onto the HSI-controlled trajectory and is never fed back into SAM memory.

Static-scene recovery and empty-mask Rauch–Tung–Striebel (RTS) smoothing [10] are supporting components rather than pri-

E-mail: ryoga.yuzawa@gmail.com, tasuku.takagi@gmail.com

mary contributions. They are included to specify the submitted system completely.

2. RELATED WORK

2.1. Hyperspectral object tracking

Hyperspectral tracking has progressed from material descriptors and correlation filters to learned spectral–spatial representations [2, 3, 1]. Recent methods transfer RGB trackers or pretrained representations through material-aware fusion and spectral prompts [4, 5]. WHISPERS entries have explored detection-driven SAM2, low-rank adaptation, SAM2-based HOT, and motion-aware memory [11, 12, 13, 14]. Our spectral component instead acts as a sequence-local, label-free decision layer: it does not retrain the tracker and intervenes only when the native tracker requests a DRM update.

2.2. Foundation models for segmentation and tracking

SAM 2 introduced promptable video segmentation with streaming memory [7], while SAM 3 unified detection, segmentation, and tracking from concept and visual prompts [8]. DAM4SAM augments the video tracker with a distractor-aware memory and an introspection-based update policy [9]. Our HSI gate tests material identity at DAM4SAM's final DRM update boundary.

2.3. Modal and amodal supervision

Modal annotations describe visible pixels, while amodal annotations represent the complete object extent behind occluders. TAO-Amodal shows that heavy occlusion remains challenging for modern trackers [15]. Our extension adds dense modal masks and mask-tight boxes to the organizer-provided HOTC 2026 competition videos. The expander learns the difference between this visible extent and the challenge-provided target box, while retaining a strict outward-only correction contract.

3. HOTC 2026 DATASET EXTENSION

HOTC 2026 provides 406 training/update videos and the 75-video Public-LB75 split with paired hyperspectral observations. The former have frame-wise amodal target boxes, whereas leaderboard ground truth is withheld. We extend all 481 videos with human-verified visible/modal segmentation masks and their mask-tight boxes, and release these annotations as HOTC2026-Modal. This annotation layer supplements rather than replaces the organizer-provided data. Paired with the official hyperspectral and false-color frames, the masks and boxes can supervise spectral adaptation of segmentation models, domain adaptation to false-color imagery, and losses that distinguish modal observations from amodal target extent. Existing HOT benchmarks are predominantly box-annotated and remain modest in scale [1]; our extension targets the missing dense modal supervision rather than introducing a new tracking split.

DAM4SAM3 predicts a visible mask and hence a modal box, while HOTC supervision specifies the object's amodal extent. Directly using the latter as a mask or visible-box target conflates observed pixels with hidden extent. Our paired modal masks and boxes separate these quantities, enabling the amodal expander to learn outward residuals from modal to amodal boxes while SAM3 and the base tracker remain frozen.

SAM3 first infers and propagates target masks. Annotators visually inspect its outputs and manually correct erroneous frames or regions; only reviewed masks enter the release. Each modal box is then computed as the tight axis-aligned box around visible foreground pixels. An empty mask denotes full occlusion. The release also records official frame identifiers, available challenge boxes, provenance, and checksums.

The release contains a 406-video primary split and a separate 75-video Public-LB75 annotation-only split. Because official leaderboard ground truth is not released, the latter annotations are not presented as challenge ground truth and are excluded from fitting, model selection, and component evaluation.

4. METHOD

The proposed system uses frozen DAM4SAM3 as its base tracker and augments it with three principal components. First, an HSI-gated memory mechanism controls whether a requested DRM update is committed, preventing spectrally inconsistent objects from contaminating the tracker state. Second, an amodal expander enlarges the native modal box toward the estimated full extent of a partially occluded object. Third, an offline Rauch–Tung–Striebel (RTS) smoother fills fully occluded frames for which the tracker produces no valid mask. We additionally exploit the prevalence of fixed-view cameras in the dataset by classifying each scene as static or dynamic. In static scenes, abrupt and spatially asymmetric box growth is treated as a likely tracking error and triggers static-scene recovery. The HSI gate is the only component that modifies the DRM update policy. Static-scene recovery may reinitialize the base tracker, whereas amodal expansion and RTS smoothing operate only on the output trajectory and never feed back into DAM4SAM3 memory. Figure 2 summarizes the complete inference pipeline.

4.1. DAM4SAM3 base tracker

Our base tracker adapts the distractor-aware DAM4SAM memory design [9] to frozen SAM3. It runs in PVS-only mode with detector inference disabled and is initialized by the official frame-zero box. SAM3 propagates the target mask and exposes alternative mask hypotheses; DAM4SAM3 uses these outputs to manage recent appearance memory and DRM. The tight box of the selected mask is the native modal trajectory. We use the entire base tracker in frozen form, with all SAM3 parameters unchanged. We initially explored LoRA adaptation using the HOTC training data [16], but the limited size and diversity of the dataset did not yield the expected generalization to held-out sequences. We therefore retain the frozen model and instead introduce the HSI-based memory gate described below.

4.2. HSI gate

DAM4SAM3 first decides when to request a DRM update. For each request, the HSI gate evaluates the proposed update mask M_t together with an expanded surrounding region $E_t = \mathrm{Dilate}(M_t) \backslash M_t$. First, the raw sensor mosaic I_t^{raw} is decoded into an aligned HSI cube $X_t \in \mathbb{R}^{H \times W \times B}$:

$$X_t = \mathcal{D}(I_t^{\mathrm{raw}}). \tag{1}$$

Here B is the number of spectral bands: $B = 16$ for VIS, $B = 25$ for NIR, and $B = 15$ for RedNIR. Let $p_0 = \mathrm{median}_{x \in M_0} X_0(x)$ be the reference spectrum from the frame-zero target mask. For each

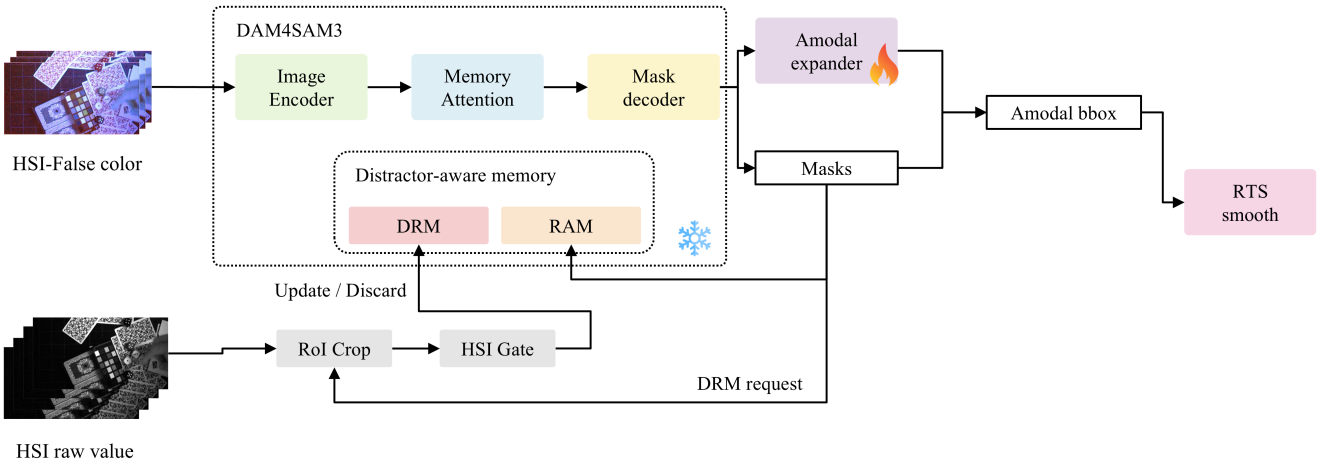


Fig. 2. Overview of the proposed HyperDAM architecture. The HSI gate filters requested DRM updates using raw hyperspectral observations; DAM4SAM3 supplies modal masks and features for amodal expansion, followed by offline RTS smoothing.

requested DRM update, the component-wise median spectra of the candidate mask and expanded region are

$$\begin{aligned} p_t^M &= \underset{x\in M_t}{\text{median}}\, X_t(x), \qquad p_t^E = \underset{x\in E_t}{\text{median}}\, X_t(x), \\ s_t^M &= \cos\left(\widehat{p_t^M}, \widehat{p_0}\right), \qquad s_t^E = \cos\left(\widehat{p_t^E}, \widehat{p_0}\right). \end{aligned} \tag{2}$$

Here $\widehat{\cdot}$ denotes mean centering across spectral bands followed by ℓ_2 normalization. The DRM update is accepted only when the candidate is similar to the initial target and its similarity exceeds that of the expanded region:

$$\begin{aligned} A_t^{\text{HSI}} &= \mathbb{1}\left[s_t^M \geq \tau_{\text{id}} \ \wedge \ s_t^M - s_t^E \geq \tau_{\text{exp}}\right], \\ W_t &= U_t \wedge A_t^{\text{HSI}}, \end{aligned} \tag{3}$$

Here U_t is DAM4SAM3's native DRM update request. The thresholds are calibrated from the frame-zero target and its expanded region. The gate may reject the requested DRM update but cannot request an update itself; rejection leaves both the current output and DRM unchanged.

No cross-sequence training or ground-truth annotation is used by the HSI gate. An unreliable frame-zero reference abstains rather than rejecting.

Figure 3 shows an example. At $t = 124$, DAM4SAM3 requests and performs a DRM update, whereas the HSI gate rejects it because the candidate mask does not separate sufficiently from its surroundings. The current boxes remain unchanged, but the different memory states produce the later divergence at $t = 236$–237.

4.3. Amodal expander for SAM

The SAM family predicts the visible segmentation mask of a target, and its bounding box is therefore the tight envelope of that modal mask. HOTC instead evaluates the amodal extent, including portions hidden by occlusion, so a SAM-based tracker requires an explicit modal-to-amodal adaptation. Motivated by the plug-in expander in TAO-Amodal [15], we preserve all pretrained SAM3 weights and attach an external amodal expander head to the detached outputs of the DAM4SAM3 mask decoder. Only this head is trained to infer outward box corrections for the hidden extent.

For a SAM mask box $B_t = (x_t, y_t, w_t, h_t)$, the expander predicts nonnegative outward residuals $\Delta_t = (\delta_{\text{L}}, \delta_{\text{T}}, \delta_{\text{R}}, \delta_{\text{B}})$ in pixel units. The amodal box is obtained by directly adding these side residuals:

$$\begin{aligned} \text{Amodal}(B_t; \Delta_t) = \big(&x_t - \delta_{\text{L}},\ y_t - \delta_{\text{T}}, \\ &w_t + \delta_{\text{L}} + \delta_{\text{R}},\ h_t + \delta_{\text{T}} + \delta_{\text{B}}\big). \end{aligned} \tag{4}$$

Training targets are the outward differences between the frozen SAM box and the official amodal box. A side is supervised as positive only when the reviewed modal box indicates genuine occlusion, preventing expansion when SAM already reaches or overshoots the official boundary.

The expander is causal and temporal. From the final 32-channel feature map of the SAM3 mask decoder, it extracts a box-aligned $32 \times 24 \times 24$ crop at each frame and stacks the latest six crops. Aligned mask logits, the 256-dimensional object query, box geometry, and an immutable frame-zero anchor are supplied alongside this feature history. Six residual 3-D convolutional blocks aggregate the spatiotemporal information, after which four side heads predict the left, top, right, and bottom residuals in Eq. (4). Confidence gates set unsupported side corrections to zero.

The external head has 9.4M parameters and is pretrained on synthetic MOVi-MC-AC amodal data [17]. All SAM3 parameters remain frozen. Frames with full occlusion are excluded from expander training, which therefore targets partial occlusion where a visible SAM box can be expanded toward the amodal box.

4.4. Full-occlusion handling

The amodal expander moves partially occluded boxes toward the amodal ground-truth extent, but learning reliable boxes under full occlusion is difficult because no visible mask remains. We therefore handle full occlusion after the full video has been processed. A frame is marked as a missing observation when the independent frozen DAM4SAM3 pass emits an empty mask; the released pipeline does not apply an additional confidence threshold. Rauch–Tung–Striebel (RTS) smoothing [10] then reconstructs only these missing boxes from valid observations before and after the gap, leaving all observed boxes unchanged.

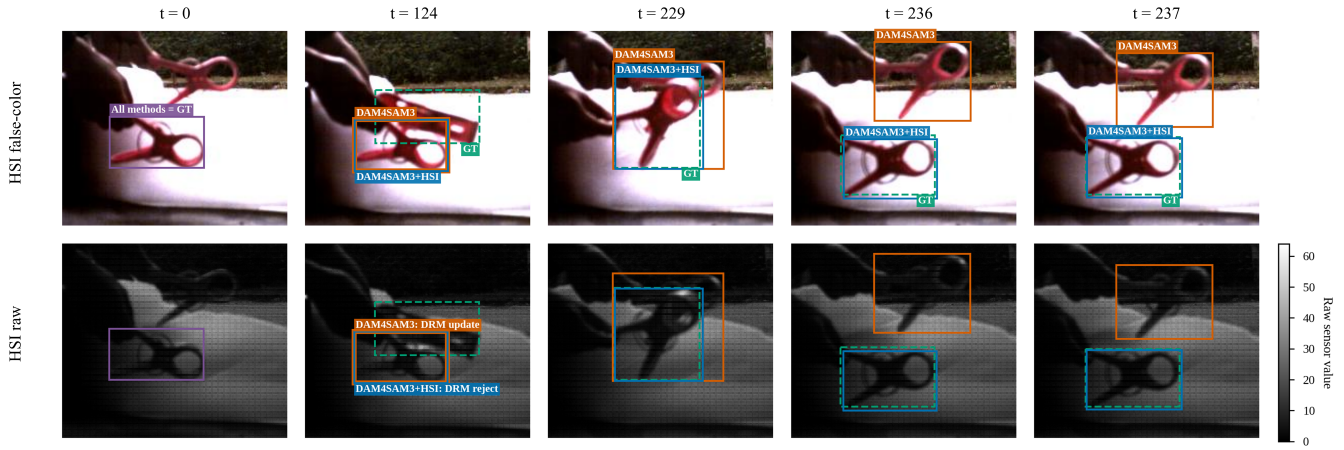


Fig. 3. Runtime trace of the HSI gate on `vis-clamp2` from the HOTC 2026 dataset. The rows show aligned false-color frames and raw 16-band mosaics. At $t = 124$, DAM4SAM3 accepts the DRM update, whereas DAM4SAM3+HSI rejects it without changing the current output. By $t = 236$–237, the native tracker follows the visually similar distractor while the gated tracker remains on the target. Orange, blue, and green dashed boxes denote DAM4SAM3, DAM4SAM3+HSI, and ground truth, respectively.

Table 1. Cumulative ablation on Public-LB75.

ID	Configuration	Public score (AUC)
B0	DAM4SAM3 native	0.69039
B1	B0 + HSI gate	0.69214
B2	B1 + static-scene recovery	0.69513
B3	B2 + amodal expander	0.69732
B4	B3 + RTS smoothing	**0.69768**
Organizer's private LB		AUC: 68.0093% DP@20: 87.7703%

Figure 1 illustrates the complementary roles of the two components. The amodal expander recovers the hidden extent while part of the target remains visible, whereas RTS smoothing supplies the box when the target is fully occluded and the SAM mask is empty.

4.5. Static/dynamic scene classification

Our analysis found that many HOTC sequences use a fixed camera and that target switches in such scenes are often accompanied by an abrupt, spatially asymmetric change in box size. We therefore estimate background motion from sparse trackable points over the initial ten frames [18], then classify each sequence as static or dynamic.

5. EXPERIMENTS

5.1. Contribution of each component

Table 1 reports a cumulative ablation on the 75-video Public-LB75 split (26,860 frames) using the same frozen DAM4SAM3 checkpoint and inference revision. Every ablation row was submitted to the organizer's public Kaggle evaluation. The final private-leaderboard result is included in the same table for reference.

5.2. Benchmark

Table 2 reports all verified results on the same Public-LB75 protocol. Each method is initialized with the provided first-frame box and produces one prediction per frame. All models are evaluated in FP32 on a single NVIDIA L4 GPU. Although HyperDAM was selected with cross-domain generalization rather than leaderboard optimization as the primary objective, it achieves the highest AUC among the compared methods and improves upon its DAM4SAM3 base. This result indicates that the proposed components improve the base tracker without sacrificing performance on the public evaluation domain.

Public-LB75 ground truth is withheld after the initial box and Kaggle returns only AUC. We therefore do not import published values obtained with different sequence sets or metrics into this benchmark. Table 2 also gives peak allocated VRAM and parameter count where available.

5.3. Discussion

The gains in Table 1 reflect complementary roles: the HSI gate prevents visually similar distractors from contaminating memory (Figure 3), static-scene recovery handles target switches, the amodal expander handles partial occlusion, and RTS fills full-occlusion gaps. Their individually modest gains therefore accumulate across distinct failure modes.

5.4. Limitations

The HSI gate yields only a modest gain because it only rejects requested DRM updates. Future work could initialize an object detector and associate multiple candidate spectral prototypes with the target by HSI correlation, moving material reasoning from memory validation to candidate selection.

6. CONCLUSION

We presented HyperDAM, extending frozen DAM4SAM3 with a frame-zero HSI update gate, outward-only amodal expansion, static-scene recovery, and empty-mask RTS, together with HOTC2026-

Table 2. Unified benchmark on Public-LB75 (75 videos; 26,860 frames). Public AUC is evaluated on the Kaggle public leaderboard. Available peak-VRAM measurements use the same NVIDIA L4 GPU. The best AUC value is shown in bold.

Method	Public-LB75 AUC ↑	Peak VRAM (MiB) ↓	Params (M) ↓	Precision	Comment
OSTrack [19]	0.49512	443.24	92.83	FP32	–
SeqTrack [20]	0.54142	1601.95	308.98	FP32	–
HIPTrack [21]	0.57374	557.88	120.41	FP32	–
ODTrack [22]	0.54442	599.99	92.83	FP32	–
SUTrack [23]	0.57169	2205.59	746.61	FP32	–
SAM2 [7]	0.65041	1634.95	224.45	FP32	–
DAM4SAM [9]	0.65464	2177.24	224.45	FP32	SAM2 base algorithm
SAM3 [8]	0.67393	4530.28	827.31	FP32	PVS
DAM4SAM3 [9, 8]	0.69039	4492.51	827.31	FP32	PVS
HyperDAM (Ours)	**0.69768**	4685.12	836.73	FP32	–

Modal annotations for all 481 HOTC 2026 videos. The system achieved the best Public-LB75 AUC among the evaluated methods while cross-domain selection limited leaderboard overfitting. Its organizer-private scores were 68.0093% AUC and 87.7703% DP@20, ranking second in HOTC 2026. Future work will explore detector-assisted spectral association and online trajectory completion.

7. REFERENCES

[1] Hanzheng Wang, et al., "BihoT: A large-scale dataset and benchmark for hyperspectral camouflaged object tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 9, pp. 16392–16406, 2025.

[2] Fengchao Xiong, et al., "Material based object tracking in hyperspectral videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3719–3733, 2020.

[3] Zhenqi Liu, et al., "Unsupervised deep hyperspectral video target tracking and high spectral-spatial-temporal resolution (H^3) benchmark dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[4] Yuzeng Chen, et al., "PHTrack: Prompting for hyperspectral video tracking," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[5] Gang He, et al., "Hyperspectral object tracking with spectral information prompt," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 12, pp. 12636–12651, 2025.

[6] Yuchao Wang, et al., "Deep feature-based hyperspectral object tracking: An experimental survey and outlook," *Remote Sensing*, vol. 17, no. 4, pp. 645, 2025.

[7] Nikhila Ravi, et al., "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[8] Nicolas Carion, et al., "SAM 3: Segment anything with concepts," in *Proc. ICLR*, 2026.

[9] Jovana Videnovic, et al., "A distractor-aware memory for visual object tracking with SAM2," in *Proc. IEEE/CVF CVPR*, 2025, pp. 24255–24264.

[10] Herbert E. Rauch, et al., "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, no. 8, pp. 1445–1450, 1965.

[11] Wangquan He, et al., "DeSAM2: Detection-driven SAM2 for hyperspectral object tracking," in *Proc. IEEE WHISPERS*, 2024.

[12] Langkun Chen, et al., "Hyperspectral object tracking with low-rank adaptation," in *Proc. IEEE WHISPERS*, 2024.

[13] Zixuan Qian, et al., "Track hyperspectral videos via segment anything model 2," in *Proc. IEEE WHISPERS*, 2024.

[14] Yuzeng Chen, et al., "HARMONY: Adapting segment anything model for hyperspectral video tracking with motion-aware memory," in *Proc. IEEE WHISPERS*, 2025.

[15] Cheng-Yen Hsieh, et al., "TAO-Amodal: A benchmark for tracking any object amodally," *arXiv preprint arXiv:2312.12433*, 2024.

[16] Edward J. Hu, et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[17] Alexander Moore, et al., "Training for X-Ray vision: Amodal segmentation, amodal content completion, and view-invariant object representation from multi-camera video," *arXiv preprint arXiv:2507.00339*, 2025.

[18] Jianbo Shi et al., "Good features to track," in *Proc. IEEE CVPR*, 1994, pp. 593–600.

[19] Botao Ye, et al., "Joint feature learning and relation modeling for tracking: A one-stream framework," in *Proc. ECCV*, 2022, pp. 341–357.

[20] Xin Chen, et al., "SeqTrack: Sequence to sequence learning for visual object tracking," in *Proc. IEEE/CVF CVPR*, 2023, pp. 14572–14581.

[21] Wenrui Cai, et al., "HIPTrack: Visual tracking with historical prompts," in *Proc. IEEE/CVF CVPR*, 2024, pp. 19258–19267.

[22] Yaozong Zheng, et al., "ODTrack: Online dense temporal token learning for visual tracking," in *Proc. AAAI*, 2024, vol. 38, pp. 7588–7596.

[23] Xin Chen, et al., "SUTrack: Towards simple and unified single object tracking," in *Proc. AAAI*, 2025, vol. 39, pp. 2239–2247.