PanoSAM2: Lightweight Distortion- and Memory-aware Adaptions of SAM2 for 360 Video Object Segmentation

Dingwen Xiao[1⋆], Weiming Zhang[1⋆], Shiqi Wen[1], and Lin Wang[2†]

[1] The Hong Kong University of Science and Technology (Guangzhou)
{dxiaoaf, wzhang915, swen750}@connect.hkust-gz.edu.cn
[2] EmPACT Lab, School of EEE, Nanyang Technological University
linwang@ntu.edu.sg

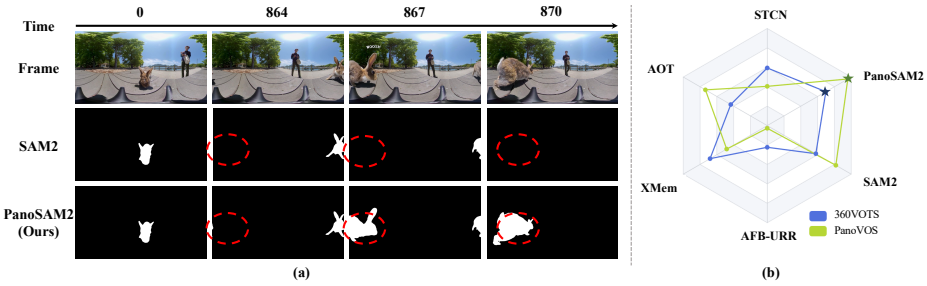

Fig. 1: Our PanoSAM2 achieves superior results on 360 Video Object Segmentation via spherical distortion and geometry adaptations of SAM2. (a) Sample results from SAM2 and PanoSAM2 on the 360 panoramic video frames, showing how PanoSAM2 better segments the target across different time frames (red dash circle). (b) Plot comparing the performance of PanoSAM2 against existing methods on 360VOTS [34] and PanoVOS [36] (star icon for the best model), demonstrating its superior video segmentation capabilities.

Abstract. 360 video object segmentation (360VOS) aims to predict temporally-consistent masks in 360 videos, offering full-scene coverage, benefiting applications, such as VR/AR and embodied AI. Learning 360VOS model is nontrivial due to the lack of high-quality labeled dataset. Recently, Segment Anything Models (SAMs), especially SAM2 – with its design of memory module – shows strong, promptable VOS capability. However, directly using SAM2 for 360VOS yields implausible results as 360 videos suffer from the projection distortion, semantic inconsistency of left-right sides, and sparse object mask information in SAM2's memory. To this end, we propose **PanoSAM2**, a novel 360VOS framework based on our lightweight distortion- and memory-aware adaptation strategies of SAM2 to achieve reliable 360VOS while retaining SAM2's user-friendly prompting design. Concretely, to tackle the projection distortion and semantic inconsistency issues, we propose a **Pano-Aware**

⋆ Equal contribution. †Corresponding author.

Decoder with seam-consistent receptive fields and iterative distortion refinement to maintain continuity across the 0°/360° boundary. Meanwhile, a **Distortion-Guided Mask Loss** is introduced to weight pixels by distortion magnitude, stressing stretched regions and boundaries. To address the object sparsity issue, we propose a **Long–Short Memory Module** to maintain a compact long-term object pointer to reinstantiate and align short-term memories, thereby enhancing temporal coherence. Extensive experiments show that PanoSAM2 yields substantial gains over SAM2: +**5.6** on 360VOTS and +**6.7** on PanoVOS, showing the effectiveness of our method. Code is available at https://github.com/Eric-Bumaro/PanoSAM2.

1 Introduction

360 video object segmentation (360VOS) aims to track and segment target objects across 360 videos, given their mask in the first frame. Unlike conventional perspective VOS with a limited field-of-view (FoV), 360 videos – with the commonly used projection type of Equirectangular projection (ERP) – can observe all entire spherical scene, maintaining awareness of objects from all directions. This property makes it particularly valuable for immersive applications such as VR/AR [8,10], autonomous driving [26,31,36,43], and embodied robotics [9,41], which demand temporally consistent tracking with stable object identity. However, an ERP image, A.K.A., panorama[3] samples pixels with a higher density at the poles compared to the equator, resulting in spherical distortions. Moreover, the left–right ERP borders correspond to adjacent longitudes, causing the semantic inconsistency at the 0°/360° seam. These render the perspective 2D imagery-based VOS research [3–5,21,32,38,39] less applicable or effective. Only recently, a few studies [34,36] consider the geometric and photometric characteristics unique to omnidirectional capture. However, the benchmarks [34,36] for 360VOS remain much under-explored compared to their 2D counterparts [24,27,29,33], and their generalization capacity is limited. Bridging this gap requires rethinking both model architectures and optimization objectives to account for the distortion and spherical continuity in 360 imagery.

Recently, Segment Anything Model (SAM), especially SAM2 [29] is a prompt-based VOS foundation model that shows strong zero-shot, promptable capabilities thanks to its user-friendly interface (points, boxes, or masks) and a memory module trained over 50K+ videos. While SAM2 has inspired extensions across domains [19,45], directly applying it to 360VOS produces implausible results (see Fig. 1) caused by the distortion, left–right semantic inconsistency at the 0°/360° seam, and small masks that often appeared in the target objects, leading to sparse object evidence in SAM2's memory. Under such sparsity and occlusion, short-term memory can drift or forget, echoing prior observations in memory-based VOS [2, 4, 7, 11, 36, 39].

To address these challenges, we explore a novel idea: lightweight **distortion- and memory-aware adaptation** that preserves SAM2's promptable design

[3] In this paper, ERP and panoramic are interchangeably used.

while making it reliable for panoramic videos. Intuitively, we propose **PanoSAM2**, a novel 360VOS framework that can achieve robust 360VOS (see Fig. 1) with strong generalization via three interconnected technical components. Firstly, to tackle the spherical continuity and projection distortion, we propose a **Pano-Aware (PA) Decoder** that reshapes the mask decoding process (see Sec. 3.1). Concretely, it performs left–right wrap concatenation to build seam-consistent receptive fields, ensuring that features at the right border attend to their true neighbors at the left border (*i.e.*, the $0°/360°$ seam). Then, the decoder conditions on the previous-frame mask to apply iterative distortion refinement, progressively correcting features near high-distortion latitudes during decoding. This geometry-aware design remarkably *reduces seam breaks and improves boundary fidelity while remaining a lightweight decoder.*

On top, we introduce a **Distortion-Guided Mask Loss**, a geometry-aware objective that weights pixels by their distortion magnitude under ERP (see Sec. 3.3). It is a spherical weighting scheme with a novel BFoV-guided [34] supervision strategy designed to cooperate with the loss function. Intuitively, pixels in highly stretched regions (and near boundaries) contribute more to the loss, encouraging projection-robust masks and sharper boundaries. *The loss is simple to compute, architecture-agnostic, and complements the PA Decoder by aligning the optimization target with the ERP.*

To enhance temporal robustness under sparsity and occlusion, we propose a novel Long–Short Memory Module (LSMM) (see Sec. 3.2). It maintains a compact long-term object pointer—an object-level summary distilled from historical observations—that periodically re-instantiates and aligns the short-term memory. By injecting this pointer into memory attention alongside recent features, PanoSAM2 resists drift, rapidly recovers from occlusions, and prevents dominance by frames where the foreground is tiny or absent. Rather than relying on motion heuristics [37], it leverages the often-overlooked occlusion score for long-term memory selection and employs the compact object pointer of SAM2's design. *This design stabilizes identity while preserving responsiveness to new inputs, thereby improving temporal coherence in challenging panoramic scenes.*

We evaluate our PanoSAM2 on the 360VOTS [34] and PanoVOS [36], observing consistent gains over SAM2. In particular, PanoSAM2 improves by $+$**5.6** on 360VOTS test set and $+$**6.7** on PanoVOS validation set, indicating that panoramic geometry and memory constraints can be effectively addressed without discarding the promptable interface. Ablation studies attribute the improvements to three innovative components. Notably, these benefits come with modest overhead, preserving the efficiency for practical applications.

In summary, our contributions are **four-fold**:

- We make the **first** attempt to leverage SAM2's zero-shot, prompt-based paradigm and propose PanoSAM2, a novel approach for the challenging 360VOS by introducing a tightly integrated, architecture-specific design.
- We propose the Pano-Aware Decoder that enforces left–right semantic continuity and mitigates ERP distortion; a Distortion-Guided Mask Loss that aligns optimization with spherical sampling;

- We propose a Long–Short Memory Module coupling long-term pointers with short-term memory to mitigate drift under sparsity and occlusion.
- We show that PanoSAM2 shows state-of-the-art (SoTA) performance on diverse benchmark 360VOS datasets ($\mathcal{J}\&\mathcal{F}$ score of **65.8** on 360VOTS test set and **78.1** on PanoVOS validation set) and strong generalization capabilities. Notably, the proposed architecture exhibits strong cross-dataset generalization, revealing principles for robust 360VOS.

2 Related Works

2.1 360 Video Object Segmentation.

Compared with perspective settings, object tracking and segmentation in 360 videos remain underexplored. PanoVOS [36] introduced the 360VOS task, releasing a dedicated dataset (PanoVOS) and the PSCFormer baseline. PSCFormer addresses equirectangular projection (ERP) distortion and semantic consistency by applying left–right padding to preserve wrap-around continuity and by restricting pixel-level attention windows, thereby reducing cross-sphere confusion while maintaining efficiency. Besides, 360VOTS [34] proposes a Bounding Field-of-View (BFoV) mechanism that handcrafts the next-frame search region from the previous prediction, effectively "windowing" the panorama so that conventional VOS models can be used without architectural changes. While BFoV enables plug-and-play reuse of mature 2D pipelines, it introduces a runtime bottleneck due to sequential localized searches and is brittle when targets are heavily occluded, leave the window, or re-enter with large appearance or viewpoint shifts. We argue that methods tailored to ERP must simultaneously handle severe projection distortion, seam consistency, and sparsely distributed target pixels, challenges that are only partially addressed by padding and local attention. To this end, *we propose PanoSAM2 that explicitly incorporates geometry-aware decoding and memory while retaining SAM2 capabilities for panoramic streams.*

2.2 SAM-Based Video Object Segmentation

The SAM family [12, 29] established segmentation foundation models whose promptable design enables broad transfer. Building on this, several works [15,19, 20,28,44,45] adapt it to video. SAM-PT [28] couples a point-tracking model with SAM: tracked points on the target are fed as prompts to segment each frame, requiring no task-specific training. SAM-I2V [19] synthesizes prompts from temporal cues by combining past masks and current-frame features, then performs inference via SAM's prompt encoder and mask decoder. SAM2Long [7] observes track drift under occlusion, and models mask uncertainty by exploring a tree of hypotheses, selecting favorable branches; the analysis further highlights the value of long-term memory for stable segmentation. CamSAM2 [45] extends SAM2 to camouflaged object tracking, injecting camouflage tokens to derive object prototypes that correct current predictions. MAPS [37] explores the effect of adding

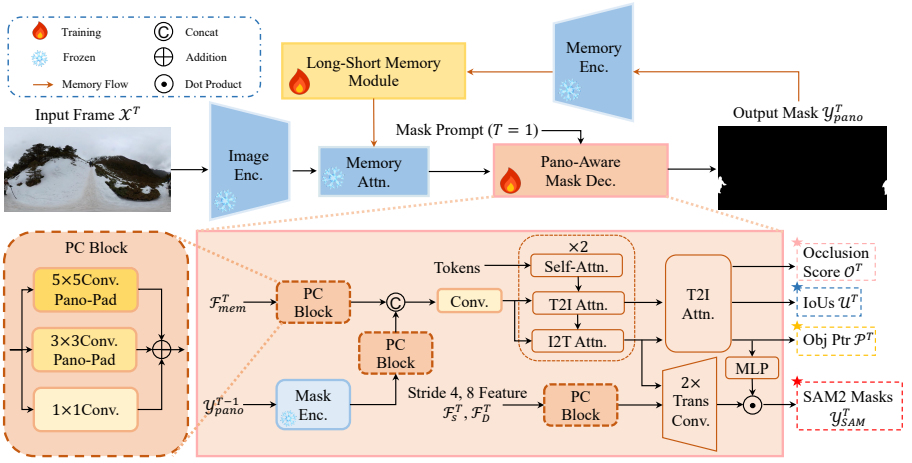

Fig. 2: Overview of our PanoSAM2 framework. Compared with SAM2 [29], it has two architectural contributions: Pano-Aware (PA) Decoder and Long-Short Memory Module (LSMM).

more representative frames to memory. While effective in perspective settings, *these approaches struggle in 360 video: affine priors become unstable across ERP distorted regions, and traditional receptive areas cannot handle semantic consistency for left and right boundaries. Consequently, direct application to panoramic streams yields degraded performance.* In contrast, our designs effectively fit the characteristics of 360 video and are integrated into SAM2's architecture, going beyond direct adaptation of existing 360 strategies.

3 Methodology

Preliminaries: SAM2. SAM2 [29] is a prompt-based foundation model for image and video object segmentation. It is trained on SA-1B [12] and SA-V [29] with over 50K+ videos and exhibits strong zero-shot capacity, allowing evaluation on unseen data without task-specific finetuning.

For the frame $\mathcal{X}^T \in \mathbb{R}^{H\times W\times 3}$ at time T, SAM2 input it into the Hiera image encoder [30] to extract feature and then memory-condition it to $\mathcal{F}^T_{mem} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 256}$ by operating cross-attention with memory from the memory bank. If it is the first frame, a prompt (point, box, or mask) is provided and encoded. $\mathcal{F}^T_{mem}$ is input into the mask decoder together with prompt information (if possible) to decode three mask logits $\mathcal{Y}^T_{SAM} \in \mathbb{R}^{H\times W\times 3}$, an object pointer $\mathcal{P}^T \in \mathbb{R}^{d_p}$ (d_p for pointer dimension) for object-level information, $\mathcal{U}^T \in \mathbb{R}^3$ that predicts the IoU scores of the three predicted masks with ground truth, and an occlusion score $\mathcal{O}^T \in \mathbb{R}$ for the probability of object being visible in this frame. In parallel, the mask with the maximum IoU score and unconditioned frame feature are combined by a memory encoder and, together with the $\mathcal{P}^T$, written

into the memory bank. By default, the bank retains only the most recent six frames' memory.

Our Idea. However, as SAM2 fails to model equirectangular distortion and left–right seam semantic continuity. Moreover, under heavy distortion or after occlusion, the short memory policy can forget the target or drift to another object, leading to identity breaks. To address these gaps, we propose PanoSAM2 for 360VOS, as depicted in Fig. 2. We elaborate a **Pano-Aware (PA) decoder** (see Sec. 3.1) and **Distortion-Guided Loss** (see Sec. 3.3) to tackle the boundary-consistent and distortion-aware prediction. The loss is simple to compute, architecture-agnostic, and complements the PA Decoder by aligning the optimization target with the ERP. Then, we articulate the **Long-Short Memory Module (LSMM)** in Sec. 3.2, augmenting temporal information with long-term memory to enhance temporal robustness under sparsity and occlusion.

3.1 Pano-Aware Decoder

Insight. Our PA decoder adapts SAM2's mask decoder to 360° FoV by building geometry awareness into the network. Unlike CamSAM2 [45] and Seg2Track-SAM2 [20], which modify SAM2's outputs, our design avoids **splitting seam-spanning objects into two identities by making the decoder itself seam-consistent and distortion-aware**.

In PA Decoder, memory-conditioned features first pass through a Pano-Consistency (PC) block composed of three convolutions with different kernel sizes. For the 3×3 and 5×5 layers, we apply left–right wrap padding: features at the left border are padded with the right border and vice versa. This PC operation $PC(\cdot)$ is formulated in Eq. (1), where s is for kernel size, Cat is for concatenating at the width dimension, and the padding p is set to 1 and 2 for $s = 3$ and $s = 5$, respectively. This preserves spatial size while allowing receptive fields to cross the 0°/360° seam, enabling the PA decoder to attend to true spherical neighbors. We apply the PC Block to features before fusion with positional embeddings, ensuring that the following attention of SAM2 preserves semantic continuity.

$$\begin{aligned} \mathcal{F}^T_{mem,l} &= \mathcal{F}^T_{mem}[:, : p] \\ \mathcal{F}^T_{mem,r} &= \mathcal{F}^T_{mem}[:, \frac{W}{16} - p :] \\ PC(\mathcal{F}^T_{mem}) &= Conv_s(Cat(\mathcal{F}^T_{mem,r}, \mathcal{F}^T_{mem}, \mathcal{F}^T_{mem,l})) \end{aligned} \tag{1}$$

For 360 videos, ERP distortion is either present from the first frame or emerges gradually with motion. The first case is guided by the initial prompt mask and needs no extra handling. For gradual changes, we fuse previous-frame mask cues: the last prediction is passed through SAM2's frozen memory-encoder mask downsampler to obtain mask features, concatenated with the output of the PC block, and fused via a convolution to stabilize features in newly distorted regions. The fused features then undergo multi-round cross-attention with tokens as in SAM2, ensuring consistent refinement across frames. During transpose convolutions, shallow features $\mathcal{F}^T_s$ and $\mathcal{F}^T_d$ from the image encoder are processed by

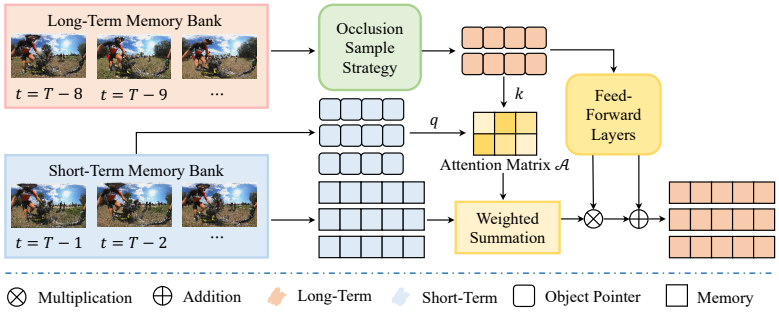

Fig. 3: Overview of our LSMM framework.

PC Blocks to retain seam consistency and further enhance spatial alignment. Finally, the mask $\mathcal{Y}^T_{pano}$ with the highest IoU score in $\mathcal{Y}^T_{SAM}$ is written to memory via the memory encoder, enabling reliable propagation in subsequent frames.

3.2 Long-Short Memory Module (LSMM)

Insight. Prior work [2, 4, 7, 37] shows that long-term memory benefits video segmentation, while SAM2 mainly preserves it via iterative updates from the prompted frame. In 360 videos, sparse object visibility further weakens long-range cues. **We introduce LSMM that fuses long- and short-term information without increasing memory footprint**, as depicted in Fig. 3.

LSMM keeps only object pointers for long-term frames in a bank and drops dense features. An Occlusion Sample Strategy selects key frames by weighted sampling with an occlusion score; pseudocode is provided in the supplementary material. Denote the sampled long-term pointers as $\mathcal{P}^T_L \in \mathbb{R}^{L\times d_p}$ and short-term pointers as $\mathcal{P}^T_S \in \mathbb{R}^{6\times d_p}$, We compute an attention matrix $\mathcal{A} \in \mathbb{R}^{L\times 6}$ that measures long–short similarity and use it to reweight the short-term memory $\mathcal{M}^T_S \in \mathbb{R}^{6\times \frac{H}{16}\times \frac{W}{16}\times d_m}$, where d_m is the embedding dimension of memory, yielding $\widetilde{\mathcal{M}}^T_S$. To inject long-range context, we apply FiLM [25]: a feed-forward network predicts per-channel scales and biases for the reweighted short-term memory, producing a pseudo long-term memory $\mathcal{M}^T_L \in \mathbb{R}^{L\times \frac{H}{16}\times \frac{W}{16}\times d_m}$:

$$\begin{aligned} \boldsymbol{\gamma}, \boldsymbol{\beta} &= \mathrm{FFN}(\mathcal{P}^T_L), \\ \mathcal{M}^T_L &= \widetilde{\mathcal{M}}^T_S \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \end{aligned} \tag{2}$$

where FFN$(\cdot)$ denotes feed-forward network, $\odot$ is element-wise multiplication, and $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_m}$ are broadcast over spatial and short-term dimensions. As in SAM2, we concatenate $\mathcal{M}^T_L$, $\mathcal{M}^T_S$, and pointer sets ($\mathcal{P}^T_L$ and $\mathcal{P}^T_S$), and pass them to the memory attention to condition image features.

3.3 Distortion-Guided Mask Loss

Insight. We propose a distortion-guided loss for 360VOS, while keeping the output head design of SAM2 [29] that outputs three mask logits $\mathcal{Y}^T_{SAM}$, three

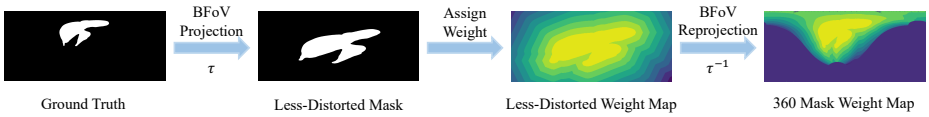

Fig. 4: Distortion-guided 360 mask weight calculation.

corresponding IoU scores $\mathcal{U}^T$, and an occlusion score $\mathcal{O}^T$ per frame. The output mask $\mathcal{Y}^T_{pano}$ is the one with the maximum IoU score. **The key challenge in 360-specific loss design is the combination of projection-induced distortion and severe class imbalance: foreground often occupies tiny regions**.

To address these, we employ the bounding field-of-view (BFoV) [34] projection τ for the ground-truth object mask $\mathcal{Y}^T_{gt}$ using camera geometry and obtain a less-distorted region inside that BFoV where projection stretching is minimized. As visualized in Fig. 4, $\tau(\mathcal{Y}^T_{gt})$ yields a robust estimate of the foreground proportion, capturing object sparsity for the current frame. Besides, it defines a spatial prior that differentiates reliable evidence from heavily distorted zones. We derive pixel-wise weights W from it. Foreground pixels receive a uniform w_f weight equal to the foreground proportion in $\tau(\mathcal{Y}^T_{gt})$, bounded by maximum value w_{max} and minimum value w_{min}, encouraging the model to learn from rare positives without overwhelming the objective. Background pixels receive spatially varying weights that are scaled by the complementary proportion and decay with normalized distance to the object boundary, as formulated in Eq. (3), where $D[i,j]$ stands for the distance of pixel $[i,j]$ to the mask boundary and D_{max} is the maximum distance. Thus, hard negatives near the contour are emphasized while far-away background contributes little. At last, we reproject the weight map to the spherical space by τ^{-1} and fill up the other area with the minimum of W. $\tau^{-1}(W)$ is utilized for weight in binary cross-entropy loss $\mathcal{L}_{BCE}$ for mask optimization.

$$W[i,j] = \frac{1}{w_f} + (w_f - \frac{1}{w_f})\sqrt{1 - \frac{D[i,j]}{D_{max}}} \tag{3}$$

For the auxiliary heads, we follow SAM2. The IoU score head is trained with mean-squared error against the true IoU value $\mathcal{U}^T_{gt}$ from $\mathcal{Y}^T_{SAM}$ and $\mathcal{Y}^T_{gt}$, and the occlusion score head with binary cross-entropy against the label $\mathcal{O}^T_{gt}$ of whether the object appears in $\mathcal{Y}^T_{gt}$. The overall objective is a weighted summation of the mask, IoU, and occlusion terms, as described in Eq. (4):

$$\begin{aligned} \mathcal{L}_{mask} &= \lambda_{bce}\mathcal{L}_{BCE}(\mathcal{Y}^T_{pano}, \mathcal{Y}^T_{gt}, \tau^{-1}(W)) + \lambda_{dice}\mathcal{L}_{dice}(\mathcal{Y}^T_{pano}, \mathcal{Y}^T_{gt}) \\ \mathcal{L} &= \lambda_{IoU}\mathcal{L}_{MSE}(\mathcal{U}^T, \mathcal{U}^T_{gt}) + \mathcal{L}_{mask} + \lambda_{occ}\mathcal{L}_{BCE}(\mathcal{O}^T, \mathcal{O}^T_{gt}) \end{aligned} \tag{4}$$

where $\mathcal{L}_{dice}$ refers to the dice loss, and $\mathcal{L}_{MSE}$ represents the mean squared error loss. The parameters λ_{bce}, λ_{dice}, λ_{IoU}, and λ_{occ} correspond to the weights applied to the mask BCE loss, mask dice loss, IoU loss, and occlusion score loss, respectively. This design does not alter inference cost, and explicitly aligns supervision with spherical geometry and the statistics of omnidirectional scenes.

Table 1: Comparisons between our method and prior arts on the 360VOTS test dataset. Our PanoSAM2 achieves a new state-of-the-art performance. The best results are shown in **bold**, and * means the method is re-implemented and reproduced.

VOS Tracker	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	Memory (GB)	GFLOPs	FPS
Perspective-Video-Specified							
CFBI [38]	ResNet-101	46.3	41.2	51.3	-	-	-
CFBI+ [40]	ResNet-101	48.0	42.8	53.2	-	-	-
TarVis [1]	Swin-L	36.8	32.5	41.1	-	-	-
GMVOS [17]	ResNet-50	47.7	43.1	52.2	-	-	-
UNICORN [35]	ConvNeXt-L	40.4	33.9	46.9	-	-	-
AFB-URR [14]	ResNet-50	43.5	38.9	48.1	-	-	-
STM [21]	ResNet-50	40.1	36.4	43.8	-	-	-
STCN [5]	ResNet-50	60.9	55.0	66.7	2.7	165.4	23.8
AOT [39]	MobileNet-V2	53.4	47.1	59.7	-	-	-
TBD [6]	DenseNet-121	53.6	47.4	59.8	-	-	-
RTS [23]	ResNet-50	59.3	54.0	64.5	-	-	-
TBD [18]	ResNet-50	53.7	48.7	58.7	-	-	-
XMem [4]	ResNet-50	65.0	59.6	70.3	3.5	361.2	22.5
SAM2 [29]	Hiera-T	59.4	53.9	64.9	3.6	686.7	42.6
SAM2 [29]	Hiera-S	60.2	56.8	63.6	4.1	751.4	39.3
SAM2Long [7]	Hiera-T	59.8	54.4	65.2	4.0	735.8	29.4
SAM2Long [7]	Hiera-S	61.1	57.5	64.7	4.4	792.3	27.4
360-Video-Specified							
PSCFormer [36]-B*	ResNet-50	61.0	57.7	64.3	2.8	274.7	18.5
PSCFormer [36]-L*	ResNet-50	62.5	58.8	66.2	-	-	-
☆**PanoSAM2**	Hiera-T	64.3↑4.9	59.2↑5.3	69.3↑4.4	3.7	728.0	34.8
	Hiera-S	**65.8**↑5.6	**59.9**↑3.1	**71.6**↑8.0	4.3	788.3	29.2

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate PanoSAM2 on two panoramic video object segmentation benchmarks: 360VOTS [34] and PanoVOS [36]. 360VOTS is a large-scale dataset that contains 290 sequences spanning 62 categories, totaling about 242K frames with an average duration of 28 seconds per sequence. It provides dense, pixel-wise ground-truth annotations. The official split assigns 170 sequences for training and 120 for testing. PanoVOS comprises 150 videos with about 14K frames and over 19K instance annotations from 35 categories, with an average duration of 20 seconds. The training set has 80 videos for training, and the validation set and test set each have 35 videos.

Implementation Details. PanoSAM2 is implemented in PyTorch [22]. All components inherited from SAM2 [29] are initialized from the released SAM2 training weights. We train with AdamW [16] optimizer ($\beta = (0.9, 0.999)$, eps=1e-8, and the weight decay is 0.01) with an initial learning rate of 2e-4 and a StepLR scheduler, for a total of 80 epochs. To exercise long-horizon memory, we cap each sampled clip at 100 frames with a batch size of 4 and sample 400 clips per epoch. To warm up, in the first two epochs, the memory encoder is fed the ground-truth mask at every frame. In subsequent epochs, we correct the memory with a GT mask every 8 frames. After 20 epochs, we introduce LSMM for training. Our

Table 2: Comparisons between our method and existing approaches on PanoVOS validation and test datasets. The best results are shown in **bold**. Evaluation metric subscript s and u denote scores in seen and unseen categories compared to the training dataset.

VOS Tracker	Backbone	PanoVOS Validation					PanoVOS Test				
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
Perspective-Video-Specified											
CFBI [38]	ResNet-101	35.8	34.6	44.8	24.2	39.7	19.1	18.2	26.1	12.2	19.8
CFBI+ [38]	ResNet-101	41.3	38.0	47.9	32.5	46.9	30.9	30.8	42.7	21.4	28.5
AFB-URR [14]	ResNet-50	34.3	34.8	42.8	24.9	34.5	34.2	28.2	38.8	32.9	36.8
STCN [5]	ResNet-50	52.0	51.2	60.8	41.5	54.5	50.8	43.6	56.5	49.3	53.7
AOTT [39]	MobileNet-V2	65.6	59.4	68.3	59.7	75.0	53.4	49.3	61.6	47.5	55.1
AOTS [39]	MobileNet-V2	67.7	61.2	70.0	62.4	77.1	55.9	53.2	65.1	48.6	57.0
AOTB [39]	MobileNet-V2	67.6	62.3	72.0	61.5	74.8	55.4	53.5	64.2	47.7	56.0
AOTL [39]	MobileNet-V2	66.6	61.4	71.1	59.4	74.3	53.8	50.0	60.3	47.8	57.1
AOTL [39]	Swin-Base	62.1	58.9	66.5	54.3	68.8	53.1	49.0	57.8	49.0	56.6
AOTL [39]	ResNet-50	65.3	61.9	71.4	56.4	71.6	54.6	52.9	63.2	47.5	54.9
RDE [13]	ResNet-50	50.5	49.7	58.4	39.2	54.9	42.5	36.9	46.6	38.5	48.2
XMem [4]	ResNet-50	55.7	54.8	63.3	45.2	59.7	53.5	49.5	62.6	47.1	54.8
SAM2 [29]	Hiera-T	74.2	64.5	79.4	68.0	85.0	70.6	62.3	79.8	63.8	76.5
SAM2 [29]	Hiera-S	71.4	62.3	77.6	65.2	80.3	69.3	62.7	78.0	62.5	74.0
SAM2Long [7]	Hiera-T	74.4	63.7	78.6	69.3	86.1	70.9	62.8	79.7	64.2	76.9
SAM2Long [7]	Hiera-S	71.9	62.6	78.0	**69.8**	81.2	69.7	63.1	78.6	62.8	74.3
360-Video-Specified											
PSCFormer [36]-B	ResNet-50	74.0	66.4	80.4	66.2	83.0	56.8	49.4	62.7	52.4	62.5
PSCFormer [36]-L	ResNet-50	77.9	70.5	85.2	69.5	86.4	59.9	54.9	69.2	53.0	62.4
☆**PanoSAM2**	Hiera-T	76.1↑1.9	67.4 ↑2.9	83.4 ↑4.0	68.6↑0.6	85.6↑0.6	71.9↑1.3	64.8↑2.5	80.5↑0.7	64.3↑0.5	**77.8** ↑1.3
	Hiera-S	**78.1**↑6.7	**71.2**↑8.9	**85.6**↑8.0	69.1↑3.9	**86.7**↑6.4	**73.4**↑4.1	**67.9**↑5.2	**84.0**↑6.0	**64.4**↑1.9	77.4↑3.4

training is conducted on two NVIDIA A800-80GB GPUs and takes about 50 hours with the Hiera-T [30] backbone. We maintain a long-term memory size L of 2. For geometric distortion augmentation, the w_{max} and w_{min} of the pixel weight are set to 2.0 and 0.5, respectively. For optimization objective, λ_{bce} is 0.5, λ_{dice} is 0.5, 1.0 for λ_{IoU}, and 0.1 for λ_{occ}.

Evaluation Metrics. We report $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$, following standard VOS metrics. $\mathcal{J}$ stands for IoU between the predicted mask and the ground truth, measuring region overlap ratio. $\mathcal{F}$ is computed on mask boundaries, assessing contour accuracy. $\mathcal{J}\&\mathcal{F}$ averages $\mathcal{J}$ and $\mathcal{F}$, providing a composite score for overall segmentation quality.

4.2 Experimental Results

Results on 360VOTS. Tab. 1 compares the performance of PanoSAM2 with existing methods on the 360VOTS [34] dataset, covering both VOS models for perspective videos and approaches tailored for 360-degree videos. The proposed PanoSAM2 demonstrates a clear and consistent improvement over previous methods, achieving state-of-the-art results across all evaluation metrics. Specifically, PanoSAM2 with the Hiera-S [30] backbone surpasses the baseline SAM2 by a significant margin, achieving a $\mathcal{J}\&\mathcal{F}$ score of 65.8, which represents an impressive improvement of 5.6 points. Also, in the 360-video-specified category, PanoSAM2 outperforms the PSCFormer method, whose $\mathcal{J}\&\mathcal{F}$ achieves 62.5, further highlighting its robustness and superiority in handling complex segmentation tasks. Moreover, Tab. 1 summarizes the computational cost of several

Table 3: Comparisons of **generalization ability** between our method pretrained on 360VOTS and prior arts on PanoVOS validation and test dataset. The best results are shown in **bold**.

Method	PanoVOS Validation			PanoVOS Test		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
PerSAM [42]	19.1	14.9	23.5	19.5	15.6	23.3
SAM-PT [28]	47.5	41.4	53.7	41.0	35.7	46.4
SAM-I2V [19]	48.9	42.8	55.0	40.5	34.9	46.1
SAM2 [29]	70.6	63.4	77.8	66.3	59.4	73.2
XMem [4]	52.1	46.9	57.2	44.4	39.6	49.1
PSCFormer [36]	69.5	62.4	76.5	64.1	58.3	69.9
☆**PanoSAM2**	**71.3**	**63.0**	**79.5**	**67.9**	**61.6**	**74.2**

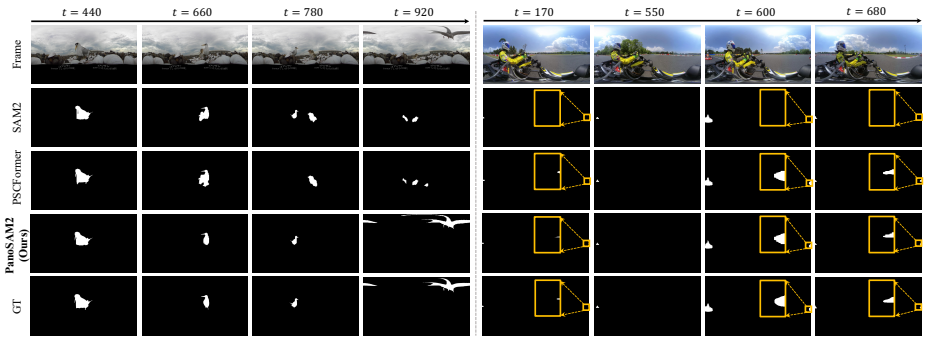

Fig. 5: Qualitative comparison between PanoSAM2 and other VOS models. The orange boxes highlight the areas where the target object straddles the 0°/360° seam, showing PanoSAM2's improved segmentation precision in complex scenes at different time steps.

representative methods. PanoSAM2 builds upon the SAM2 [29] framework with minimal architectural modification. Despite integrating panoramic-aware modules, the overall memory cost of PanoSAM2 remains close to its base SAM2 counterparts, increasing only modestly from 3.6G to 3.7G for Hiera-T and from 4.1G to 4.3G for Hiera-S. These results underline PanoSAM2's substantial enhancement in 360 video segmentation and tracking, positioning it as a powerful and efficient solution for the 360VOS.

Results on PanoVOS. Tab. 2 presents a comparison between PanoSAM2 and existing methods on the PanoVOS validation and test datasets. PanoSAM2 demonstrates substantial improvements over prior techniques, achieving state-of-the-art performance across multiple evaluation metrics. On the PanoVOS validation dataset, PanoSAM2 (Hiera-S) achieves a $\mathcal{J}\&\mathcal{F}$ score of 78.1, surpassing the SAM2 baseline by 6.7 points and indicating stronger spatial stability. In the PanoVOS test dataset, PanoSAM2 (Hiera-S) achieves an outstanding $\mathcal{J}\&\mathcal{F}$ score of 73.4, improving by 4.1 points over SAM2. Furthermore, PanoSAM2 also excels SAM2 in the unseen category tracking in terms of $\mathcal{J}_u$ and $\mathcal{F}_u$. When compared with other 360VOS methods, PanoSAM2 also outperforms PSCFormer.

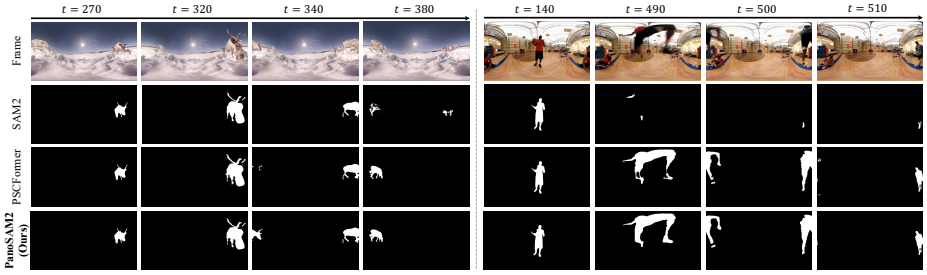

Fig. 6: Visualization of **zero-shot results** from PanoSAM2 and other VOS models on the PanoVOS dataset.

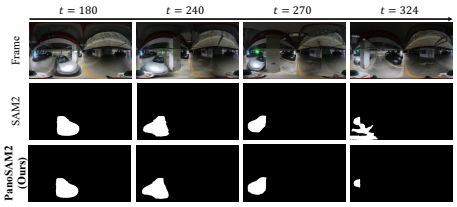

Fig. 7: Visual results on self-captured **open-world** 360 scene.

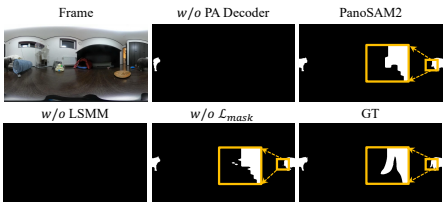

Fig. 8: Ablation visualizations of proposed modules.

PanoSAM2's result is an improvement over PSCFormer-Base and PSCFormer-Large in most evaluation metrics. *Similar to the findings in the SAM2 paper, our experiments show that SAM2 (Hiera-T) outperforms SAM2 (Hiera-S), which is consistent with the observation that smaller models can sometimes outperform larger ones, as shown in experiments of SAM2's paper.* These results confirm the effectiveness of PanoSAM2 in handling 360VOS tasks. The performance gains across multiple metrics demonstrate that PanoSAM2 sets new benchmarks for 360VOS, offering a significant advancement over current methods in the 360VOS challenge.

Qualitative Comparison. As visualized in Fig. 5, PanoSAM2 shows a clear advantage on 360VOS task. Unlike SAM2 and PSCFormer, it isolates the target without distraction from similar objects. PanoSAM2 also maintains accurate segmentation even in complex scenarios when only a small part of the object crosses the boundary. This is evident in the highlighted regions, where it successfully tracks and segments the intended target, demonstrating robustness to the unique challenges of panoramic images.

Zero-Shot Comparison. Tab. 3 compares the generalization ability of our PanoSAM2 pretrained on the 360VOTS training dataset with other VOS models on the PanoVOS validation and test datasets. Our model consistently demonstrates superior performance across evaluation metrics, clearly showing that PanoSAM2 possesses strong zero-shot capability and robust transferability, as further visualized in Fig. 6. We also show the result of PanoSAM2 on the open-world 360 video taken by ourselves in Fig. 7.

Table 4: Ablation study on key components of PanoSAM2. The best result is shown in **bold**.

PA Decoder	LSMM	$\mathcal{L}_{mask}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
			59.4	53.9	64.9
✓			62.7	57.6	67.8
	✓		61.6	57.4	65.8
		✓	60.1	54.7	65.5
✓	✓	✓	**64.3**	**59.2**	**69.3**

Table 5: Ablation study on innovative elements of PA Decoder. The best result is shown in **bold**.

PC Block	$\mathcal{Y}^{T-1}_{pano}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
		62.2	56.3	68.1
✓		63.8	58.7	68.9
	✓	62.9	57.4	68.6
✓	✓	**64.3**	**59.2**	**69.3**

Table 6: Impact of using different long-term memory sizes in LSMM. The best result is shown in **bold**.

Long-Term Memory Size L	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
w/o	63.4	58.8	68.0
1	64.0	59.1	68.9
2	**64.3**	**59.2**	**69.3**
3	63.7	59.1	68.3

4.3 Ablation Studies

We conduct a series of ablations on the 360VOTS test set using the Hiera-T backbone. We vary one factor at a time to assess the contributions of the PA decoder, LSMM, and the distortion-aware mask loss $\mathcal{L}_{\text{mask}}$. We also examine model sensitivity to hyperparameters.

Effectiveness of Key Components. Tab. 4 reports performance gains of our components and Fig. 8 visualizes the effect of these ablation studies. Starting from the SAM2 baseline, the PA decoder provides the largest single boost, adding +**3.3** in terms of $\mathcal{J}\&\mathcal{F}$, and keeps the tracking seam-consistent. LSMM alone yields a modest +**2.2** and prevents object drift. Replacing the naive cross-entropy mask loss with the distortion-aware $\mathcal{L}_{mask}$ lifts $\mathcal{J}\&\mathcal{F}$ from 59.4 to 60.1 (+**0.7**) and produces a mask with more precise boundary. Enabling all three components yields a total +**4.9**. All other settings are held fixed to isolate effects. These findings highlight specific roles of different components in improving tracker performance.

Influence of PA Decoder Elements. As shown in Tab. 5, the incorporation of PC Block alone improves $\mathcal{J}\&\mathcal{F}$ from 62.2 to 63.8 (+**1.6**), while the incorporation of the mask of the last time step $\mathcal{Y}^{T-1}pano$ alone raises it to 62.9 (+**0.7**). When combined, performance increases to 64.3 (+**2.1**), demonstrating their complementarity. This indicates that the PC Block enhances seam-consistent feature aggregation, while $\mathcal{Y}^{T-1}pano$ provides temporal guidance, jointly improving panoramic understanding and boundary consistency.

Impact of Long-Term Memory Size L. Tab. 6 indicates that introducing long-term memory notably enhances performance, confirming its importance for stable temporal reasoning. Compared with no LSMM, the best configuration

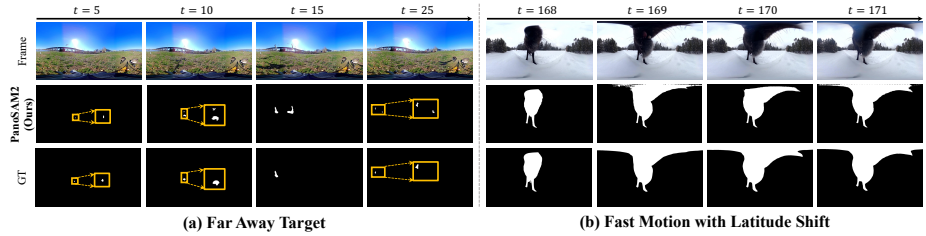

Fig. 9: Examples of failure cases. Orange bounding boxes highlight and zoom in on the small mask region.

at $L = 2$ improves $\mathcal{J}\&\mathcal{F}$ from 63.4 to **64.3** (+**0.9**), with $\mathcal{J}$ and $\mathcal{F}$ also rising to **59.2** and **69.3**, respectively. However, when L grows larger ($L = 3$), the gain diminishes slightly, suggesting that excessive long-term memory may dilute short-term information critical for precise memory conditioning.

4.4 Failure Cases

Far Away Object. The first example of Fig. 9 demonstrates a particularly challenging scenario in which distant objects with highly similar appearances interact or partially occlude one another. When the target pigeon and another overlap or move close together, PanoSAM2 occasionally misassigns masks or swaps identities, as their visual cues provide limited discriminative information.
Fast Motion with Latitude Shift. As illustrated in the second example of Fig. 9, when the animal rapidly moves upward toward the camera, its position shifts abruptly from low to high latitudes, causing severe projection distortion. The previous-frame mask becomes heavily stretched and no longer serves as a reliable reference, resulting in fragmented predictions of PanoSAM2.

5 Conclusion and Future Work

We introduced PanoSAM2, a novel 360VOS framework based on our distortion- and memory-aware adaptation strategies of SAM2 to achieve reliable 360VOS while retaining SAM2's user-friendly prompting design. Extensive experiments demonstrated that PanoSAM2 delivered robust and consistent performance across diverse 360 datasets. Overall, this work underscored the importance of jointly modeling geometric distortion, temporal coherence, and memory dynamics when adapting foundation VOS models to omnidirectional perception, paving the way toward more reliable and generalizable panoramic video understanding.
Future Work. We will extend PanoSAM2 toward multi-object tracking and diverse prompt understanding. While proposed PanoSAM2 handles single-object segmentation effectively, reasoning over multiple interacting targets remains a challenging yet valuable direction for real-world scenes. Incorporating richer prompt types – such as points, boxes, or scribbles – could enable more flexible interaction, enhancing adaptability across tasks and environments.

Acknowledgements

This work was supported by the MOE AcRF Tier 1 (Call 2/2025) Grant under Grant No. RG160/25, NTU Start-Up Grant, and NTU EEE Internal Funding.

References

1. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.: Tarvis: A unified approach for target-based video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18738–18748 (2023)
2. Bekuzarov, M., Bermudez, A., Lee, J.Y., Li, H.: Xmem++: Production-level video segmentation from few annotated frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 635–644 (2023)
3. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
4. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European conference on computer vision. pp. 640–658. Springer (2022)
5. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in neural information processing systems **34**, 11781–11794 (2021)
6. Cho, S., Lee, H., Lee, M., Park, C., Jang, S., Kim, M., Lee, S.: Tackling background distraction in video object segmentation. In: European conference on computer vision. pp. 446–462. Springer (2022)
7. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13614–13624 (2025)
8. Eger Passos, D., Jung, B.: Measuring the accuracy of inside-out tracking in xr devices using a high-precision robotic arm. In: Stephanidis, C., Antona, M. (eds.) HCI International 2020 - Posters. pp. 19–26. Springer International Publishing, Cham (2020)
9. Huang, H., Yeung, S.K.: 360vo: Visual odometry using a single 360 camera. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 5594–5600. IEEE (2022)
10. Jost, T.A., Nelson, B., Rylander, J.: Quantitative analysis of the oculus rift s in controlled movement. Disability and Rehabilitation: Assistive Technology **16**(6), 632–636 (2021)
11. Khoreva, A., Perazzi, F., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. arXiv preprint arXiv:1612.02646 (2016)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
13. Li, M., Hu, L., Xiong, Z., Zhang, B., Pan, P., Liu, D.: Recurrent dynamic embedding for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1332–1341 (2022)

14. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 3430–3441. Curran Associates, Inc. (2020)
15. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. arXiv preprint arXiv:2408.07931 (2024)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L.: Video object segmentation with episodic graph memory networks. In: European conference on computer vision. pp. 661–679. Springer (2020)
18. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9670–9679 (2021)
19. Mei, H., Zhang, P., Shou, M.Z.: Sam-i2v: Upgrading sam to support promptable video segmentation with less than 0.2% training cost. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3417–3426 (2025)
20. Mendonça, D., Barros, T., Premebida, C., Nunes, U.J.: Seg2track-sam2: Sam2-based multi-object tracking and segmentation for zero-shot generalization. arXiv preprint arXiv:2509.11772 (2025)
21. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9226–9235 (2019)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
23. Paul, M., Danelljan, M., Mayer, C., Van Gool, L.: Robust visual tracking by segmentation. In: European Conference on Computer Vision. pp. 571–588. Springer (2022)
24. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 724–732 (2016)
25. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
26. Petrovai, A., Nedevschi, S.: Semantic cameras for 360-degree environment perception in automated urban driving. IEEE Transactions on Intelligent Transportation Systems **23**(10), 17271–17283 (2022)
27. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
28. Rajič, F., Ke, L., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Segment anything meets point tracking. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9302–9311. IEEE (2025)
29. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

30. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: International conference on machine learning. pp. 29441–29454. PMLR (2023)
31. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6902–6912 (2024)
32. Xu, G., Udupa, J.K., Yu, Y., Shao, H.C., Zhao, S., Liu, W., Zhang, Y.: Segment anything for video: A comprehensive review of video object segmentation and tracking from past to future. arXiv preprint arXiv:2507.22792 (2025)
33. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
34. Xu, Y., Huang, H., Chen, Y., Yeung, S.K.: 360vots: Visual object tracking and segmentation in omnidirectional videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
35. Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H.: Towards grand unification of object tracking. In: European conference on computer vision. pp. 733–751. Springer (2022)
36. Yan, S., Xu, X., Zhang, R., Hong, L., Chen, W., Zhang, W., Zhang, W.: Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation. In: European Conference on Computer Vision. pp. 346–365. Springer (2024)
37. Yang, W., Anwar, S., Park, B., Yuan, S., Sarcevic, A., Linguraru, M.G., Burd, R.S., Marsic, I.: Maps: A morphology-aware ppe segmentation framework for healthcare settings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4383–4391 (2025)
38. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: European Conference on Computer Vision. pp. 332–348. Springer (2020)
39. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
40. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by multi-scale foreground-background integration. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4701–4712 (2021)
41. Zhang, J., Langbehn, E., Krupke, D., Katzakis, N., Steinicke, F.: Detection thresholds for rotation and translation gains in 360 video-based telepresence systems. IEEE transactions on visualization and computer graphics **24**(4), 1671–1680 (2018)
42. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
43. Zhang, W., Liu, Y., Zheng, X., Wang, L.: Goodsam: Bridging domain and capacity gaps via segment anything model for distortion-aware panoramic semantic segmentation. arXiv preprint arXiv:2403.16370 (2024)
44. Zhang, W., Xiao, D., Dai, A., Liu, Y., Pan, T., Wen, S., Chen, L., Wang, L.: Leader360v: The large-scale, real-world 360 video dataset for multi-task learning in diverse environment. arXiv preprint arXiv:2506.14271 (2025)
45. Zhou, Y., Sun, G., Li, Y., Fu, Y., Benini, L., Konukoglu, E.: Camsam2: Segment anything accurately in camouflaged videos. arXiv preprint arXiv:2503.19730 (2025)