Sparsely Supervised Surgical Video Segmentation with Reliable Asymmetric Dual Memory

Nishchal Sapkota, Yejia Zhang, Bofang Zheng, Xianshi Ma, Haoyan Shi, Yohannes Mariam, and Danny Z. Chen

Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
{nsapkota, yzhang46, bzheng2, xma3, hshi23, ymariam, dchen}@nd.edu

Abstract. Surgical video segmentation often operates on long image sequences for which dense annotation is prohibitively expensive, making sparse supervision more practical. Existing approaches suffer from temporal drift due to error accumulation, while formulations based on video object segmentation (VOS) or independent per-frame processing rely on restrictive assumptions or dense labels. Recent VOS foundation models offer strong visual representations, which are often per-object and tightly coupled with memory, limiting scalability and flexibility under domain shift in surgical videos. We propose a reliability-gated asymmetric dual-memory framework built on self-supervised visual representations of DINOv3. Our model accounts for both short-term temporal continuity and long-term semantic stability using a gated transient memory with bounded capacity and an evolving anchor memory that incrementally builds semantic representations without requiring complete first-frame class coverage. By decoupling memory from the encoder, our framework enables direct use of pre-trained vision foundation encoders and facilitates data-efficient adaptation under surgical video domain shift. Experiments on multiple surgical video datasets demonstrate improved temporal consistency and robustness compared to state-of-the-art baselines. Code is publicly available at: https://github.com/nsapkota417/D-GEM

Keywords: Surgical video segmentation · Temporal memory · Sparse supervision · Foundation models

1 Introduction

Surgical video segmentation is critical for tool tracking [6], workflow analysis [19], intraoperative scene understanding [16], data-driven education systems [7], and building automated surgical systems [5]. It operates on long image sequences for which dense annotation is prohibitively expensive. Although the global scene remains largely static (e.g., in fixed-camera recordings), surgical videos exhibit frequent object appearance/disappearance (see Fig:1), occlusions, visual corruption by liquor or smoke, etc. [8]. This creates significant challenges for surgical video segmentation due to error accumulation over time and temporal drift.

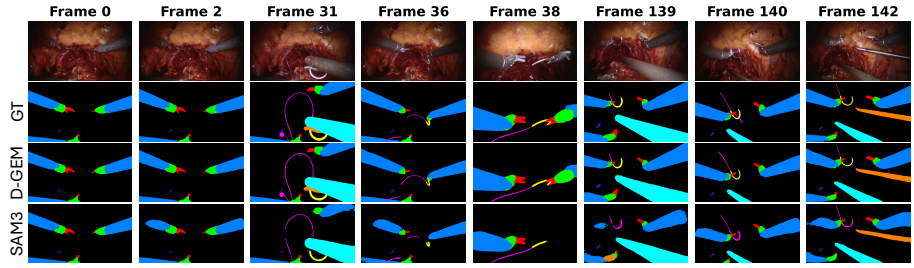

Fig. 1. Video frames sampled throughout a long surgical video (SAR-RARP50), illustrating substantial temporal variations in scene configuration, camera viewpoint, instrument appearance/disappearance, and suture geometry. Our model (D-GEM) maintains accurate and temporally consistent segmentations across these challenging transitions, while SAM3 exhibits more frequent errors.

Due to the temporal nature of video data, surgical video semantic segmentation (SVSS) is studied [10,18] through the lens of video object segmentation (VOS) [1,4,12,21]. But, several core assumptions underlying VOS do not transfer to surgical settings. Unlike VOS that typically focuses on tracking one or a few persistent foreground objects, SVSS requires dense per-pixel semantic predictions across multiple classes. Moreover, VOS formulations are often per-object by design, which limits their ability to fully exploit joint semantic relationships and contextual dependencies across multiple classes in surgical scenes. Besides, short-term temporal continuity and long-term semantic consistency play distinct and equally critical roles in surgical videos, but were often treated uniformly in VOS formulations. Severe class imbalance further limits the effectiveness of first-frame–only (support frame) annotation schemes, as many key objects may be absent or underrepresented in the labeled frames. At the same time, state-of-the-art (SOTA) foundational VOS models that tightly couple temporal states with feature extraction (e.g., memory-conditioned encoder designs such as SAM2/SAM3 [2,13] video models) can reduce flexibility for domain adaptation, as pretrained representations are less easily reused or adapted independently under domain shift and limited supervision. Together, these limitations suggest that SVSS must be modeled beyond conventional VOS formulations, as first-frame–only annotation schemes are ill-suited. Conversely, formulating SVSS as per-frame segmentation [11,20] without explicit temporal modeling is also insufficient. Although such approaches can be resilient to occlusions and abrupt visual changes [1,9], they do not exploit temporal consistency and rely heavily on dense annotations, which are impractical for long and complex surgical videos.

Consequently, sparse frame-level supervision over long temporal sequences becomes the most practical setting for SVSS. However, sparse supervision creates large temporal gaps in labeled information, requiring asymmetric temporal reasoning that jointly addresses short-term continuity and long-term semantic stability. Therefore, we adopt a dual memory design that serves these complementary roles. **Our short-term memory** employs reliability-gated updates,

building on concepts of constrained-memory from VOS models [4,21] and extending them to sparse supervision. Complementary to this, **our long-term memory** maintains a set of semantic anchors that capture persistent class-level information across extended temporal horizons. Unlike prior object-centric representations in VOS [3,17], these anchors are not assumed to be known a priori, but are instead discovered, refined, and selectively updated as reliable evidence becomes available over time. As a natural consequence of these "evolving" anchors, our method does not require the first annotated frame to contain all semantic classes. Building on these observations, we introduce **D-GEM**, a new sparsely supervised SVSS framework designed to operate over long temporal horizons.

Our approach mitigates temporal drift by explicitly separating short-term temporal continuity from long-term semantic stability within a reliability-aware and adaptive memory formulation. Additionally, we decouple memory from the encoder, enabling the direct use of powerful pretrained visual models without conditioning them on complex or task-specific memory mechanisms. This design enables pretrained encoders to be adapted to surgical domain shift with minimal modification, and empirically shows improved adaptability compared to propagation-based models (e.g., SAM2/SAM3). Our main contributions are:

1. **Asymmetric dual memory.** Our primary contribution is a principled adaptation of complementary short- and long-term memory mechanisms to the unique temporal characteristics of surgical videos. This asymmetric design combines short-term propagation with long-term semantics for robust surgical video segmentation.
2. **Reliable and evolving memory banks.** We introduce reliability-based gating and evolving anchor memory banks under bounded memory budgets to support less error-prone temporal context and incremental adaptation of semantic representations without complete first-frame class coverage.
3. **Architecture-agnostic and data-efficient domain adaptation.** We decouple memory from the architecture, enabling the use of powerful and extremely versatile pretrained DINOv3 Vision Transformers [15] and facilitating data-efficient surgical domain adaptation with limited supervision.
4. **Strong empirical performance.** Extensive experiments demonstrate consistent improvements over SOTA baselines across multiple surgical video datasets, under sparse supervision and long-range temporal evaluation.

2 Method

We propose a sparsely supervised SVSS model with a Reliability-Gated Asymmetric Dual Memory, called **D-GEM**: **D**INOv3-powered video segmentation with **G**ated Transient & **E**volving Anchor **M**emory. The asymmetric dual-memory design of D-GEM consists of a ***Gated Transient Memory (GTM)*** for reliable short-term temporal consistency and an ***Evolving Anchor Memory (EAM)*** for adaptive long-term semantics. Temporal context from both memories is retrieved via lightweight ***memory attention*** and fused using learnable

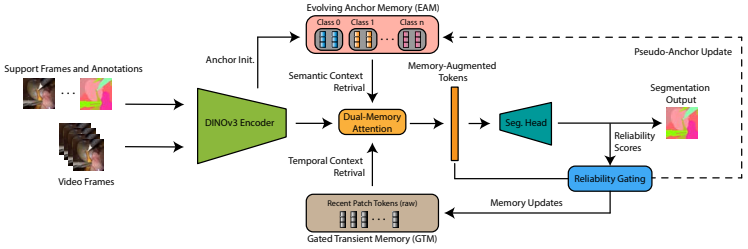

Fig. 2. An overview of our D-GEM model.

weights. It uses a pretrained DINOv3 ViT encoder with a lightweight decoder augmented by our memory components (see Fig. 2).

Let a video be a sequence of 2D image frames, $\mathcal{V} = \{I_t\}_{t=1}^{T}$, with $I_t \in \mathbb{R}^{H\times W\times 3}$. Pixel-wise semantic segmentation annotations $Y_t \in \{0,\ldots,C\}^{H\times W}$ are available only for a small set of frames in $\mathcal{V}$. Given the annotated frames to inform object semantics, we aim to predict dense semantic segmentation masks $\hat{Y}_t \in \{0,\ldots,C\}^{H\times W}$ for all frames in $\mathcal{V}$. When only the first frame is labeled, this formulation reduces to the standard first-frame–supervised VOS problem. Training and inference details under this formulation are provided in Section 2.3.

2.1 Architecture

Our D-GEM follows an encoder-decoder architecture augmented with a dual-memory module for temporal reasoning. **Frame Encoder and Decoder:** Each frame is processed independently by a pretrained DINOv3 ViT encoder that produces patch-level features. A lightweight FPN-style decoder [14] aggregates multi-scale features and outputs dense per-pixel segmentation logits. **Asymmetrical Dual Memory:** Patch-level features interface with two asymmetric memory banks: Gated Transient Memory (GTM) and Evolving Anchor Memory (EAM). For each frame, patch tokens retrieve temporal context from both memories via lightweight memory attention before decoding, and the memory is updated sequentially, one frame at a time. Critically, GTM and EAM differ not only in temporal scope but also in their update policies, capacities, and reliability constraints, enforcing an asymmetric division between short-term dynamics and long-term semantics. These are explained in the following subsections.

Gated Transient Memory. The *Gated Transient Memory* (GTM) models short-term temporal dynamics while remaining robust to local prediction noise. GTM maintains a compact set of patch-level tokens from the most recent T frames, ensuring constant memory usage and bounded per-frame computation. For each frame I_t at time t, patch features are decoded into segmentation logits, from which a reliability score is computed using both prediction confidence and uncertainty. Specifically, confidence is measured as the mean maximum softmax probability, while uncertainty is captured by the entropy of the predicted distribution. Memory updates are permitted only when these statistics satisfy a

permissive gating criterion, reflecting the importance of recent visual context under sparse supervision. When accepted, a frame-level prototype is formed by spatial pooling, and a top-k subset of patch tokens is selected based on cosine similarity to this prototype and written to memory, limiting redundancy. This top-k operation is applied only during memory *write*, while all patch tokens participate during memory *read*. To enable recency-aware retrieval, each memory token is augmented with a temporal position encoding representing its relative offset Δt to the current frame. When the memory budget is exceeded, the oldest entries are evicted using a FIFO policy, enforcing a sliding temporal window over recent frames. Together, these design choices allow GTM to provide effective short-term temporal coherence while remaining lightweight and memory-bound.

Evolving Anchor Memory. The *Evolving Anchor Memory* (EAM) provides long-term semantic stability by maintaining a compact set of class-specific anchor representations that persist across extended temporal horizons. EAM encodes object-level semantics and allows new classes and appearance modes to be discovered and refined over time, without requiring complete object coverage in any initial frames (e.g., I_1). Given annotated frames, patch-level features are extracted and aggregated into class-conditioned prototypes, which serve as candidate anchors. For each class c, EAM maintains a small set of anchors $\{\mathbf{a}_c^i\}$, each encoding a distinct appearance mode. When a new candidate prototype is observed, it is compared against existing anchors of the same class using cosine similarity. If the similarity exceeds a threshold s_{thr}, the candidate is merged with the closest anchor; otherwise, it is added as a new anchor, subject to the per-class and global memory budgets. In addition to ground-truth labels, EAM can optionally incorporate **pseudo-anchors** derived from highly confident model predictions. Such pseudo-anchors are written only when strict confidence and consistency criteria are satisfied and are downweighted relative to annotated anchors to prevent error accumulation. To provide coarse temporal awareness, anchors are augmented with temporal position encodings corresponding to their observation time. Memory usage is bounded by a fixed global budget, and eviction prioritizes removing low-confidence or redundant anchors while protecting anchors derived from annotated frames. Candidate anchors are further pruned via top-k selection prior to writing. Together, these mechanisms allow EAM to adapt to gradual appearance changes while maintaining compact, long-term semantic representations.

2.2 Memory Fusion

The current frame patch tokens retrieve long-term semantic context from EAM and short-term temporal context from GTM via lightweight cosine-similarity-based attention. Frame tokens and memory tokens are ℓ_2-normalized, and attention is computed over a bounded Top-K set of memory tokens to control computation cost, with a larger budget allocated to EAM to reflect its semantic role. For EAM, memory tokens may additionally carry confidence- or mask-derived weights that bias retrieval toward more reliable anchors. The retrieved

contexts are fused with the frame features via residual addition:

$$\mathbf{Z} = \mathbf{Q} + s_A(t)\,\alpha_A\,\mathbf{C}_A + s_T(t)\,\alpha_T\,\mathbf{C}_T, \tag{1}$$

where $\mathbf{Q}$ denotes the current frame patch tokens, $\mathbf{C}_A$ and $\mathbf{C}_T$ are the attention contexts from EAM and GTM, $\alpha_A, \alpha_T \in [0, 1]$ are learnable fusion coefficients, and the scaling terms $s_A(t)$ and $s_T(t)$ introduce a gradual ramp-in of memory contributions, allowing semantic anchors to influence predictions early while transient memory is incorporated more conservatively. This asymmetric fusion enables adaptive balancing of long-term semantics and short-term temporal continuity without unbounded memory growth.

2.3 Training and Supervision

Following the standard video model training procedure, we process the frames of $\mathcal{V}$ in a continuous temporal sliding window of length W, ensuring the presence of at least one annotated frame in the window. Annotated frames within the window are partitioned into two disjoint sets: a *support* set ($\mathcal{S}$) providing semantic cues, and a *supervision* set ($\mathcal{Q}$) used for model optimization. EAM is initialized using frames from $\mathcal{S}$, whereas GTM is progressively populated using reliable model predictions from recently processed frames, up to a fixed temporal budget. Both memories are accessed and potentially updated using features from all W frames, while supervision is applied only to frames in $\mathcal{Q}$. Each frame I_t is independently encoded into patch tokens $X_t = E(I_t) \in \mathbb{R}^{N\times D}$, where N is the number of patch tokens per frame and D is the token embedding dimension. Temporal reasoning is performed by conditioning the current-frame representation on a dual-memory state $\mathcal{M}_t = (\mathcal{M}_t^A, \mathcal{M}_t^T)$, which provides long-term semantic and short-term temporal context, respectively. The memory-augmented tokens are then decoded to produce per-pixel predictions, and the memory state is updated sequentially after processing each frame I_t.

Our segmentation loss $\mathcal{L}_{\text{seg}}$ combines Cross Entropy $\mathcal{L}_{\text{CE}}$ and Dice loss $\mathcal{L}_{\text{D}}$, applied to memory-fused predictions ($\hat{Y}_t^{\text{fused}}$) and memory-free predictions ($\hat{Y}_t^{\text{raw}}$):

$$\mathcal{L} = \frac{1}{|\mathcal{Q}|}\sum_{t\in\mathcal{Q}}\Big[\mathcal{L}_{\text{seg}}\Big(\hat{Y}_t^{\text{fused}}, Y_t\Big) + \lambda_{\text{raw}}\,\mathcal{L}_{\text{seg}}\Big(\hat{Y}_t^{\text{raw}}, Y_t\Big)\Big], \quad \mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{D}}\,\mathcal{L}_{\text{D}}. \tag{2}$$

Here, λ_{D} weights the Dice loss, while λ_{raw} weights the auxiliary supervision on the raw (memory-free) predictions. The primary loss on $\hat{Y}_t^{\text{fused}}$ directly supervises memory-enhanced predictions, whereas the auxiliary loss regularizes the encoder–decoder structure and stabilizes training under sparse supervision. During segmentation, frames of the entire video are processed sequentially, with the first frame always included in the *support* set; additional annotated frames, if available, may also be incorporated as support. Memory updates follow the same procedure as during training, and predictions are produced for the entire video, one frame at a time.

3 Experiments and Results

We evaluate our method on three public surgical video segmentation benchmarks. **CholecSeg8k** contains laparoscopic cholecystectomy clips of 80 frames each with 13

Table 1. Main Results: Results on three public datasets (ordered by increasing average video length from left to right). Average mIoU (%) is reported.

Method	Total Param. (Trainable)	CholecSeg8k	Endovis 2018	SAR-RARP50	Avg.
1 label-support (only the first frame labeled)					
SAM 2	224.5 M (-)	72.4	50.1	33.6	52.0
SAM 3	840.5 M (-)	76.7	54.8	36.3	55.9
XMem	62.2M (2.2K)	70.3	45.8	41.2	52.4
DINOv3Seg	34.1M (5.4M)	75.4	53.1	50.9	59.8
DINOv3Seg-V	34.1M (5.4M)	76.0	53.6	51.1	60.2
D-GEM (our)	34.1M (5.4M)	**78.9**	**56.3**	**52.5**	**62.6**
Adaptation: Using 10 labeled frames (for fine-tuning or refreshing the prompt)					
SAM 2 [R]	224.5 M (-)	76.3	52.7	44.6	57.9
SAM 3 [R]	840.5 M (-)	81.1	56.4	54.3	63.9
XMem	62.2M (2.2K)	74.6	50.2	58.6	61.1
DINOv3Seg	34.1M (34.1M)	85.4	57.2	62.0	68.2
DINOv3Seg-V	34.1M (34.1M)	85.6	59.6	67.7	71.0
D-GEM (our)	34.1M (34.1M)	**86.5**	**64.4**	**71.6**	**74.2**

anatomical and tool classes. **EndoVis2018** contains robotic surgical videos with 12 classes, each with 149 frames and resized to 640×800 due to memory constraints. **SAR-RARP50** comprises long robotic prostatectomy videos with 10 tool classes, averaging ∼300 frames (up to 706), resized to 720 × 1280. Results are reported on the standard validation/test splits, except for EndoVis2018, where 4 videos are randomly selected.

Baselines: We compare with foundational VOS models **SAM2** [13] and **SAM3** [2], and further adapt them to SVSS by injecting additional annotated mask prompts at fixed temporal intervals (denoted as **SAM 2 [R]**, **SAM 3 [R]**), which are the same labels used by D-GEM. As a representative hybrid memory-based approach, we adopt **XMem** [4] to the SVSS setting by attaching a lightweight semantic segmentation head to enable dense predictions under sparse supervision. In addition, we construct DINOv3-based baselines using pretrained encoders: (1) **DINOv3Seg**, a memory-free image model with a lightweight segmentation decoder, and (2) **DINOv3Seg-V**, a video model equipped with a lightweight decoder and a SAM2-style memory mechanism. For XMem, DINOv3-based baselines, and our model, segmentation decoders (and encoders where explicitly noted) are optimized on the annotated frames held out from the video (typically 1 or 10) that is being segmented.

Implementation Details for D-GEM: The best performance is obtained with a GTM temporal window of T=20 frames and an EAM budget of 2 anchors per class. These memory budgets balance temporal persistence, compactness, and robustness to drift. These settings perform well across all datasets considered in this work, and may be adjusted as needed for applications with substantially different video dynamics using the metrics mentioned in Sec. 3.2). Memory retrieval attends to the top-256 anchor tokens and top-128 transient tokens, fused via learnable weights initialized to α_{A}=0.2 and α_{T}=0.15 with asymmetric temporal ramping. Transient memory updates are enabled after a 3-step warm-up, pseudo-anchor refreshes are performed every 10 frames, and anchor updates use a cosine similarity threshold s_{thr}=0.95. Transient updates are gated by confidence (0.12) and normalized entropy (0.99). Experiments use a single A10 GPU. Additional implementation details will be released with the code.

Table 2. Results on temporal drift and mitigation with additional support frames. We report mIoU (%) performance differences for 1st 20% vs the last 20% frames.

Method	Labels	CholecSeg8k	EndoVis 2018	SAR-RARP50
SAM 3	1	82.1 → 72.3 (↓11.9%)	56.3 → 49.8 (↓11.5%)	47.2 → 29.4 (↓37.7%)
SAM 3 [R]	3	82.8 → 78.8 (↓ 4.8%)	58.1 → 52.6 (↓ 9.5%)	59.7 → 46.6 (↓21.9%)
SAM 3 [R]	20	86.9 → 85.4 (↓ 1.7%)	62.3 → 59.7 (↓ 4.2%)	67.1 → 56.6 (↓15.6%)
D-GEM (our)	1	82.4 → 78.1 (↓ 5.2%)	60.2 → 56.0 (↓ 7.0%)	56.5 → 51.2 (↓ 9.4%)
D-GEM (our)	3	82.6 → 79.3 (↓ 3.9%)	61.2 → 56.9 (↓ 7.0%)	61.4 → 56.3 (↓ 8.3%)
D-GEM (our)	20	88.1 → 87.0 (↓ 1.2%)	67.6 → 65.3 (↓ 3.4%)	68.2 → 63.4 (↓ 7.0%)

3.1 Main Results

Table 1 reports two settings: a single label-support and adaptation using 10 labeled frames. Fine-tuning denotes optimizing model parameters using a held-out annotation set within the same video (not used as support), with the total number of support and supervision frames matched to prompt-based methods (SAM2/SAM3) for fairness. For models where fine-tuning is impractical due to prompt-centric designs and tightly coupled encoders, we instead apply re-prompting at fixed temporal intervals as a practical adaptation strategy. On CholecSeg8k, which consists of short clips, most methods already perform well and temporal modeling yields modest gains (DINOv3Seg: 85.4; XMem: 74.6). In contrast, on the long-video SAR-RARP50, prompt-based and memory-free models degrade substantially even after adaptation (SAM3[R]: 63.9; DINOv3Seg: 68.2), whereas memory-augmented methods show clear advantages. Notably, D-GEM achieves the largest improvement, reaching 74.2, underscoring the benefit of asymmetric memory design. With only 10 labeled frames, D-GEM improves by +11.6 mIoU, while tightly coupled or prompt-based models saturate at 64 mIoU.

3.2 Temporal Drift and Mitigation with Additional Supports

In this section, we investigate temporal drift in long surgical videos and analyze how additional support frames mitigate its effects. To motivate the design of D-GEM, we first characterize each dataset using simple measures of visual diversity (pairwise cosine distance) and frame stability (SSIM), as these properties guide the tradeoff between temporal persistence, compactness, and robustness to drift when choosing memory budgets and update heuristics. **SAR-RARP50** exhibits the highest visual diversity and class turnover (0.17, 0.32) together with the lowest frame stability (SSIM 0.63), while **CholecSeg8k** is substantially more stable (0.04, 0.03; SSIM 0.91), with **EndoVis2018** lying in between (0.15, 0.21; SSIM 0.60). Consistent with these characteristics, temporal drift is minimal on CholecSeg8k but becomes severe on the longer, less stable SAR-RARP50 videos (Table 2). Although re-prompting in **SAM 3 [R]** reduces drift as additional support frames are provided, substantial degradation persists even under dense supervision, as also illustrated in Fig. 1. In contrast, **D-GEM** consistently exhibits lower drift across all settings, achieving comparable or better temporal stability with only 1–3 labeled frames than SAM 3 [R] with 20 labeled frames.

3.3 Ablation Study

We analyze encoder adaptation and memory components starting from a frozen DINOv3 baseline (see 3). Fine-tuning the encoder with only 10 labeled frames yields the

Table 3. Ablation study on memory components in D-GEM. We report mIoU (%).

Component	CholecSeg8k	EndoVis 2018	SAR-RARP50
DINOv3 weights (frozen)	75.4	53.1	50.9
+ Fine Tuned Encoder (10 labeled frames)	↑10.0	↑ 4.1	↑11.1
+ Transient Memory	↑ 0.2	↑ 2.4	↑ 5.7
+ Gated Transient Memory (GTM)	↑ 0.1	↑ 1.7	↑ 2.0
+ Evolving Anchor Memory (EAM)	↑ 0.9	↑ 3.3	↑ 2.2

largest gains (+4.1 to +11.1 mIoU), demonstrating data-efficient adaptation through our modular design. Transient Memory further improves long-video performance, GTM reduces temporal error accumulation, and EAM provides complementary long-term semantic modeling. Together, our memory modules yield up to 4.2 mIoU improvement on long videos over a standard memory module.

4 Conclusions

D-GEM presents an asymmetric, reliable, and evolving dual memory framework for surgical video semantic segmentation under sparse supervision. Reliability-based gating limits error accumulation, while evolving anchors enable incremental semantic adaptation over time under bounded memory constraints. By decoupling memory from the encoder, the framework flexibly integrates powerful pretrained visual models and facilitates efficient adaptation under surgical domain shift. Together, these design choices highlight the importance of reliability-aware, evolving, and architecture-agnostic memory mechanisms for scalable surgical video segmentation.

Acknowledgments. This research was funded in part by the Advanced Research Project Agency for Health (ARPA-H) Award 1AY2AX000049. The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR. pp. 221–230 (2017)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: CVPR. pp. 3151–3161 (2024)
4. Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In: ECCV. pp. 640–658. Springer (2022)
5. Haidegger, T., Speidel, S., Stoyanov, D., Satava, R.M.: Robot-assisted minimally invasive surgery – Surgical robotics in the data age. Proceedings of the IEEE **110**(7), 835–846 (2022)

6. Huang, B., Nguyen, A., Wang, S., Wang, Z., Mayer, E., Tuch, D., Vyas, K., Giannarou, S., Elson, D.S.: Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. IEEE Transactions on Medical Robotics and Bionics **4**(2), 335–338 (2022)
7. Jian, Z., Yue, W., Wu, Q., Li, W., Wang, Z., Lam, V.: Multitask learning for video-based surgical skill assessment. In: 2020 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2020)
8. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: MICCAI. pp. 440–448. Springer (2019)
9. Khoreva, A., Perazzi, F., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. arXiv preprint arXiv:1612.02646 (2016)
10. 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:2408.07931 (2024), https://openreview.net/forum?id=WSDrF5mKVp, neurIPS 2024 Workshop on Advancements in Medical Foundation Models (AIM-FM)
11. Ni, Z.L., Bian, G.B., Wang, G.A., Zhou, X.H., Hou, Z.G., Chen, H.B., Xie, X.L.: Pyramid attention aggregation network for semantic segmentation of surgical instruments. In: AAAI. vol. 34, pp. 11782–11790 (2020)
12. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV. pp. 9226–9235 (2019)
13. 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)
14. Sapkota, N., Shi, H., Zhang, Y., Ma, X., Zheng, B., Vazquez, F., Gu, P., Chen, D.Z.: When Swin Transformer meets KANs: An improved Transformer architecture for medical image segmentation. In: 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2026)
15. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
16. Valderrama, N., Ruiz Puentes, P., Hernández, I., Ayobi, N., Verlyck, M., Santander, J., Caicedo, J., Fernández, N., Arbeláez, P.: Towards holistic surgical scene understanding. In: MICCAI. pp. 442–452. Springer (2022)
17. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
18. Yin, M., Wang, F., Ye, X., Meng, Y., Fu, Z.: Memory-augmented SAM2 for training-free surgical video segmentation. In: MICCAI. pp. 328–337. Springer (2025)
19. Yue, W., Liao, H., Xia, Y., Lam, V., Luo, J., Wang, Z.: Cascade multi-level Transformer network for surgical workflow analysis. IEEE Transactions on Medical Imaging **42**(10), 2817–2831 (2023)
20. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: SurgicalSAM: Efficient class promptable surgical instrument segmentation (2024)
21. Zhou, J., Pang, Z., Wang, Y.X.: RMem: Restricted memory banks improve video object segmentation. In: CVPR. pp. 18602–18611 (2024)