OCS-MAMBA: Object-Centric, Logic-Guided Approach for Surgical Action Understanding

Md Rezowan H. F. Shuvo[0000−0002−4186−2817], M. S. Mekala[0000−0002−1313−285X], and Eyad Elyan[0000−0002−8342−9026]

Robert Gordon University, Garthdee Rd, Aberdeen AB10 7AQ m.shuvo@rgu.ac.uk

Abstract. Surgical action recognition and anticipation from video are essential for robotic assistance and decision support, yet remain challenging due to long procedural durations and complex Instrument–Verb–Target (IVT) interactions. Existing transformer-based methods are limited by quadratic attention complexity and pixel-centric representations that can produce anatomically implausible predictions. To address these challenges, we propose Object Centric Surgical Graph Mamba (OCS-Mamba), a three-stage logic-constrained, object-centric framework for surgical action understanding that integrates spatial, structural, and temporal reasoning: (1) decomposing videos into semantic entities using a hybrid DINOv2 – Slot Attention backbone; (2) modelling long-range temporal dependencies with a linear-complexity Mamba State Space Model; and (3) enforcing valid surgical relationships through a semantic grammar and binary logic matrix complemented by an inference-time RAG safety head for procedural guideline validation. This design enables efficient long-term reasoning while supporting anticipatory prediction of future actions. On public datasets, OCS-Mamba achieves a triplet mAP of 48.5%, outperforming ensemble methods by 7.8%, with comparable performance to state-of-the-art approaches at lower computational cost. Improvements in short-horizon action anticipation are also observed. These results demonstrate the feasibility of safe, anticipatory surgical action understanding, paving the way for advances in robotic surgery and surgical training

Keywords: Robotic Surgery · Surgical Action Anticipation · Action Recognition · Triplet Recognition · Neuro-Symbolic Reasoning

1 Introduction

Minimally Invasive Surgery reduces patient trauma and recovery time; however, it increases surgeons' cognitive load due to limited visibility and tactile feedback. Evidence shows that Intraoperative Adverse Events often arise from failures in situational awareness and judgment rather than technical skill [2, 17]. As a result, surgical AI is evolving from retrospective tasks toward real-time Surgical Action Planning (SAP) and adverse event anticipation [24]. For clinical deployment, such systems must maintain long-term context, reason about tool–tissue interactions, and align decisions with established safety guidelines [24, 11].

Existing methods address phase recognition [23], instrument segmentation [10], and recognition of surgical triplets [18]. Vision-Language Models (VLMs) already achieve high accuracy (up to 90%+) in basic surgical object detection, such as vessel clips and retrieval bags [24], and increasingly support planning and decision-making by aligning visual inputs with language goals to generate step-by-step surgical plans. SurgRAW [14] extends this capability through a multi-agent reasoning framework with hierarchical task orchestration across specialised VLM agents. Task-specific Chain-of-Thought prompts enhance interpretability and reliability, while Retrieval Augmented Generation(RAG) grounds decisions in external medical knowledge. An Action Evaluator agent further ensures logical consistency and adherence to valid instrument–action relationships.

Despite recent advances, three key issues prevent current models from working reliably in surgery. Architectures like Future Transformer (FUTR) [12] and VideoLLaMA [3] rely on self-attention mechanisms that scale quadratically with sequence length. For long procedures like cholecystectomy (45—90 minutes), maintaining a complete history is computationally intractable. Models are forced to truncate context to short windows (e.g., 32–64 frames), resulting in vanishing context where critical past events, such as a clip applied 20 minutes prior, are lost [23]. Pixel-based architectures process video as flattened sequences of patch tokens. As argued in Future Slot Prediction [12], this bag-of-tokens representation lacks explicit object-centric structure. It treats a grasper holding a cystic duct as a statistical pattern of texture rather than a discrete physical interaction between two entities. This makes models prone to hallucinations in the presence of visual noise (smoke, bleeding) and prevents the enforcement of logical safety constraints, such as chirality or anatomical safety zones [9].

Current planning frameworks like LLM-SAP and CSAP-Assist attempt to mitigate memory constraints by summarising distant history into text labels [24, 27]. However, this compression is lossy; fine-grained details of previous tissue manipulation are discarded. The FACTS framework [16] suggests that a robust world model requires factored state-space independent tracks for different objects rather than a single monolithic summary. While RAG has been introduced to ground AI in medical knowledge, current implementations often perform static retrieval based on image-text similarity. They fail to track the evolving state of the surgery, often retrieving generic procedural steps rather than context-specific safety warnings triggered by real-time deviations in the workflow [22].

To address these challenges, we introduce **Neuro Symbolic Surgical Graph Mamba (OCS-Mamba)**, a novel framework that fundamentally alters the surgical video understanding pipeline by fusing spatial segmentation, semantic reasoning, and efficient temporal modeling. Our approach creates a rigorous three-layer abstraction hierarchy. Leveraging principles from Future Slot Prediction [12], we employ a generic visual backbone (DINOv2-reg [5]) whose dense feature maps are decomposed into discrete representations using Adaptive Slot Attention [6]. By incorporating a Gumbel-Softmax gating mechanism, we dynamically filter out absent entities, shifting the input representation from unstructured pixel grids to a sparse, efficient set of trackable semantic nodes. We replace standard

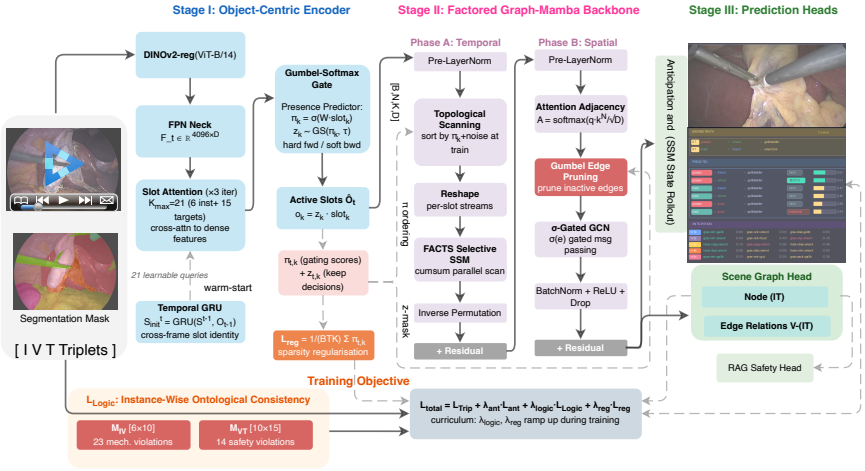

Fig. 1. Overview of OCS-Mamba. Adaptive object-centric perception extracts sparse semantic slots, which are processed by an efficient spatio-temporal Graph-Mamba backbone, with final predictions regularised by surgical logic constraints and monitored by an inference-time RAG guardrail.

Transformer-based decoding with a Graph-Factored Mamba State Space Model (SSM). By scanning dynamic scene graphs based on physical interaction topologies, Mamba achieves linear computational complexity. This allows the model to process multi-instance, long-range procedural history without conflating independent tool trajectories, maintaining a bounded inference memory footprint [23, 16]. We project ontological constraints (e.g., Clipping cannot occur on the Liver) directly onto the dynamic scene graph [9]. This acts as a differentiable logic regularizer ($\mathcal{L}_{Logic}$) during training, explicitly penalizing anatomically impossible predictions and suppressing spurious visual correlations without restricting valid multi-tool affordances. We adopt Mamba for efficient temporal state propagation over long surgical sequences, where linear state-space modelling scales better than quadratic self-attention. Main contributions are outlined below:

- We propose the first object-centric Graph-Mamba for surgical video understanding, reducing memory usage by 70% versus patch-based Transformers.
- We design a task-specific semantic grammar and a logic-constrained loss ($\mathcal{L}_{Logic}$) to penalise physically impossible instrument interactions.
- We add a RAG-based safety head that provides interpretable *Stop/Proceed* guidance from surgical guidelines. OCS-Mamba surpasses CurConMix-Ens [11] by +7.8 mAP on the public benchmark.

2 Methodology

2.1 Problem Definition

Given a surgical video $V = \{I_1, \ldots, I_{\mathcal{N}}\}$, our objective is to predict at each time step the set of active triplets $\mathcal{T}_t = \{\tau_1, \ldots, \tau_N\}$, where each triplet $\tau = \langle i, v, o \rangle$ represents a distinct interaction (Instrument-Verb-Target). Additionally, we anticipate the future triplet set at a fixed horizon $\Delta t = 1,5$s, while enforcing physical and procedural safety constraints.

2.2 Dataset Preparation and Ontological Alignment

We harmonise three datasets to construct our surgical knowledge base, *CholecT50* [18] for multi-instance triplet grounding ($\sim$1.5 concurrent $\langle I, V, T \rangle$ per frame), *CholecSeg8k* [10] to spatially align object slots with I, T entities, and *SSG-VQA* [26] to define normative scene graph connectivity. To penalise anatomically or mechanically invalid predictions, we introduce a Logic-Constrained Objective ($\mathcal{L}_{Logic}$). We derived two binary constraint matrices by thresholding empirical training co-occurrences ($P < \epsilon$) and validating against clinical safety standards, The Invalid Affordance Matrix ($\mathbf{M}_{IV}$): This matrix targets 23 pairs representing mechanical or functional impossibilities. For example, pairs such as ⟨Hook, Clip⟩ are constrained because the Hook lacks the mechanical jaws to apply clips. The Unsafe Interaction Matrix ($\mathbf{M}_{VT}$): This matrix targets 14 pairs representing clinical safety violations. While physically possible, actions such as ⟨Clip, Liver⟩ or ⟨Cut, Gallbladder⟩ represent catastrophic surgical errors (e.g., organ perforation) that the model must be penalised for predicting. Crucially, interactions that are physically valid but rare, such as ⟨Grasper, Cut⟩ were excluded from the constraint matrices. This ensures the model is not penalised for detecting valid, albeit non-standard, surgical techniques. To address the multi-triplet nature of CholecT50, we do not apply these constraints globally. Instead, $\mathcal{L}_{Logic}$ is applied element-wise to individual instrument nodes within the scene graph. This spatial disentanglement ensures that a valid action by one tool cannot shield a co-occurring invalid action by another from penalisation.

2.3 OCS-MAMBA

As depicted in Figure 1, we initialise $K_{max} = 21$ learnable slot queries. The choice $K_{max} = 6 + 15$ matches the size of the CholecT50 instrument and target vocabularies, but no slot is hard-coded to a specific class: slot semantics emerge through joint supervision from CholecSeg8k masks and the presence predictor. Following Adaptive Slot Attention [6], we use slot attention as a structural inductive bias for object-like decomposition into K competing representations, replacing its unsupervised reconstruction objective with task-specific surgical supervision. Each frame is encoded via a frozen *DINOv2-reg* [5] backbone and an FPN neck, generating high-resolution features $F_t \in \mathbb{R}^{4096 \times D}$ to ensure thin

surgical tools are accurately resolved. To preserve object identity across occlusions, these queries are first warm-started using a GRU from the previous frame's state: $S^t_{init} = \text{GRU}(S^{t-1}_{init}, O_{t-1})$. We avoid processing all K_{max} slots by using adaptive routing to filter absent entities. A Presence Predictor estimates activity probabilities $\pi_{t,k}$ and applies the Gumbel-Softmax trick [6] to sample keep/drop decisions $z_{t,k} \in \{0, 1\}$. Slots with $z_{t,k} = 0$ are pruned, yielding a sparse active set $\hat{O}_t = \{o_{t,k} \mid z_{t,k} = 1\}$. Redundant activations are penalised with a complexity regularisation loss:

$$\mathcal{L}_{reg} = \frac{1}{B \cdot \mathcal{N} \cdot K_{max}} \sum_{b,t,k} \pi_{t,k}. \tag{1}$$

Standard SSMs process flat 1D sequences, conflating parallel multi-object dynamics, while isolated temporal streams per slot ignore scene topology. We address this with a *Factored Graph-Mamba* backbone of L=4 stacked blocks, each alternating a temporal and a spatial phase.

Topological Scanning. Before each temporal phase, we apply an input-dependent node permutation: slots are reordered by their gating scores $\pi_{t,k}$ (averaged across time) so that highly-active entity pairs are memory-adjacent for the subsequent SSM scan. The inverse permutation restores the original ordering after temporal processing. This reordering is critical because the Mamba's local convolution and recurrent scan are sensitive to input ordering, placing interacting instrument–target pairs adjacently allows the model to capture their co-evolution more effectively. **Temporal Phase (Factored Mamba).** Each slot k is treated as an independent temporal stream. We reshape from $[B, \mathcal{N}, K, D]$ to $[B{\cdot}K, \mathcal{N}, D]$ and apply the Mamba selective scan [16] across the *time* dimension independently per slot. This factored design achieves linear complexity enabling full-procedure reasoning, and independent per-object hidden states $h_{\mathcal{N},k}$ that maintain identity-specific memory (e.g., recalling a clip applied to the cystic duct minutes earlier) without cross-contamination between unrelated entities. **Spatial Phase (Importance-Gated Graph Convolution).** After restoring slot ordering, we apply a GatedGCN [1] at each timestep for inter-slot message passing. Our key modification is *importance-gated adjacency* where we compute the adjacency via scaled dot-product attention between slot features, then apply the Gumbel-derived importance mask to prune edges involving inactive entities:

$$\mathbf{A}^{sp}_{kj} = \text{softmax}\left(\frac{\mathbf{q}_k^\top \mathbf{k}_j}{\sqrt{D}}\right) \cdot \mathbb{1}[\pi_k > \theta] \ \cdot \mathbb{1}[\pi_j > \theta], \tag{2}$$

where θ is the importance threshold. This ensures edges are only instantiated between entities deemed present by the adaptive gate, preventing absent-entity hallucinations from propagating through the graph. Both phases use pre-LayerNorm and residual connections.

Anticipation via State Rollout. Temporal modelling is performed over sequential frame groups simultaneously: each group of $\mathcal{N}$ observed frames is encoded by the Factored Graph-Mamba, and the final per-slot states $\{h_{\mathcal{N},k}\}_{k=1}^{K}$ are recursively propagated to produce future slot states before IVT decoding. For action anticipation, we exploit the generative property of the SSM hidden state.

Given the final states $\{h_{\mathcal{N},k}\}_{k=1}^{K}$ after processing $\mathcal{N}$ frames, we autoregressively unroll the Mamba state transition for H future horizons:

$$\hat{h}_{\mathcal{N}+\delta,k} = \bar{\mathbf{A}}\,\hat{h}_{\mathcal{N}+\delta-1,k}, \quad \delta \in \{1,\ldots,H\}, \tag{3}$$

using only the autonomous transition ($\bar{\mathbf{B}}$ is omitted since future observations are unavailable) and the temporal phase only (no spatial interaction, since future inter-object topology is unknown). Per-horizon MLP heads then classify I, V, T from mean-pooled future slots $\bar{o}_\delta = \frac{1}{K}\sum_k \hat{h}_{\mathcal{N}+\delta,k}$. Unlike Transformer-based anticipation heads that require re-encoding future context, this rollout reuses the learned SSM dynamics at negligible computational cost. The refined last-frame slots $\{\tilde{o}_{\mathcal{N},k}\}$ are routed to dedicated Presence, Relation, and Triplet prediction heads for multi-label IVT classification.

Conventional video models are prone to hallucinating physically impossible actions due to spurious statistical correlations. To address this, we introduce a direct, differentiable Ontological Consistency Loss that explicitly grounds our model in the surgical physics defined by our semantic grammar (Section 2.2). Importantly, the matrices $\mathbf{M}_{IV}$ and $\mathbf{M}_{VT}$ encode *negative* constraints rather than positive rules [15]. This ensures that while impossible actions are penalised, the model remains entirely free to use its spatio-temporal visual features to disambiguate valid affordances, or to correctly predict a low joint probability for tools that are present but currently idle.

Because CholecT50 contains multiple simultaneous triplets per frame, we must ensure that a valid action by one tool does not mask an invalid action by another. Therefore, we apply the logic penalty element-wise to the predicted graph edges. Given the set of predicted triplets $\hat{\mathcal{T}}$, the Instance-Wise Consistency Loss ($\mathcal{L}_{Logic}$) is computed as:

$$\mathcal{L}_{Logic} = \sum_{\tau \in \hat{\mathcal{T}}} \left(\underbrace{P(\tau) \cdot \mathbf{M}_{IV}[i_\tau, v_\tau]}_{\text{Mechanical Violation}} + \underbrace{P(\tau) \cdot \mathbf{M}_{VT}[v_\tau, o_\tau]}_{\text{Anatomical Violation}} \right) \tag{4}$$

where $P(\tau) = P(i_\tau) \cdot P(v_\tau) \cdot P(o_\tau)$ represents the factored joint probability of the predicted triplet. *Crucially, this formulation operates on soft probabilities rather than discrete hard predictions(i.e., argmax), making the loss fully differentiable.* This allows gradients to flow back through the Graph-Mamba backbone, suppressing forbidden classes and modeling the latent manifold to respect strict surgical affordances.

Training Strategy and Implementation Details. The overall training objective integrates triplet prediction, anticipation, ontological consistency, and adaptive slot sparsity:

$$\mathcal{L}_{total} = \mathcal{L}_{Trip}^{(t)} + \lambda_{ant} \sum_{\Delta t} \mathcal{L}_{Trip}^{(t+\Delta t)} + \lambda_{logic}\,\mathcal{L}_{Logic} + \lambda_{reg}\,\mathcal{L}_{reg}. \tag{5}$$

Here, $\mathcal{L}_{Trip}$ denotes the Asymmetric Focal Loss for triplet prediction, while $\mathcal{L}_{reg}$ penalises redundant activations in the adaptive Gumbel-Softmax slot routing [6].

Table 1. Performance comparison on CholecT50 for (IVT) recognition (mAP).

Method	Backbone	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RDV [19]	ResNet-18	89.3 ± 2.1	62.0 ± 1.3	40.0 ± 1.4	34.0 ± 3.3	30.8 ± 2.1	29.4 ± 2.8
RiT [21]	ResNet-18	88.6 ± 2.6	64.0 ± 2.5	43.4 ± 1.4	38.3 ± 3.5	36.9 ± 1.0	29.7 ± 2.6
TDN [4]	ResNet-50	91.2 ± 1.9	65.3 ± 2.8	43.7 ± 1.6	–	–	33.8 ± 2.5
MT4M. [8]	Swin-L	93.1 ± 2.1	71.8 ± 3.4	48.8 ± 3.8	44.9 ± 2.4	43.1 ± 2.0	37.1 ± 0.5
SelfD [25]	Swin-B	90.3 ± 2.3	67.4 ± 1.5	47.9 ± 1.8	43.7 ± 4.1	42.9 ± 1.6	37.1 ± 1.9
TERL [7]	Ensemble	$\mathbf{94.6 \pm 1.9}$	73.5 ± 1.9	50.8 ± 3.3	47.3 ± 4.1	45.3 ± 1.9	38.5 ± 1.1
CurCon. [11]	Ensemble	91.7 ± 2.2	69.5 ± 0.4	51.3 ± 2.9	46.3 ± 5.0	47.1 ± 1.6	40.7 ± 2.1
Ours	**DINOv2**	92.9 ± 2.4	$\mathbf{74.0 \pm 1.7}$	$\mathbf{52.8 \pm 0.7}$	$\mathbf{56.7 \pm 1.1}$	$\mathbf{52.1 \pm 1.3}$	$\mathbf{48.5 \pm 1.0}$

Training is performed in two phases. In Phase 1, the segmentation backbone is jointly supervised on CholecSeg8k using multi-task heads to provide spatial priors for object-centric slot learning. This per-frame warm-up uses batch size 32 for 60 epochs. In Phase 2, the full temporal Graph-Mamba model is trained end-to-end on CholecT50. We use AdamW [13] with learning rate 10^{-4}, weight decay 10^{-2}, and cosine annealing with $\eta_{\min} = 10^{-6}$. The loss weights are $\lambda_{ant} = 0.5$, $\lambda_{logic} = 0.1$, and $\lambda_{reg} = 0.01$, with λ_{logic} and λ_{reg} linearly ramped from 0 over the first 10 epochs to avoid over-constraining early representation learning. We set the Gumbel-Softmax temperature to $\tau = 1.0$, the importance threshold to $\theta = 0.5$, and use three Slot Attention iterations. ASL parameters are $\gamma^+ = 1.0$, $\gamma^- = 4.0$, and clip margin $m = 0.05$. Experiments follow the official CholecT50 split and are run on a single NVIDIA A100-80 GB GPU.

RAG Safety Module (Inference-Time). The RAG safety module operates at inference time only. It is not used during training, contributes no gradient signal, and does not factor into any reported result. Its role is to provide post-hoc, interpretable safety verification by querying a vectorised SAGES knowledge base with the predicted scene graph; Med-Gemma [20] then evaluates the retrieved constraints to produce a binary *Stop/Proceed* signal.

3 Results

Triplet Recognition: We evaluate OCS-Mamba on fine-grained surgical action recognition. Table 1 compares state-of-the-art methods on CholecT50. Our model achieves **48.5%** mAP on complete IVT triplets, outperforming the strongest ensemble baseline [11] by **7.8%**. The largest gain appears in the challenging Instrument-Verb (IV) category (+10.4%), which requires distinguishing subtle mechanical interactions, validating the effectiveness of Ontological Consistency Regularisation for learning instrument affordances.

Triplet Anticipation: We evaluate future triplet prediction with a 2-layer LSTM. Table 2 shows that OCS-Mamba achieves state-of-the-art forecasting at all time horizons, with the performance gap over baselines growing at longer horizons (from +2.3% IVT-mAP at Δt=1s to +5.7% at Δt=5s). The Factored Mamba model maintains memory better than LSTMs or Transformers, and large

Table 2. Triplet anticipation on CholecT50 at Δt=1s and Δt=5s (mAP, %) L=$LSTM$. OCS-Mamba leads at both horizons, highlighting superior long-range modelling.

	$\Delta t = 1$s						$\Delta t = 5$s					
Method	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RDV [19] + L	78.4	48.2	31.6	24.1	21.3	19.8	61.2	33.8	23.1	15.6	13.4	11.7
RiT [21] + L	77.9	50.1	34.2	27.5	25.9	20.4	60.5	35.4	25.0	17.8	16.5	12.2
FUTR [12]	83.6	57.7	36.4	35.8	33.0	30.5	66.1	40.5	29.3	20.2	17.1	15.0
MT4M. [8] + L	83.7	57.4	39.5	33.2	31.0	27.2	67.4	41.2	29.4	22.1	20.3	17.4
TERL [7] + L	**85.1**	58.9	41.2	35.4	33.1	28.7	**69.0**	42.8	30.7	23.8	21.6	18.4
CurCon[11]+ L	82.4	55.6	41.8	34.6	34.8	30.5	66.3	40.1	31.2	22.9	22.8	19.6
OCS-Mamba	84.7	**60.8**	**42.9**	**39.3**	**35.7**	**32.8**	61.7	**43.5**	**33.2**	**30.8**	**28.1**	**25.3**

Table 3. Comprehensive ablation of OCS-Mamba.

Variant / Model	**Temp. Module**	**Complexity**	**Slot**	**Ont.**	Δt=1s	Δt=5s
FUTR	Self-Attn	$O(\mathcal{N}^2)$	–	–	24.5	15.0
CurConMix-Ens + LSTM	Trans. + LSTM	$O(\mathcal{N}^2)$	–	–	30.5	19.6
Base Model (ResNet50)	Mamba	$O(\mathcal{N})$	–	–	30.7	20.8
+ DINOv2	Mamba	$O(\mathcal{N})$	–	–	31.1	21.2
+ Slot Attention	Graph-Mamba	$O(K^2\mathcal{N})$	✓	–	32.2	22.9
+ GatedGCN	Graph-Mamba	$O(K^2\mathcal{N})$	✓	–	32.5	23.8
+ $\mathcal{L}_{Logic}$ **(Full OCS-Mam.)**	**Graph-Mamba**	$O(K^2\mathcal{N})$	✓	✓	**32.8**	**25.3**
Δ (Full − CurConMix-Ens)					*+2.3*	*+5.7*

gains in composite relation metrics (e.g., AP_{IV} from 23.8% to 30.8%) show improved anticipation of complex interactions.

Ablation Study: We perform stepwise ablation to quantify each module's contribution (Table 3). *Backbone (ResNet → DINOv2):* Upgrading the base feature extractor yields modest anticipation gains (+0.4% at Δt=5s) alongside broader recognition improvements, demonstrating DINOv2's [5] robustness to surgical visual noise. *Slot Attention:* Transitioning from global features to object-centric slot decomposition yields a performance leap (+1.1% at 1s; +1.7% at 5s). This confirms that isolating surgical tools into distinct semantic states is a strict prerequisite for accurate long-range reasoning. *GatedGCN:* Introducing spatial graph reasoning adds explicit relational context between interacting slots, providing a robust +0.9% boost at the challenging Δt=5s horizon.

SSM Rollout vs. LSTM: To isolate the contribution of Mamba for future prediction, we replaced the SSM rollout head with a standard LSTM. While short-term performance remained stable, Δt=5s anticipation dropped severely from 25.3% to 22.5%, indicating OCS-MAMBA better maintains long-range temporal memory. Finally, integrating ontological constraints $\mathcal{L}_{Logic}$ with $\mathbf{M}_{IV}/\mathbf{M}_{VT}$ provides the strongest long-horizon gain (+1.5% at Δt=5s), showing that penalising anatomically invalid configurations improves surgical affordance learning and stabilises anticipation. Statistical analysis on per-video AP confirms that our per-

formance gains over [11], with a substantial effect for AP_{IVT} (d=0.89, p=0.008) and significant improvement for AP_{IV} (p=0.017).

4 Conclusion

We introduce **OCS-Mamba** to address the context–compute bottleneck in surgical video understanding, enabling efficient long-term procedural reasoning. On CholecT50, it outperforms prior methods in recognition and multi-horizon anticipation. By incorporating symbolic constraints, the model penalises anatomically implausible interactions, thereby promoting safer, more realistic predictions and supporting context-aware intraoperative decision-making. Current constraints are tailored to cholecystectomy and require adaptation for other procedures, and the fixed slot budget may not generalise to scenes with more entity types. Our evaluation is restricted to publicly available datasets, and the safety claims of the RAG module have not been validated in a prospective clinical setting; deployment in real surgical workflows will require prospective validation, regulatory review, and surgeon-in-the-loop assessment beyond the scope of this work. Future work will explore procedure-agnostic constraint learning, dynamic slot allocation, real-time edge optimisation, and evolution toward an interactive multimodal assistant.

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

References

1. Behrouz, A., Hashemi, F.: Graph mamba: Towards learning on graphs with state space models. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. pp. 119–130 (2024)
2. Calland, J.F., Guerlain, S., Adams, R., Tribble, C., Foley, E., Chekan, E.: A systems approach to surgical safety. Surgical Endoscopy And Other Interventional Techniques **16**(6), 1005–1014 (2002)
3. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18407–18418 (2024)
4. Chen, Y., He, S., Jin, Y., Qin, J.: Surgical activity triplet recognition via triplet disentanglement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 451–461. Springer (2023)
5. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2dnO3LLiJ1
6. Fan, K., Bai, Z., Xiao, T., He, T., Horn, M., Fu, Y., Locatello, F., Zhang, Z.: Adaptive slot attention: Object discovery with dynamic slot number. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23062–23071 (2024)

7. Gui, S., Wang, Z.: Tail-enhanced representation learning for surgical triplet recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 689–699. Springer (2024)
8. Gui, S., Wang, Z., Chen, J., Zhou, X., Zhang, C., Cao, Y.: Mt4mtl-kd: A multi-teacher knowledge distillation framework for triplet recognition. IEEE Transactions on Medical Imaging **43**(4), 1628–1639 (2023)
9. Holm, F., Ünver, G., Ghazaei, G., Navab, N.: Cat-sg: A large dynamic scene graph dataset for fine-grained understanding of cataract surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 96–106. Springer (2025)
10. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
11. Jeon, Y., Shin, J., Park, S., Kim, B., Park, K., Oh, N., Jung, K.H.: Curconmix: A curriculum contrastive learning framework for enhancing surgical action triplet recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 149–158. Springer (2025)
12. Liao, G., Jogan, M., Hussing, M., Zhang, E., Eaton, E., Hashimoto, D.A.: Future slot prediction for unsupervised object discovery in surgical video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 219–229. Springer (2025)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
14. Low, C.H., Wang, Z., Zhang, T., Zeng, Z., Zhuo, Z., Mazomenos, E.B., Jin, Y.: Surgraw: Multi-agent workflow with chain-of-thought reasoning for surgical intelligence. arXiv preprint arXiv:2503.10265 (2025)
15. Mittal, H., Agarwal, N., Lo, S.Y., Lee, K.: Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18580–18590 (2024)
16. Nanbo, L., Laakom, F., XU, Y., Wang, W., Schmidhuber, J.: FACTS: A factored state-space framework for world modelling. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=dmCGjPFVhF
17. Neri, A., Haouchine, N., Penza, V., Mattos, L.S.: Towards patient-specific deformable registration in laparoscopic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 638–647. Springer (2025)
18. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
19. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
20. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)

21. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. International Journal of Computer Assisted Radiology and Surgery **18**(6), 1053–1059 (2023)
22. Su, C., Wang, Y., Xu, B., Feng, R., Du, L., Liu, H., Meng, G.: Medicl: In-context learning for semantically enhanced aki prediction in cardiac surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 402–411. Springer (2025)
23. Wang, Z., Pan, S., Fang, M., Zhang, R., Tian, J., Dong, D.: Cholecmamba: A mamba-based multimodal reasoning model for cholecystectomy surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–116. Springer (2025)
24. Xu, M., Huang, Z., Zhang, J., Zhang, X., Dou, Q.: Surgical action planning with large language models. arXiv preprint arXiv:2503.18296 (2025)
25. Yamlahi, A., Tran, T.N., Godau, P., Schellenberg, M., Michael, D., Smidt, F.H., Nölke, J.H., Adler, T.J., Tizabi, M.D., Nwoye, C.I., et al.: Self-distillation for surgical action recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 637–646. Springer (2023)
26. Yuan, K., Kattel, M., Lavanchy, J.L., Navab, N., Srivastav, V., Padoy, N.: Advancing surgical vqa with scene graph knowledge. International journal of computer assisted radiology and surgery **19**(7), 1409–1417 (2024)
27. Zhang, J., Xu, M., Wang, Y., Dou, Q.: Csap-assist: Instrument-agent dialogue empowered vision-language models for collaborative surgical action planning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 139–148. Springer (2025)