Supplementary Material

1. Other model structures

1.1. Spatial feature extraction module

Inductive Bias of Feature Extractors. The inductive bias of spatial encoders strongly affects both attention distribution and global semantic perception. Swin Transformer [1, 2] are often activated around surgical instruments and their contours, but fail to capture global contextual cues. This arises from Swin's window-based self-attention and hierarchical design, which emphasize local detail modeling while limiting long-range semantic reasoning. For instrument–verb–target (IVT) recognition, verbs require fine-grained spatial cues, whereas targets rely heavily on global context, making it difficult for single-scale features to satisfy both requirements.

Model structure. To mitigate the limitations of single encoders (e.g., ResNet or Swin Transformer), we employ a **dual-branch spatial encoding** architecture:

- Global branch: extracts self-supervised global representations using DINOv3, $\mathbf{F}^{\text{Di}} = \mathcal{E}_{\text{DINOv3}}(x)$.
- Local branch: extracts hierarchical window-based features using Swin Transformer, $\mathbf{F}^{\text{Sw}} = \mathcal{E}_{\text{Swin}}(x)$.

The features are concatenated to form a complementary spatial representation:

$$\mathbf{F} = \text{Concat}\big(\mathbf{F}^{\text{Di}}, \mathbf{F}^{\text{Sw}}\big), \quad \mathbf{F} \in \mathbb{R}^{D\times H'\times W'}.$$

1.2. Dual-Memory Attention Model

Surgical recordings often span long durations and are sparsely sampled (typically $\sim$1 FPS), making it challenging to capture consistent spatial-temporal transitions of surgical instruments. To address this, we introduce a lightweight yet effective memory attention module that fuses historical semantic and decision representations, enhancing temporal reasoning and prediction stability under low-frame-rate conditions.

Let the global semantic descriptor at time t be $\mathbf{d}_t \in \mathbb{R}^D$. We maintain two temporal memories: a memory $\mathbf{M}^d_{t-1} \in \mathbb{R}^{L\times D}$ that stores the latest L descriptors $\{\mathbf{d}_{t-L}, \ldots, \mathbf{d}_{t-1}\}$, and a decision memory $\mathbf{M}^o_{t-1} \in \mathbb{R}^{L\times C_{\text{all}}}$ that stores their corresponding output embeddings $\{\mathbf{e}_{t-L}, \ldots, \mathbf{e}_{t-1}\}$, where $\mathbf{e}_\tau = \hat{\mathbf{y}}^{\text{all}}_\tau$ denotes the predicted multi-label vector.

Temporal fusion is achieved via a Transformer-based decoder, where the current semantic descriptor serves as the query and both memories jointly provide key-value context:

$$\mathbf{K}_t = [\mathbf{M}^d_{t-1}; \mathbf{M}^o_{t-1}], \quad \mathbf{V}_t = [\mathbf{M}^d_{t-1}; \mathbf{M}^o_{t-1}], \tag{1}$$

$$\mathbf{F}^T = \text{Dec}(\mathbf{q} = \mathbf{d}_t, \mathbf{K} = \mathbf{K}_t, \mathbf{V} = \mathbf{V}_t), \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation along the feature dimension. Learnable projections and positional encodings align both modalities within a unified attention space. After each step, the memories are updated in a sliding-window manner.

This dual-memory attention effectively integrates historical semantics and decision embeddings, facilitating long-range temporal dependency modeling and improving coherence in *triplet-based multi-label recognition* for long, sparsely sampled surgical videos.

2. Model Efficiency

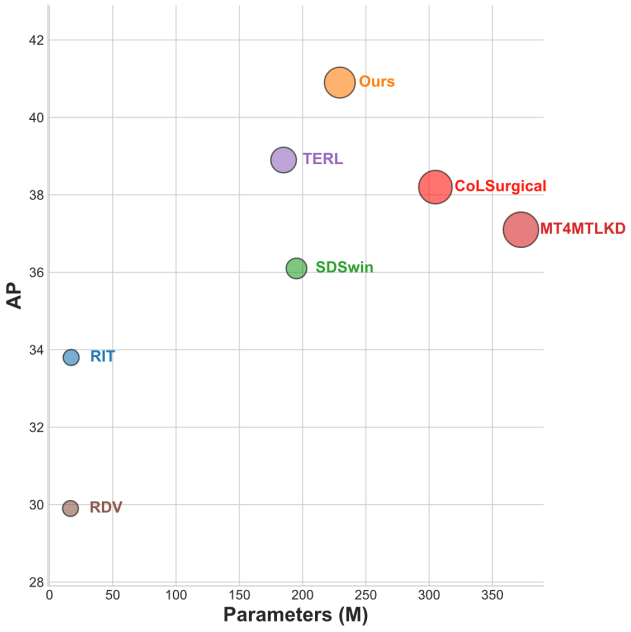

Figure 1. Comparison of model efficiency and performance. The horizontal and vertical axes denote the number of parameters and AP_{IVT} scores, respectively; circle size represents computational cost (FLOPs).

Figure 1 illustrates the trade-off between performance (AP_{IVT}) and model complexity. High-performing models (e.g., CoL, MT4MTLKD) achieve competitive accuracy but require large architectures (>300M parameters), incurring high computational costs. Lightweight models (e.g., TEFL, SDS) reduce complexity but suffer performance degradation. Our model (**Ours**) achieves the highest AP_{IVT} with moderate parameter size, demonstrating an optimal balance between accuracy and efficiency.

3. Detailed configuration

3.1. First Stage

The model is trained with a batch size of 32, using an initial learning rate of 3e-4. Exponential Moving Average (EMA) is enabled to stabilize training, with a decay factor of 0.99. We apply a warmup ratio of 0.05 and a weight decay of 1e-5. The learning-rate schedule follows a polynomial decay strategy with a decay rate of 0.99 and a power parameter of 0.1. The minimum learning rate is set to 1e-6.

3.2. Second Stage

Dual-memory model employs a 2-layer architecture with 8 attention heads and a sequence length of 8. The hidden feature dimension for the GCN module is set to 768, and a dropout rate of 0.1 is applied during training. The batch size is 32, and the model is trained for 10 epochs with an initial learning rate of 2e-5. Exponential Moving Average (EMA) is enabled with a decay factor of 0.99. For optimization, we adopt a warmup ratio of 0.05, weight decay of 1e-5 and a decay rate of 0.99. The polynomial learning-rate scheduler uses a power value of 0.1, with the minimum learning rate clipped at 1e-6.

4. More ablation experiments

4.1. Ablation on spatial feature extraction module

To assess the contribution of different spatial feature extractors, we compared several backbone configurations on the CholecT45 dataset (fold 1). As shown in Table 1, transformer-based models generally outperform CNN-based architectures. Among the single backbones, *Swin_Large* achieves the strongest performance with an AP_{IVT} of 0.329.

Furthermore, combining complementary backbones leads to noticeable improvements. In particular, the hybrid configurations *Swin_Base + DINOv2_Base* and *Swin_Base + DINOv3_Base* both reach an AP_{IVT} of 0.332, indicating that fusing representations from heterogeneous architectures provides richer spatial features. These results validate the importance of backbone diversity and demonstrate that hybrid feature extractors can offer superior representational strength compared to single-model designs.

4.2. Ablation on dual-memory attention module

We further evaluate the effectiveness of the proposed dual-memory attention mechanism. As shown in Table 2, incorporating the decision module yields consistent improvements across both evaluation metrics. Specifically, AP_{IVT} increases from 0.388 to 0.392, and the Top_{20} score rises from 95.5 to 95.8.

Although the absolute gains are modest, the improvement across two complementary metrics indicates that the decision module enhances the model's ability to select and integrate informative temporal cues. This confirms that the dual-memory attention framework contributes to more stable and discriminative feature aggregation.

Table 1. Ablation study on different backbone combinations on CholecT45 Dataset fold 1.

Model	**Params**	$\mathbf{AP_{IVT}}$
DINOv3_Base	87.7M	0.312
Swin_Base	86.9M	0.325
DINOv2_base	304.5M	0.317
DINOv3_large	196.4M	0.323
Swin_Large	197.2M	0.329
Swin_Base + DINOv2_Base	391.3M	0.332
DINOv2_Base + DINOv3_Base	392.2M	0.320
Swin_Base + DINOv3_Base	174.6M	0.332

Table 2. Dual-memory attention Ablation on CholecT45 Dataset fold 1.

Setting	$\mathbf{AP_{IVT}}$	Top_{20}
Without Decision Module	0.388	95.5
With Decision Module	0.392	95.8

References

[1] Shuangchun Gui, Zhenkun Wang, Jixiang Chen, Xun Zhou, Chen Zhang, and Yi Cao. Mt4mtl-kd: A multi-teacher knowledge distillation framework for triplet recognition. *IEEE Transactions on Medical Imaging*, 43(4):1628–1639, 2023. 1

[2] Amine Yamlahi, Thuy Nuong Tran, Patrick Godau, Melanie Schellenberg, Dominik Michael, Finn-Henri Smidt, Jan-Hinrich Nölke, Tim J Adler, Minu Dietlinde Tizabi, Chinedu Innocent Nwoye, et al. Self-distillation for surgical action recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–646. Springer, 2023. 1