RIVET: MULTI-EVIDENCE RELATIONAL INFERENCE FOR RARE SURGICAL TRIPLETS IN SINUS ENDOSCOPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognizing surgical interactions as instrument–verb–target (IVT) triplets is hindered by long-tailed class distributions and visual ambiguity among triplets that share components. These challenges are particularly pronounced in sinus endoscopy, where narrow operative fields, occlusions, and visually similar tissues obscure discriminative local cues. Here we introduce SinusTriplet-95, a dual-center dataset of 324 videos from 15 patients, comprising 670,381 frames and 2,428 segments annotated by multiple senior attending surgeons across 95 valid IVT triplets. We further propose RIVET (Relational Inference with Visual Evidence for Tail Triplets), which refines long-tailed triplet prediction at inference by integrating structural, historical, and local evidence. Its Dynamic Structural Relation module bidirectionally calibrates component- and triplet-level predictions. To improve the recognition of underrepresented triplets, the Tail-Class Memory Bank stores observations of rare triplet classes across surgical process, while the Tail-Class Prototype Query searches the current frame for local visual evidence associated with each rare triplet candidate. These complementary sources of evidence are integrated through frame-adaptive gating. On SinusTriplet-95, the RIVET ensemble achieves an $\mathrm{AP}_{\mathrm{IVT}}$ of 50.80%, surpassing the strongest ensemble baseline by 5.71 percentage points and ranking first among the compared methods across all six AP metrics. Progressive ablations show incremental gains from each evidence source, while class-level analysis shows improvements over the corresponding baselines in 46 of 50 tail classes. Together, SinusTriplet-95 and RIVET provide a benchmark and an evidence-guided framework for surgical interaction recognition in sinus endoscopy.

1 INTRODUCTION

Surgical triplet recognition represents an $\langle$instrument, verb, target$\rangle$ (IVT) triplet, specifying which instrument performs which action on which tissue or object (Nwoye et al., 2022). Unlike coarse-grained phase recognition or instrument detection alone, this task requires both identifying the components and correctly associating them within a frame. This compositional requirement makes visually similar interactions particularly challenging to distinguish, especially when their identifying cues are subtle (Nwoye et al., 2023).

Existing surgical triplet benchmarks, such as CholecT45 and the associated CholecTriplet2021 challenge, focus on laparoscopic cholecystectomy (Nwoye et al., 2022; 2023). Their procedure-specific label spaces and imaging conditions do not provide a direct benchmark for sinus endoscopy, which involves different instruments and anatomical targets within narrow, frequently occluded operative fields. Visually similar adjacent tissues and fluid or smoke artifacts further obscure the local cues needed to distinguish interactions. To address this domain-coverage gap, we introduce **SinusTriplet-95**, a dual-center dataset comprising 324 video clips from 15 patients and 2,428 activity segments annotated by multiple senior surgeons across 95 valid IVT classes. The dataset provides interaction annotations for sinus endoscopy, enabling the evaluation of fine-grained surgical activity recognition.

SinusTriplet-95 also presents a pronounced long-tailed distribution: 50 of its 95 triplet classes contain no more than five annotated segments. Sparse positive observations make it difficult to learn reliable class references for rare interactions. This difficulty is compounded by triplets that share two components but differ in the third. For example, $\langle$forceps, remove, packing$\rangle$ and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

⟨forceps, cauterize, packing⟩ share the same instrument and target but differ in the verb, so distinguishing them depends on subtle, action-specific visual cues within the current frame. The challenge therefore extends beyond unequal class frequencies to fine-grained discrimination between visually similar triplet candidates. Reliable recognition of rare interactions calls for stable historical class references, discriminative local evidence from the current frame, and consistency between component-level and triplet-level predictions.

Existing relation-modeling methods capture label dependencies through semantic representations or empirical co-occurrence statistics (Xi et al., 2022; Li et al., 2026). For sparsely observed triplets, however, statistical relations may be poorly supported or sensitive to the data distribution, whereas the compositional definitions linking each triplet to its instrument, verb, and target remain available regardless of class frequency. Long-tail approaches improve supervision and feature discrimination through self-distillation, prototype-enhanced learning, and hard-sample training (Yamlahi et al., 2023; Gui & Wang, 2024; Jeon et al., 2025). These training-side improvements do not by themselves provide reliability-weighted historical comparisons during inference, motivating the retention and direct reuse of scarce positive features as class references rather than relying solely on their influence on learned representations. Localization-based methods associate instruments with actions and tissues through explicit spatial prediction (Sharma et al., 2023a; Pei et al., 2025). However, locating an instrument or interaction region does not by itself establish which of several similar triplets is present, leaving a need to verify the subtle local cues supporting each complete rare-triplet candidate. To address these complementary requirements, we propose RIVET (Relational Inference with Visual Evidence for Tail Triplets), which combines structural, historical, and local evidence to refine rare-triplet predictions. Its **Dynamic Structural Relation** module converts taxonomy-defined component–triplet mappings into bidirectional calibration signals, allowing component predictions and holistic triplet scores to refine one another without estimating relational priors from class co-occurrence frequencies. The **Tail-Class Memory Bank** retains scarce positive features across training batches and maintains momentum-updated class prototypes for direct prediction calibration, weighting prototype-matching evidence by memory confidence derived from the number and diversity of stored observations. **Tail-Class Prototype Query** combines component semantics with learned classifier weights to construct complete-triplet queries, retrieves candidate-specific evidence from current-frame spatial features, and supplies positive local support for tail-class predictions without introducing a separate detection head. Adaptive gates regulate these residual corrections for each frame, integrating compositional agreement, historical similarity, and local visual evidence within a unified framework for prediction refinement.

Our main contributions are as follows:

- **SinusTriplet-95:** We introduce a dataset collected from two centers and annotated by experts, extending surgical triplet benchmarking to sinus endoscopy and enabling the study of fine-grained interaction recognition under severe class imbalance and visual ambiguity characteristic of this domain.
- **RIVET:** We propose a unified framework for prediction refinement that combines bidirectional calibration, confidence-aware historical prototype matching, and local visual evidence for tail classes. Adaptive gates regulate these complementary corrections for each frame.
- **Empirical evaluation:** On SinusTriplet-95, RIVET achieves 47.47% $\mathrm{AP}_{\mathrm{IVT}}$ with a single Swin-B model and 50.80% with an ensemble. The ensemble exceeds the strongest compared ensemble baseline by 5.71 percentage points. Progressive ablations and improvements on 46 of the 50 analyzed tail classes over both matched baselines support the effectiveness of the proposed refinement for rare-triplet recognition.

2 Related Work

2.1 Surgical Triplet Recognition and Benchmarks

Surgical triplet recognition represents fine-grained interactions as instrument–verb–target compositions. The CholecTriplet2021 benchmark established standardized evaluation protocols for this task (Nwoye et al., 2023), and has supported the development of a series of triplet recognition methods. RDV uses instrument features to guide verb and target prediction (Nwoye et al., 2022), while

SIRNet adaptively balances the contributions of the three components according to the current visual content (Li et al., 2022). Rendezvous in Time further incorporates temporal information across frames (Sharma et al., 2023b).

Existing benchmarks and methods, however, have been developed mainly for laparoscopic cholecystectomy. Sinus endoscopy involves a different interaction taxonomy and more challenging visual conditions, including narrow operative fields, frequent occlusion, and visually similar anatomical structures. These differences limit the direct use of existing benchmarks for studying fine-grained sinus interactions. To fill this gap, we introduce SinusTriplet-95, a dual-center, expert-annotated benchmark for instrument–verb–target recognition in sinus endoscopy.

2.2 Relational and Long-Tailed Triplet Recognition

Relational modeling is widely used to improve consistency among triplet components. Forest GCN performs graph-based message passing among instruments, verbs, and targets (Xi et al., 2022), while TRCoRSurg jointly models intra-frame label relations and temporal dependencies across frames (Li et al., 2026). These methods show the value of exploiting label dependencies, but learned relations can be less reliable when triplet observations are sparse. RIVET therefore uses the deterministic mappings defined by the IVT taxonomy to calibrate component and triplet predictions without relying on empirical co-occurrence frequencies.

Long-tailed methods further improve recognition of rare surgical interactions. SelfD uses self-distillation to reduce bias toward frequent classes (Yamlahi et al., 2023), MT4MTL-KD transfers knowledge through multiple teacher networks (Gui et al., 2024), and TERL enhances tail representations with contrastive learning and class prototypes (Gui & Wang, 2024). CurConMix further improves fine-grained discrimination through curriculum learning and hard-sample synthesis (Jeon et al., 2025). These approaches mainly strengthen tail representations during training. For extremely rare triplets, however, scarce positive observations are not explicitly reused when predicting a new frame. RIVET complements these methods with a Tail-Class Memory Bank, which preserves historical tail-class features as confidence-weighted references. Together with taxonomy-based structural calibration and current-frame local evidence, these historical references provide complementary cues for refining rare triplet predictions.

Local visual evidence provides another cue for distinguishing surgical interactions. MCIT-IG explicitly localizes instruments and associates them with verbs and tissues (Sharma et al., 2023a), while ITG-Trip combines spatial, textual, and temporal cues in a detection-based framework (Pei et al., 2025). Localization identifies relevant interaction regions, but visually similar triplets may still be difficult to separate when they share multiple components. RIVET addresses this issue through Tail-Class Prototype Query, which retrieves candidate-specific evidence directly from current-frame spatial features without requiring a separate detection head.

3 Method

3.1 Overall Framework of RIVET

Figure 1 presents the overall architecture of RIVET, which refines surgical triplet predictions with structural, historical, and local evidence. A shared encoder and head–tail classifiers produce base instrument, verb, target, and IVT predictions, which the Dynamic Structural Relation module bidirectionally calibrates. However, structural relations alone do not provide visual references for rare triplets; the Tail-Class Memory Bank supplies these through historical prototypes. These references offer global visual support, but global similarity alone may miss subtle differences between triplets sharing components. Tail-Class Prototype Query therefore retrieves local evidence from the input image. A shared adaptive gate controls the memory and query residuals added to the structurally calibrated IVT predictions.

Let $\mathcal{R} = \{I, V, T\}$. For each task $r \in \mathcal{R} \cup \{\mathrm{IVT}\}$, let $\mathcal{C}_r$ denote its label space, $C_r = |\mathcal{C}_r|$, and $\mathbf{y}^r \in \{0,1\}^{C_r}$ its multi-label annotation. A Swin Transformer (Liu et al., 2021) encodes a frame x into $F \in \mathbb{R}^{D\times H\times W}$. Global average pooling gives $\mathbf{z} \in \mathbb{R}^D$ for gating and historical matching. Spatial flattening yields $X \in \mathbb{R}^{L\times D}$, where $L = HW$, for candidate-conditioned matching.

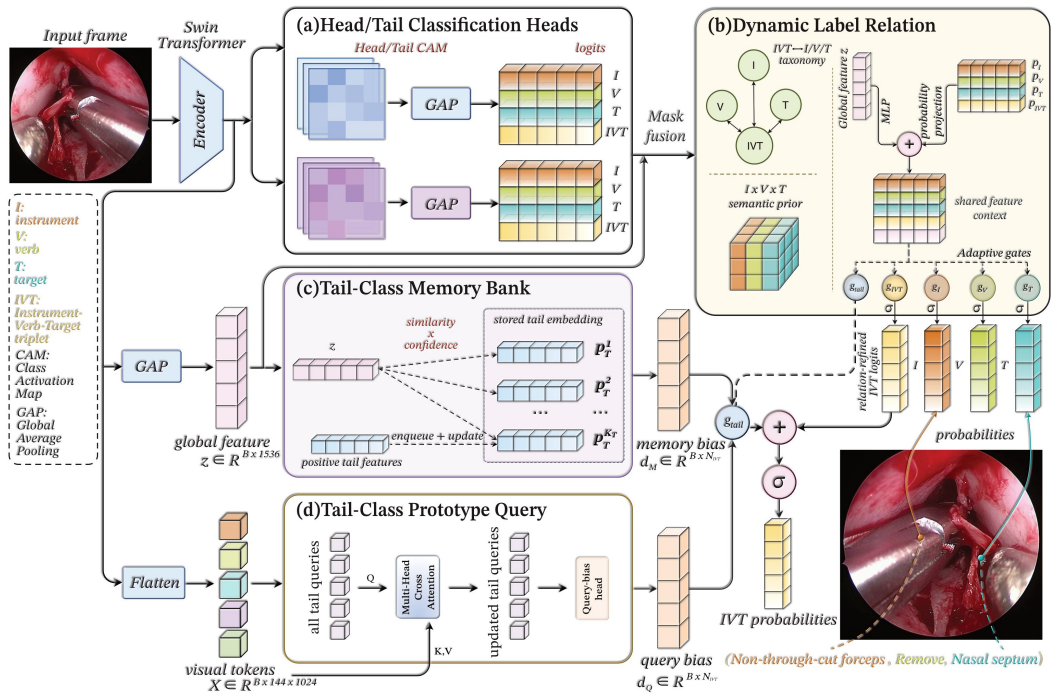


Figure 1: Overview of RIVET. (a) Separate head and tail classifiers first generate class activation maps (CAMs), which are globally pooled to obtain base predictions for surgical components and IVT triplets. (b) Dynamic Structural Relation refines predictions across all classes using deterministic mappings between components and triplets. For tail triplets, (c) the Tail-Class Memory Bank provides global visual references from historical positive samples, while (d) the Tail-Class Prototype Query extracts spatial evidence specific to each candidate from the current frame. The two visual residuals are computed independently and fused with the structurally refined IVT logits through a shared adaptive gate.

3.2 Head-Tail Dual-Branch Classification

Following Gui & Wang (2024), we partition each task's classes by frame-level frequency. The top K_{IVT} triplet classes form the head set; the remainder form the tail set $\mathcal{T}_+$, with $|\mathcal{T}_+| = K_{\text{tail}}$. Component classes are partitioned analogously, with complementary masks $\mathbf{m}_H^r$ and $\mathbf{m}_{\text{tail}}^r$ for each task. Two 1×1 classifiers operate on F, and their globally averaged outputs give the base logits $\boldsymbol{\ell}_0^r = \mathbf{m}_H^r \odot \boldsymbol{\ell}_H^r + \mathbf{m}_{\text{tail}}^r \odot \boldsymbol{\ell}_{\text{tail}}^r$, with probabilities $\mathbf{p}_0^r = \sigma(\boldsymbol{\ell}_0^r)$. The corresponding mask-selected classifier weight $\widehat{\mathbf{w}}_c^r \in \mathbb{R}^D$ is reused in tail-query construction.

3.3 Dynamic Structural Relation

Dynamic Structural Relation exchanges evidence through known component–triplet mappings rather than frequency-estimated relational priors. These mappings remain available for rare classes, while reference scores and correction gates depend on the current frame.

Bidirectional structural references. Let $\pi_r(q)$ identify the component of triplet q in task r. Define $A_{qc}^r = \mathbf{1}[\pi_r(q) = c]$, with $A^r \in \{0,1\}^{C_{\text{IVT}} \times C_r}$, and $\mathcal{N}_r(c) = \{q : A_{qc}^r = 1\}$. Component-to-triplet support is

$$c_q = \left(\prod_{r \in \mathcal{R}} p_{0,\pi_r(q)}^r \right)^{1/3}, \qquad s_q = \frac{1}{2} \left(c_q + p_{0,q}^{\text{IVT}} \right). \tag{1}$$

Here, c_q is compositional support, not a joint probability; s_q balances component-derived and direct IVT evidence. Conversely, triplet probabilities aggregated over $\mathcal{N}_r(c)$ are averaged with $p^r_{0,c}$ to form the component reference u^r_c.

Frame-adaptive refinement. Let $\mathbf{p}_0$ concatenate the four base-probability vectors. The scene feature and predictions determine a shared context and five scalar gates:

$$\mathbf{h} = \phi_h([\phi_z(\mathbf{z}), \phi_p(\mathbf{p}_0)]), \qquad [g_I, g_V, g_T, g_{\mathrm{IVT}}, g_{\mathrm{tail}}] = \sigma(W_g \mathbf{h}). \tag{2}$$

Brackets denote concatenation, and the ϕ functions are learned mappings. With $\boldsymbol{\rho}^r = \mathbf{u}^r$ for component tasks and $\boldsymbol{\rho}^{\mathrm{IVT}} = \mathbf{s}$, structural refinement is

$$\widetilde{\boldsymbol{\ell}}^r_{\mathrm{str}} = \boldsymbol{\ell}^r_0 + g_r \left[\mathbf{b}_r(\mathbf{h}) + \lambda_s \tanh(\alpha_r) \operatorname{logit}(\boldsymbol{\rho}^r)\right], \quad r \in \mathcal{R} \cup \{\mathrm{IVT}\}. \tag{3}$$

The scene-conditioned residual $\mathbf{b}_r(\mathbf{h}) \in \mathbb{R}^{C_r}$ complements the logit-transformed structural reference. The signed coefficient $\lambda_s \tanh(\alpha_r)$ scales the reference term, while g_r scales the full residual across classes. We use $\lambda_s = 0.24$ and initialize $\alpha_r = 0.1$. This is a soft refinement, not a hard consistency constraint. Component outputs are $\widetilde{\boldsymbol{\ell}}^r = \widetilde{\boldsymbol{\ell}}^r_{\mathrm{str}}$; g_{tail} controls the additional IVT residuals below.

3.4 Tail-Class Memory Bank

Positive tail-class examples can be scarce within a mini-batch. The Tail-Class Memory Bank retains their features across training batches to provide visual references beyond taxonomy-based support.

For each $q \in \mathcal{T}_+$, a first-in-first-out queue $\mathcal{M}_q$ stores normalized positive global features, with capacity $K = 8$. A momentum prototype $\boldsymbol{\mu}_q$ is initialized from the first normalized positive feature and updated on subsequent positive observations:

$$\boldsymbol{\mu}_q \leftarrow \mathrm{Norm}\left(m\boldsymbol{\mu}_q + (1-m)\,\mathrm{Norm}(\mathbf{z})\right), \qquad m = 0.9. \tag{4}$$

Here, Norm denotes vector normalization. Queue observations determine a confidence weight $\kappa_q \in [0, 1]$ based on their number and diversity, with lower weights for sparse or redundant observations. For an initialized prototype, the historical residual is

$$d^M_q = \frac{\lambda_M \kappa_q}{\tau_M} \cos(\mathbf{z}, \boldsymbol{\mu}_q), \qquad q \in \mathcal{T}_+, \qquad \lambda_M = 0.28, \quad \tau_M = 0.07. \tag{5}$$

The confidence weight scales agreement with the historical reference. Signed similarities allow this residual to reinforce or suppress a candidate.

3.5 Tail-Class Prototype Query

Tail-Class Prototype Query (Tail Query) retrieves candidate-specific spatial evidence to distinguish tail triplets that share components. Its queries use component embeddings and classifier weights rather than memory prototypes, without requiring a separate detection head.

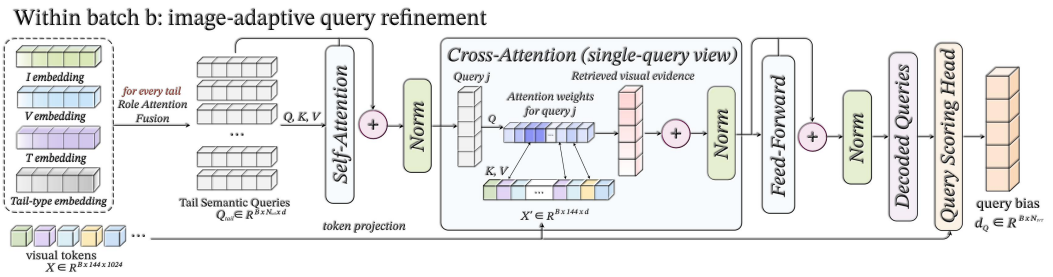


Figure 2: Tail Query refinement. Role-aware triplet queries are matched to current-frame spatial tokens, and the resulting evidence is used for positive tail-class calibration.

Role-aware query construction. For $q = (i, v, t)$, each component token combines its embedding, projected mask-selected classifier weight, and role embedding. A tail-class type embedding forms a fourth token. Self-attention yields role representations $S_{q,n} \in \mathbb{R}^d$ and normalized weights $\beta_{q,n}$. The initial query is

$$\mathbf{u}_q^{(0)} = f_{\text{fuse}} \left(\sum_{n=1}^{4} \beta_{q,n} S_{q,n} + \phi_{\text{IVT}} \left(\widehat{\mathbf{w}}_q^{\text{IVT}} \right) \right). \tag{6}$$

This combines the weighted role representations with the complete-triplet classifier representation.

Spatial matching and refinement. Stacked queries $U^{(0)} \in \mathbb{R}^{K_{\text{tail}} \times d}$ undergo self-attention, cross-attention to projected spatial tokens $X_c \in \mathbb{R}^{L \times d}$, and feed-forward transformations (Figure 2). Each sublayer uses residual addition followed by normalization.

For each decoded query, $\mathbf{e}_q$ contains the strongest local match, global match, attended-region match, and attention concentration. A scoring head produces s_q^{TQ} from these signals, with learnable scale and bias applied to peak responses of the first three matches. The query residual is

$$d_q^{\text{TQ}} = \lambda_{\text{TQ}} \, \text{ReLU}\big(s_q^{\text{TQ}}\big), \qquad q \in \mathcal{T}_+. \tag{7}$$

Here, λ_{TQ} scales the correction; non-positive scores yield no residual.

Query supervision. $\mathcal{L}_{\text{match}}$ supervises query–frame consistency. $\mathcal{L}_{\text{dn}}$ penalizes negative queries obtained by component perturbation and random sampling. $\mathcal{L}_{\text{rank}}$ encourages positive queries to outscore competing tail-class queries.

3.6 Prediction Fusion and Training Objective

Prediction fusion. The final IVT logits combine the three evidence sources:

$$\widetilde{\boldsymbol{\ell}}^{\text{IVT}} = \widetilde{\boldsymbol{\ell}}_{\text{str}}^{\text{IVT}} + g_{\text{tail}} \left(\mathbf{d}^M + \mathbf{d}^{\text{TQ}} \right). \tag{8}$$

Both tail residuals have C_{IVT} entries and are zero outside $\mathcal{T}_+$. The memory and query modules retain separate states, while g_{tail} scales their summed correction. A sigmoid gives the final probabilities.

Classification supervision. We use class-weighted asymmetric loss (ASL) (Ridnik et al., 2021), with $\gamma_+ = 0$, $\gamma_- = 2$, and ASL probability clipping disabled. Class weights are based on inverse effective positive-sample counts, normalized to unit mean over classes with observed positives, and bounded to limit extreme values.

The strongest associated IVT response, $\ell_c^{\text{IVT}\to r} = \max_{q \in \mathcal{N}_r(c)} \widetilde{\ell}_q^{\text{IVT}}$, provides additional supervision for each component. The classification objective is

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{ASL}} \left(\widetilde{\boldsymbol{\ell}}^{\text{IVT}}, \mathbf{y}^{\text{IVT}} \right) + \sum_{r \in \mathcal{R}} \left[\mathcal{L}_{\text{ASL}} \left(\widetilde{\boldsymbol{\ell}}^{r}, \mathbf{y}^{r} \right) + \mathcal{L}_{\text{ASL}} \left(\boldsymbol{\ell}^{\text{IVT}\to r}, \mathbf{y}^{r} \right) \right]. \tag{9}$$

Relation supervision and total objective. Let $\bar{\mathbf{p}}_{\text{ref}}^r$ denote the IVT-derived component reference and $\mathbf{s}$ the triplet reference in Equation (1). Relation supervision uses these references as detached targets:

$$\mathcal{L}_{\text{rel}} = \frac{1}{3} \sum_{r \in \mathcal{R}} \text{MSE} \left(\sigma(\widetilde{\boldsymbol{\ell}}^{r}), \text{sg}(\bar{\mathbf{p}}_{\text{ref}}^r) \right) + \text{BCEWithLogits} \left(\widetilde{\boldsymbol{\ell}}^{\text{IVT}}, \text{sg}(\mathbf{s}) \right), \tag{10}$$

where sg stops gradients through the targets. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{rel}} \mathcal{L}_{\text{rel}} + \lambda_{\text{TQ}}^{\text{loss}} \left(\mathcal{L}_{\text{match}} + \mathcal{L}_{\text{dn}} + \lambda_R \mathcal{L}_{\text{rank}} \right). \tag{11}$$

The coefficients λ_{rel}, $\lambda_{\text{TQ}}^{\text{loss}}$, and λ_R weight relation supervision, query supervision, and query ranking, respectively.

4 EXPERIMENTS

4.1 DATASET AND EXPERIMENTAL SETUP

SinusTriplet-95 Dataset. SinusTriplet-95 contains 324 surgical video clips from 15 patients, totaling 670,381 frames and 2,428 annotated intraoperative activity segments. The annotations were provided by multiple senior surgeons. The dataset includes 12 instrument classes, 13 verb classes, 17 anatomical target classes, and 95 valid IVT triplet classes. Sinus endoscopy presents several challenges for fine-grained interaction recognition, including closely spaced anatomical structures, similar tissue appearance, frequent instrument occlusion, and interference from fluid or smoke. As shown in Figure 3, the statistics of SinusTriplet-95 reveal a highly imbalanced label distribution: 50 of the 95 IVT classes contain no more than five annotated segments and a small number of frequent classes account for a large proportion of the annotations.

Experimental Protocol, Metrics, and Implementation Details. The dataset is split at the patient level into training, validation, and test sets using a fixed 6:2:2 ratio. The same split is used in all experiments, and the checkpoint with the highest validation $\mathrm{AP}_{\mathrm{IVT}}$ is used for test evaluation. We report AP for instruments, verbs, and targets (AP_{I}, AP_{V}, and AP_{T}), for instrument–verb and instrument–target pairs ($\mathrm{AP}_{\mathrm{IV}}$ and $\mathrm{AP}_{\mathrm{IT}}$), and for complete IVT triplets ($\mathrm{AP}_{\mathrm{IVT}}$). $\mathrm{AP}_{\mathrm{IVT}}$ is the primary metric. As reported in Table 1, RIVET-T uses Swin-T with 224×224 inputs; RIVET-B uses Swin-B with either 224×224 or 384×384 inputs. RIVET-Ens combines RIVET-T with the Swin-B(384) configuration of RIVET-B. The backbones are pretrained on ImageNet-22K. RIVET is trained for 20 epochs with a batch size of 24 using SGD and a weight decay of 10^{-5}. The loss formulation and hyperparameters are provided in Sec. 3.6. All methods are evaluated on the same fixed patient-level split using the same evaluation metrics.

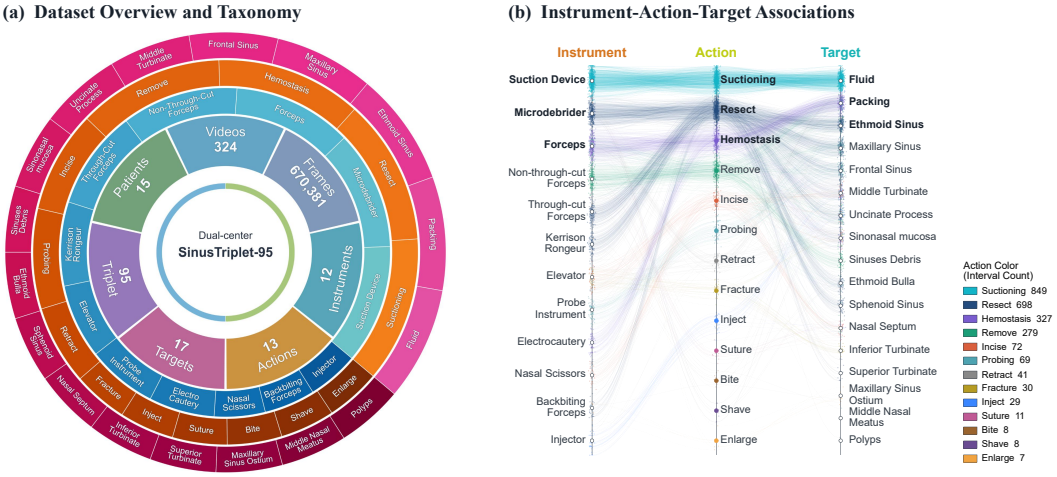


Figure 3: Statistical overview of SinusTriplet-95. (a) Dataset scale and taxonomy. The radial plot summarizes the dataset composition and the distributions of instrument, verb, and target categories, with sector widths proportional to their occurrence frequencies. (b) Interval-level Instrument–Verb–Target Associations. Each link corresponds to one annotated segment, with colors denoting verb categories and the legend reporting the number of segments associated with each verb.

4.2 COMPARISON WITH EXISTING METHODS

RIVET delivers consistent gains in complete-triplet recognition across backbones and surgical settings. On SinusTriplet-95 (Table 1), RIVET-B with Swin-B(384) reaches 47.47% $\mathrm{AP}_{\mathrm{IVT}}$, exceeding the matched CurConMix baseline by 6.05 percentage points and even CurConMix-Ens by 2.38 points, showing that the advantage does not depend on ensembling. RIVET-Ens further achieves 50.80% $\mathrm{AP}_{\mathrm{IVT}}$ (+5.71 points over CurConMix-Ens) and leads all six AP metrics. The benefit also extends beyond sinus endoscopy: on CholecT45 (Appendix A.1), RIVET outperforms TERL and CurConMix in $\mathrm{AP}_{\mathrm{IVT}}$ across matched backbones, with its ensemble reaching 43.9% (+3.2 points

over CurConMix-Ens). These results support the applicability of the proposed evidence-guided refinement in both evaluated surgical settings.

Table 1: Benchmarking surgical triplet recognition methods on SinusTriplet-95. We report AP for instrument, verb, target, instrument–verb, instrument–target, and complete IVT triplet recognition. The best result in each column is highlighted in bold.

Method	Backbone	$\mathrm{AP_I}$	$\mathrm{AP_V}$	$\mathrm{AP_T}$	$\mathrm{AP_{IV}}$	$\mathrm{AP_{IT}}$	$\mathrm{AP_{IVT}}$
RDV	Res18	0.4842	0.4579	0.3419	0.4735	0.3287	0.3172
SelfD	SwinB(224)	0.5754	0.5608	0.4053	0.5405	0.3754	0.3602
TERL-T	SwinT(224)	0.6141	0.5954	0.4254	0.5602	0.3925	0.3707
TERL-B	SwinB(224)	0.6208	0.6092	0.4403	0.5751	0.4111	0.3904
TERL-B	SwinB(384)	0.6435	0.6229	0.4758	0.6001	0.4307	0.4106
TERL-Ens	Ensemble	0.6559	0.6251	0.4802	0.6157	0.4589	0.4403
CurConMix-T	SwinT(224)	0.5906	0.5721	0.4205	0.5551	0.4056	0.3825
CurConMix-B	SwinB(224)	0.6001	0.5807	0.4352	0.5656	0.4251	0.4002
CurConMix-B	SwinB(384)	0.6072	0.5842	0.4412	0.5703	0.4386	0.4142
CurConMix-Ens	Ensemble	0.6355	0.6101	0.4728	0.6181	0.4645	0.4509
RIVET-T	SwinT(224)	0.6518	0.6078	0.4661	0.5926	0.4462	0.4372
RIVET-B	SwinB(224)	0.6609	0.6214	0.4857	0.6096	0.4664	0.4525
RIVET-B	SwinB(384)	0.6741	0.6314	0.5055	0.6245	0.4826	0.4747
RIVET-Ens	Ensemble	**0.6923**	**0.6504**	**0.5256**	**0.6529**	**0.5112**	**0.5080**

4.3 Ablation Study

The head–tail evaluation protocol and metrics are detailed in Appendix A.2. Table 2 shows that the proposed evidence sources provide complementary gains. SR raises $\mathrm{AP_{IVT}}$ from 35.17% to 41.52%. Adding TMB to SR further improves $\mathrm{AP_{IVT}}$ to 45.80%, compared with 43.28% when TQ is added instead. Combining all three modules yields the best $\mathrm{AP_{IVT}}$ of 47.47%, indicating that TQ provides additional benefits alongside historical prototype evidence. Both head and tail classes benefit substantially: $\mathrm{AP_{Head}}$ increases from 43.33% to 55.07%, while $\mathrm{AP_{Tail}}$ rises from 27.83% to 40.63%, corresponding to gains of 11.74 and 12.80 percentage points, respectively. Meanwhile, H improves from 33.89% to 46.76%.

Table 2: Ablation results of core components. All values are reported in percentage; H is the harmonic mean of $\mathrm{AP_{Head}}$ and $\mathrm{AP_{Tail}}$. SR, TMB, and TQ denote Structural Relation, Tail Memory Bank, and Tail Query, respectively.

Modules			Overall AP						Head–Tail AP		
SR	TMB	TQ	$\mathrm{AP_I}$	$\mathrm{AP_V}$	$\mathrm{AP_T}$	$\mathrm{AP_{IV}}$	$\mathrm{AP_{IT}}$	$\mathrm{AP_{IVT}}$	$\mathrm{AP_{Head}}$	$\mathrm{AP_{Tail}}$	H
			59.78	54.36	41.94	52.22	38.99	35.17	43.33	27.83	33.89
✓			61.40	56.87	43.92	56.56	42.62	41.52	49.71	34.15	40.49
✓		✓	63.72	59.61	46.58	58.47	44.71	43.28	51.46	35.92	42.31
✓	✓		64.88	60.29	47.76	60.27	46.08	45.80	53.30	39.05	45.08
✓	✓	✓	**67.41**	**63.10**	**50.55**	**62.45**	**48.26**	**47.47**	**55.07**	**40.63**	**46.76**

4.4 Further Analysis and Visualization

RIVET improves recognition of underrepresented triplets: it achieves higher mean tail AP than CurConMix and TERL on CholecT45 and SinusTriplet-95, outperforming both in 46 of 50 SinusTriplet-95 tail classes (Figure 4). Relative to the ablation baseline, RIVET improves $\mathrm{AP_{Tail}}$ by 12.80 percentage points and $\mathrm{AP_{Head}}$ by 11.74 percentage points (Table 2), demonstrating gains across both tail and head triplets. Qualitative cases (Appendix A.3, Figure 5) illustrate that relevant memories can reinforce or correct predictions. When retrieval is ambiguous, refinement can still narrow the probability gap between the ground truth and the top incorrect prediction without reversing their ranking, motivating more discriminative evidence from the current frame.

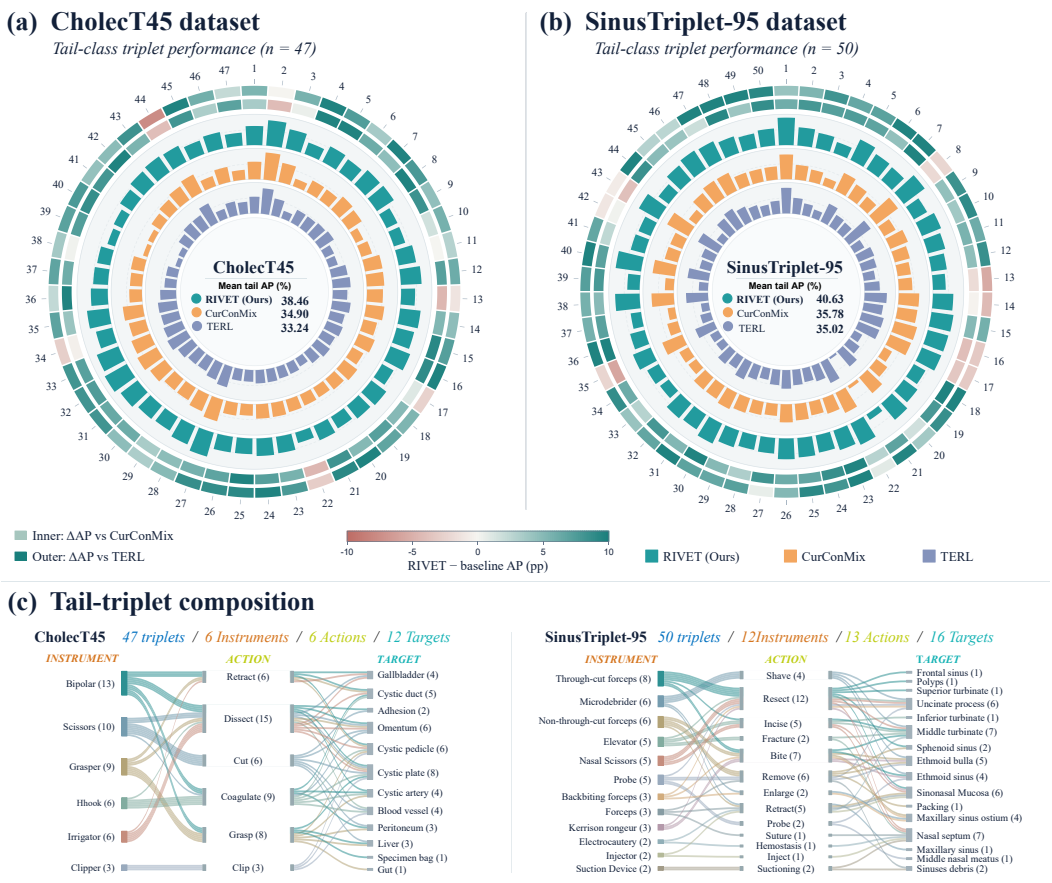


Figure 4: Tail-class performance on CholecT45 ($n = 47$) and SinusTriplet-95 ($n = 50$). Radial bars show class-wise AP, while the inner and outer rings show RIVET's AP differences relative to CurConMix and TERL, respectively; green indicates higher AP for RIVET and red indicates lower AP, with color intensity reflecting the magnitude of the difference. Instrument–action–target composition of the tail-triplet classes in both datasets.

5 Conclusion

We introduced SinusTriplet-95 and RIVET, demonstrating the value of structural, historical, and local evidence for long-tailed surgical triplet recognition. The dual-center benchmark extends evaluation to sinus endoscopy; RIVET integrates bidirectional calibration, historical prototypes weighted by confidence, and spatial queries tailored to each candidate through a gating mechanism that adapts to each frame. On SinusTriplet-95, RIVET achieves 50.80% ensemble $\mathrm{AP}_{\mathrm{IVT}}$, with improvements in 46 of 50 tail classes over both matched baselines. Ablations and CholecT45 results support the framework across both evaluated procedures. However, ambiguous memory retrieval can leave errors unresolved, indicating that evidence selection remains an open challenge. By establishing a benchmark for triplet recognition in sinus endoscopy, this work opens a new application setting beyond cholecystectomy-centered benchmarks and provides a foundation for future evidence-guided recognition across procedures, anatomies, and institutions.

AI Use Statement

Generative AI tools were used solely for LaTeX formatting and typesetting and were not used to generate the scientific content of this work. The authors take full responsibility for the final content.

REFERENCES

Y. Chen, S. He, Y. Jin, and J. Qin. Surgical activity triplet recognition via triplet disentanglement. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 451–461, 2023.

S. Gui and Z. Wang. Tail-enhanced representation learning for surgical triplet recognition. In *Medical Image Computing and Computer Assisted Intervention*, pp. 689–699, 2024.

S. Gui, Z. Wang, J. Chen, X. Zhou, C. Zhang, and Y. Cao. MT4MTL-KD: A multi-teacher knowledge distillation framework for triplet recognition. *IEEE Transactions on Medical Imaging*, 43 (4):1628–1639, 2024.

Y. Jeon, J. Shin, S. Park, B. Kim, K. Park, N. Oh, and K.-H. Jung. CurConMix: A curriculum contrastive learning framework for enhancing surgical action triplet recognition. In *Medical Image Computing and Computer Assisted Intervention*, pp. 149–158, 2025.

F. Li, S. Zou, W. Si, Y. Gao, S. Li, and A. Hao. TRCoRSurg: Temporal-relational co-reasoning for surgical video triplet recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2811–2820, 2026.

L. Li, X. Li, S. Ding, Z. Fang, M. Xu, H. Ren, and S. Yang. SIRNet: Fine-grained surgical interaction recognition. *IEEE Robotics and Automation Letters*, 7(2):4212–4219, 2022.

Z. Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

C. I. Nwoye et al. CholecTriplet2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis*, 86:102803, 2023.

J. Pei, J. Zhang, G. Qin, K. Wang, Y. Jin, and P.-A. Heng. Instrument-tissue-guided surgical action triplet detection via textual-temporal trail exploration. *IEEE Transactions on Medical Imaging*, 44(12):5278–5289, 2025.

T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

S. Sharma, C. I. Nwoye, D. Mutter, and N. Padoy. Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 505–514, 2023a.

S. Sharma, C. I. Nwoye, D. Mutter, and N. Padoy. 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, 2023b.

N. Xi, J. Meng, and J. Yuan. Forest graph convolutional network for surgical action triplet recognition in endoscopic videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (12):8550–8561, 2022.

A. Yamlahi, T. N. Tran, P. Godau, et al. Self-distillation for surgical action recognition. In *Medical Image Computing and Computer Assisted Intervention*, pp. 637–646, 2023.

A Appendix

A.1 Comparison on CholecT45

Evaluation scope and baselines. We evaluate RIVET on CholecT45 to assess its performance in laparoscopic cholecystectomy, extending the evaluation beyond sinus endoscopy. The comparison includes RDV (Nwoye et al., 2022), RiT (Sharma et al., 2023b), TDN (Chen et al., 2023), MT4MTL-KD (Gui et al., 2024), SelfD (Yamlahi et al., 2023), TERL (Gui & Wang, 2024), and CurConMix (Jeon et al., 2025).

Table 3: Comparison on CholecT45. Results are reported as mean ± standard deviation of AP (%). The best result in each column is highlighted in bold.

Method	Backbone	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RDV	Res18	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	Res18	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	Res50	91.2±1.9	65.3±2.8	43.7±1.6	39.6±4.0	39.2±2.9	33.8±2.5
MT4MTL-KD	SwinL(384)	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	SwinB(224)	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-T	SwinT(224)	91.8±2.2	70.5±1.4	47.2±2.6	44.7±3.8	42.0±2.4	35.7±1.6
TERL-B	SwinB(224)	92.0±2.0	71.4±2.1	49.4±4.7	43.9±3.4	43.6±2.6	35.6±1.4
TERL-B	SwinB(384)	92.3±1.8	71.9±2.5	51.1±3.8	46.5±4.9	44.9±1.8	37.7±1.5
TERL-Ens	Ensemble	92.9±1.7	72.7±1.3	50.8±3.3	47.3±4.1	45.3±1.9	38.5±1.1
CurConMix-T	SwinT(224)	90.4±2.1	67.8±1.8	48.3±3.4	43.3±2.9	43.3±1.8	37.7±2.1
CurConMix-B	SwinB(224)	90.4±3.0	68.2±1.5	49.7±2.5	44.8±5.4	45.3±2.4	38.8±2.8
CurConMix-B	SwinB(384)	90.9±2.0	68.3±1.3	49.8±3.2	45.2±4.2	45.1±1.1	39.1±2.0
CurConMix-Ens	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
RIVET (ours)							
RIVET-T	SwinT(224)	93.5±1.5	71.4±2.2	52.4±2.3	48.1±3.7	47.2±1.7	39.2±1.9
RIVET-B	SwinB(224)	93.9±2.0	70.8±2.3	54.1±2.9	50.2±3.9	48.3±1.9	40.8±2.1
RIVET-B	SwinB(384)	94.1±2.3	73.0±1.4	55.6±3.4	**52.1±4.0**	49.4±2.0	42.4±2.2
RIVET-Ens	Ensemble	**94.6±1.9**	**73.5±1.9**	**56.2±2.5**	51.2±4.2	**51.7±1.8**	**43.9±1.6**

A.2 Head–Tail Evaluation

To further assess performance under the pronounced long-tailed distribution of SinusTriplet-95, we separately evaluate the frequent head IVT classes and the underrepresented tail IVT classes in the ablation study. The tail group consists of the 50 rare triplets identified in Sec. 4.1, each with no more than five annotated segments. These sparsely observed interactions provide limited examples for learning reliable visual references. In contrast, the head group consists of frequently occurring triplets, providing a complementary evaluation of performance on well-represented interactions.

We report AP_{Head} and AP_{Tail} by averaging the class-wise IVT AP over the head and tail groups, respectively. Their harmonic mean is additionally reported to characterize the balance between performance on frequent and rare triplets:

$$H = \frac{2\,AP_{Head}AP_{Tail}}{AP_{Head} + AP_{Tail}}.$$

A.3 Qualitative Analysis of Memory-Guided Prediction Refinement

Figure 5 illustrates three representative outcomes of memory guided prediction refinement. In Case 1, the ground truth triplet is clearly distinguished from the other retrieved candidates, providing reliable historical evidence that enables RIVET to correct the baseline prediction by promoting the ground truth from rank 2 to rank 1. In Case 2, the baseline already identifies the correct triplet, and the retrieved memory further reinforces this decision by increasing its prediction probability while

suppressing the competing predictions. Case 3 reveals a more challenging situation. Although the ground truth receives the highest retrieval similarity, its score of 0.61 is nearly identical to the 0.60 similarity of an incorrect candidate, making the retrieved evidence inherently ambiguous. RIVET nevertheless shifts the prediction in the correct direction, increasing the ground truth probability from 0.22 to 0.26 while reducing the top incorrect prediction from 0.29 to 0.27. This narrows the probability gap from 0.07 to 0.01, but the refinement is not sufficient to reverse their ranking. This case suggests that memory retrieval can still provide useful corrective evidence even when the final prediction remains incorrect, but its effectiveness depends on how clearly the retrieved memories distinguish the ground truth from competing candidates. When multiple triplets receive nearly identical retrieval similarities, the historical evidence alone may be insufficient to resolve the ambiguity, highlighting the need for more discriminative evidence from the current frame.



Figure 5: Qualitative examples of memory bank guided prediction refinement. For each case, the left panel shows the input frame and its ground truth IVT triplet, while the middle panel presents the top three triplets retrieved from the memory bank with their similarity scores. The right panel compares the top three predictions of CurConMix and RIVET, with #1 to #3 indicating their ranks and the corresponding values representing IVT prediction probabilities. Green boxes indicate the ground truth triplet, while orange boxes indicate incorrect predictions.