Chain-of-Look Prompting for Verb-centric Surgical Triplet Recognition in Endoscopic Videos

Nan Xi
State University of New York at Buffalo
Buffalo, USA
nanxi@buffalo.edu

Jingjing Meng
Amazon
USA
jingjing.meng1@gmail.com

Junsong Yuan
State University of New York at Buffalo
Buffalo, USA
jsyuan@buffalo.edu

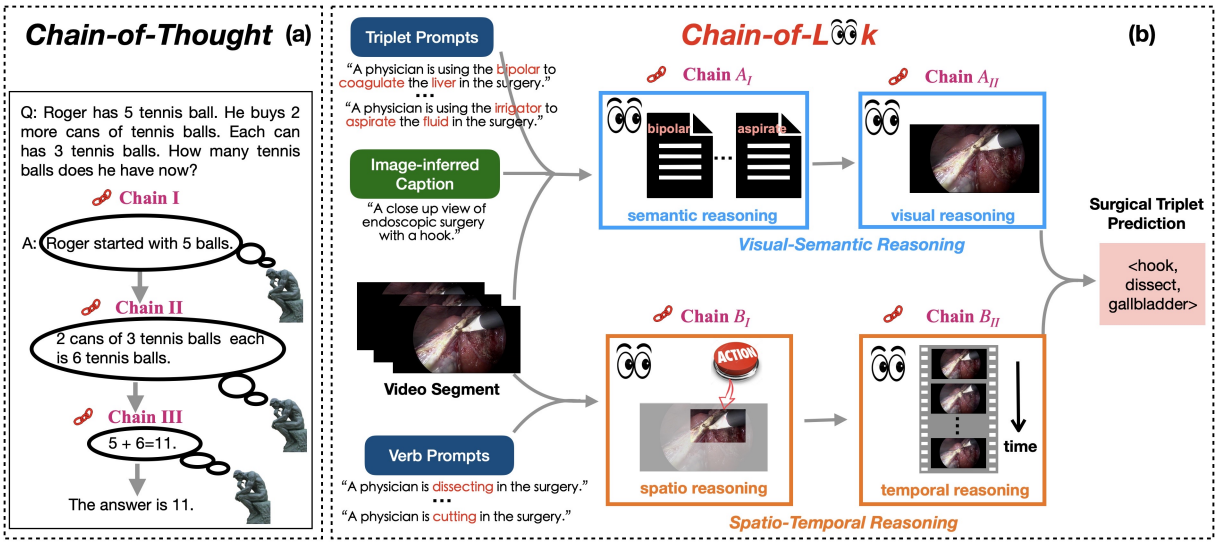

Figure 1: Chain-of-thought prompting in natural language processing (left) and chain-of-look prompting in surgical triplet recognition (right). (a) Chain-of-thought prompting helps answer a complicated question via a series of intermediate reasoning steps, enabling transparancy and explainability. (b) Our proposed chain-of-look prompting helps multi-modal video reasoning via (1) visual-semantic reasoning process that focuses on understanding semantics from the visual information and (2) spatio-temporal reasoning process that leverages whole video context to understand the activity.

ABSTRACT

Surgical triplet recognition aims to recognize surgical activities as triplets (i.e., <*instrument, verb, target*>), which provides fine-grained information essential for surgical scene understanding. Existing methods for surgical triplet recognition rely on compositional methods that recognize the instrument, verb, and target simultaneously. In contrast, our method, called chain-of-look prompting, casts the problem of surgical triplet recognition as visual prompt generation from large-scale vision-language (VL) models, and explicitly decomposes the task into a series of video reasoning processes. **Chain-of-Look** prompting is inspired by: (1) the chain-of-thought prompting in natural language processing, which divides a problem into a sequence of intermediate reasoning steps; (2) the interdependency between motion and visual appearance in the human vision system. Since surgical activities are conveyed by the actions of physicians, we regard the **verbs** as the carrier of semantics in surgical endoscopic videos. Additionally, we utilize the BioMed large language model to calibrate the generated visual prompt features for surgical scenarios. Our approach captures the visual reasoning processes underlying surgical activities and achieves better performance compared to the state-of-the-art methods on the largest surgical triplet recognition dataset, CholecT50. The code is available at https://github.com/southnx/CoLSurgical.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611898

CCS CONCEPTS

• **Computing methodologies** → Activity recognition and understanding.

KEYWORDS

Surgical Triplet Recognition, Chain-of-Look Prompting, Verb-centric, Endoscopic Videos

ACM Reference Format:
Nan Xi, Jingjing Meng, and Junsong Yuan. 2023. Chain-of-Look Prompting for Verb-centric Surgical Triplet Recognition in Endoscopic Videos. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611898

1 INTRODUCTION

More than 4.8 billion people worldwide have no access to surgery [1]. However, the need for surgery keeps growing worldwide [29]. The urgency of developing surgical AI systems that are highly adaptable has become more pressing. These systems are essential not only for assisting and intervening during surgical procedures, but also for educating and training physicians.

One of the most commonly performed surgical procedures is laparoscopic cholecystectomy (LC) [20] through endoscopic videos, as it provides better clinical outcomes with less pain and faster recovery. To build a surgical assistance system for LC, a fundamental step is to recognize fine-grained surgical workflows in endoscopic videos, e.g. in the form of triplets of <*instrument, verb, target*>. These triplets provide a fine-grained and holistic surgical activity understanding for building surgical assistance systems. Existing works on surgical triplet recognition are developed based on compositional methods, which recognize each individual elements in surgical triplets separately. Triplet recognition results are then generated by combining the predictions of individual elements. For example, Tripnet [18] and Rendezvous (RDV) [19] utilize a multi-task learning framework to recognize instrument, verb and target separately from a single video frame. The state-of-the-art Forest GCN [30] recognizes surgical triplets by decomposing triplet recognition into recognizing instruments, verbs and targets respectively with the classification forest and Graph Convolutional Network (GCN). Directly applying existing works for surgical triplet recognition has two major drawbacks: (1) compositional methods model triplet recognition by the three individual elements, lacking an explicit reasoning process to guide the video understanding; (2) existing methods exert equal emphasis on instrument, verb and target, neglecting that the intent of surgical activities are mainly conveyed by the action of the physicians (i.e., verbs).

To address the above drawbacks, we propose a novel ***chain-of-look*** visual prompting scheme for ***verb-centric*** surgical triplet recognition. Firstly, since existing compositional methods lack explicit reasoning processes to guide surgical triplet recognition, we propose a chain-of-look prompting scheme to explicitly the decompose surgical triplet recognition task into a series of video reasoning processes and cast surgical triplet recognition task as visual prompt generation. Our chain-of-look visual prompting scheme draws inspiration from two main sources: **(I)** The chain-of-thought prompting [28] in natural language processing (NLP), which is designed to "prompt" the model with input-output reasoning steps. In light of this, as shown in Fig. 1, we construct the first multi-modal reasoning process by decomposing the generation of triplet prompt features into two visual reasoning steps - the first step (*chain* A_I) with global semantic information and the second step (*chain* A_{II}) with visual information; **(II)** It has been identified in neuroscience that there are two major processing systems in human vision system, one for visual recognition of objects (known as "what" stream) and the other for motion integration (known as "where" stream) [15]. These two systems work in sequence to process different visual attributes: motion integration stream is in favor of sensitivity to rapid temporal change, while visual recognition stream for coding fine details. Secondly, actions of physicians during surgery convey the intent of surgical activities, while instruments and targets are chosen to accomplish the intended actions accordingly. Therefore, verbs in surgical triplets carry key information of the surgical activities happening in endoscopic videos. To this end, we introduce the verb-centric modeling scheme for video reasoning by first modeling endoscopic videos with verb prompts (*chain* B_I) and then modeling the temporal dynamics of those verb sequences among neighboring frames(*chain* B_{II}).

Specifically, we first generate frame captions from large-scale VL model and further generate context-calibrated frame caption features with pre-trained BioMed Language Model [4] based on the original frame captions. Then we introduce visual-semantic reasoning network (VSR) and spatio-temporal reasoning network (STR) for surgical triplet prompting and verb prompting, respectively. Each network contains two steps of chains of reasoning for visual prompt featrue generation, which are akin to the chain-of-thought prompting in NLP. Specifically, VSR includes CaptionTrip Prompting (CTP) and VideoTrip Prompting (VTP). CTP is designed to incorporate global semantic information into individual surgical triplet prompts, serving as the first reasoning chain. VTP follows CTP as the other reasoning chain with video information for surgical triplet prompting. Similarly, STR also contains two reasoning chains: Verb Prompting (VP) and Dynamic GNN (D-GNN). VP is introduced to abstract visual information of each frame into a fixed number of verb prompts. D-GNN is then employed to model the temporal dynamics across frames, thus enabling verb prompt features to be more semantic-aware. Finally, triplet prediction is conducted under the guidance of verb prompt features, which is less noisy than only utilizing surgical triplet prompt features.

Our main contributions on the surgical triplet recognition in endoscopic videos are summarized as follows:

- We introduce the chain-of-look prompting and design the underlying visual reasoning processes in endoscopic videos for surgical triplet recognition, generating visual-semantic aware and spatio-temporal aware prompt features from VL models. BioMed language model is further employed to calibrate semantic features for surgical scenarios.
- We present an verb-centric surgical triplet recognition modeling scheme, which can reliably capture the most central semantic information in surgical endoscopic videos.
- Our model achieves substantial improvements compared with state-of-the-art methods for surgical triplet recognition.

2 RELATED WORK

2.1 Surgical Triplet Recognition

A surgical action in surgical ontology is described as a triplet with instrument, verb and the target that the instrument is acting upon. Surgical triplet recognition aims to recognize fine-grained surgical actions from surgery videos. In early studies, retinal microsurgery [22–25] and laparoscopic surgery [9, 25, 26] are the most concentrated fields for surgical triplet recognition. Other early methods for

surgical triplet recognition include using detector-tracker framework [24] and optical flow tracker [14]. Information of triplet annotation for surgical phase recognition [9, 26] has also been employed. However, those line of works do not produce fine-grained surgical triplet recognition results. Recent solutions for fine-grained surgical triplet recognition include Tripnet [18] and RDV [19], which are both compositional methods by modeling instrument, verb and target individually. Another recent work Forest GCN [30] employs classification forest and Graph Convolutional Network for surgical triplet recognition.

2.2 Large Scale Visual-Language (VL) Models.

Recent pretrained large-scale VL models with a representative work of CLIP [21] and BLIP [11] bridge visual and language information by jointly learning two encoders. Follow-up studies employing the pretrained VL models on downstream tasks have achieved remarkable progress, including CLIP-Adapter [5] and PointCLIP [32]. Two types of mainstream VL model structures exist currently: (1) Single-stream (1-stream) VL models [12] by directly fusing the initial language/visual representation by utilizing the joint cross-modal encoder, and (2) Double-stream (2-stream) VL models [16], which separately apply the intra-modality processing to two modalities along with a shared cross-modal encoder. Most single-stream and parts of the double-stream VL models are regarded as self-attention-based VL models because they directly perform cross-modal modeling by applying single-stream self-attention module to the modality representations. Comparatively, co-attention-based VL models decouple the intra- and cross-modal modeling processes.

2.3 Prompt Learning

Prompt learning was first introduced in NLP area [6], aiming to produce a task-specific template for language models. Common prompt learning scheme involves hard prompt learning [5] and soft prompt learning [13]. Hard prompt learning searches for a specific word for the predesigned template, such as "`I [MASK] running.`" in sentiment analysis, where the mask placeholder will be replaced with either "love" or "hate". Different from hard prompt learning, soft prompt learning is designed to tune masked tokens into learnable vectors. We employ the idea of soft prompting, proposing chain-of-look prompting modules for verbs and surgical triplets in endoscopic videos. Motivated by the well performance of prompt learning on NLP, recently researchers begin to apply it into the vision-language models. CLIP [21] uses a manually designed prompt on the text encoder, which enables the zero-shot image classification of vision-language model. To avoid human efforts on prompt design, CoOp [34] proposes a continuous prompts learning method and two implementations that can be applied on different recognition tasks. Yet CoOp [34] seems over-fitting the base classes in the training, resulting in inferior performance on unseen classes even within the same dataset. To cure this problem, CoCoOp [33] propose to generate an input-conditional vector for each image by a lightweight neural network, which boosts the classifier performance on new classes. Although CoOp and CoCoOp achieve promising improvements, they requires supervised data from the target datasets which may restrict the model scalability. In the contrary, Huang et al. [8] propose the unsuper- vised prompt learning (UPL) method which improves transfer performance of CLIP-like VL models.

2.4 Chain-of-Thought (CoT)

Chain-of-Thought (CoT) prompting is designed for enhancing LLMs by prompting them to generate a sequence of intermediate reasoning steps, generating the final answer of a multi-step problem. Those intermediate reasoning steps significantly improve the reasoning ability of LLMs to perform complex reasoning [17, 27, 28]. In addition, fine-tuning with CoT exhibit more harmless compared with no CoT [2]. It has been regarded that CoT prompting is an emergent property of model scale, suggesting the larger and more powerful language models lead to better CoT performance. In order to enhance the CoT ability and stimulate better explainability, fine-tuning models on CoT reasoning dataset would also be a feasible approach.

An example of CoT is shown in Fig. 1 (a). To solve a multi-step math world problem containing complicated reasoning task, decomposing the problem into multiple intermediate steps is commonly employed. The final answer is generated by solving each intermediate steps. LLMs is thus endowed with the ability to generate a similar chain of thought to result in the final answer of a problem. In this paper, we propose Chain-of-Look prompting similar with CoT prompting by explicitly decomposing surgical triplet recognition into a sequence of visual reasoning processes.

3 METHODOLOGY

In this part, we illustrate the architecture of our model in detail. We first present the problem formulation of surgical triplet recognition task. Then we introduce the two visual reasoning networks: Visual-Semantic Reasoning (VSR) network and Spatio-Temporal Reasoning (STR) network. These two visual reasoning networks explicitly decompose surgical triplet recognition into a sequence of visual prompting generation steps, constructing the chain-of-look prompting scheme for surgical triplet recognition.

3.1 Formulation

Denote an endoscopic video dataset with $(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ represents the set of video frames and $\mathcal{Y}$ indicates triplet labels <*instrument, verb, target*>. The number of all possible triplets in the dataset is N and the number of all possible verbs in the dataset is K. Our objective is to identify all the triplets that occur in each video frame $x \in \mathcal{X}$. The number of triplets occur in x varies, meaning there could be no triplets in x, or there could be several triplets in x simultaneously. We aim to learn a prediction model $\boldsymbol{f}_\theta : x_i \rightarrow y_i$, where θ is the model parameter, x_i denotes the input video frame and $y_i \in \{0, 1\}^{l \times 1}$ ($l = |\mathcal{Y}|$) indicates the binary triplet prediction vector.

3.2 Visual-Semantic Reasoning (VSR) Network

In this section, we illustrate VSR network in detail, which is designed as a two-step chain-of-look reasoning process for triplet prompt feature generation from pretrained large-scale VL model (BLIP [11] and CLIP [21]). Concretely, as shown in Fig. 2, the first chain-of-look reasoning process CaptionTrip Prompting (CTP) employs global semantic information of each frame for triplet prompt

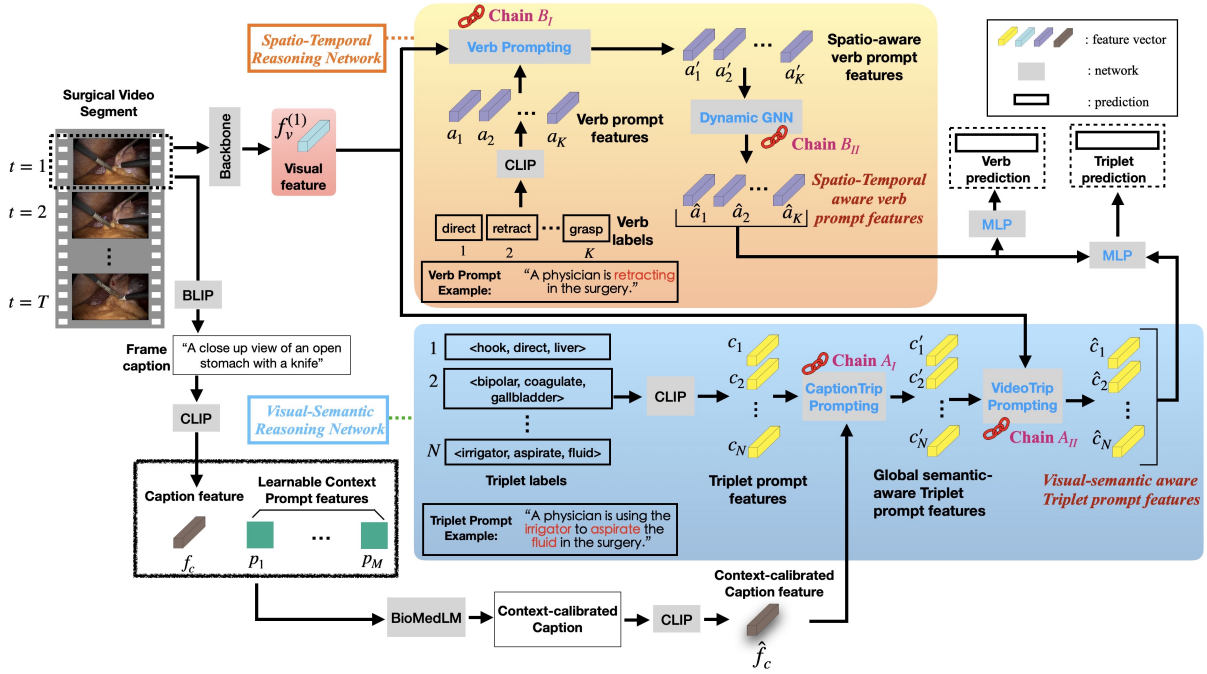

Figure 2: Our model is constructed with the visual-semantic reasoning (VSR) network and the spatio-temporal reasoning (STR) network. Each of the two networks consists of two chain-of-look reasoning steps (VTR: CaptionTrip Prompting + VideoTrip Prompting; STR: Verb Prompting + Dynamic GNN). Only the networks highlighted in blue are optimized during training. For a surgical video segment, BLIP model produces frame captions of each frame, followed by CLIP model to generate initial caption feature f_c. Learnable context prompt features $\{p_i\}_{i=1}^{M}$ are then combined with caption feature f_c and applied with a BioMed language model to calibrate frame captions to be feasible in surgical scenarios. CLIP model is further utilized to produce calibrated caption feature $\hat{f}_c$ and initial triplet prompt features $\{c_n\}_{n=1}^{N}$ as well as surgical triplet prompt features $\{a_k\}_{k=1}^{K}$. The two reasoning networks generate verb prediction and triplet prediction, which are further combined for final surgical triplet prediction.

feature generation. Follow-up reasoning process VideoTrip Prompting (VTP) employs visual information from frames. These two complementary reasoning steps enable the final triplet prompt features to be visual-semantic aware of the activities happening in endoscopic videos.

In an endoscopic video dataset containing N possible surgical triplets $\{g_n\}_{n=1}^{N}$ of <*instrument, verb, target*>, we pre-define a triplet prompt template $t(g_n)$ ="A physician is using the [Instrument] to [Verb] the [Target] in the surgery." for each triplet, where "[Instrument]", "[Verb]" and "[Target]" are replaced with their corresponding class names in each triplet. Each template is applied with a pretrained large-scale VL model CLIP (shown in Fig. 2) to generate triplet prompt feature c_n:

$$c_n = CLIP(t(g_n)) \in \mathbb{R}^d, n \in [1, N], \tag{1}$$

where d denotes the prompt feature dimension. The first chain-of-look reasoning process CTP incorporates global semantic information of each frame into triplet prompt features for semantic reasoning. To extract semantic information in video frames, we employ the current state-of-the-art image caption model BLIP to generate frame caption from frame x and further generate caption feature f_c with CLIP model:

$$f_c = CLIP(BLIP(x)) \in \mathbb{R}^d. \tag{2}$$

As shown in Fig. 2, since BLIP model is not trained on datasets in medical domain, the generated frame captions from BLIP model are not exactly coherent with the semantic meaning of the surgical scene in the frame. Therefore, we introduce M learnable context prompt features $\{p_i\}_{i=1}^{M}, p_i \in \mathbb{R}^d$ into caption feature f_c and apply BioMed Large Language Model (BioMedLM) to calibrate global semantic information with BioMed domain knowledge, where the value of M equals the sum of class numbers of instrument, verb and target, respectively. $\{p_i\}_{i=1}^{M}$ are initialized with word embeddings of class names of instrument, verb and target. Thus, we generate context-calibrated caption feature $\hat{f}_c$:

$$\hat{f}_c = CLIP(BioMedLM(Avg(f_c, p_1, \cdots, p_M))) \in \mathbb{R}^d, \tag{3}$$

where Avg indicates average operation. The overall caption features are denoted as $\hat{\mathbf{f}}_c \in \mathbb{R}^{T \times d}$, where T is the length of a video segment.

CaptionTrip Prompting (CTP). The **first chain-of-look prompting** scheme CTP in VSR network is designed for semantic reasoning of triplet prompt features by "looking at" global semantic information of endoscopic video frames. CTP module takes triplet

prompt features $\{c_i\}_{i=1}^{N}$ and context-calibrated caption feature $\hat{f}_c$ as inputs, generating global semantic-aware triplet prompt features $\{c_i'\}_{i=1}^{N}, c_i' \in \mathbb{R}^d$. Concretely, CTP module consists of a multi-head attention (MHA), where the query is triplet prompt feature c_i while the key and value are both context-calibrated caption feature $\hat{f}_c$. A feed-forward network (FFN) is followed to learn video-specific prompt feature c_i',

$$\bar{c}_i = MHA(c_i, \hat{\mathbf{f}}_c) + c_i, \tag{4}$$

$$c_i' = FFN(\bar{c}_i) + \bar{c}_i. \tag{5}$$

VideoTrip Prompting (VTP). The **second chain-of-look prompting** scheme in VSR network is VTP, extending the previous semantic reasoning process to visual reasoning process by incorporating visual information into triplet prompt features. Consequently, VTP generates visual-semantic aware triplet prompt features $\{\hat{c}_i\}_{i=1}^{N}, \hat{c}_i \in \mathbb{R}^d$. VTP holds the same structure with CTP, with only the inputs changed to be the global semantic-aware triplet prompt feature c_i' and visual feature $f_v \in \mathbb{R}^d$ of video frame. For a video frame I_t at time t, visual feature is extracted from backbone model ResNet18 [7]. Similar to Eq. 4 and Eq. 5, the visual-semantic aware triplet prompt features $\hat{c}_i$ generated from VHP is formulated as

$$\bar{c}_i' = MHA(c_i', \mathbf{f}_v) + c_i', \tag{6}$$

$$\hat{c}_i = FFN(\bar{c}_i') + \bar{c}_i', \tag{7}$$

where $\mathbf{f}_v \in \mathbb{R}^{T\times d}$ represents all visual features of a video segment.

3.3 Spatio-Temporal Reasoning (STR) Network

In a surgical scene, the intention of surgical activities are determined by the actions (**verbs**) occurred in that scene, while instruments and targets serve as participants to accomplish different activities. Namely, the verbs distinguish the uniqueness of different surgical scenes, since the same instruments and targets could result in different surgical activities. Therefore, modeling temporal dynamics of verbs in endoscopic videos provides fundamental semantic information of surgical scenes. To this end, we structurize the visual feature of video frames into a fixed number of verb prompt features, where each verb prompt feature represents a specific verb class.

For all the K verb labels $\{u_k\}_{k=1}^{K}$ in the dataset, the verb prompt template $t(u_k)$="A physician is [Verb]ing in the surgery." is predefined for each verb, where "[Verb]" represents each verb name in the dataset. Then we generate verb prompt features $\{a_k\}_{k=1}^{K}$ with CLIP text encoder:

$$a_k = CLIP(t(u_k)) \in \mathbb{R}^d, k \in [1, K]. \tag{8}$$

To endow the verb prompt features with spatio-aware reasoning capabilities, we design a novel Verb Prompting (VP) module as the **first chain-of-look prompting** scheme in STR network shown in Fig. 3. VP takes visual feature f_v and verb prompt features $\{a_k\}_{k=1}^{K}$ as inputs and outputs spatio-aware verb prompt features $\{a_k'\}_{k=1}^{K}, a_k' \in \mathbb{R}^d$. The VP module is divided into two stages: (I) In the first stage, to align the verb prompt features and visual features to the same embedding space, each visual feature f_v is applied with a Visual Prompting Network (VPN) against all the K verb prompt features $\{a_k\}_{k=1}^{K}$. The VPN first consists of a Multi-Head Attention (MHA), where verb prompt feature a_k serves as query, while visual feature f_v serves as key and value. MHA is followed by a Layer Norm (LN) module, Feed-Forward Network (FFN) and another LN. In this way, we generate learned visual prompt feature f_v':

$$f_v' = VPN(a_k, f_v) \in \mathbb{R}^d, \tag{9}$$

which is aligned to the same embedding space with respect to each verb prompt feature a_k. (II) The second stage of VP module employs a lightweight attention module

$$Atten(Q, K, V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{10}$$

to compute the relative importance of the learned visual prompt feature f_v' to a specific verb prompt feature a_k, where $Q, K, V, \sqrt{d}$ represent query, key, value and scaling factor, respectively. This attention module takes f_v' as query, while a_k as key and value, outputting the spatio-aware verb prompt features $\{a_k'\}_{k=1}^{K}$: $a_k' = Atten(f_v', a_k)$. Thus the combined formulation of a_k' can be expressed as:

$$a_k' = Atten(Q = VPN(a_k, f_v), K = a_k, V = a_k). \tag{11}$$

The above visual reasoning step happens on spatio domain, focusing on what actions (verbs) are happening in a particular frame. Since a video segment consists of a sequence of interdependent frames, the next visual reasoning step should extend from one frame to a sequence of frames along the timeline. In light of this, we design the **second chain-of-look prompting** module to extend the reasoning process from spatio domain to temporal domain. Now that we have semantic-aware verb prompt features $\{a_k'\}_{k=1}^{K}$ of each frame, we construct a fully-connected graph $\mathcal{G}$ at each time t, whose nodes are the K verbs with node features $\{a_k'\}_{k=1}^{K}$. Motivated by the ROLAND model [31] of dynamic GNN, we capture the temporal dynamics of semantic-aware verb prompt features by recurrently updating node features over time. As shown in Fig. 4, at time t, dynamic GNN takes into a_k', followed by GNN Layer 1 to generate updated level 1 node state $H_t^{(1)}$:

$$\tilde{H}_t^{(l)} = GNN^{(l)}(H_t^{(l-1)}), \tag{12}$$

$$H_t^{(l)} = Update^{(l)}(H_{t-1}^{(l)}, \tilde{H}_t^{(l)}), \tag{13}$$

where $l = \{1, 2\}$ indicates the number of GNN layer. Then node embedding update is employed by taking $\tilde{H}_t^{(l)}$ and historical node state $H_{t-1}^{(l)}$. Following ROLAND, we take GRU (Gated Recurrent Unit) cell [3] for node embedding updating:

$$H_t^{(l)} = GRU(H_{t-1}^{(l)}, \tilde{H}_t^{(l)}). \tag{14}$$

With generated $H_t^{(l)}$, the other stacked GNN Layer and Embedding Update layer is applied to generate final node embedding H_t^{L}, where $L = 2$ in our architecture. With the second chain-of-look prompting module of dynamic GNN, the spatio-temporal aware verb prompt features $\{\hat{a}_k\}_{k=1}^{K}$ at time t is generated, which equals to H_t^{L}.

3.4 Surgical Triplet Prediction and Inference

Surgical triplet prediction is performed by employing the above computed spatio-temporal aware verb prompt features $\{\hat{a}_k\}_{k=1}^{K}$ and visual-semantic aware triplet prompt features $\{\hat{c}_n\}_{n=1}^{N}$. For each $\hat{a}_k \in \mathbb{R}^d$ of frame x at time t, we concatenate it with $\hat{c}_n \in \mathbb{R}^d$ that contains the same verb class within that triplet. Then we apply a

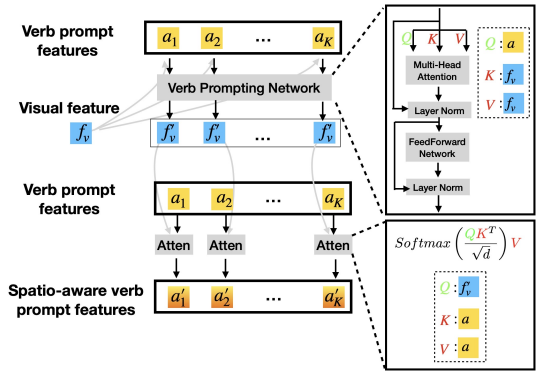

Figure 3: Verb Prompting. Visual feature f_v and verb prompt features $\{a_k\}_{k=1}^K$ are applied with verb prompting network, where $\{a_k\}_{k=1}^K$ serve as query and f_v serves as key and value. The generated visual feature f_v' is further employed as query in an attention module with verb prompt features $\{a_k\}_{k=1}^K$ perform as key and value. The spatio-aware verb prompt features $\{a_k'\}_{k=1}^K$ are finally generated from the attention module.

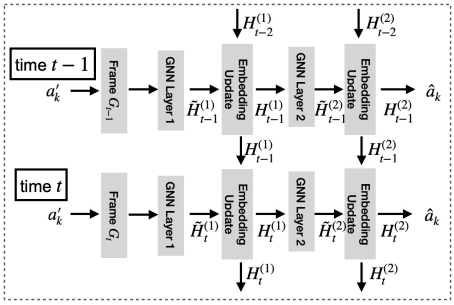

Figure 4: The structure of dynamic GNN for temporal modeling of verb prompt features across frames. For the k−th verb ($k \in [1, K]$), semantic-aware verb prompt features a_k' at time t and time $t-1$ are taken as inputs for dynamic GNN. The embedding update at each time is based on both the node states $H_t^{(l)}$ and $\tilde{H}_t^{(l)}$ at time t as well as node states $H_{t-1}^{(l)}$ and $\tilde{H}_{t-1}^{(l)}$ at time $t-1$. Two embedding update layer and GNN layer pairs are stacked for the generation of final node state $H_t^{(2)}$, which equals to the saptio-temporal aware verb prompt feature $\hat{a}_k$.

multi-layer perceptron (MLP) followed by a sigmoid function to generate predicted logits $p_n^{trip} \in [0, 1]$ of the triplet corresponding to $\hat{c}_n$: $p_n^{trip} = Sigmoid(MLP([\hat{a}_k, \hat{c}_n]))$, where $[\cdot, \cdot]$ indicates concatenating operation. Thus, the triplet prediction loss $\mathcal{L}_{trip}$ for each frame can be generated by computing the binary cross-entropy (BCE) between total triplet prediction logits $\{p_n^{trip}\}_{n=1}^N$ and triplet ground truth $\{y_n^{trip}\}_{n=1}^N$, where $y_n^{trip} \in \{0, 1\}$:

$$\mathcal{L}_{trip} = \frac{1}{N}\sum_{n=1}^{N} BCE(p_n^{trip}, y_n^{trip}). \tag{15}$$

At the same time, we apply another MLP to for verb prediction by employing semantic-temporal aware verb prompt features $\{\hat{a}_k\}_{k=1}^K$. The verb prediction logit $p_k^{verb} \in [0, 1]$ are formulated as: $p_k^{verb} = Sigmoid(MLP(a_k))$. Then the verb prediction loss $\mathcal{L}_{verb}$ is computed with the BCE between total verb prediction logits $\{p_n^{verb}\}_{i=1}^N$ and verb ground truth $\{y_n^{verb}\}_{i=1}^N$, where $y_n^{verb} \in \{0, 1\}$:

$$\mathcal{L}_{verb} = \frac{1}{K}\sum_{n=1}^{K} BCE(p_n^{verb}, y_n^{verb}). \tag{16}$$

The overall loss function $\mathcal{L}$ to be optimized in the training phase is formulated as:

$$\mathcal{L} = \mathcal{L}_{trip} + \lambda \mathcal{L}_{verb}, \tag{17}$$

where λ is the weight for verb loss.

During inference, for each frame in a given endoscopic video segment, we generate triplet prediction logits $\{p_n^{trip}\}_{n=1}^N$. A threshold for triplet predicted logits is further set for obtaining the binary triplet prediction.

4 EXPERIMENTS

4.1 Experiment Settings

4.1.1 Dataset. CholecT50 dataset [18][19] contains 50 endoscopic videos of laparoscopic cholecystectomy surgery. Following the practice of [19], among all the videos, 35 videos are chosen for training, 5 videos for validation and the remaining 10 videos for testing. There are 100 triplet classes presented as <*instrument, verb, target*> in the dataset in total. Every single frame in a video could contain one label, multiple labels or without any label.

4.1.2 Evaluation Metrics. We follow previous work [18][19] and use average precision (AP) to measure triplet classes prediction ability in the form of <*instrument, verb, target*> to evaluate the performance of the model. Top-N recognition performance is further employed to measure the ability of predicting the exact triplets within its top N outputs.

- **Average Precision**, For a given video during testing, AP score for per triplet class is computed across all frames in this video. The AP score for a given video is then obtained by averaging all AP scores of triplet classes occur in this video. The final mean AP (mAP) is calculated by averaging those AP scores over all test videos. For the AP computation of triplet classes recognition, a prediction is counted as correct only when all of the three elements of the triplet are correctly identified. We measure AP scores for instrument (AP_I), verb (AP_V), target (AP_T), instrument-verb (AP_{IV}), instrument-target (AP_{IT}) and instrument-verb-target (AP_{IVT}). Among the six AP scores, AP_{IVT} serves as the main metric for evaluating the surgical triplet prediction ability.
- **Top-N Accuracy**, Given test sample x_i, a model makes a correct prediction if the ground-truth appears in its top N prediction scores for the sample. We present the top-5, top-10 and top-20 accuracies for triplet recognition.

4.1.3 Implementation Details. All video frames are resized to a unified dimension of 256×448 and no data augmentation operations are employed during training. ResNet18 is used to extract visual features. The prompt feature dimension d is set to be 1024. GNN

Table 1: Quantitative results of comparisons between SOTA methods and our proposed model on CholecT50 dataset. Δ CTP: no CaptionTrip Prompting module; Δ VTP: no VideoTrip Prompting module; Δ VP: no Verb Prompting module; Δ D-GNN: no Dynamic GNN module; Δ BioMedLM: no BioMed Language module; Δ verb loss: $\mathcal{L}_{verb}$ is not utilized for optimization. $AP_I, AP_V, AP_T, AP_{IV}, AP_{IT}, AP_{IVT}$ represent the mean average precision of instrument, verb, targe, instrument-verb, instrument-target and instrument-verb-target across all test videos.

	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
Naive CNN [18]	57.7	39.2	28.3	21.7	18.0	13.6
TCN [19]	48.9	29.4	21.4	17.7	15.5	12.4
MLT [19]	84.5	28.4	28.2	26.6	21.2	17.6
Tripnet [18]	92.1	54.5	33.2	29.7	26.4	20.0
RDV [19]	92.0	60.7	38.3	39.4	36.9	29.9
Forest GCN [30]	93.1	60.1	40.2	36.2	37.5	36.7
Ours (Δ CTP)	85.3	56.8	37.4	32.5	34.2	32.8
Ours (Δ VTP)	84.8	57.1	35.4	32.0	35.1	31.9
Ours (Δ VP)	92.4	59.3	38.7	34.8	36.3	34.5
Ours (Δ D-GNN)	91.9	59.8	38.8	35.7	36.1	34.6
Ours (Δ BioMedLM)	92.6	60.5	39.2	36.7	38.6	36.3
Ours (Δ verb loss)	93.7	61.6	41.4	40.3	39.1	37.3
Ours	**94.1**	**62.5**	**41.9**	**41.7**	**39.5**	**38.2**

layers in Dynamic GNN is implemented as Graph Convolutional Networks (GCN) [10]. MLPs in the model consists of 3 layers, with ReLU activation function and LayerNorm at the end of each layer except the last one. We employ the pretrained CLIP [21] model as text encoder for extracting text features of 768-dim, which are further projected to 1024-dim. The parameters of CLIP and BLIP [11] are frozen during training. The number of heads in MHA module in Sec. 3.2 is set to 8. λ in Eq. 17 is 0.2. During training, the initial learning rate is 0.0002, decaying the learning rate of each parameter group by 0.1 every 40 epochs. We use Adam as the optimizer for network optimization and the weight decay is set to be 0.001.

4.2 Comparison with state-of-the-art methods

4.2.1 Quantitative Results. We compare our proposed model with current state-of-the-art (SOTA) methods. Results in Tab. 1 show that our method outperforms existing SOTA methods Forest GCN [30] and RDV [19] in terms of the main metric AP_{IVT} (1.5 mAP increase against Forest GCN and 8.3 mAP increase against RDV) in test set. Our method also surpasses competing methods in all of the five remaining AP scores ($AP_I, AP_V, AP_T, AP_{IV}, AP_{IT}$). Among these five AP scores, AP_V (from 60.1 to 62.5 compared to Forest GCN) and AP_{IV} (from 36.2 to 41.7 compared to Forest GCN) achieve the largest improvements, indicating our verb-centric modeling scheme generates more precise verb predictions compared to previous methods. For the Top-N accuracy of triplet prediction, results in Tab. 2 indicates that our method achieves higher accuracy on all of the three top-N metrics compared with SOTA methods. We further show the top 10 predicted triplet classes from different methods in Tab. 3. Results indicate the prediction pattern is different from existing methods.

4.2.2 Qualitative Results. We qualitatively compare our method with two SOTA methods RDV [19] and Forest GCN [30]. The top-5 triplet predictions of six test samples from each method are presented in Fig. 5. The qualitative results show that our model predicts most of the ground-truth surgical triplets as top-5 confident output, while RDV and Forest GCN only predict part of the ground truth triplets as top-5 outputs. There are also some failure cases in the predictions from our model, including the sample on the second row and the third column of Fig. 5. This sample contains three surgical triplets, but our model only predicts two of them in the top 5 predictions. The missing prediction is possibly caused by the severe occlusion between two instruments.

Table 2: Quantitative comparison of Top-N accuracy of triplet classes prediction of different models. Notations are the same with Tab. 1.

Method	Top-5	Top-10	Top-20
CNN [18]	67.0	80.0	90.2
TCN [19]	54.5	69.4	84.3
MTL [19]	70.2	80.2	89.5
Tripnet [18]	70.5	81.9	91.4
RDV [19]	76.3	88.7	95.9
Forest GCN [30]	83.2	91.8	97.0
Ours (Δ CTP)	72.1	79.5	91.2
Ours (Δ VTP)	71.8	82.2	90.8
Ours (Δ VP)	77.4	84.7	93.3
Ours (Δ D-GNN)	78.6	87.9	95.1
Ours (Δ BioMedLM)	82.0	88.5	94.3
Ours (Δ verb loss)	82.4	90.3	96.8
Ours	**84.5**	**92.4**	**97.2**

4.3 Ablation Studies

In this section, we analyze different design choices in our model. For both visual-semantic reasoning (VSR) network and spatio-temporal reasoning network (STR), we ablate each individual visual reasoning chain to validate the effectiveness of that chain on the whole model. From the results of Tab. 1, ablating every single reasoning

Table 3: Top 10 predicted triplet classes comparison with different methods. AP_{IVT} indicates the average precision score for every single triplet class prediction. The average triplet AP_{IVT} of top 10 predictions from each method is presented at the bottom of of the table.

Tripnet[18]		RDV[19]		Forest GCN [30]		**Ours**	
Triplet	AP_{IVT}	Triplet	AP_{IVT}	Triplet	AP_{IVT}	Triplet	AP_{IVT}
grasper,retract,gallbladder	77.3	grasper,retract,gallbladder	85.34	hook,dissect,gallbladder	84.48	hook,dissect,gallbladder	87.45
grasper,grasp,specimen-bag	76.5	grasper,grasp,specimen-bag	81.75	grasper,grasp,specimen-bag	83.61	bipolar,coagulate,liver	84.32
bipolar,coagulate,liver	67.39	hook,dissect,gallbladder	75.90	grasper,retract,gallbladder	82.35	grasper,grasp,specimen-bag	83.14
hook,dissect,gallbladder	57.54	grasper,retract,liver	66.70	bipolar,coagulate,liver	80.92	clipper,clip,cystic-duct	69.45
irrigator,aspirate,fluid	57.51	bipolar,coagulate,liver	63.12	grasper,retract,liver	70.23	hook,dissect,cystic-artery	67.84
grasper,retract,liver	54.25	clipper,clip,cystic-duct	59.68	hook,dissect,cystic-artery	65.46	irrigator,aspirate,fluid	64.33
clipper,clip,cystic-artery	47.44	bipolar,coagulate,blood-vessel	57.18	clipper,clip,cystic-duct	56.82	grasper,retract,gallbladder	61.95
scissors,cut,cystic-duct	42.57	scissors,cut,cystic-artery	53.84	grasper,retract,gallbladder	56.48	scissors,cut,cystic-duct	57.04
scissors,cut,cystic-artery	40.37	irrigator,aspirate,fluid	51.95	hook,dissect,peritoneum	49.50	grasper, retract, cystic-plane	53.67
clipper,clip,cystic-duct	39.62	clipper,clip,cystic-artery	51.52	bipolar,coagulate,gallbladder	40.24	clipper,clip,cystic-artery	47.96
mean	56.05		64.70		67.01		67.71

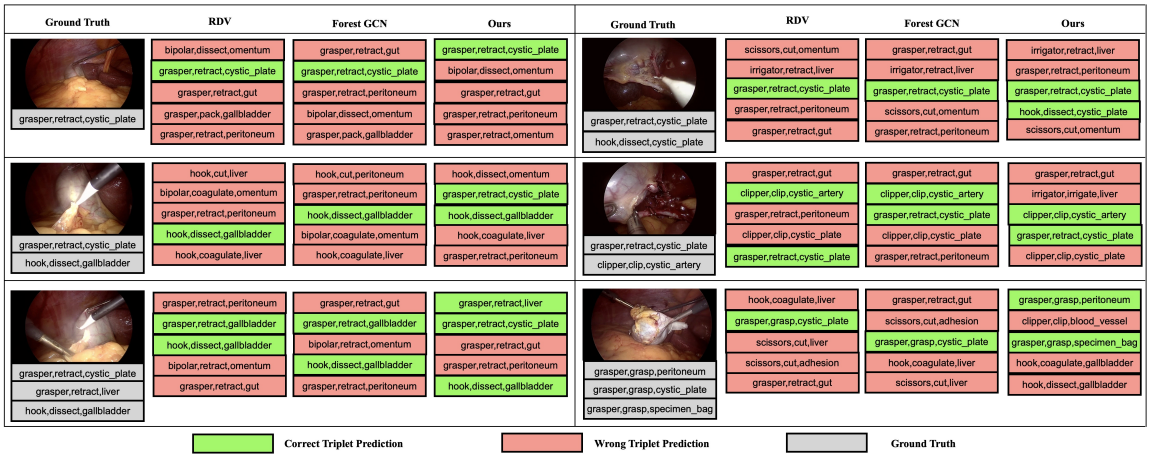

Figure 5: Comparison of qualitative results of Top-5 predictions from RDV [18], Forest GCN [30] and our model. Green boxes represent correct triplet predictions, while red boxes and gray boxes represent wrong triplet predictions and ground truth, respectively. Best view in screen.

chain in VSR network or STR network will result in the drop of triplet recognition mAP performance, validating the effectiveness of the two reasoning networks. We also notice that ablating the CTP module and VTP module in VSR network results in more severe mAP drop than ablating AP module and D-GNN module in STR network. Meanwhile, ablating BioMedLM decreases triplet mAP performance, suggesting that calibrating caption features with domain knowledge is necessary for surgical triplet recognition. Ablating the verb loss term in Eq. 17 results in 0.9 AP_{IVT} drop compared to optimizing two loss terms. This result suggests that the verb prediction guides the triplet prediction to be more accurate, possibly because the verb prediction is easier than triplet prediction since the class number of verbs are much smller than triplet class number. Similarly, Tab. 2 show the ablation results on Top-N accuracy, presenting the same tendency shown in Tab. 1.

5 CONCLUSION

In this paper, we propose chain-of-look prompting scheme to explicitly decompose the surgical triplet recognition into a series of visual reasoning processes in endoscopic videos and cast the surgical triplet recognition task as a visual prompt generation task. We further utilize BioMed language model to endow the generated visual prompts to be applicable in surgical scene. We also regard verbs as the central semantic carriers in surgical triplets and present a verb-centric scheme to model surgical triplets. Extensive experiments validate the effectiveness of our method for surgical triplet recognition.

Acknowledgement This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

REFERENCES

[1] Blake C Alkire, Nakul P Raykar, Mark G Shrime, Thomas G Weiser, Stephen W Bickler, John A Rose, Cameron T Nutt, Sarah LM Greenberg, Meera Kotagal, Johanna N Riesel, et al. 2015. Global access to surgical care: a modelling study. *The Lancet Global Health* 3, 6 (2015), e316–e323.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[4] Stanford CRFM. 2022. BioMed Language Model. *https://huggingface.co/stanford-crfm/BioMedLM* (2022).

[5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* (2021).

[6] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Tony Huang, Jack Chu, and Fangyun Wei. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649* (2022).

[9] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. 2018. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 691–699.

[10] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[12] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 121–137.

[13] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[14] Benny PL Lo, Ara Darzi, and Guang-Zhong Yang. 2003. Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In *International conference on medical image computing and computer-assisted intervention*. Springer, 230–237.

[15] George Mather, Andrea Pavan, Rosilari Bellacosa Marotti, Gianluca Campana, and Clara Casco. 2013. Interactions between motion and form processing in the human visual system. *Frontiers in Computational Neuroscience* 7 (2013), 65.

[16] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII*. Springer, 336–352.

[17] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599* (2019).

[18] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2020. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 364–374.

[19] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2022. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* 78 (2022), 102433.

[20] Philip H Pucher, L Michael Brunt, Neil Davies, Ali Linsk, Amani Munshi, H Alejandro Rodriguez, Abe Fingerhut, Robert D Fanelli, Horacio Asbun, Rajesh Aggarwal, et al. 2018. Outcome trends and safety measures after 30 years of laparoscopic cholecystectomy: a systematic review and pooled data analysis. *Surgical endoscopy* 32 (2018), 2175–2183.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[22] Rogério Richa, Marcin Balicki, Eric Meisner, Raphael Sznitman, Russell Taylor, and Gregory Hager. 2011. Visual tracking of surgical tools for proximity detection in retinal surgery. In *International Conference on Information Processing in Computer-Assisted Interventions*. Springer, 55–66.

[23] Nicola Rieke, David Joseph Tan, Chiara Amat di San Filippo, Federico Tombari, Mohamed Alsheakhali, Vasileios Belagiannis, Abouzar Eslami, and Nassir Navab. 2016. Real-time localization of articulated surgical instruments in retinal microsurgery. *Medical image analysis* 34 (2016), 82–100.

[24] Raphael Sznitman, Anasuya Basu, Rogerio Richa, Jim Handa, Peter Gehlbach, Russell H Taylor, Bruno Jedynak, and Gregory D Hager. 2011. Unified detection and tracking in retinal microsurgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 1–8.

[25] Raphael Sznitman, Carlos Becker, and Pascal Fua. 2014. Fast part-based classification for instrument detection in minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 692–699.

[26] Armine Vardazaryan, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2018. Weakly-supervised learning for tool localization in laparoscopic videos. In *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis*. Springer, 169–179.

[27] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[29] Thomas G Weiser, Scott E Regenbogen, Katherine D Thompson, Alex B Haynes, Stuart R Lipsitz, William R Berry, and Atul A Gawande. 2008. An estimation of the global volume of surgery: a modelling strategy based on available data. *The Lancet* 372, 9633 (2008), 139–144.

[30] Nan Xi, Jingjing Meng, and Junsong Yuan. 2022. Forest Graph Convolutional Network for Surgical Action Triplet Recognition in Endoscopic Videos. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 12 (2022), 8550–8561.

[31] Jiaxuan You, Tianyu Du, and Jure Leskovec. 2022. ROLAND: graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2358–2366.

[32] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. 2017. Relationship proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5678–5686.

[33] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

Appendix A MORE QUANTITATIVE RESULTS

We show the per-class detection results of instrument, verb and target in Tab. 4, Tab. 5 and Tab. 6, respectively. The instrument names and verb names are listed on the top row of Tab. 4 and Tab. 5. Target classes are denoted as numbers from 1 to 14 in the top row of Tab. 6, corresponding to *gallbladder, cystic-plate, cystic-duct, cystic-artery, cystic-pedicle, blood-vessel, fluid, abdominal-wall-cavity, liver, omentum, peritoneum, gut, specimen-bag, null* respectively. Results indicate that our method outperforms state-of-the-art methods in most of those individual per-class recognition (highlighted in the last row of each table).

Table 4: Results of Per-Class Instrument Detection (AP_I).

Method	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	mAP
CNN [18]	91.4	47.9	89.1	24.0	50.2	43.2	57.7
TCN [19]	90.5	37.6	86.2	15.9	33.3	29.6	48.9
MTL [19]	95.5	85.8	96.6	74.8	85.8	68.2	84.5
Tripnet [18]	97.8	91.2	98.1	90.7	92.1	82.7	92.1
RDV [19]	**97.7**	89.4	98.1	92.0	92.2	82.7	92.0
Ours	97.3	**92.4**	**98.6**	**93.1**	**93.5**	**89.2**	**94.1**

Table 5: Results of Per-Class Verb Detection (AP_V).

Method	Grasp	Retract	Dissect	Coagulate	Clip	Cut	Aspirate	Irrigate	Pack	Null	mAP
CNN [18]	48.6	82.1	80.5	30.5	49.5	23.8	32.4	16.0	9.2	15.9	39.2
TCN [19]	24.9	80.2	66.4	27.4	31.9	14.7	14.8	13.9	2.0	15.4	29.4
MTL [19]	47.9	85.0	84.8	55.0	79.1	44.1	35.4	13.4	18.0	17.0	48.4
Tripnet [18]	45.8	88.1	86.7	66.3	85.1	68.3	44.9	12.2	22.5	20.1	54.5
RDV [19]	60.4	90.5	**89.5**	68.7	86.7	87.8	50.4	**17.4**	**30.5**	21.0	60.7
Ours	**60.9**	**92.4**	88.6	**73.5**	**89.2**	**88.6**	**60.2**	16.8	30.2	**24.6**	**62.5**

Table 6: Results of Per-Class Target Detection (AP_T).

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	mAP
CNN [18]	84.2	14.8	26.3	18.7	14.3	3.6	32.4	10.1	49.8	35.2	8.4	8.4	69.3	15.9	28.3
TCN [19]	79.9	10.0	21.4	19.6	7.0	1.3	14.8	6.9	43.1	27.9	1.9	9.0	37.4	15.4	21.4
MTL [19]	85.1	12.2	29.3	18.6	6.5	6.4	30.6	9.8	55.7	35.8	2.1	8.4	71.1	17.5	28.2
Tripnet [18]	87.0	22.5	29.7	21.9	4.7	15.0	42.9	32.3	57.5	36.7	2.0	11.9	74.1	20.9	33.2
RDV [19]	89.1	15.3	35.2	34.5	**22.7**	11.4	53.7	40.6	59.3	46.6	4.3	12.5	84.0	**25.0**	38.3
Ours	**91.4**	**15.6**	**39.6**	**34.9**	21.8	**17.7**	**58.6**	**41.7**	**73.9**	**49.8**	**14.6**	**15.2**	**88.4**	23.4	**41.9**