Procedure-Aware Hierarchical Alignment for Open Surgery Video-Language Pretraining

Boqiang Xu, Jinlin Wu, Jian Liang, *Member, IEEE*, Zhenan Sun, *Senior Member, IEEE*, Hongbin Liu, Jiebo Luo, *Fellow, IEEE*, and Zhen Lei, *Fellow, IEEE*

***Abstract*—Recent advances in surgical robotics and computer vision have greatly improved intelligent systems' autonomy and perception in the operating room (OR), especially in endoscopic and minimally invasive surgeries. However, for open surgery, which is still the predominant form of surgical intervention worldwide, there has been relatively limited exploration due to its inherent complexity and the lack of large-scale, diverse datasets. To close this gap, we present OpenSurgery, by far the largest video–text pretraining and evaluation dataset for open surgery understanding. OpenSurgery consists of two subsets: OpenSurgery-Pretrain and OpenSurgery-EVAL. OpenSurgery-Pretrain consists of 843 publicly available open surgery videos for pretraining, spanning 102 hours and encompassing over 20 distinct surgical types. OpenSurgery-EVAL is a benchmark dataset for evaluating model performance in open surgery understanding, comprising 280 training and 120 test videos, totaling 49 hours. Each video in OpenSurgery is meticulously annotated by expert surgeons at three hierarchical levels of video, operation, and frame to ensure both high quality and strong clinical applicability. Next, we propose the Hierarchical Surgical Knowledge Pretraining (HierSKP) framework to facilitate large-scale multimodal representation learning for open surgery understanding. HierSKP leverages a granularity-aware contrastive learning strategy and enhances procedural comprehension by constructing hard negative samples and incorporating a Dynamic Time Warping (DTW)-based loss to capture fine-grained temporal alignment of visual semantics. Extensive experiments show that HierSKP achieves state-of-the-art performance on OpenSurgegy-EVAL across multiple tasks, including operation recognition, temporal action localization, and zero-shot cross-modal retrieval. This demonstrates its strong generalizability for further advances in open surgery understanding.**

***Index Terms*—Multimodal pretraining, dataset, cross-modal alignment, open surgery, surgical scene understanding.**

I. Introduction

AS SURGICAL robotic platforms like the Da Vinci® system continue to advance in sophistication, there is a growing interest in incorporating enhanced intelligence into operating room (OR) environments [19], [25], [44]. Advances in computer vision have significantly enhanced the ability of robotic systems to autonomously perceive and adapt to the complexities of surgical environments. In recent years, numerous studies have focused on endoscopic and microscopic surgeries, demonstrating the potential of deep learning in enhancing autonomous capabilities. These advances span various tasks, including surgical workflow analysis [20], [44], [49], instrument and anatomical structure segmentation [15], [31], [38], and depth estimation [53], among others.

However, the majority of surgical procedures worldwide are still performed as open surgeries without the aid of minimally invasive camera systems [7], [37], and the successful development and application of computer vision techniques in open surgery have remained limited. Open surgical videos capture intricate interactions among multiple operators and surgical instruments, accompanied by significant variation in operative environments, which present more complex and dynamic visual scenes compared to the more constrained and structured imagery of minimally invasive procedures. The development of computer vision techniques for open surgery has been further hindered by the absence of large, diverse datasets [8]. In contrast, dataset curation for endoscopic surgery is relatively simpler, as the routine clinical use of intracorporeal fiber-optic cameras enables rapid and high-quality video acquisition. However, video recording is not a standard practice in open surgery, making data collection and curation significantly more challenging. Previous studies [13], [55] have been limited to small datasets generated in simulated open surgical settings, which lack the complexity of real-world cases, or have relied on specialized equipment such as instrumented gloves that may not be applicable in actual clinical environments. A large, diverse, and representative dataset is

Received 24 August 2025; revised 11 December 2025; accepted 17 January 2026. Date of publication 6 February 2026; date of current version 18 February 2026. This work was supported in part by the InnoHK Program and in part by the National Natural Science Foundation of China under Grant U2441251 and Grant U1836217. The associate editor coordinating the review of this article and approving it for publication was Dr. Yu Liu. *(Corresponding author: Zhen Lei.)*

Boqiang Xu, Jinlin Wu, and Hongbin Liu are with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China (e-mail: boqiang.xu@cripac.ia.ac.cn; jinlin.wu@nlpr.ia.ac.cn; liuhongbin@ia.ac.cn).

Jian Liang and Zhenan Sun are with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: liangjian92@gmail.com; znsun@nlpr.ia.ac.cn).

Jiebo Luo is with Hong Kong Institute of Science and Innovation, Hong Kong, SAR, China (e-mail: jiebo@ieee.org).

Zhen Lei is with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China (e-mail: zhen.lei@ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2026.3659752

essential for enabling the development of AI models that can effectively adapt to the complex and variable surgical scenes encountered in open procedures.

To address this challenge, we introduce OpenSurgery, a large-scale and expert-annotated video–text pretraining and evaluation dataset designed to advance the understanding of open surgery. The main advantages of OpenSurgery include:

- **Largest scale and diversity:** To the best of our knowledge, OpenSurgery is currently the largest and most comprehensively annotated video-text pretraining and evaluation dataset available for open surgery understanding. OpenSurgery consists of two subsets: OpenSurgery-Pretrain and OpenSurgery-EVAL. OpenSurgery-Pretrain contains 843 publicly available open surgery videos collected from YouTube, with a total duration of 102 hours. Additionally, OpenSurgery-Pretrain encompasses the greatest diversity of surgical types, including over 20 different surgeries such as anal sphincterotomy, radical neck dissection, branchial fistula, and so on. The videos span a 15-year period and originate from more than 50 countries, capturing a wide spectrum of clinical practices and reflecting the global diversity of open surgical procedures. OpenSurgery-EVAL is currently the largest benchmark for open surgery understanding. OpenSurgery-EVAL supports multiple tasks, including operation recognition, temporal action localization, and cross-modal retrieval. Its training set consists of 280 videos with a total duration of 35 hours, while the test set includes 120 videos totaling 14 hours. The large scale and broad diversity of OpenSurgery ensure its superiority for advancing the understanding of open surgery.
- **Hierarchically structured, fine-grained annotation:** We provide meticulously designed hierarchical annotations at the video-level, operation-level, and frame-level, effectively supporting the training of models tailored to specific challenges. This annotation framework is designed to facilitate a multifaceted understanding of open surgery while enhancing the usability and interpretability of the dataset.
- **High-quality annotations by expert surgeons:** The annotation process of OpenSurgery was undertaken by expert surgeons and comprised a comprehensive set of procedures, including but not limited to the standardization of operation label definitions, video filtering, localization annotations, and secondary validation. The quality and professionalism of OpenSurgery are ensured through expert-level annotation conducted by highly experienced surgical professionals.

We further propose the **Hie**rarchical **S**urgical **K**nowledge **P**retraining (HierSKP) framework to enable large-scale multimodal representation learning for open surgery understanding. The advantages of HierSKP reside in two key aspects. First, we propose a granularity-wise learning strategy to effectively exploit the hierarchical surgical knowledge embedded within the dataset. We perform contrastive learning at three levels of granularity (video, operation, and frame levels) for each type of hierarchical video–text pair. Second, we enhance the model's semantic understanding of surgical procedures by enforcing temporal alignment using a Dynamic Time Warping [28], [29] (DTW)-based loss function. Specifically, we construct pairs of video sequences and introduce hard negative samples by reversing the temporal order of one sequence within each pair. Finally, a dynamic time warping (DTW)-based loss function is applied to learn fine-grained temporal correspondences between videos, facilitating semantic alignment of visual features and improving procedural understanding during the pretraining phase.

The contributions of this paper are summarized as follows:

- We present OpenSurgery, currently the largest pretraining and evaluation dataset for open surgery understanding. The OpenSurgery-Pretrain subset contains 843 videos with a total duration of 102 hours. The OpenSurgery-OR subset comprises 280 training videos spanning 35 hours and 120 test videos totaling 14 hours.
- We propose HierSKP, a pretraining framework designed to facilitate the learning and understanding of surgical knowledge through hierarchical contrastive learning and temporal alignment using a DTW-based loss function.
- We conduct extensive experiments to show that HierSKP, after being pretrained on OpenSurgery-Pretrain, achieves state-of-the-art transferability and visual representation capabilities across various downstream surgical scene understanding tasks in OpenSurgery-EVAL, including operation recognition, temporal action localization, and cross-modal retrieval.

II. Related Work

A. Surgical Video-Language Pretraining

Numerous studies have demonstrated the effectiveness of learning visual representations through natural language supervision provided by associated text descriptions [1], [46], [52]. These approaches typically employ contrastive learning techniques [34] to align video clips or images with their corresponding narrations or captions. Similarly, in the surgical domain, recent works have begun curating large-scale multimodal datasets in areas such as endoscopic [50], [51] and ophthalmic surgeries [18] to facilitate vision-language pertaining. In [51], speech from surgical video lectures is transcribed to construct a video-text pretraining dataset for endoscopic surgery. PeskaVLP [50] integrates language supervision with visual self-supervision and employs temporal alignment to effectively facilitate cross-modal understanding of surgical procedures. OphNet [18] is a large-scale video benchmark with expert annotations for advancing ophthalmic surgical workflow understanding. However, for open surgeries, which remain the most widely performed type of surgery worldwide, there is still no dedicated dataset for pretraining and downstream task evaluation.

B. Surgical Workflow Recognition

Methods originating from the broader domain of video content analysis are increasingly being adapted and integrated into the field of surgery [2], [26], [49]. Comprehensive surgical workflow recognition includes both phase recognition and operation recognition. Although both focus on detecting surgical events, they operate at different temporal resolutions, capturing events at varying levels of granularity. SV-RCNet [20] introduced an end-to-end framework by combining Residual Networks (ResNets) [17] for spatial feature extraction with LSTMs to model temporal dynamics. Building on this, Gao et al. [11] enhanced sequence modeling by integrating a tree search algorithm into the LSTM structure, allowing the network to leverage future contextual information. TMRNet [21] proposed a non-local bank operator to establish connections between the current frame and LSTM-generated features. With the successful application of Transformers [42] in the field of computer vision, surgical workflow recognition has also witnessed related innovations. OperA [5] focuses on anticipating future phases. Trans-SVNet [12] aimed to overcome the limitations of TCNs highlighted in TeCNO [4] by incorporating Transformers to enable multiscale temporal feature fusion. Overall, gaining a comprehensive understanding of the surgical workflow demands a thorough analysis of its temporal, spatial, and contextual aspects.

III. Dataset Construction

In this section, we detail the construction of our dataset, which involved rigorous data collection and annotation processes. The dataset is derived from the Annotated Videos of Open Surgery (AVOS) [14], a collection of videos sourced from YouTube to mitigate privacy concerns while capturing a diverse range of open surgical procedures. To accurately reflect the complexity of open surgeries, we conducted comprehensive, hierarchical annotations at the video, operation, and frame levels. These annotations were performed by a dedicated team of surgeons to ensure both accuracy and clinical relevance.

A. Dataset Collection and Preprocessing

1) Dataset Collection: Our data is sourced from the Annotated Videos of Open Surgery (AVOS) [14], which includes 1,997 videos covering 23 types of open surgical procedures, uploaded to YouTube from 50 countries over a span of 15 years. The AVOS dataset provides YouTube links for all videos and includes annotations of commonly used surgical instruments for 343 of them. However, the majority of the videos remain unannotated and lack hierarchical and rich textual annotations.

2) Dataset Preprocessing: Some of the video links provided by the AVOS dataset are no longer accessible, and several videos lack descriptions of the surgical content, making it difficult to determine the type of surgery and posing challenges for subsequent annotation. Therefore, during the preprocessing stage, we filtered out videos with invalid links or missing surgical descriptions, ultimately collecting 1,243 videos to form our dataset. We then randomly selected 843 of these videos to construct our open surgery pretraining dataset, OpenSurgery-Pretrain, and used the remaining 400 videos to build our open surgery understanding benchmark, OpenSurgery-EVAL. Although the videos used in our work originate from publicly available YouTube sources, they were manually screened during the construction of the AVOS dataset, and we further performed an additional round of manual verification during our annotation process. As a result, all videos included in our datasets are high-quality clips that have undergone multiple stages of human screening.

B. Hierarchical Annotation

As shown in Fig. 1, we performed hierarchical annotations on the dataset at three levels: video-level, operation-level, and frame-level. The video-level annotation provides an abstract of the entire video, introducing the type of surgical procedure. The operation-level annotation describes individual video segments; we defined 13 types of surgical operations and assigned a corresponding operation label to each segment. The frame-level annotation focuses on individual frames, identifying the surgical instruments and the actions present. We introduce the annotation methods for each level separately:

1) Video-Level: We leveraged Qwen2-VL-72B [43] to generate video-level annotations from the original YouTube video introductions. We employ the following prompt: “This is a surgical video, and its description is *[Introduction]*. Please refine and summarize the surgery-related information contained in the description.” In this context, *[Introduction]* denotes the video description provided on YouTube. The model was prompted to describe the type of surgical procedure and filter out any information unrelated to the surgery.

2) Operation-Level: In collaboration with surgeons, we defined 13 types of surgical operations. Based on this taxonomy, the surgeons performed operation localization on each video, producing segmented clips corresponding to specific surgical operations, each annotated with the appropriate operation label.

3) Frame-Level: We employed Qwen2-VL-72B [43] for frame-level annotation. Specifically, we employ the following prompt: “This image corresponds to a single frame extracted from a surgical video, specifically from the *[operation]* segment, and the video introduction is *[Introduction]*. Based on the visual content of this frame, the surgical operation, and the provided video introduction, please provide a concise description of the image. The description should include the surgical instruments visible and the ongoing surgical actions.” In this context, *[operation]* denotes the operation-level annotation, and *[Introduction]* denotes the video-level annotation. The model was prompted to generate detailed descriptions of the image content, including the surgical instruments and actions present. These generated descriptions serve as our frame-level annotations.

C. Dataset Statistics

OpenSurgery consists of two subsets: OpenSurgery-Pretrain and OpenSurgery-EVAL. OpenSurgery-Pretrain consists of 843 videos with a total duration of 102 hours, covering more than 20 different types of surgical procedures and 13 types of surgical operations. The videos span a 15-year period and originate from over 50 countries. OpenSurgery-EVAL consists of 400 videos covering 13 types of surgical operations. The training set includes 280 videos with a total duration of

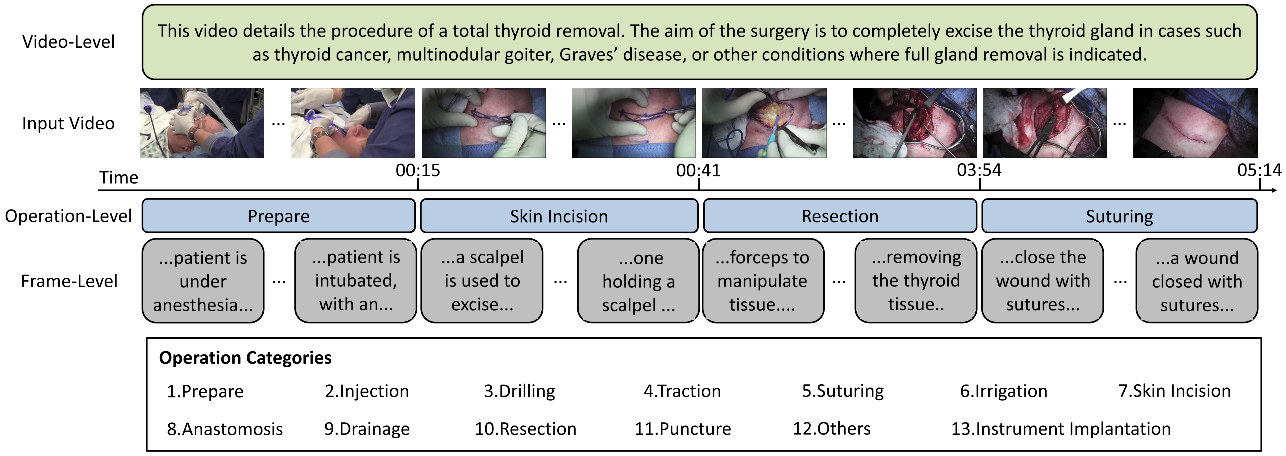
Video-Level
This video details the procedure of a total thyroid removal. The aim of the surgery is to completely excise the thyroid gland in cases such as thyroid cancer, multinodular goiter, Graves' disease, or other conditions where full gland removal is indicated.
Input Video
Time
00:15
00:41
03:54
05:14
Operation-Level
Prepare
Skin Incision
Resection
Suturing
Frame-Level
...patient is under anesthesia...
...patient is intubated, with an...
...a scalpel is used to excise...
...one holding a scalpel ...
...forceps to manipulate tissue....
...removing the thyroid tissue..
...close the wound with sutures...
...a wound closed with sutures...
Operation Categories
1.Prepare
2.Injection
3.Drilling
4.Traction
5.Suturing
6.Irrigation
7.Skin Incision
8.Anastomosis
9.Drainage
10.Resection
11.Puncture
12.Others
13.Instrument Implantation

Fig. 1. Examples of our OpenSurgery dataset. The dataset consists of more than 1,200 videos spanning a wide variety of open surgeries and includes rich hierarchical annotations at the video, operation, and frame levels. It covers 13 distinct surgical operations, each annotated with precise temporal boundaries. The datasets support a range of tasks, including operation recognition, temporal action localization, and cross-modal retrieval. The comprehensive, large-scale, and hierarchical annotations make the dataset particularly valuable for both pretraining open surgery models and evaluating fine-grained understanding of surgical scenes in complex open surgical environments.

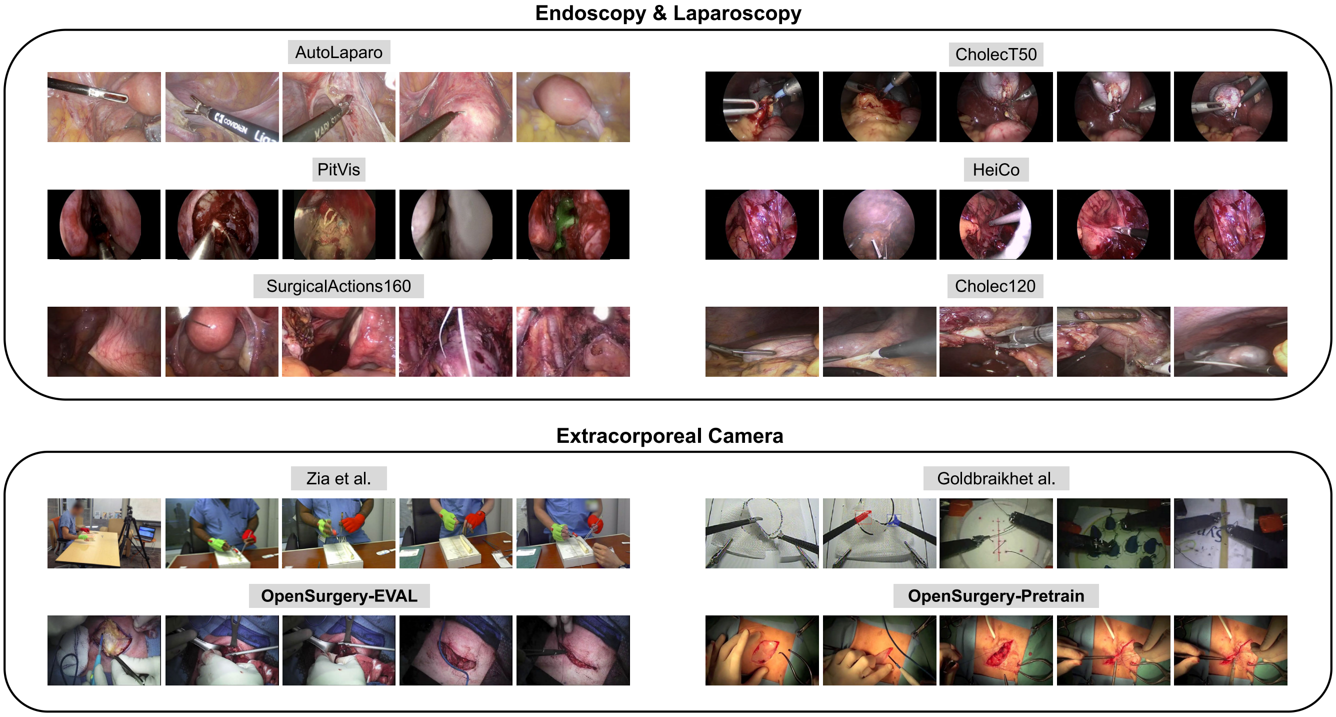

Fig. 2. The comparison among existing surgical datasets and our proposed OpenSurgery-EVAL and OpenSurgery-Pretrain datasets. While most existing datasets focus on endoscopy and laparoscopy, datasets for open surgery remain scarce and are predominantly simulated, thus diverging from real-world scenarios. In contrast, OpenSurgery-EVAL and OpenSurgery-Pretrain are collected from real surgical environments and constitute the largest-scale datasets to date for open surgery understanding.

35 hours, while the test set comprises 120 videos totaling 14 hours. A comparison between existing surgical datasets and our proposed OpenSurgery-EVAL and OpenSurgery-Pretrain datasets is illustrated in Fig. 2, with the corresponding statistical summaries provided in Tab. I. As shown, the majority of existing surgical datasets focus on endoscopic and laparoscopic procedures, whereas datasets dedicated to open surgery are scarce and predominantly simulation-based, thus limiting their applicability to real-world scenarios. In contrast, OpenSurgery-EVAL and OpenSurgery-Pretrain are collected from real surgical environments and represent the largest-scale datasets to date for advancing the understanding of open surgery.

IV. Method

We introduce the **Hier**archical **S**urgical **K**nowledge **P**retraining (HierSKP) framework, which effectively learns multi-modal embeddings by hierarchically modeling

TABLE I

THE STATISTICS COMPARISON AMONG EXISTING SURGICAL DATASETS AND OUR PROPOSED OPENSURGERY-EVAL AND OPENSURGERY-PRETRAIN DATASETS. COMPARED TO OTHER DATASETS, OPENSURGERY-EVAL AND OPENSURGERY-PRETRAIN PROVIDE MORE COMPREHENSIVE COVERAGE OF DIVERSE SURGERIES AND OPERATION CATEGORIES, OFFERING A SIGNIFICANTLY LARGER VIDEO COLLECTION, TOTALLING 49 HOURS AND 102 HOURS, RESPECTIVELY

Protocol	Datasets Properties						
	Datasets	No. of Videos	No. of Operation Segmentation	No. of Surgery Categories	No. of Operation Categories	Total Duration	Real
Endoscopy & Laparoscopy	Cholec120 [32]	120	-	1	7	76.2h	✓
	SurgicalActions160 [39]	160	160	1	16	0.2h	✓
	HeiCo [27]	30	-	3	14	2.8h	✓
	PitVis [6]	25	287	1	17	33.3h	✓
	CholecT50 [33]	50	-	1	10	44.7h	✓
	AutoLaparo [45]	21	300	1	7	23.1h	✓
Extracorporeal Camera	Zia et al. [55]	71	-	1	2	1.7h	✗
	Goldbraikh et al. [13]	100	-	1	1	10.0h	✗
	OpenSurgery-EVAL (Ours)	**400**	**4469**	**20+**	**13**	**49.0h**	✓
	OpenSurgery-Pretrain (Ours)	**843**	**8166**	**20+**	**13**	**102.0h**	✓

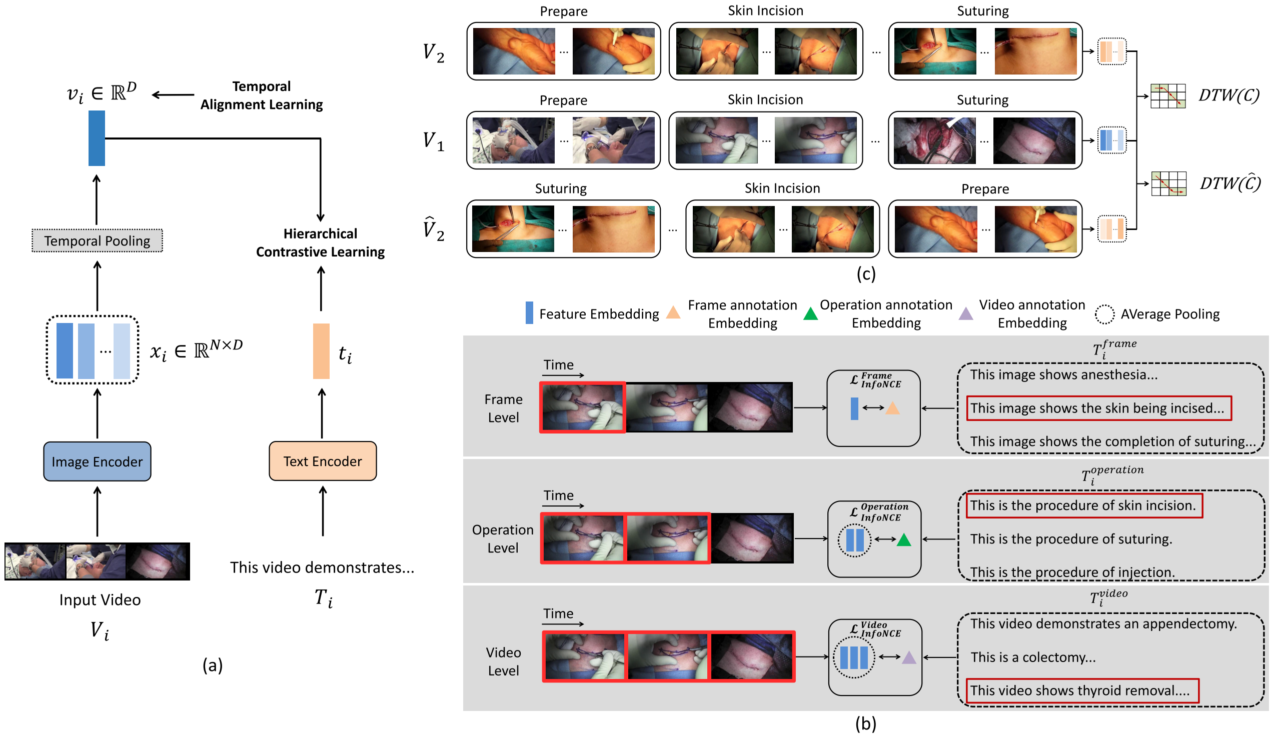
$v_i \in \mathbb{R}^D$
Temporal Alignment Learning
Temporal Pooling
Hierarchical Contrastive Learning
$x_i \in \mathbb{R}^{N \times D}$
t_i
Image Encoder
Text Encoder
This video demonstrates...
Input Video
V_i
T_i
(a)
Prepare
Skin Incision
Suturing
V_2
V_1
$\hat{V}_2$
DTW(C)
DTW($\hat{C}$)
(c)
Feature Embedding
Frame annotation Embedding
Operation annotation Embedding
Video annotation Embedding
AVerage Pooling
Frame Level
Operation Level
Video Level
Time
$\mathcal{L}_{InfoNCE}^{Frame}$
$\mathcal{L}_{InfoNCE}^{Operation}$
$\mathcal{L}_{InfoNCE}^{Video}$
T_i^{frame}
This image shows anesthesia...
This image shows the skin being incised...
This image shows the completion of suturing...
$T_i^{operation}$
This is the procedure of skin incision.
This is the procedure of suturing.
This is the procedure of injection.
T_i^{video}
This video demonstrates an appendectomy.
This is a colectomy...
This video shows thyroid removal....
(b)

Fig. 3. (a) The architecture of our proposed HierSKP framework. (b) Hierarchical contrastive learning at the video, operation, and frame levels. (c) Temporal alignment learning with a Dynamic Time Warping (DTW)-based loss function.

video–text pairs at the video, operation, and frame levels from surgical video data. Sec. IV-A describes the architecture design of HierSKP. Sec. IV-B and Sec. IV-C formalize the hierarchical contrastive learning and temporal alignment learning, respectively. Finally, Sec. IV-D presents the training objectives of HierSKP.

A. Architecture Design

Fig. 3(a) provides an overview of the proposed HierSKP framework. To bridge the gap between images, videos, and texts, we adopt the Video Fine-tuned CLIP (ViFi-CLIP) [36] baseline as the backbone architecture of HierSKP.

Given a video sample $V_i \in \mathbb{R}^{N \times H \times W \times C}$, consisting of N frames, along with its corresponding text label T_i, each frame is independently encoded by the CLIP image encoder as part of a batch. This yields a sequence of frame-level embeddings $x_i \in \mathbb{R}^{N \times D}$. To derive a holistic video-level representation, these frame embeddings are average-pooled along the temporal dimension, resulting in $v_i \in \mathbb{R}^D$. This process, referred to as *temporal pooling*, enables implicit modeling of temporal dynamics by aggregating information across multiple frames.

The CLIP text encoder processes the text label T_i, which is embedded within a prompt template (e.g., *"a photo of a < category>"*), to generate a corresponding text embedding $t_i \in \mathbb{R}^D$. Finally, the hierarchical contrastive learning and temporal alignment learning are used to train HierSKP.

B. Hierarchical Contrastive Learning

As shown in Fig. 1, the Opensurgery is designed as a hierarchically annotated video-text dataset, denoted as $\mathcal{D} = \{(V_i, T_i)\}_{i=1}^{|\mathcal{D}|}$, where V_i represents a long-form surgical video composed of a sequence of short-term segments and $T_i = (T_i^{frame}, T_i^{operation}, T_i^{video})$. For each video V_i, three levels of textual annotations are provided, capturing semantic information at different levels of granularity—from fine to coarse. These annotations include: frame-level descriptive texts (T_i^{frame}), operation-level phrase annotations ($T_i^{operation}$), and video-level abstract summaries (T_i^{video}).

The frame-level descriptive texts (T_i^{frame}) consist of sequences of narrations that describe individual frames, including visual elements such as surgical instruments and actions. The operation-level phrase annotations ($T_i^{operation}$) provide concise, conceptual summaries of high-level surgical operations, corresponding to temporal segments of the video. The video-level abstract summaries (T_i^{video}) are paragraph-length textual descriptions that encapsulate the overall content of the surgical video, including the type of surgery and its clinical background. This hierarchical structure enables HierSKP to effectively model the rich, multi-scale semantic alignment between surgical videos and their textual descriptions.

As shown in Fig. 3(b), we employ the InfoNCE [34] for contrastive video-language pretraining. Specifically, for each of the aforementioned hierarchical levels, the visual features v_i and their corresponding annotation text embeddings t_i are treated as positive pairs P_n, while the unpaired ones are treated as negative pairs N_n. Then, the contrastive training loss InfoNCE [34] can be formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\sum\limits_{(v_i,t_i)\in\mathcal{P}_i^n}\exp\left(v_i^\top t_i/\tau\right)}{\sum\limits_{(v_i,t_i)\in\mathcal{P}_i^n\cup\mathcal{N}_i^n}\exp\left(v_i^\top t_i/\tau\right)}\right), \quad (1)$$

where B denotes the batch size and τ is a temperature hyperparameter. The loss function facilitates cross-modal alignment by maximizing the cosine similarity between paired video and text representations, while minimizing the similarity between unpaired ones. We apply contrastive training to the frame level, operation level, and video level, respectively:

$$\mathcal{L}_{InfoNCE} = \mathcal{L}_{InfoNCE}^{Frame} + \mathcal{L}_{InfoNCE}^{Operation} + \mathcal{L}_{InfoNCE}^{Video}. \quad (2)$$

Through hierarchical contrastive learning, the model gradually captures fine- to coarse-grained semantic correspondences across modalities.

C. Temporal Alignment Learning

Contrastive loss leverages image–text pairs for cross-modal learning. In addition, we propose $\mathcal{L}_{DTW}$, which employs temporal alignment to semantically pretrain the model using visual features.

Algorithm 1 DTW to Align Video Sequences

```
Input: Cost matrix $C$, sequence $V_1$ of length $N_1$,
       sequence $V_2$ of length $N_2$
Output: Total alignment distance $DTW(C)$
procedure AlignSequences($C$, $V_1$, $V_2$):
    Set $i$ to $N_1$ and $j$ to $N_2$;
    Initialize $DTW(C)$ to 0;
    while $i > 0$ and $j > 0$ do
        $DTW(C) = DTW(C) + C[i][j]$;
        if $i > 1$ and $j > 1$ and
         $C[i-1][j-1] \le C[i-1][j]$ and
         $C[i-1][j-1] \le C[i][j-1]$ then
            $i \leftarrow i - 1$;
            $j \leftarrow j - 1$;
        else
            if $i > 1$ and $C[i-1][j] \le C[i][j-1]$ then
                $i \leftarrow i - 1$;
            else
                $j \leftarrow j - 1$;
    return $DTW(C)$;
```

As shown in Fig. 3(c), in each batch, we randomly sample K surgical operations. For each operation, we randomly sample a different number of video segments in two separate passes, constructing two video sequences: $V_1 = [v_1^1, v_2^1, \ldots, v_B^1]$ and $V_2 = [v_1^2, v_2^2, \ldots, v_B^2]$, where B denotes the batch size. We use $\mathcal{L}_{\text{DTW}}$ to learn the latent correspondences between the video sequences V_1 and V_2. Specifically, we construct a cost matrix $C \in \mathbb{R}^{B\times K}$ between the video sequences based on their embeddings, where each element $c_{i,j}$ is computed using a distance function D. We adopt the same distance function as defined in [16]:

$$c_{i,j} = \mathcal{D}(v_i^1, v_j^2) = -\log\left(\frac{\exp\left(\mathbf{v}_i^{1\top}\mathbf{v}_j^2/\beta\right)}{\sum_{k=1}^{B}\exp\left(\mathbf{v}_i^{1\top}\mathbf{v}_k^2/\beta\right)}\right), \quad (3)$$

where β is a temperature hyperparameter. After obtaining the cost matrix C, we apply Dynamic Time Warping (DTW) to identify the minimum cost path that aligns the video segments in the sequences V_1 and V_2, as shown in Algorithm 1. We adopt the approach proposed in [47] to make the DTW function differentiable, thereby enabling gradient backpropagation during training. A significant advantage of using DTW is that it eliminates the need for additional temporal modeling modules, such as recurrent neural networks or attention mechanisms for capturing temporal dependencies. This simplification allows the model to focus on learning more effective representations by directly aligning video segments based on their semantics.

Furthermore, we make the intuitive assumption that the global semantics of a video sequence and its reversed counterpart are inherently different. Consequently, aligning video segments V_1 with V_2 should result in a lower alignment cost compared to aligning V_1 with the reversed sequence of V_2. Based on this assumption, we construct a temporally reversed version of V_2, denoted as $\hat{V}_2 = [v_B^2, v_{B-1}^2, \ldots, v_1^2]$, and compute the corresponding cost matrix $\hat{C}$ between the video sequence

V_1 and $\hat{V}_2$. The minimum alignment cost is then obtained as $\text{DTW}(\hat{C})$. To encourage correct temporal alignment, we introduce a DTW-based contrastive regularization formulated as a hinge loss:

$$\mathcal{L}_{\text{DTW}} = \max\left(\text{DTW}(C) - \text{DTW}(\hat{C}), \phi\right). \tag{4}$$

The $\mathcal{L}_{\text{DTW}}$ learns generalizable surgical visual features by performing temporal alignment between different video segments of the same surgical operation.

D. Learning Objectives

We train our model using hierarchical contrastive learning and temporal alignment learning. The final learning objectives can be formulated as:

$$\mathcal{L} = \mathcal{L}_{InfoNCE} + \lambda \mathcal{L}_{DTW}, \tag{5}$$

where λ is the hyperparameter to scale two losses.

V. Experiments

Datasets. Our pertaining is conducted on the OpenSurgery-Pretrain dataset. The pertaining dataset includes hierarchical textual annotations of open surgery videos. We evaluate our model on the OpenSurgery-EVAL dataset for operation recognition and temporal action localization under the fully supervised training (SFT) setting. Additionally, we evaluate our pretrained model on a zero-shot cross-modal retrieval task spanning multiple hierarchical levels.

Training Parameters. We utilize the CLIP-ViT-B-16 [35] as our backbone. We train the model with a batch size of 512/32/16 for frame-/operation-/video-level, respectively. We sample 16/32 frames for the videos of operation-/video level. We use Adam optimizer [23] with a learning rate of $2.2e-5$. We set the weighting coefficient λ in Equation 5 to 1. We pretrain our model on the OpenSurgery-Pretrain dataset using 4 80 GB NVIDIA A100 GPUs for 200 epochs, which took approximately 70 hours, and finetune it on the OpenSurgery-EVAL dataset for 100 epochs, which took about 20 hours.

A. Zero-Shot Cross-Modal Retrieval

1) Task Description: We evaluate the cross-modal alignment capability of the pretrained models by performing zero-shot image-to-text and text-to-image retrieval tasks on the OpenSurgery-EVAL dataset across the frame, operation, and video levels. We provide hierarchical labels for the test set of OpenSurgery-EVAL.

2) Results: The results of zero-shot cross-model retrieval are presented in Tab. II. We compare our method with a cross-encoder approach, ALBEF [24], and two CLIP-based methods, ViFi-CLIP [36] and X-CLIP [30]. ViFi-CLIP [36] and X-CLIP [30] are evaluated in two versions: 1) using the pretrained ViT-B/16 model, and 2) pretraining with weights from the Kinetics 400 [22] dataset. The results show that HierSKP consistently outperforms other methods across the frame, operation, and video levels, demonstrating that our hierarchical pretraining scheme significantly enhances the model's generalization ability for zero-shot cross-modal retrieval in open surgery. Additionally, we find that for both X-CLIP [30] and ViFi-CLIP [36], the versions pretrained on ViT-B/16 achieve higher zero-shot cross-modal retrieval accuracy than those pretrained on Kinetics 400 [22], suggesting that ViT-B/16 pretraining provides better generalization than Kinetics 400 [22] pretraining in surgical scenarios.

B. Temporal Action Localization

1) Task Description: Temporal action localization (TAL) is the task of identifying and localizing all action instances within an untrimmed video by accurately predicting their start and end timestamps, as well as assigning the correct action category to each instance. Unlike action classification, which assumes temporally trimmed inputs, TAL must operate on continuous video streams that may contain multiple actions interspersed with irrelevant or background content, making it substantially more challenging.

2) Results: The results of temporal action localization are shown in Tab. III. TAL pipelines commonly first extract video features using diverse action classification networks, and then perform action localization based on the extracted features using different methods. To assess the generalization capability of the features extracted by our proposed approach, we employ X-CLIP [30], ViFi-CLIP [36], and HierSKP to extract video features, followed by applying various TAL methods to perform action localization.

We report the standard mean Average Precision (mAP) at various temporal Intersection over Union (tIoU) thresholds, which are commonly used to evaluate TAL methods. The tIoU metric measures the overlap between two temporal segments, defined as the 1D Jaccard index. For a given tIoU threshold, mAP is computed by averaging the average precision (AP) over all action categories. Additionally, we report the average mAP, which is the mean of mAP scores across multiple tIoU thresholds. Experimental results show that the video features extracted by our method achieve the best performance across all TAL methods. Specifically, when used with ActionFormer, DyFADet, TriDet, and TemporalMaxer, our features improve the average mAP by at least 4.9%, 6.5%, 6.8%, and 5.9%, respectively, compared to other feature extractors. The qualitative results are presented in Fig. 4. The results demonstrate that by pretraining on the OpenSurgery-Pretrain dataset, our method achieves the best generalization performance on the temporal action localization task.

C. Surgical Operation Recognition

1) Task Description: Surgical operation recognition serves as a proxy task for evaluating a model's ability to comprehend surgical scenes. The objective is to classify trimmed surgical video segments into predefined operation categories, which necessitates the model's understanding of the presence and interaction of surgical instruments and anatomical structures by capturing meaningful visual features throughout the procedure.

2) Results: The results of surgical operation recognition are shown in Tab. IV. We compare our method against C2D [17], SlowFast [10], X3D [9], and I3D [3], as well as two

TABLE II

ZERO-SHOT CROSS-MODAL RETRIEVAL RESULTS (%) ON THE OPENSURGERY-EVAL DATASET. IN EACH SETTING, THE BEST RESULT IS MARKED IN BOLD AND THE SECOND-BEST RESULT IS MARKED WITH AN UNDERLINE

Pretrain	Method	Frame-Level			Operation-Level			Video-Level		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
		Image-to-Text (%)								
ViT-B/16	X-CLIP [30]	0.1	0.2	0.4	13.3	79.1	93.0	14.2	35.8	44.2
	ViFi-CLIP [36]	0.2	0.6	1.0	34.3	80.4	92.3	25.0	42.5	50.8
Knetics 400	X-CLIP [30]	-	-	-	22.7	79.1	93.2	13.3	35.0	46.7
	ViFi-CLIP [36]	0.1	0.2	0.3	18.8	78.6	87.3	19.2	36.7	44.2
Conceptual Captions, SBU Captions, COCO, Visual Genome	ALBEF [24]	0.1	0.2	0.4	11.2	74.3	84.2	10.2	28.7	38.5
OpenSurge	HierSKP(ours)	**5.9**	**19.8**	**30.5**	**64.6**	**94.3**	**98.3**	**29.2**	**58.3**	**66.7**
		Text-to-Image (%)								
ViT-B/16	X-CLIP [30]	0.0	0.0	0.1	7.7	15.4	23.1	0.8	10.0	12.5
	ViFi-CLIP [36]	0.2	0.5	0.7	38.5	46.2	46.2	12.5	27.5	31.9
Knetics 400	X-CLIP [30]	-	-	-	23.1	30.8	30.8	10.8	25.8	34.2
	ViFi-CLIP [36]	0.1	0.2	0.3	38.3	51.5	61.2	10.8	25.8	31.7
Conceptual Captions, SBU Captions, COCO, Visual Genome	ALBEF [24]	0.1	0.2	0.3	5.2	9.8	18.5	1.2	8.9	12.6
OpenSurge	HierSKP(ours)	**4.6**	**16.4**	**26.1**	**38.7**	**53.8**	**61.5**	**31.7**	**50.8**	**58.3**

TABLE III

TEMPORAL ACTION LOCALIZATION RESULTS (%) ON THE OPENSURGERY-EVAL DATASET. WE REPORT mAP ATDIFFERENT tIoU THRESHOLDS. AVERAGE mAP IN [0.3:0.1:0.7] IS REPORTED. IN EACH SETTING, THE BEST RESULT IS MARKED IN BOLD, AND THE SECOND-BEST RESULT IS MARKED WITH AN UNDERLINE

Method	Feature	0.3	0.4	0.5	0.6	0.7	Avg.
ActionFormer [59]	X-CLIP [30]	13.4	11.1	8.6	5.2	1.9	8.0
	ViFi-CLIP [36]	14.4	12.6	9.6	6.9	4.0	9.5
	HierSKP(ours)	**21.3**	**19.2**	**16.4**	**8.9**	**6.3**	**14.4**
DyFADet [52]	X-CLIP [30]	13.2	11.1	7.7	4.6	2.2	7.8
	ViFi-CLIP [36]	11.6	9.2	6.9	4.4	2.8	7.0
	HierSKP(ours)	**22.1**	**18.8**	**14.6**	**10.2**	**5.6**	**14.3**
TriDet [43]	X-CLIP [30]	13.1	11.3	7.3	5.2	3.9	8.2
	ViFi-CLIP [36]	13.9	11.4	9.7	6.5	5.9	9.5
	HierSKP(ours)	**24.8**	**20.2**	**16.5**	**13.2**	**7.0**	**16.3**
TemporalMaxer [44]	X-CLIP [30]	18.2	14.6	9.0	5.1	1.7	9.7
	ViFi-CLIP [36]	18.3	14.8	12.6	9.9	4.0	11.9
	HierSKP(ours)	**28.1**	**22.5**	**18.0**	**11.6**	**8.7**	**17.8**

CLIP-based models, ViFi-CLIP [36] and X-CLIP [30]. These models are evaluated in three versions: 1) training from random initialization, 2) using the pretrained ViT-B/16 model, and 3) pretraining with weights from Kinetics 400 [22], which is a human action recognition dataset. We report the $Top - N$ accuracy by retrieving the nearest neighbors for each query and checking whether the corresponding ground-truth element appears among the top N results. The results show that HierSKP achieves the highest Top-1 and Top-5 accuracies, outperforming other methods by at least 4.2% and 2.5%, respectively. The results demonstrate that by pretraining on the OpenSurgery-Pretrain dataset, our method achieves the best generalization performance on the operation recognition task.

D. Ablation Studies on Different Learning Objectives

We conduct ablation studies on various learning objectives for the operation recognition task on the OpenSurgery-EVAL dataset. Table V presents the results of different configurations. The first row, in which no loss functions are applied, corresponds to the result obtained by fine-tuning the ViFi-CLIP [36] model on the OpenSurgery-EVAL dataset using pretrained weights from Kinetics-400, without any

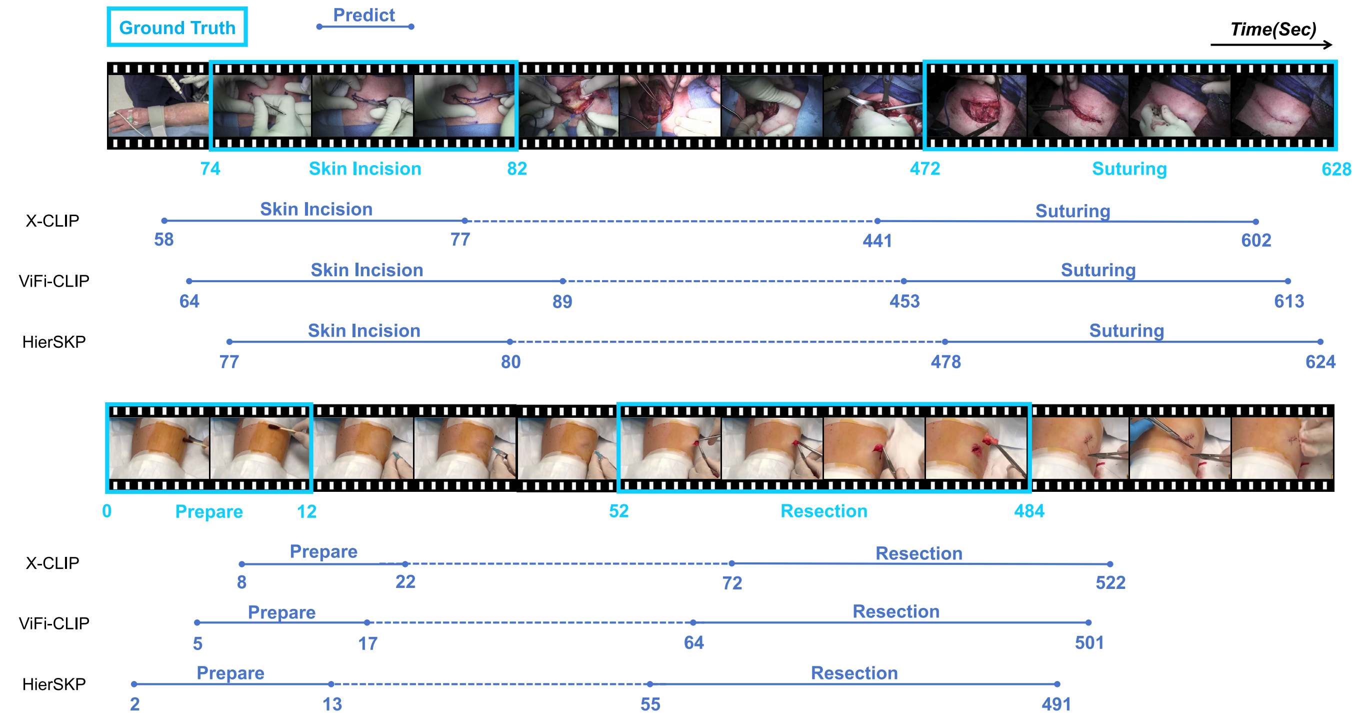

Fig. 4. Qualitative results of TemporalMaxer [41] in the temporal action localization task, obtained with video features extracted by X-CLIP [30], ViFi-CLIP [36], and HierSKP, respectively.

TABLE IV

TOP-1 AND TOP-5 ACCURACY (%) FOR SURGICAL OPERATION RECOGNITION ON THE OPENSURGERY-EVAL DATASET. THE BEST RESULT IS MARKED IN BOLD, AND THE SECOND-BEST RESULT IS MARKED WITH AN UNDERLINE

Pretrain	Method	Top-1	Top-5
Random Initialization	C2D [17]	40.6	87.5
	Slowfast [10]	40	89.4
	X3D [9]	45.5	89.3
	I3D [3]	36.4	88.1
ViT-B/16	X-CLIP [30]	68.7	92.5
	ViFi-CLIP [36]	67.5	91.5
Knetics 400	C2D [17]	53.3	89.7
	Slowfast [10]	60.5	92.1
	X3D [9]	38.6	89.2
	I3D [3]	54	89.7
	X-CLIP [30]	<u>70.2</u>	<u>93.9</u>
	ViFi-CLIP [36]	69.7	93.8
OpenSurgery-Pretrain	HierSKP(ours)	**74.4**	**96.4**

TABLE V

ABLATION STUDY OF DIFFERENT TRAINING OBJECTIVES THROUGH OPERATION RECOGNITION EXPERIMENTS ON THE OPENSURGERY-EVAL DATASET. WE REPORT TOP-1 AND TOP-5 ACCURACY IN THIS TABLE

$\mathcal{L}^{Frame}_{InfoNCE}$	$\mathcal{L}^{Operation}_{InfoNCE}$	$\mathcal{L}^{Video}_{InfoNCE}$	$\mathcal{L}_{DTW}$	Top-1	Top-5
✗	✗	✗	✗	69.7	93.8
✓	✗	✗	✗	70.5	94.5
✓	✓	✗	✗	72.8	95.7
✓	✓	✓	✗	73.3	96.1
✓	✓	✓	✓	74.4	96.4

pretraining on the OpenSurgery-Pretrain dataset. From the results, we can find that applying $\mathcal{L}_{InfoNCE}$ at different levels consistently improves the model's accuracy. Specifically, incorporating $\mathcal{L}^{Frame}_{InfoNCE}$ yields a Top-1 accuracy improvement of 0.8%. Building upon this, adding $\mathcal{L}^{Operation}_{InfoNCE}$ further increases the performance by 2.3%, and applying $\mathcal{L}^{Video}_{InfoNCE}$ on top of both yields an additional 0.5% gain, demonstrating the cumulative benefit of our hierarchical contrastive learning scheme. Furthermore, when we append $\mathcal{L}_{\text{DTW}}$ for temporal alignment learning, the Top-1 accuracy is further improved by 1.1%. The experimental results have validated the effectiveness of our proposed learning objectives for surgical video-language pretraining.

E. Ablation Studies on Hyperparameters

We conduct ablation studies on hyperparameters for the operation recognition task on the OpenSurgery-EVAL dataset. Fig 5(a) presents the results under different settings of the weighting coefficient λ. The results show that HierSKP achieves the best performance when $\lambda = 1$. Fig. 5(b) illustrates the results under different values of the number of selected surgical operations K per batch for $\mathcal{L}_{\text{DTW}}$. The results show that as K increases, the model performance initially improves,

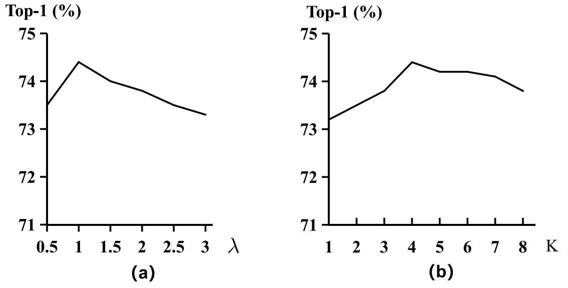

Fig. 5. (a) Ablation study on the weighting coefficient λ. (b) Ablation study on the number of selected surgical operations K.

achieving optimal performance at $K = 4$, and then gradually declines as K continues to grow.

F. Quantitative Assessment of Frame-Level Annotations Against Human Annotations

We conducted a quantitative comparison between the frame-level annotations generated by Qwen2-VL and manually curated annotations on 100 randomly sampled frames. We evaluated the agreement using the standard caption similarity metric BERTScore. Specifically, we sampled 100 frames from the OpenSurgery dataset and invited five annotators to provide manual annotations for these frames. We then computed the BERTScore for each annotator's description and reported the average score. The results show an average BERTScore of 0.76, indicating strong alignment between the automated and manual annotations and suggesting that the noise introduced by the automated annotation is limited, and that the generated annotations are sufficiently reliable for model training.

VI. Conclusion

Open surgery understanding remains underexplored, primarily due to its inherent complexity and the lack of large-scale, diverse datasets. To address these challenges, we first introduce OpenSurgery, the largest and most diverse video–text dataset to date for open surgery understanding. OpenSurgery consists of two subsets: OpenSurgery-Pretrain and OpenSurgery-EVAL. OpenSurgery-Pretrain consists of 843 publicly available open surgery videos for pretraining, spanning 102 hours and encompassing over 20 distinct surgical types. OpenSurgery-EVAL is a benchmark dataset for evaluating model performance in open surgery understanding, comprising 280 training and 120 test videos, with a total duration of 49 hours. Each video is meticulously annotated by expert surgeons across three hierarchical levels: video, operation, and frame. Building on this foundation, we propose HierSKP, a hierarchical multimodal pretraining framework that integrates granularity-aware contrastive learning with a Dynamic Time Warping (DTW)-based temporal alignment objective. Extensive experiments on operation recognition, temporal action localization, and zero-shot cross-modal retrieval demonstrate that our proposed HierSKP framework achieves state-of-the-art performance in generalized zero-shot learning and offers a strong visual representation that serves as a general initialization for a wide variety of open surgery understanding tasks.

References

[1] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1728–1738.

[2] O. Bar et al., "Impact of data on generalization of AI for surgical intelligence applications," *Sci. Rep.*, vol. 10, no. 1, p. 22208, Dec. 2020.

[3] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[4] T. Czempiel et al., "TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Lima, Peru. Cham, Switzerland: Springer, 2020, pp. 343–352.

[5] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam, and N. Navab, "OperA: Attention-regularized transformers for surgical phase recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Strasbourg, France. Cham, Switzerland: Springer, 2021, pp. 604–614.

[6] A. Das et al., "PitVis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery," 2024, *arXiv:2409.01184*.

[7] G. P. Dobson, "Trauma of major surgery: A global problem that is not going away," *Int. J. Surg.*, vol. 81, pp. 47–54, Sep. 2020.

[8] A. Esteva et al., "Deep learning-enabled medical computer vision," *NPJ Digit. Med.*, vol. 4, no. 1, p.5, 2021.

[9] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 203–213.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[11] X. Gao, Y. Jin, Q. Dou, and P.-A. Heng, "Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8440–8446.

[12] X. Gao, Y. Jin, Y. Long, Q. Dou, and P.-A. Heng, "Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Strasbourg, France. Cham, Switzerland: Springer, 2021, pp. 593–603.

[13] A. Goldbraikh, A.-L. D'Angelo, C. M. Pugh, and S. Laufer, "Video-based fully automatic assessment of open surgery suturing skills," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 3, pp. 437–448, Mar. 2022.

[14] E. D. Goodman et al., "Analyzing surgical technique in diverse open surgical videos with multitask machine learning," *JAMA Surg.*, vol. 159, pp. 185–192, Mar. 2024.

[15] M. Grammatikopoulou et al., "CaDIS: Cataract dataset for image segmentation," 2019, *arXiv:1906.11586*.

[16] I. Hadji, K. G. Derpanis, and A. D. Jepson, "Representation learning via global temporal alignment and cycle-consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11063–11072.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[18] M. Hu et al., "OphNet: A large-scale video benchmark for ophthalmic surgical workflow understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, pp. 481–500.

[19] M. F. Jacobsen, L. Konge, M. Alberti, M. la Cour, Y. S. Park, and A. S. S. Thomsen, "Robot-assisted vitreoretinal surgery improves surgical accuracy compared with manual surgery: A randomized trial in a simulated setting," *Retina*, vol. 40, no. 11, pp. 2091–2098, Nov. 2020.

[20] Y. Jin et al., "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1114–1126, May 2018.

[21] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng, "Temporal memory relation network for workflow recognition from surgical video," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1911–1923, Jul. 2021.

[22] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[24] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9694–9705.

[25] S. Lin et al., "Semantic-SuPer: A semantic-aware surgical perception framework for endoscopic tissue identification, reconstruction, and tracking," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4739–4746.

[26] C. Loukas, "Video content analysis of surgical procedures," *Surgical Endoscopy*, vol. 32, no. 2, pp. 553–568, Feb. 2018.

[27] L. Maier-Hein et al., "Heidelberg colorectal data set for surgical data science in the sensor operating room," *Sci. Data*, vol. 8, no. 1, p. 101, Apr. 2021.

[28] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 6, pp. 623–635, Dec. 1980.

[29] I. Naim, Y. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea, "Unsupervised alignment of natural language instructions with video segments," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014, pp. 1558–1564.

[30] B. Ni et al., "Expanding language-image pretrained models for general video recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 1–18.

[31] Z. L. Ni et al., "RAUNet: Residual attention u-net for semantic segmentation of cataract surgical instruments," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2019, pp. 139–149.

[32] C. I. Nwoye and N. Padoy, "Data splits and metrics for method benchmarking on surgical action triplet datasets," 2022, *arXiv:2204.05235*.

[33] C. I. Nwoye et al., "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102433.

[34] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[35] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[36] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned CLIP models are efficient video learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6545–6554.

[37] M. K. Richards, J. P. McAteer, F. T. Drake, A. B. Goldin, S. Khandelwal, and K. W. Gow, "A national review of the frequency of minimally invasive surgery among general surgery residents: Assessment of ACGME case logs during 2 decades of general surgery resident training," *JAMA Surgery*, vol. 150, no. 2, pp. 169–172, Feb. 2015.

[38] F. Ritter et al., "Medical image analysis," *IEEE Pulse*, vol. 2, no. 6, pp. 60–70, Nov. 2011.

[39] K. Schoeffmann, H. Husslein, S. Kletz, S. Petscharnig, B. Muenzer, and C. Beecks, "Video retrieval in laparoscopic video recordings with dynamic content descriptors," *Multimedia Tools Appl.*, vol. 77, no. 13, pp. 16813–16832, Jul. 2018.

[40] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Lit, and D. Tao, "TriDet: Temporal action detection with relative boundary modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18857–18866.

[41] T. N. Tang, K. Kim, and K. Sohn, "TemporalMaxer: Maximize temporal context with only max pooling for temporal action localization," 2023, *arXiv:2303.09055*.

[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[43] P. Wang et al., "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," 2024, *arXiv:2409.12191*.

[44] T. Wang, H. Li, T. Pu, and L. Yang, "Microsurgery robots: Applications, design, and development," *Sensors*, vol. 23, no. 20, p. 8503, Oct. 2023.

[45] Z. Wang et al., "AutoLaparo: A new dataset of integrated multitasks for image-guided surgical automation in laparoscopic hysterectomy," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 486–496.

[46] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," 2021, *arXiv:2109.14084*.

[47] Z. S. Xue and K. Grauman, "Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 53688–53710.

[48] L. Yang et al., "DyFADet: Dynamic feature aggregation for temporal action detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2024, pp. 305–322.

[49] F. Yu et al., "Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques," *JAMA Netw. Open*, vol. 2, Jan. 2019, Art. no. e191860.

[50] N. Navab, N. Padoy, V. Srivastav, and K. Yuan, "Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation," in *Proc. Adv. Neural Inf. Process. Syst. 37*, 2024, pp. 122952–122983.

[51] K. Yuan et al., "Learning multi-modal representations by watching hundreds of surgical video lectures," 2023, *arXiv:2307.15220*.

[52] X. Yuan et al., "Multimodal contrastive training for visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6991–7000.

[53] R. Zha, X. Cheng, H. Li, M. Harandi, and Z. Ge, "EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2023, pp. 13–23.

[54] C. L. Zhang, J. Wu, and Y. Li, "ActionFormer: Localizing moments of actions with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 492–510.

[55] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, "Automated assessment of surgical skills using frequency analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 430–438.