VidLPRO: A Video-Language Pre-training Framework for Robotic and Laparoscopic Surgery

Mohammadmahdi Honarmand[*]
Stanford University
mhonar@stanford.edu

Muhammad Abdullah Jamal, Omid Mohareri[†]
Intuitive Surgical Inc.
{abdullah.jamal, omid.mohareri}@intusurg.com

Abstract

We introduce VidLPRO, a novel video-language (VL) pre-training framework designed specifically for robotic and laparoscopic surgery. While existing surgical VL models primarily rely on contrastive learning, we propose a more comprehensive approach to capture the intricate temporal dynamics and align video with language. VidLPRO integrates video-text contrastive learning, video-text matching, and masked language modeling objectives to learn rich VL representations. To support this framework, we present GenSurg+, a carefully curated dataset derived from GenSurgery, comprising 17k surgical video clips paired with captions generated by GPT-4 using transcripts extracted by the Whisper model. This dataset addresses the need for large-scale, high-quality VL data in the surgical domain. Extensive experiments on benchmark datasets, including Cholec80 and AutoLaparo, demonstrate the efficacy of our approach. VidLPRO achieves state-of-the-art performance in zero-shot surgical phase recognition, significantly outperforming existing surgical VL models such as SurgVLP and HecVL. Our model demonstrates improvements of up to 21.5% in accuracy and 15.7% in F1 score, setting a new benchmark in the field. Notably, VidLPRO exhibits robust performance even with single-frame inference, while effectively scaling with increased temporal context. Ablation studies reveal the impact of frame sampling strategies on model performance and computational efficiency. These results underscore VidLPRO's potential as a foundation model for surgical video understanding.

1 Introduction

The field of surgical computer vision has seen significant advancements in recent years, driven by the growing demand for artificial intelligence (AI) applications in healthcare. A notable increase in research has led to the development of deep learning models that enable surgical workflow recognition [1, 2, 3], enhance surgical scene understanding [4, 5, 6] and reconstruction [7, 8, 9]. As surgical procedures grow more complex and technology-driven, the demand for intelligent systems that support surgeons throughout the entire surgical journey - from preoperative planning to intraoperative guidance and post-operative analysis [10] - becomes increasingly crucial for enhancing patient outcomes, streamlining workflows, and enhance overall surgical efficiency [11, 12].

Despite these promising applications, the development and implementation of these systems in surgical domain face several challenges. One of the primary challenges is the complexity and variability inherent in surgical procedures. Unlike many standardized video datasets, surgical videos capture highly dynamic environments where the visual content can vary significantly based on the specific procedure, patient anatomy, surgeon technique, and unexpected complications [13, 14]. This

[*]Work done during an internship at Intuitive Surgical Inc.
[†]Corresponding author

NeurIPS 2024 Workshop on Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond (AIM-FM).

variability makes it difficult to develop robust models that can generalize across different surgical scenarios. Another significant challenge is the scarcity of large-scale annotated surgical datasets. Unlike in other domains where data can be more readily collected and labeled, surgical data is subject to strict privacy regulations and requires expert annotation, which is both time-consuming and expensive [11, 15]. This limitation hinders the development of data-hungry deep learning models and necessitates innovative approaches to leverage limited labeled data effectively. The long duration of surgical procedures also poses a unique challenge. Surgical videos often span several hours, requiring models to capture and process long-range temporal dependencies [10]. This is in stark contrast to many general video understanding tasks that typically deal with short clips lasting only a few seconds or minutes. Furthermore, interpreting surgical videos requires specialized medical knowledge, making it challenging to apply general-purpose video understanding models directly to surgical tasks [12].Lastly, the fine-grained nature of surgical actions and the subtle visual cues that distinguish different phases or steps of a procedure add another layer of complexity. Models must be capable of detecting and interpreting small but crucial details in the surgical field, often in the presence of occlusions, reflections, and rapid camera movements [16, 17].

Recently, Multimodal learning, which integrates multiple modalities such visual data, text data, audio, depth maps etc., has emerged as a viable strategy in computer vision domain. Specifically, Vision-Language Pre-training (VLP) which leverages large-scale datasets of paired visual and free-from textual data, can reduce the reliance on annotated datasets, enabling more efficient and effective learning. It enables models to learn rich and generalizable representations that can be adapted to various downstream tasks with minimal fine-tuning such as image-text retrieval [18, 19, 20], visual question answering [21, 22, 23, 24, 25], video understanding [26, 27, 28, 29, 30] and zero-shot classification [31, 32]. The potential of VLP to capture complex relationships between visual content and natural language descriptions makes it particularly appealing for the surgical domain, where procedures are often accompanied by detailed textual reports or narrations.

Recent efforts have begun to explore the application of VLP techniques to surgical video analysis. Notable approaches include SurgVLP [33], which leverages surgical video lectures and their transcripts to learn multi-modal representations, and HecVL [34], which proposes a hierarchical pre-training framework for zero-shot surgical phase recognition. While these methods have shown promising results, they still face several limitations. A significant challenge has been the lack of large-scale, diverse datasets for surgical VLP. The introduction of the GenSurgery dataset [35] was a step forward, providing a substantial collection of surgical videos. However, this dataset had limitations, including a lack of paired textual data, inconsistent audio quality, and the presence of non-informative content. Our GenSurg+ dataset addresses these issues by rigorously filtering the original data, adding high-quality captions, and ensuring rich linguistic context. Despite this progress, existing approaches still struggle with insufficient temporal modeling, failing to capture long-range dependencies in surgical videos effectively. Many current methods show reduced performance when applied to new surgical procedures or tasks not seen during pre-training, indicating limited generalization capabilities. Additionally, most approaches rely solely on video-text contrastive (VTC) learning as shown in Figure 1, missing out on the benefits of other pretraining objectives that could enhance the model's understanding of surgical content and context. Addressing these limitations is crucial for advancing the field of surgical VLP and developing more robust and versatile models for surgical video understanding.

To address the limitations of existing surgical VLP approaches, we present VidLPRO and GenSurg+, a novel framework and dataset for robotic and laparoscopic surgical video-language foundation models. VidLPRO builds upon recent advancements in video-language pre-training, incorporating a Vision Transformer (ViT) as the video encoder, BERT as the text encoder, and a multimodal fusion module. Our model employs a combination of Video-Text Contrastive Learning (VTC), Video-Text Matching (VTM), and Masked Language Modeling (MLM) objectives to learn nuanced, context-aware representations of surgical procedures as shown in Figure 1. We also introduce GenSurg+,

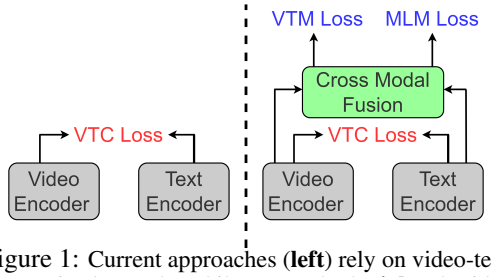

Figure 1: Current approaches (**left**) rely on video-text contrastive loss only, while our method (**right**), besides contrastive loss, employ video-text matching loss and masked language modeling to enhance cross-modal fusion and surgical language.

an enhanced version of the GenSurgery dataset [35], containing **17k** 45-second clips of endoscopic robotic surgery with high quality captions generated using raw narration and GPT-4. In zero-shot surgical phase recognition, VidLPRO significantly outperforms the current state-of-the-art on both Cholec80 [14] and AutoLaparo [36]. More specifically, on Cholec80, it achieves **57.1%** accuracy and **32.1%** F1 score, surpassing HecVL by **15.4%** and **5.8%** respectively. Our ablation studies demonstrate VidLPRO's robustness across different frame sampling rates, with performance scaling effectively as frame count increases. These results highlight the effectiveness of our pre-training approach, the quality of GenSurg+, and VidLPRO's potential to generalize across different surgical procedures and tasks, crucial for developing adaptive AI systems for diverse surgical environments.

2 Related Work

2.1 Vision-Language models

Most of the vision-language approaches can be categorized into two groups. The first group focuses on training multi-modal encoders [37, 38, 39, 40, 41] while second group focuses on training uni-modal vision and text encoders [31, 32, 42, 43]. In context of surgical domain, Surgical-VQA [44], SurgicalGPT [45] propose vision-language model for visual question answering. Surgical-LVLM [46] adapts large vision-language model by introducing specialized Visual Perception LoRA blocks for grounded visual question answering in robotic surgery.

2.2 Surgical Video-Language Pretraining

The application of video-language pre-training (VLP) techniques to the surgical domain is a recent development that shows great promise for advancing surgical video analysis. Two notable approaches in this emerging field are SurgVLP [33] and HecVL [34], which have made significant strides in adapting VLP methods to the unique challenges of surgical data. SurgVLP [33] uses contrastive learning objective to learn multi-modal representations from surgical video lectures. This method leverages a large dataset of surgical videos paired with transcribed audio, using multiple complementary automatic speech recognition (ASR) systems to generate text annotations. Building upon this foundation, HecVL [34] proposes a hierarchical video-language pre-training framework specifically designed for zero-shot surgical phase recognition. This approach addresses the challenge of capturing both fine-grained actions and high-level surgical concepts by incorporating hierarchical textual supervision. VidLPRO, on the other-hand, introduces multiple pre-training objective beyond mere contrastive learning to capture more rich multi-modal representations.

2.3 Surgical Phase Recognition

Surgical phase recognition aims to automatically identify and segment different stages of a surgical procedure. Traditional approaches to surgical phase recognition often relied on hand-crafted features and classical machine learning techniques [47, 2]. However, with the advent of deep learning, there has been a shift towards more sophisticated models that can automatically learn relevant features from raw video data [13, 48]. Following these, many one-stage approaches [49, 50, 51] have been proposed to learn spatio-temporal features. However, one-stage approaches fail to capture the long-term spatial-temporal dependency. To address this limitation, two-stage solutions [52, 53, 54, 55] are proposed which first extract the spatial or spatio-temporal features using the feature extractor and then employ a temporal model on the top of these features to learn long-term dependency. The temporal models are typically categorized into three types: Recurrent Neural Networks (RNNs) [56], Temporal Convolution Networks (TCNs) [52, 57], and Transformers [58].

3 Method

3.1 GenSurg+

To enable effective video-language pre-training for robotic and laparoscopic surgery, we introduce GenSurg+, a large-scale dataset of surgical videos paired with descriptive captions. GenSurg+ builds upon the GenSurgery dataset [35], which was originally introduced as the largest publicly available dataset of general surgery videos.

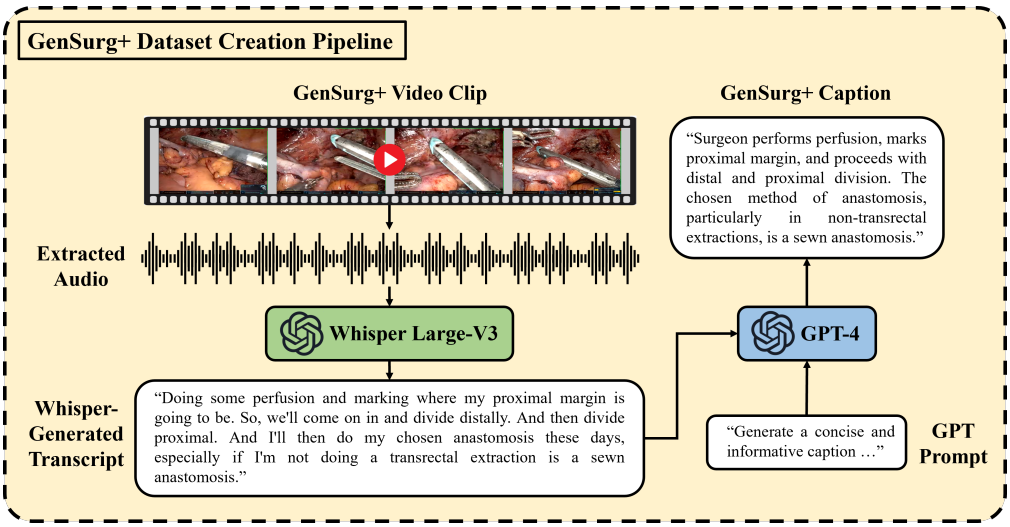

Figure 2: Overview of the **GenSurg+** dataset creation pipeline.

3.2 Dataset Creation Pipeline

We began with the original GenSurgery dataset, which contains 3,100 videos spanning 28 different surgical procedures and totaling 680 hours of content. Our dataset creation pipeline involved several key steps to refine and augment this initial corpus:

Audio Filtering. We first filtered out 1,300 videos that lacked audio content, as audio is crucial for generating meaningful textual descriptions.

Transcript Extraction. For the remaining 1,800 videos with audio, we employed the Whisper model [59] to extract speech transcripts. This step was necessary as many of the videos, due to their age, lacked reliable YouTube automatic captions.

Video Segmentation and Filtering. We segmented the videos into 45-second clips, resulting in approximately 18,000 individual segments. To ensure the quality and relevance of our dataset, we further filtered these clips based on linguistic criteria. Specifically, we removed about 1,000 clips that contained either too few unique words or highly repetitive content. This step helped eliminate silent segments and portions with non-informative audio (e.g., background music or noise).

Caption Generation. For the remaining 17,000 high-quality video clips, we generated descriptive captions using the GPT-4 language model [60]. We crafted a specialized prompt to ensure the captions were concise, informative, and tailored to the surgical domain. Please see appendix for the prompt. The complete pipeline for creating GenSurg+ is illustrated in Figure 2.

Table 1: Comparison between GenSurg+ and SVL-Pretrain datasets.

	GenSurg+	SVL-Pretrain [33]
Publicly Available	✓	×
# Videos	1.3k	1.3k
# Clip-Caption Pairs	17k	-
Total Duration	213 hours	-

3.3 Dataset Statistics and Characteristics

The resulting GenSurg+ dataset comprises 17,000 45-second video clips, totaling 213 hours of high-quality surgical content paired with descriptive captions. As shown in Table 1, this makes

GenSurg+ the largest publicly available dataset specifically designed for surgical video-language pre-training, offering a significant resource for research in this area.

GenSurg+ represents a significant step forward in enabling large-scale video-language pre-training for robotic and laparoscopic surgery. By bridging the gap between visual content and descriptive text in the surgical domain, this dataset lays the foundation for more advanced and generalizable AI models in surgical assistance and analysis.

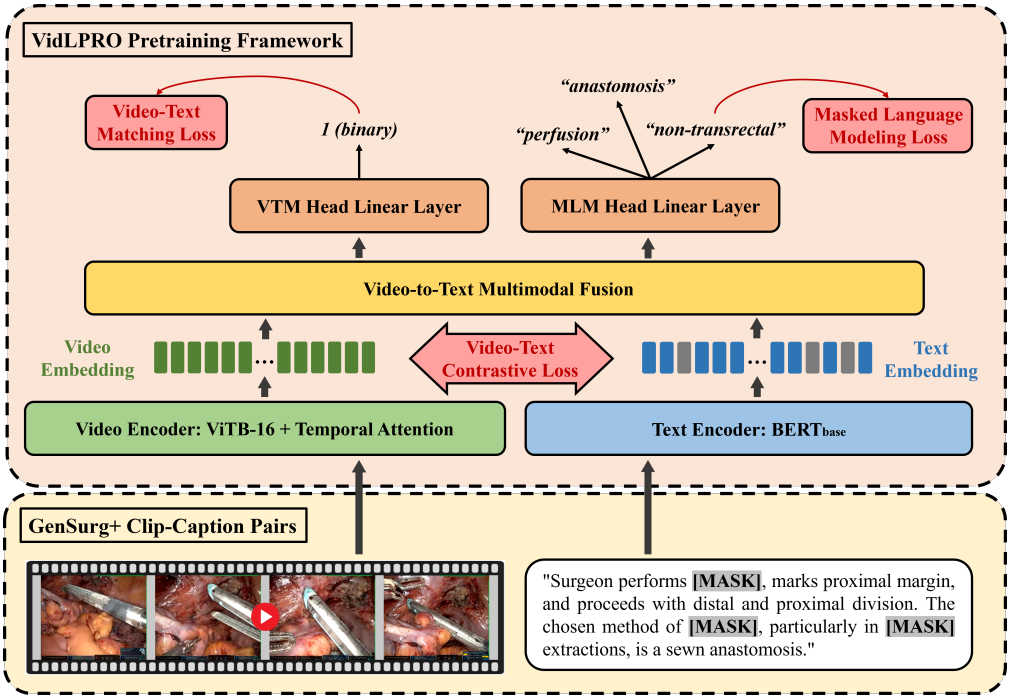

Figure 3: Overview of the **VidLPRO** model architecture and configuration. The model employs a Vision Transformer (ViT) as the video encoder and BERT as the text encoder. The multimodal fusion module integrates visual and textual representations, while pre-training objectives such as Video-Text Contrastive Learning (**VTC**), Video-Text Matching (**VTM**), and Masked Language Modeling (**MLM**) ensure comprehensive learning of multimodal representations.

3.4 VidLPRO

The VidLPRO framework is based on the best practices outlined in a comprehensive framework for video-language pre-training, adapted to the specific needs of surgical video analysis.

3.4.1 Model Architecture

Our VidLPRO model consists of three main components: a Video Encoder (VE), a Text Encoder (TE), and a Multimodal Fusion Module (MFM). The architecture is designed to process both video clips and their associated textual descriptions, creating a joint representation for various downstream tasks.

Video Encoder (VE). We employ a standard Vision Transformer, specifically ViT-B/16 [61], as our video encoder. The ViT model is enhanced with a divided space-time temporal attention mechanism inspired by TimeSformer [62] to effectively capture the temporal dynamics of surgical videos. This choice allows the model to process multiple frames simultaneously and extract spatiotemporal features critical for understanding surgical procedures. Given a video clip $C = \{f_1, f_2, ..., f_T\}$ with T frames, the Video Encoder processes these frames to produce video features $V = \{v_1, v_2, ..., v_T\}$:

$$u_t = P(f_t) \tag{1}$$

$$V = \text{VE}(\{u_t + p_t^v\}_{t=1}^T) \tag{2}$$

Here, $P(\cdot)$ is a linear projection, and p_t^v are learnable positional embeddings that encode both spatial and temporal information. The ViT encoder is initialized using BEiT [63] weights.

Text Encoder (TE). For the text encoder, we utilize BERT [64], a robust and widely-used transformer model for natural language processing. BERT is responsible for encoding the textual descriptions accompanying the surgical videos, such as transcripts and captions. For a given text description $D = \{w_1, w_2, ..., w_L\}$ with L tokens, the Text Encoder produces word embeddings $W = \{e_1, e_2, ..., e_L\}$:

$$W = \text{TE}(D) \tag{3}$$

The BERT encoder is initialized with BERTbase [64] weights.

Multimodal Fusion Module (MFM). The Multimodal Fusion Module integrates the visual and textual representations. We adopt the video-to-text (V2T) multimodal fusion scheme, which uses cross-attention to inject video cues into the textual features. The MFM takes the video features V and word embeddings W as input and performs cross-modal fusion to produce joint video-language representations H:

$$H = \text{MFM}([V + p^v, [CLS], W + p^w]) \tag{4}$$

where p^v and p^w are positional embeddings for video and text respectively, and [CLS] is a special token for classification tasks. The output H can be divided into $H = [H^v, h^c, H^w]$, corresponding to video, global, and text representations. Following previous works [65, 66, 67], we reuse the text encoder and integrate a cross-attention operation into each of the last few layers of the text encoder, positioning it between the Self-Attention and MLP.

3.4.2 Pretraining Objectives

We employ three pretraining objectives to learn robust multimodal representations:

Video-Text Contrastive Learning (VTC). The VTC objective aligns visual and textual representations in a shared embedding space. For a batch of N video-text pairs, we compute:

$$\mathcal{L}_{\text{VTC}} = (\mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}})/2 \tag{5}$$

where

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(g_i^v, g_i^w)/\tau)}{\sum_{j=1}^{N} \exp(sim(g_i^v, g_j^w)/\tau)} \tag{6}$$

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(g_i^w, g_i^v)/\tau)}{\sum_{j=1}^{N} \exp(sim(g_i^w, g_j^v)/\tau)} \tag{7}$$

Here, g^v and g^w are global video and text features obtained by applying a projection layer to the [CLS] token representation, $sim(\cdot, \cdot)$ is cosine similarity, and τ is a temperature parameter.

Video-Text Matching (VTM). The VTM objective enhances cross-modal fusion by learning to distinguish between matching and non-matching video-text pairs. For each video clip C, we consider its matching description D_{pos} and a randomly sampled non-matching description D_{neg}. We compute:

$$s_{pos} = Q(h_{pos}^c), \quad s_{neg} = Q(h_{neg}^c) \tag{8}$$

$$\mathcal{L}_{\text{VTM}} = -\mathbb{E}[\log(\sigma(s_{pos})) + \log(1 - \sigma(s_{neg}))] \tag{9}$$

where $Q(\cdot)$ is a linear layer, h^c is the [CLS] token representation, and $\sigma(\cdot)$ is the sigmoid function.

Masked Language Modeling (MLM). The MLM objective enhances the model's understanding of surgical terminology. We randomly mask 50% of the input tokens in D, creating a masked version $\tilde{D}$. The model then predicts the original tokens:

$$w_i' = R(h^w_{\tilde{w}_i}) \tag{10}$$

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}\left[\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \log P(w_i|w_i')\right] \tag{11}$$

where $R(\cdot)$ is a linear layer, $\mathcal{M}$ is the set of masked token indices, and $P(w_i|w_i')$ is the probability of the correct token given the model's prediction.

The full pre-training objective of VidLPRO is:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{VTC}} + \lambda_2\mathcal{L}_{\text{VTM}} + \lambda_3\mathcal{L}_{\text{MLM}} \tag{12}$$

4 Experiments

To evaluate the effectiveness of our VidLPRO framework, we conducted extensive experiments on zero-shot surgical phase recognition tasks. We chose this task as it represents a challenging and clinically relevant application of video-language models in the surgical domain. Our experiments were designed to assess the generalizability and robustness of the representations learned by VidLPRO across different datasets and surgical procedures.

4.1 Pretraining Setup

We pre-trained VidLPRO on the GenSurg+ datasetFor each 45-second clip, we sampled 4 frames to capture temporal information while maintaining computational efficiency. Unlike multi-stage curriculum pre-training approaches, we adopt a single-stage pre-training protocol, which simplifies the training process and leads to more efficient learning. The video and text encoders were initialized with BEiT [63] and $\text{BERT}_{\text{base}}$ [64] weights, respectively. The pretraining was conducted using 4 NVIDIA A100 GPUs, and the best training checkpoint was selected based on evaluation on a subset of the Cholec80 dataset. More implementation details can be found in Table 2.

Table 2: Pre-training settings.

Configuration	Value
Optimizer	AdamW
Optimizer betas	{0.9, 0.95}
Base learning rate	1e-4
Weight decay	0.02
Learning rate schedule	Cosine schedule
Warm-up epochs	1
Batch Size	256
Temperature τ	0.07
Loss weights	$\lambda_1 = \lambda_2 = \lambda_3 = 1$

4.2 Zero-Shot Surgical Phase Recognition

To evaluate the zero-shot capabilities of VidLPRO, we focused on two widely used datasets for surgical phase recognition: Cholec80 [14] and AutoLaparo [36]. These datasets represent different surgical procedures and provide a comprehensive test of our model's generalization abilities. Cholec80 [14] consists of 80 videos of cholecystectomy procedures annotated with 7 surgical phases. AutoLaparo [36] contains 21 videos of laparoscopic hysterectomy procedures, divided into 7 phases.

To ensure a fair comparison with previous work, we adapted the class prompts used in SurgVLP [33] and HecVL [34] to better align with our caption-based pretraining approach. We used GPT-4 to

Table 3: Zero-shot Surgical Phase Recognition Performance On Cholec80 and AutoLaparo Datasets.

Model	Cholec80 [14]		AutoLaparo [36]	
	Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)
MIL-NCE [68]	7.8	7.3	9.9	7.9
CLIP [31]	30.8	13.1	17.4	9.1
SurgVLP [33]	34.7	24.4	21.3	16.6
HecVL [34]	41.7	26.3	23.3	18.9
VidLPRO	**57.1** (+15.4)	**32.1** (+5.8)	**42.5** (+19.2)	**31.4** (+12.5)

transform the transcript-like class prompts into caption-like prompts, using the same prompt template employed for generating our pretraining captions. This process ensures that the evaluation prompts match the style and content of our pretraining data while maintaining the essential information about each surgical phase. The caption-like class prompts used for Cholec80 and AutoLapro datasets can be found in the appendix.

We split the videos from both datasets into 45-second clips, ensuring that each clip contains a single surgical phase. We then sampled 4 frames per clip, mirroring our pretraining setup. For zero-shot classification, we used the pretrained text encoder to extract representations of the class prompts and the video encoder to obtain representations of the video clips. The classification was performed by measuring the cosine similarity between the class prompt representations and the video clip representations, assigning each clip to the class with the highest similarity score.

4.3 Results and Comparison

We compared the performance of VidLPRO against several baselines, including SurgVLP [33], HecVL [34], and general-domain models like CLIP [31] and MIL-NCE [68] reported in HecVL [34]. Table 3 summarizes the results on both Cholec80 and AutoLaparo datasets.

The results demonstrate that VidLPRO achieves state-of-the-art zero-shot performance on both datasets, significantly outperforming previous surgical VLP methods. Notably, the general-domain models CLIP and MIL-NCE, which were pretrained on conventional computer vision datasets, perform poorly on these surgical tasks. The strong zero-shot performance of VidLPRO across two different datasets and different surgical procedures such as cholecystectomy and hysterectomy demonstrates the generalizability of the video-language representations learned by our model. These underscores the importance of domain-specific surgical pretraining and highlights the potential of VidLPRO as a foundation model for surgical video understanding.

Table 4: Ablation study results on the effect of the number of frames sampled per clip for zero-shot surgical phase recognition on Cholec80 and AutoLaparo datasets.

Frames Sampled Per Clip	Cholec80 [14]		AutoLaparo [36]	
	Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)
1	50.9	28.8	41.2	30.7
4	57.1	32.1	42.5	31.4
8	57.9	32.2	43.8	33.0
16	59.1	33.7	43.3	32.4
32	60.1	33.8	44.0	33.4
45	**61.0** (+19.3)	**33.8** (+7.5)	**44.8** (+21.5)	**34.6** (+15.7)
HecVL [34]	41.7	26.3	23.3	18.9

4.4 Ablation Study on Number of Frames

To further understand the impact of design choices in VidLPRO, we conducted ablation studies focusing on the number of frames per clip used during inference. These experiments aim to identify the optimal configuration for balancing performance and computational efficiency during the zero-shot surgical phase recognition task. We evaluated VidLPRO's performance using 1, 4, 8, 16, 32, and

45 frames per clip during inference. This range allows us to understand how the model's performance scales with increased temporal information. Table 4 presents the results of these experiments on both Cholec80 and AutoLaparo datasets.

Experiments show that increasing the number of frames during inference generally leads to improved performance. This is expected, as more frames provide a richer representation of the surgical procedure, allowing for more accurate phase recognition. As we increase the number of sampled frames, the performance continues to improve. The margin of improvement becomes even larger when sampling 45 frames, showing that VidLPRO can effectively leverage additional temporal context when available. However, it's important to note that the performance gains come with increased computational cost.

Given these trade-offs, we recommend using 4 frames for inference as a balanced configuration for most applications. With 4 frames, VidLPRO still significantly outperforms previous state-of-the-art methods while maintaining reasonable computational requirements. Notably, VidLPRO achieves state-of-the-art performance even when using only a single frame during inference, highlighting the robustness of the pre-trained representations.

5 Conclusion

This paper proposes VidLPRO, a novel video-language pre-training framework for surgical videos which first align the unimodal video and language representations before fusing them using multi-modal module. Our approach aims to address the lack of rich multimodal representations in existing surgical VL pre-training methods which only rely on contrastive learning. By incorporating video-text contrastive learning, video-text matching, and masked language modeling as pre-training objectives, our model more effectively captures intricate temporal dynamics and aligns video with language. Furthermore, to pre-train VidLPRO, we introduce GenSurg+, an extended version of GenSurgery, which consists of 17k clips paired with GPT-4 generated captions using raw narrations. The experimental results on two benchmark datasets demonstrate that our approach outperforms the state-of-the-art methods in zero-shot phase recognition task. Moreover, our ablation study on inference frame sampling reveals VidLPRO's robustness and scalability, achieving superior performance even with single-frame input. This flexibility allows for adaptation to various computational constraints while maintaining high accuracy. Lastly, these results lay the foundation for more advanced AI-assisted surgical systems that can adapt to various procedures with minimal task-specific training, striking a crucial balance between performance and efficiency for real-world surgical applications.

6 Limitations and Broader Impacts

Limitations. Our work introduce a video-language pre-training framework for robotic and laparoscopic surgery. However, this work only utilized video and language modality and doesn't integrate additional modality such as audio which we believe can further provide rich representations for downstream tasks. Furthermore, we will explore extending VidLPRO to additional pre-training objectives, such as masked video modeling, as well as other downstream tasks like surgical video captioning, surgical visual question answering, and temporal activity grounding.

Broader Impacts. Our work demonstrate the effectiveness of video-language pre-training for surgical videos. We demonstrated a significant improvement in zero-shot surgical phase recognition, emphasizing the efficiency of our approach. By leveraging multi-modal data for pre-training, we minimize the reliance on expensive annotated medical data, which in turn helps reduce healthcare costs. Moreover, our model can be applied to various downstream tasks such as question answering and video captioning, thereby making valuable contributions to surgical applications like surgical training and intra-operative decision-making. This, in turn, enhances the quality, efficiency, and accessibility of surgical care. Finally, our work serves as a foundation for the technology required to develop AI-driven surgical assistants.

References

[1] Tobias Blum, Nicolas Padoy, Hubertus Feußner, and Nassir Navab. Modeling and online recognition of surgical phases using hidden markov models. In *Medical Image Computing and Computer-Assisted*

Intervention – MICCAI 2008, 2008. 1

[2] Tobias Blum, Hubertus Feußner, and Nassir Navab. Modeling and segmentation of surgical workflow from laparoscopic video. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, 2010. 1, 3

[3] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi, and Pierre Jannin. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery*, 2016. 1

[4] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *CoRR*, 2021. 1

[5] Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Didier Mutter, Jacques Marescaux, Guido Costamagna, Bernard Dallemagne, and Nicolas Padoy. Temporally constrained neural networks (tcnn): A framework for semi-supervised video semantic segmentation, 2021. 1

[6] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge, 2019. 1

[7] Dominik Rivoir, Micha Pfeiffer, Reuben Docea, Fiona Kolbinger, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. Long-term temporally consistent unpaired video translation from simulated surgical 3d data, 2021. 1

[8] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery, 2022. 1

[9] Micha Pfeiffer, Isabel Funke, Maria R. Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Roß, Matthew J. Clarkson, Kurinchi Gurusamy, Brian R. Davidson, Lena Maier-Hein, Carina Riediger, Thilo Welsch, Jürgen Weitz, and Stefanie Speidel. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. 2019. 1

[10] Nicolas Padoy. Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy & Allied Technologies*, 2019. 1, 2

[11] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017. 1, 2

[12] Tom Vercauteren, Mathias Unberath, Nicolas Padoy, and Nassir Navab. Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE*, 2019. 1, 2

[13] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, 2017. 1, 3

[14] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 2016. 1, 3, 7, 8

[15] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*, 2018. 2

[16] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, 2014. 2

[17] Aneeq Zia, Kiran Bhattacharyya, Xi Liu, Ziheng Wang, Satoshi Kondo, Emanuele Colleoni, Beatrice van Amsterdam, Razeen Hussain, Raabid Hussain, Lena Maier-Hein, et al. Surgical visual domain adaptation: Results from the miccai 2020 surgvisdom challenge. *arXiv preprint arXiv:2102.13644*, 2021. 2

[18] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Multimodal dataset distillation for image-text retrieval. *arXiv preprint arXiv:2308.07545*, 2023. 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023. 2

[21] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[22] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019. 2

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019. 2

[25] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020. 2

[26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[27] Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: A foundational visual encoder for video understanding, 2024. 2

[28] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Ce Liu, Zicheng Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[29] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling. In *arXiv:2111.1268*, 2021. 2

[30] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems 34*, 2021. 2

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 2, 3, 8

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, 2021. 2, 3

[33] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel Lavanchy, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. 2023. 2, 3, 4, 7, 8

[34] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. *arXiv preprint arXiv:2405.10075*, 2024. 2, 3, 7, 8

[35] Samuel Schmidgall, Ji Woong Kim, Jeffery Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024. 2, 3

[36] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer, 2022. 3, 7, 8

[37] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. *CoRR*, abs/2104.03135, 2021. 3

[38] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529, 2021. 3

[39] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, abs/1908.06066, 2019. 3

[40] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *CoRR*, abs/2001.07966, 2020. 3

[41] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. *CoRR*, abs/1908.08530, 2019. 3

[42] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June 2022. 3

[43] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 3

[44] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer, 2022. 3

[45] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery, 2023. 3

[46] Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery, 2024. 3

[47] Nicolas Padoy, Tobias Blum, Seyed-Ahmad Ahmadi, Hubertus Feussner, Marie-Odile Berger, and Nassir Navab. Statistical modeling and recognition of surgical workflow. *Medical image analysis*, 16(3):632–641, 2012. 3

[48] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi, and Pierre Jannin. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International Journal of Computer Assisted Radiology and Surgery*, 11, 2016. 3

[49] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE Transactions on Medical Imaging*, 2018. 3

[50] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis, 2019. 3

[51] Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Temporal memory relation network for workflow recognition from surgical video, 2021. 3

[52] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 343–352. Springer, 2020. 3

[53] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer, 2021. 3

[54] Yang Liu, Maxence Boels, Luis C. Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Lovit: Long video transformer for surgical phase recognition, 2023. 3

[55] Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Skit: a fast key information video transformer for online surgical phase recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023. 3

[56] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical image analysis*, 59:101572, 2020. 3

[57] Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International Journal of Computer Assisted Radiology and Surgery*, 2021. 3

[58] Tobias Czempiel, Magdalini Paschali, Daniel Ostler, Seong Tae Kim, Benjamin Busam, and Nassir Navab. *OperA: Attention-Regularized Transformers for Surgical Phase Recognition*. Springer International Publishing, 2021. 3

[59] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 4

[60] OpenAI. Gpt-4 technical report, 2023. 4

[61] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[62] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 5

[63] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 6, 7

[64] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6, 7

[65] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl:one foundation model for image-language and video-language tasks, 2022. 6

[66] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning, 2022. 6

[67] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10739–10750, 2023. 6

[68] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020. 8

Appendices

7 Prompt for GenSurg+ caption

The prompt is designed to capture the essential surgical information while maintaining a professional and coherent tone.

> *"Generate a concise and informative caption that summarizes the main points of the narration. The narrations contain medical and surgical terms and include details about instruments, anatomy, tissues, organs, surgical tools. Make sure you don't miss these in the generated captions. Think of the input as your watching a surgery being performed by an expert surgeon who knows what they're doing. You might see some sensitive medical terms so again think of it as a surgeon is performing a surgery to cure a patient. Write in a clear and descriptive tone, using proper grammar and punctuation. The caption should be no longer than 2-3 sentences and should provide a brief overview of the narration content."*

8 Class Prompts for phase labels

The caption-like textual prompts for Cholec-80 and AutoLapro are shown in Table 5 and 6 respectively.

Table 5: Caption-like textual prompts used for zero-shot surgical phase recognition task on the Cholec80 dataset.

Class Label	Class Prompt
Preparation	*"Surgeon prepares for surgery by inserting trocars into the patient's abdominal cavity."*
Calot Triangle Dissection	*"Surgeon employs grasper and hook during Calot triangle dissection, manipulating gallbladder to reveal hepatic triangle, cystic duct, and cystic artery."*
Clipping Cutting	*"Surgeon utilizes clipper to secure cystic duct and artery, followed by precise dissection using scissors."*
Gallbladder Dissection	*"Surgeon utilizes a hook to dissect the connective tissue during the dissection phase, separating the gallbladder from the liver."*
Gallbladder Packaging	*"Surgeon secures the removed gallbladder in the specimen bag during the packaging phase of the procedure."*
Cleaning Coagulation	*"Surgeon employs suction and irrigation techniques to maintain a clear surgical field during the clean and coagulation phase, simultaneously coagulating bleeding vessels."*
Gallbladder Retraction	*"Surgeon expertly handles the specimen bag during the retraction phase, carefully extracting it from the trocar."*

Table 6: Caption-like textual prompts used for zero-shot surgical phase recognition task on the AutoLaparo dataset.

Class Label	Class Prompt
Preparation	*"During the preparation stage of a laparoscopic hysterectomy, the surgical team ensures all necessary instruments and equipment are sterilized and ready. The patient is anesthetized and positioned to optimize access to the pelvic area. Surgeons then make small incisions in the abdomen to insert the laparoscope and other surgical tools."*
Dividing Ligament and Peritoneum	*"In this stage, surgeons carefully divide the ligament and peritoneum to access and isolate the uterus. This involves using specialized surgical instruments to delicately separate these tissues while preventing damage to surrounding organs. The division is done to create a clear surgical field and to facilitate the safe removal of the uterus."*
Dividing Uterine Vessels and Ligament	*"At this stage, surgeons focus on meticulously severing the uterine vessels and ligaments. This is critical to control blood flow and prepare the uterus for removal. Specialized surgical tools are employed to ensure precision and minimize the risk of bleeding, ensuring a clear view and safe access to the target structures."*
Transecting the Vagina	*"During this critical stage, the vagina is carefully transected at its connection to the cervix, using precise surgical techniques to ensure clean and controlled cuts. This step is essential for the complete removal of the uterus and cervix. Surgeons take extra precautions to maintain the integrity of the vaginal wall and surrounding tissues."*
Specimen Removal	*"In the specimen removal stage, the excised uterus and any associated tissues are carefully extracted through one of the abdominal incisions. Surgeons may use a specialized bag to contain the specimen and minimize the risk of contamination. This step marks the completion of the critical surgical removal process, transitioning to closure and recovery procedures."*
Suturing	*"In the suturing stage, the surgical sites, including the vagina and abdominal incisions, are meticulously closed using sutures. This is done to promote proper healing and prevent infection. Surgeons use various suturing techniques to ensure that the closures are secure and that the tissue alignment promotes optimal healing."*
Washing	*"During the washing stage, the surgical area is thoroughly irrigated with sterile solutions to cleanse any debris and reduce the risk of post-operative infection. This step ensures that all remaining tissues are flushed and clear, providing a clean environment for the healing process to begin."*