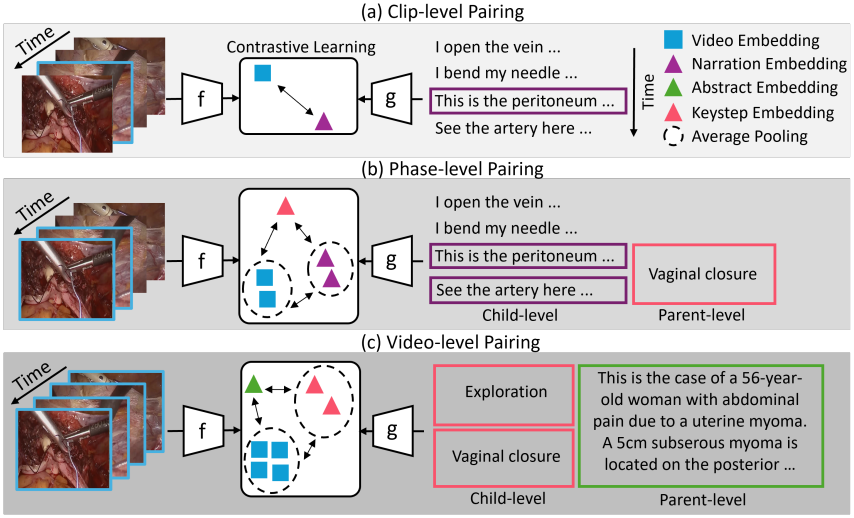
PeskaVLP: Procedure-Aware Surgical Video-language Pretraining with Hierarchical Knowledge Augmentation

Kun Yuan, Vinkle Srivastav, Nassir Navab, Nicolas Padoy

NEURAL INFORMATION PROCESSING SYSTEMS

(a) Hierarchical Surgical Video-language Pretraining



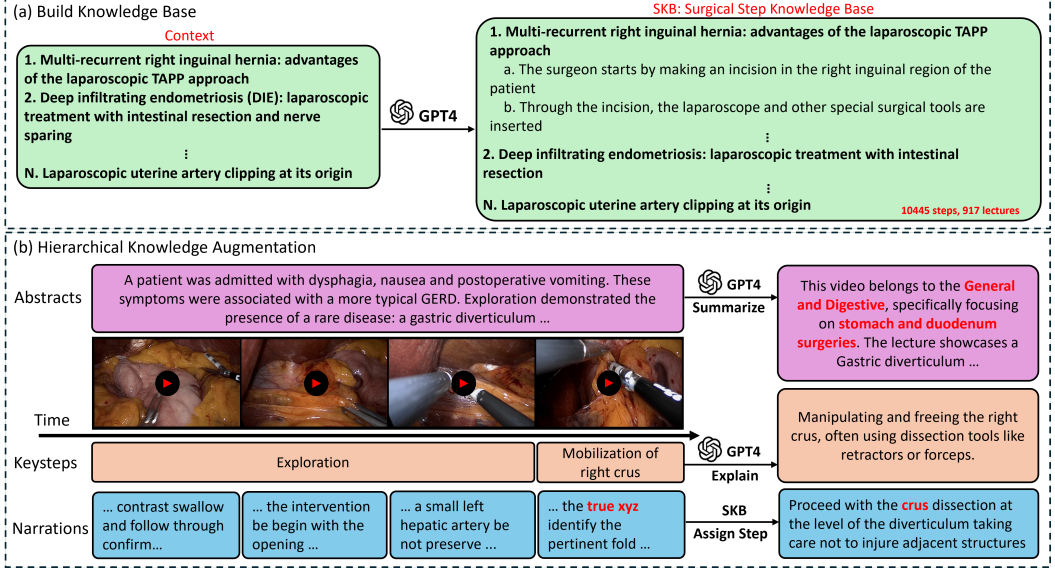
Challenges

Textual Information Loss: Transcribed narrations from surgical lectures often contain errors, missing information, and inconsistencies.

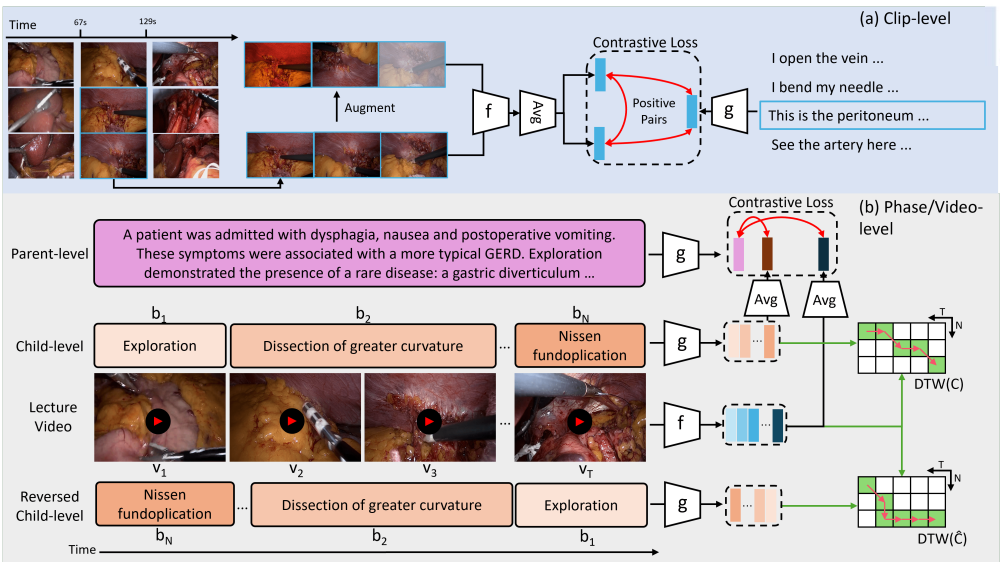
Spatial-Temporal Complexity: Surgical procedures involve complex sequences of actions and interactions between instruments and anatomy, requiring models to capture both fine-grained details and long-term dependencies.

PeskaVLP: This novel framework tackles the challenges of textual information loss and spatial-temporal complexity in surgical VLP. It uses hierarchical knowledge augmentation and a procedure-aware contrastive learning objective (LecNCE) to learn robust visual representations for surgical scene understanding, from language supervision.

(b) Hierarchical Knowledge Textual Augmentation



(c) Cross-modal Alignment Procedure Awareness



Learning Objectives

Procedure-Aware Contrastive Learning (LecNCE):

- LecNCE clip: Combines language supervision with visual self-supervision at the clip level to improve data efficiency and visual understanding.
- LecNCE phase/video: Employs Dynamic Time Warping (DTW) loss to explicitly model procedural alignment between hierarchical video-text pairs.

Benefits:

- Captures long-term dependencies and temporal order of surgical steps.
- Handles variations in procedure execution and temporal alignment across videos.

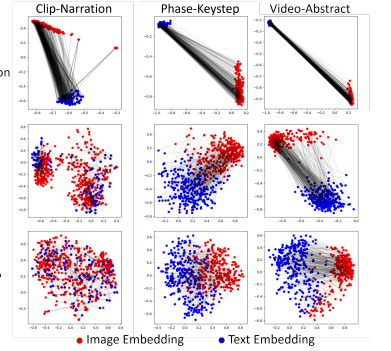
(d) Results

Zero-shot Text-based Video Retrieval

method	Clip-Narration R@1	Clip-Narration R@5	Clip-Narration R@10	Phase-Keystep R@1	Phase-Keystep R@5	Phase-Keystep R@10	Video-Abstract R@1	Video-Abstract R@5	Video-Abstract R@10
Text-to-Image (%)									
CLIP [50]	2.9	5.2	6.7	1.7	3.2	6.3	1.2	11.7	25.8
SurgVLP [73]	2.8	11.8	16.1	1.6	6.8	11.6	1.3	8.2	15.5
HecVL [72]	2.7	11.3	17.2	3.9	13.7	21.3	28.2	74.1	82.3
PeskaVLP	**3.2**	**13.2**	**23.3**	**6.1**	**21.0**	**35.4**	**38.8**	**75.3**	**85.9**
Image-to-Text (%)									
CLIP [50]	1.8	3.9	6.0	0.3	1.2	2.7	0	7.0	16.4
SurgVLP [73]	1.3	8.6	13.5	1.0	4.1	7.3	1.3	8.6	14.6
HecVL [72]	2.1	9.0	16.2	1.9	8.3	14.8	21.2	65.9	71.8
PeskaVLP	**2.4**	**13.1**	**21.3**	**3.4**	**14.9**	**24.8**	**38.8**	**75.3**	**81.1**

Zero-shot Surgical Phase Recognition

Model	Dataset	Cholec80	Autolaparo	StrasBypass70	BernBypass70	Average
MIL-NCE [44]	Howto100M	7.8 / 7.3	9.9 / 7.9	5.6 / 3.1	2.4 / 2.1	6.4 / 5.1
CLIP [50]	CLIP400M	30.8 / 13.1	17.4 / 9.1	16.9 / 5.5	14.8 / 4.1	19.9 / 8.0
	Scratch	29.4 / 10.4	15.3 / 10.9	6.3 / 3.5	4.9 / 2.3	14.0 / 6.8
	SVL	33.8 / 19.6	18.9 / 16.2	15.8 / 8.6	17.8 / 7.1	21.6 / 12.9
SurgVLP [73]	SVL	34.7 / 24.4	21.3 / 16.6	10.8 / 6.9	11.4 / 7.2	19.6 / 13.8
HecVL [72]	SVL	41.7 / 26.3	23.3 / 18.9	26.9 / 18.3	22.8 / 13.6	28.7 / 19.3
PeskaVLP	SVL	**45.1 / 34.2**	**26.5 / 23.6**	**46.7 / 28.6**	**45.7 / 22.6**	**41.0 / 27.1**



Conclusion

Zero-Shot Transferability: Pretrained PeskaVLP models can be directly applied to downstream tasks without finetuning, showcasing its generalizability and versatility. **Strong Visual Representation:** PeskaVLP learns robust visual representations that generalize well to various surgical scene understanding tasks.