Automated Surgical Report Generation Using In-context Learning with Scene Labels from Surgical Videos

Keisuke Ueda[1], Mineto Fujisawa[1, 2], Daiki Hiratsuka, MD[1], Giovanni Ruggeri, MD[3], Takashi Matsumoto[1], Michael Mueller, MD[3]

[1]Medical Dataway AG,
[2]Faculty of Medicine, The University of Tokyo
[3]Department of Gynecology and Obstetrics, University of Bern and Bern University Hospital

Introduction

The primary objective of this research is to develop software that automatically generates surgical reports from surgical video scene labels by utilizing an existing open-source language model. Writing surgical reports is a significant administrative burden for surgeons, limiting their time for patient care. Automating this process could allow surgeons to dedicate more time to surgeries and patient interactions. Recent advancements in language models present a promising solution for this automation. Our goal is to develop software that generates surgical reports from video scene labels and to verify its effectiveness and optimal experimental conditions.

Materials and Methods

We collected ten videos of total laparoscopic hysterectomy (TLH) surgeries performed at Inselspital from 2022 to 2024. For each video, a gynecological surgeon annotated scene labels and wrote surgical reports, as illustrated in **Figure 1**. Scene labels, recorded every second, detailed the surgeon's actions and objects of interest. Three videos were used as examples within the in-context learning (ICL) prompts, and one video was used for designing the evaluation method, while the remaining six were used for evaluation. We investigated the impact of the number of examples in the ICL prompt (1-shot, 2-shot, 3-shot) on the performance of the report generation using the llama-3-8b model. The generated reports (**Figure 2**) were evaluated by comparing the number of discrepancies with the surgeon-written reports. Specifically, we counted the total number of omissions (items present in the reference reports but absent in the generated reports) and hallucinations (items present in the generated reports but absent in the reference reports).

Time Range	Activity
0:00 - 0:42	Others (Camera port insertion)
0:42 – 7:30	Others (Intra-abdominal observation)
7:30 – 11:32	Others (Insertion of trocars for operation)
11:32 - 23:30	Others (Intra-abdominal observation: Pelvic adhesion dissection)
23:30 – 25:50	Dissection - Right round ligament
25:50 - 32:50	Dissection - Right ovarian ligament, fallopian tube, and right parametrium
32:50 – 35:30	Dissection - Left round ligament
35:30 – 44:15	Incision - Left ovarian ligament, fallopian tube, and left parametrium
44:15 – 51:45	Dissection - Vesicouterine peritoneum and upper vaginal wall
51:45 – 59:53	Dissection - Peritoneum near the left sacrouterine ligament
59:53 - 62:03	Others (Hemostasis)
62:03 - 64:04	Dissection - Left parametrium
64:04 - 67:12	Dissection - Lower vaginal wall (partial)
67:12 – 70:25	Dissection - Left parametrium including the left uterine artery
70:25 - 77:30	Dissection - Right parametrium including the right uterine artery, right vaginal wall to lower vaginal wall
77:30 - 79:30	Incision - Lower vaginal wall
79:30 - 82:30	Uterus removal
82:30 - 89:30	Others (Hemostasis confirmation)
89:30 – 102:40	Closure - Vaginal stump
102:40 - 108:50	Biopsy (Peritoneum)
108:50 – 112:10	Others (Hemostasis, irrigation)

Surgery Date:
Age/Sex:
Diagnosis:
Procedure: Total laparoscopic hysterectomy + Bilateral salpingectomy
Surgeon:
Assistant:
Anesthesiologist:
Anesthesia Time:
Surgery Time:

Intra-abdominal Findings:
***

Surgical Course:

1. Positioned the patient in lithotomy position, disinfected, and inserted the manipulator (no video).
2. Inserted a 12mm port at the umbilicus and 5mm ports on the left, right, and midline of the lower abdomen.
3. Observed the abdominal cavity and confirmed the above findings.
4. Dissected the right round ligament, fallopian tube, and ovarian ligament.
5. Dissected the left round ligament, fallopian tube, and ovarian ligament.
6. Incised the upper and lower vaginal wall.
7. Dissected the parametrium, including both uterine arteries, and incised the left and right vaginal walls.
8. Removed the uterus transvaginally.
9. Confirmed hemostasis and closed the vaginal stump with continuous sutures.
10. Biopsied the suspected endometriosis peritoneal lesions.
11. Reconfirmed hemostasis and irrigated the abdominal cavity.
12. Removed the trocars.
13. Closed the incisions.

Figure 1. An example of scene labels (left) and surgical report by human (right).

Surgery Date:
Age/Gender:
Diagnosis:
Procedure: Total laparoscopic hysterectomy + Bilateral salpingectomy
Surgeon:
Assistant:
Anesthesiologist:
Anesthesia Time:
Surgery Time:

Intra-abdominal Findings:
***

Surgical Course:

1. Positioned in lithotomy position, disinfected, and inserted the manipulator (no video).
2. Inserted a 12mm port at the umbilicus using the open technique, and 5 mm ports on the left and right lower abdomen and midline.
3. Observed the abdominal cavity and confirmed the above findings.
4. Dissected the right round ligament, ovarian ligament, fallopian tube, and parametrium.
5. Dissected the left round ligament, ovarian ligament, fallopian tube, and parametrium.
6. Incised the left ovarian ligament, fallopian tube, and parametrium.
7. Dissected the vesicouterine peritoneum and upper vaginal wall.
8. Dissected the peritoneum near the left sacrouterine ligament.
9. Confirmed hemostasis.
10. Dissected the left parametrium.
11. Incised the lower vaginal wall.
12. Removed the uterus transvaginally.
13. Confirmed hemostasis and closed the vaginal stump with continuous sutures.
14. Biopsied the peritoneum.
15. Confirmed hemostasis and irrigated the abdominal cavity.
16. Removed the trocars.
17. Closed the incisions.

Figure 2. An example of surgical reports generated by Llama-3-8b.

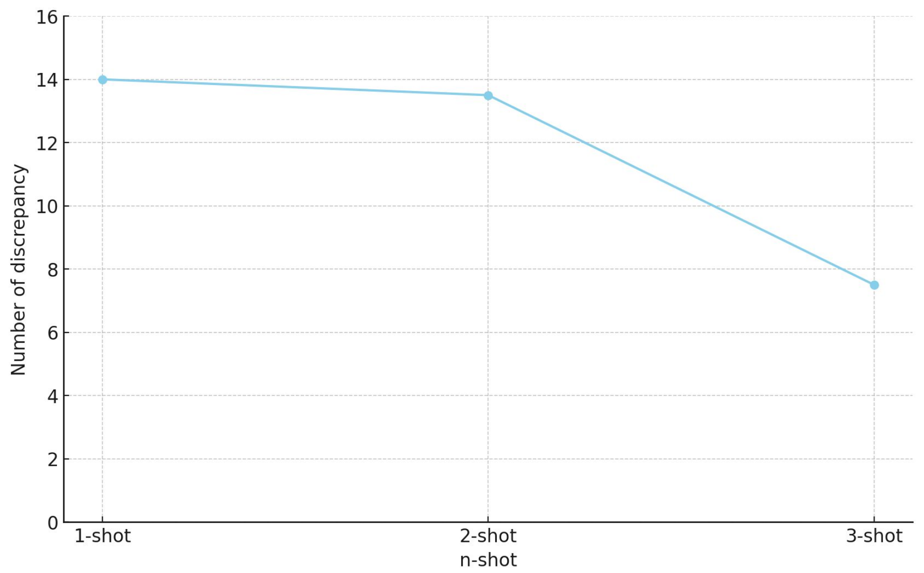

Figure 3. The numbers of discrepancies between AI-generated vs. surgeon-written reports

Results and Discussions

Figure 3 describes the numbers of discrepancies between AI-generated reports and surgeon-written reports for six surgical videos. The average number of discrepancies across the six evaluation reports was 14 (1-shot setting), 13.5 (2-shot setting), and 7.5 (3-shot setting). These results indicate that increasing the number of examples in the ICL prompts improves the accuracy of the generated reports, demonstrating the effectiveness of ICL in surgical report generation. However, even with the 3-shot setting, the average discrepancy count remains at 7.5, indicating a need for further refinement to reduce errors.

Conclusion

This study demonstrates the effectiveness of in-context learning in generating surgical reports from surgical video scene labels. Despite the improvements shown with an increased number of examples in ICL, there remains a considerable number of errors. Future research should explore the impact of model size on performance and investigate whether fine-tuning the model can further enhance report generation accuracy.