Transferability of General Large Multimodal Model ChatGPT-4-Vision to Surgical Video Domain with Minimal Training

Garrett Skinner MD[1,2*], Emily M Hannah BS[1*], Joseph L'Huillier MD[1], Christopher McCulloh MD[2], Steven D Schwaitzberg MD[1], Gene Yang MD[1], Peter CW Kim MD PhD[1,2]

[1] Jacobs School of Medicine and Biomedical Sciences at the University at Buffalo, Department of Surgery, [2] Activ Surgical Inc., Boston, MA, * Contributed equally.

Introduction

Foundational large multimodal models like ChatGPT-4-Vision (GPT-4V) promise opportunity to rethink surgical AI development which historically uses highly data intensive training.

We hypothesize that a foundational LMM could be transferred to the surgical video using one-shot transfer learning and prompt engineering with minimal training.

Methodology

GPT-4V was asked to classify frames (n=731) containing a single laparoscopic instrument sourced randomly from the validated Cholec80 test data split. (**Fig. 1**).

1. Low-data transfer learning was applied to GPT-4V.

- **Zero-shot:** no example images included provided.
- **One-shot:** 1 example of each image of each instrument class provided.

1

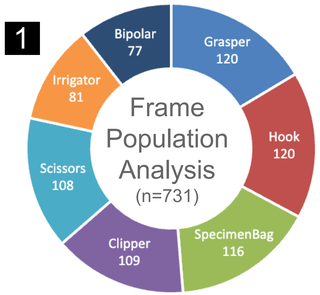


Figure 1. Frame Population Analysis

2. Prompt engineering was done by generating two learning contexts:

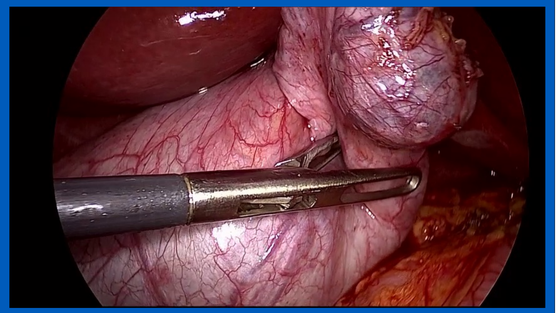
- **Self-taught:** GPT-4V was asked to visually describe the instrument to itself (**Fig. 2A**).
- **Human-taught:** A surgical resident crafted a conversation (7 messages in length) about each surgical instrument to "teach" GPT-4V (**Fig. 2B**).

Performance of the model was measured using the harmonic mean of precision and recall (F1).

2A SELF-TAUGHT

<user> : Here is an image of a laparoscopic grasper instrument [...]

<system> : Here is a description of a laparoscopic grasper. Visually, the tip of the grasper consists of two opposing jaws that come together to grip tissue. These jaws are typically made of a high-grade stainless steel. . . *[196 words]*



<End of exchange>

2B HUMAN-TAUGHT

<user> : Here is an image of a laparoscopic grasper instrument [...]

<system> : I understand that you want me to describe the visual appearance of the laparoscopic grasper instrument and you would like me to reference the information in this discussion when identifying this information in the future. A laparoscopic grasper is composed of . . . *[28 words]*

<user> : That is correct. A grasper is a grey, metallic instrument [...] How would you differentiate the grasper from a laparoscopic scissors, another instrument with two jaws?

<Continues for total of 7 messages exchanged *[180 words]*>

Figure 2. Sample Conversations for Grasper Instrument: 2A) One-Shot-Self-Taught Prompt Engineering, GPT-4V is prompted by user then self-describes grasper; 2B) One-Shot-Human-Taught Prompt Engineering, GPT-4V is prompted by user, GPT-4V response is crafted by human.

Results

Compared to zero-shot, model performance improved for one-shot learning context by 41% (+0.11), one-shot-self-taught improved by 38% (+0.10), and one-shot-human-taught improved by 70% (+0.19). (**Fig. 3**).

3

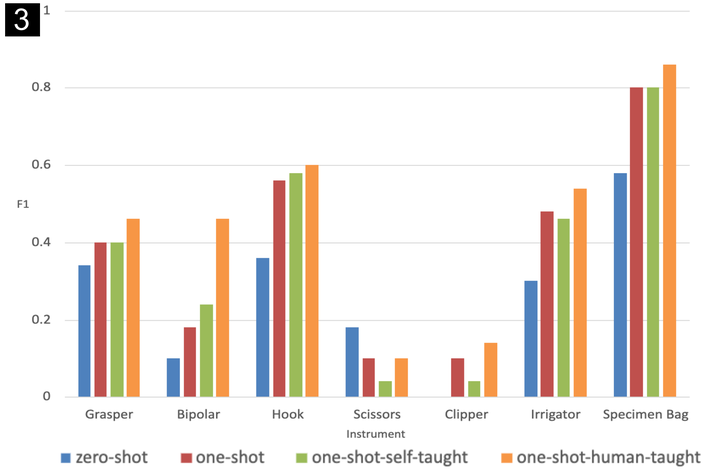


Figure 3. Instrument classification performance by transfer learning and prompt engineering methods.

Conclusion

Despite improvements using one-shot transfer learning and prompt engineering, classification of laparoscopic instruments was lacking compared to fully-supervised benchmarks.

Improving transferability of fundamental LMM models to surgical discipline could allow for economical scaling of surgical-specific computer vision tasks.