Multi-Modal Self-Supervised Learning for Surgical Feedback Effectiveness Assessment

Arushi Gupta[*]   **Rafal Kocielnik**[*]   **Jiayun Wang**[*]   {AGUPTA5,RAFALKO,PETERW}@CALTECH.EDU

Firdavs Nasriddinov   FIRDAVS@CALTECH.EDU
California Institute of Technology, USA

Cherine Yang   **Elyssa Wong**   {CYANG,EYWONG}@USC.EDU
University of Southern California, USA

Anima Anandkumar   ANIMA@CALTECH.EDU
California Institute of Technology, USA

Andrew J. Hung   ANDREW.HUNG@CSHS.ORG
Cedars-Sinai Medical Center, USA

Abstract

During surgical training, real-time feedback from trainers to trainees is important for preventing errors and enhancing long-term skill acquisition. Accurately predicting the effectiveness of this feedback, specifically whether it leads to a change in trainee behavior, is crucial for developing methods for improving surgical training and education. However, relying on human annotations to assess feedback effectiveness is laborious and prone to biases, underscoring the need for an automated, scalable, and objective method. Creating such an automated system poses challenges, as it requires an understanding of both the verbal feedback delivered by the trainer and the visual context of the real-time surgical scene. To address this, we propose a method that integrates information from transcribed verbal feedback and corresponding surgical video to predict feedback effectiveness. Our findings show that both transcribed feedback and surgical video are individually predictive of trainee behavior changes, and their combination achieves an AUROC of 0.70 ± 0.02, improving prediction accuracy by up to 6.6%. Additionally, we introduce self-supervised fine-tuning as a strategy for enhancing surgical video representation learning, which is scalable and further enhances prediction performance. Our results demonstrate the potential of multi-modal learning to advance the automated assessment of surgical feedback.

[*] These authors contributed equally

Keywords: Robotic surgery, Surgical feedback, Self-supervised learning, Fine-tuning, Multimodality, Video understanding

Data and Code Availability   Due to privacy restrictions, data is available upon request. The code can be accessed here.

Institutional Review Board (IRB)   The data used was collected under IRB. IRB information will be provided if the paper is accepted.

1. Introduction

Importance: During surgical training, real-time feedback from trainers to trainees with the intention of modifying trainee behavior is crucial for avoiding errors (Wong et al., 2023) and enhancing long-term learning outcomes (Agha et al., 2015). The quality of such feedback has been shown to significantly influence a trainee's intraoperative performance (Bonrath et al., 2015), development of technical skills (Ma et al., 2022), and sense of autonomy (Haglund et al., 2021). The quality of communication has also been directly linked to operative outcomes (D'Angelo et al., 2020). Given its impact, understanding what makes feedback effective, specifically in terms of leading to trainee behavior change, is important. However, manually assessing feedback is unsustainable and prone to human biases (Chinh et al., 2019). As a result, there is a need for an objective, scalable, and automated approach for predicting feedback effectiveness (Figure 1).
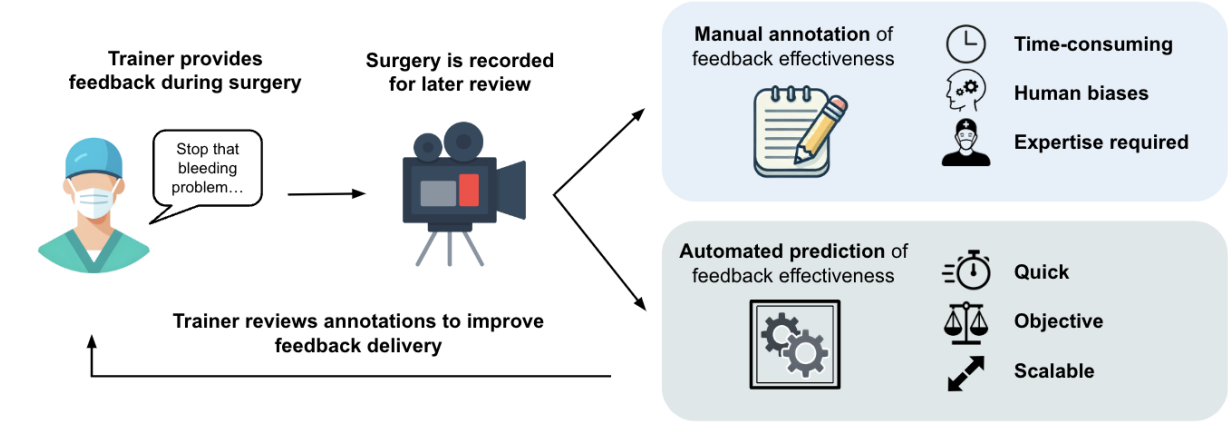
Figure 1: **Motivation for automating the prediction of feedback effectiveness in surgical training.** Manual annotation is time-consuming, prone to human biases, and requires expertise, while automated prediction is quick, objective, and scalable. Trainers can use these assessments to improve their feedback delivery in the operating room, enhancing the quality of surgical training.
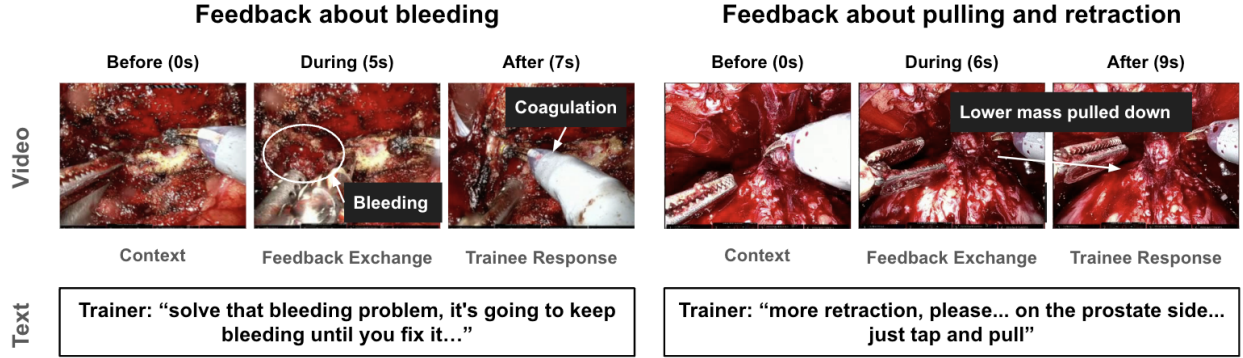

Figure 2: **Examples of feedback instances that are effective (result in trainee behavior change).** The left example depicts feedback that describes addressing bleeding which causes the trainee to perform coagulation. The right example depicts feedback about pulling and retraction which causes the trainee to pull down the lower mass.

Challenges: Analyzing feedback effectiveness is challenging due to its inherently unstructured and multimodal nature (Kocielnik et al., 2023; Wong et al., 2023). Surgical feedback delivery involves a combination of verbal conversations, visual cues, and non-verbal appraisals, all of which can vary significantly in form. Thus, evaluating feedback effectiveness requires a comprehensive understanding of the verbal feedback content and the visual surgical context (Wong et al., 2023). Feedback content provides information about the quality and nature of the trainer's guidance, while visual context provides insight into trainee behavior during feedback delivery, relevant complexities of the surgical scene (i.e., in-strument usage, presence of bleeding, etc.), and post-feedback trainee behavior (Figure 2).

Our Approach: In this work, we propose a multi-modal approach that integrates textual and visual information to predict feedback effectiveness. We transcribe audio recordings of feedback during surgery and use Sentence Transformers (SBERT) to extract embeddings from the transcriptions, capturing the semantic meaning of the trainer's guidance (Reimers and Gurevych, 2019). On the visual side, we use video masked autoencoders (VideoMAE) to extract information from the surgical video corresponding to when feedback is delivered (Tong et al., 2022). To enhance model performance, we employ self-supervised

2

1. Self-supervised learning (SSL) using unlabeled data
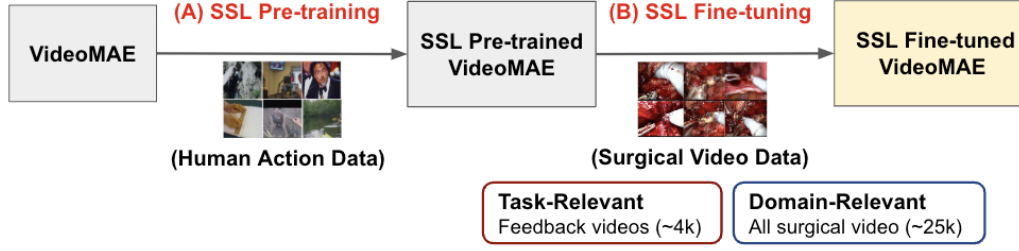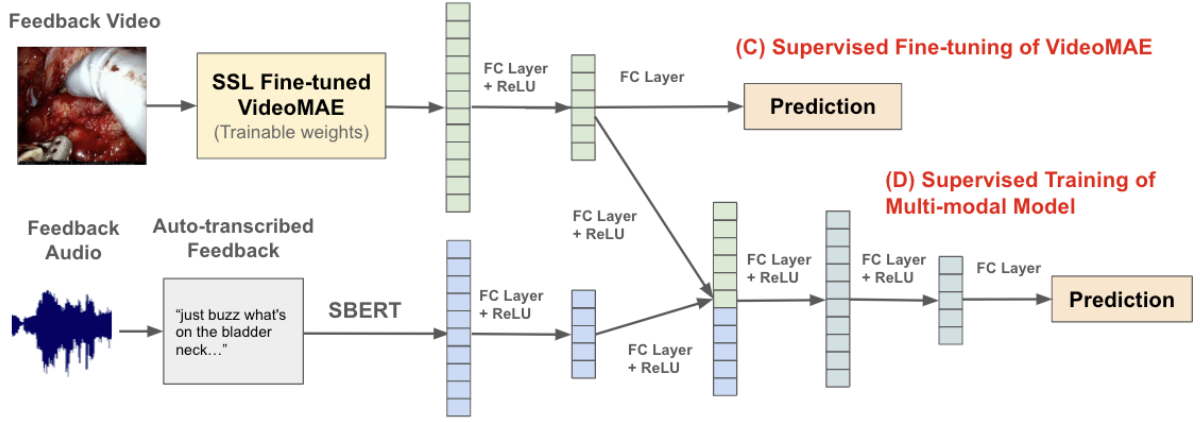

2. Supervised learning using labeled data

Figure 3: **The model pipeline** for predicting trainee behavior change from feedback audio and video, composed of four steps: **(A) Self-Supervised Pre-training** of VideoMAE using Kinetics-400 (human action data); **(B) Self-Supervised Fine-tuning** of VideoMAE using either **Task-Relevant** fine-tuning on feedback-specific videos or **Domain-Relevant** fine-tuning on all surgical video; **(C) Supervised Fine-tuning of VideoMAE** for predicting trainee behavior change; and **(D) Supervised Training of Multi-modal Model** which combines video features from VideoMAE and text features from SBERT for the prediction task.

learning (SSL) fine-tuning of VideoMAE on surgical data. We do SSL fine-tuning for the following reasons: **1) SSL is easily scalable:** SSL does not require labeled data, making it easy to continue SSL pre-training on additional domain-specific data as it is acquired (Caron et al., 2020; Kaplan et al., 2020; Sorscher et al., 2022). **2) Trainee behavior change is a weak label:** There is not always a clear link between surgical video corresponding to feedback and trainee behavior change, and even when there is, visual features indicative of trainee behavior vary across different feedback instances. SSL fine-tuning enables the model to learn more generalizable visual features from surgical video, which enhances the effectiveness of supervised training for the prediction task.

We explore two SSL fine-tuning strategies: task-relevant SSL fine-tuning on task-relevant feedback video instances and domain-relevant SSL fine-tuning on all surgical videos. Both approaches result in improved predictive accuracy, while task-relevant SSL achieves comparable accuracy with 16.8% of the data.

Findings: We summarize our findings as follows:

- We achieve an Area under the ROC Curve (AUC) of 0.70 $\pm$ 0.02 for the novel task of predicting trainee behavior change in surgical training.

- We show that self-supervised fine-tuning on surgical video data even with a limited amount of data significantly helps improve prediction performance.

- We show that verbal feedback content and surgical video are predictive of trainee behavior, and the fusion of both improves prediction by up to 6.6%.

Contributions:

- To the best of our knowledge, we are the first to show the feasibility of an automated system to predict the effectiveness of surgical feedback in changing trainee behavior from multi-modal features of surgical video and audio.

- We introduce self-supervised fine-tuning strategies to improve representation learning of surgical video.

- We provide novel insights into the importance of textual and visual features for evaluating feedback effectiveness in real-world surgeries.

2. Related Work

Feedback in Robot-Assisted Surgery. Prior work on analyzing surgical feedback in robot-assisted surgery has primarily focused on categorization. Wong et al. (2023) most recently developed a clinically validated system to manually categorize feedback into three core components (Anatomic, Procedural, Technical) and auxiliary aspects (Praise, Criticism, Visual Aid). The study found some associations between these categories and trainee behavior change. However, its analysis was limited by the broadness of the categories, which do not capture the finer details necessary for a more accurate understanding of feedback effectiveness. Furthermore, the study relied heavily on manual effort to determine these categories, making the approach vulnerable to human limitations in recognizing complex patterns (Chan and Mazzocco, 2021), inherent cognitive biases (Chinh et al., 2019), and attention span constraints (Baralt, 2011).

To the best of our knowledge, there exists no prior work on the automated prediction of feedback effectiveness.

Self-Supervised Video Understanding. The success of deep learning has relied heavily on the availability of large-scale annotated datasets, but obtaining these annotations is costly and labor-intensive, especially for surgical feedback, which additionally requires specialized expertise. Self-supervised learning (SSL) addresses this issue by enabling models to learn from unlabeled data (Schiappa et al., 2023; Tong et al., 2022). SSL trains models using objectives derived from the data itself, and the model is then later fine-tuned (using supervised learning) for specific downstream tasks (Schiappa et al., 2023). This approach is well-suited for feedback effectiveness prediction, where only a subset of surgery video data is tied to feedback instances.

A notable SSL technique for video is video masked autoencoders (VideoMAE), recognized for its data efficiency (Tong et al., 2022). VideoMAE masks parts of the video and trains the model to reconstruct missing segments, capturing essential contextual and temporal information (Tong et al., 2022). VideoMAE pre-trained on human action data has been effective in surgical tasks like phase recognition in thoracic surgery (Mateen et al., 2024) and surgical feedback classification (Kocielnik et al., 2023). We use VideoMAE to predict feedback effectiveness and explore further improvements through self-supervised fine-tuning.

Video/Multi-modal Models for Surgery. Most surgical computer vision research has focused on deep learning models for specific, highly visual tasks such as segmentation, tool identification, anatomic part recognition, and surgical scene reconstruction (Ramesh et al., 2023; Twinanda et al., 2016; Allan et al., 2020; Jin et al., 2020; Liu et al., 2024; Wang et al., 2022). However, these models have mainly been tested on limited, mono-centric, procedure-specific datasets that do not fully capture real-world surgical workflows (Ramesh et al., 2023; Schmidgall et al., 2024).

There has been limited progress in developing general or foundational surgical video models, as general models face challenges when applied to specific downstream tasks. For example, Schmidgall et al. (2024) introduced GSViT, a General Surgery Vision Transformer pre-trained on 680 hours of surgery videos using next-frame reconstruction. However, its focus on single-frame prediction limits its applicability for feedback prediction, where integrating information across multiple frames is key. Similarly, Yuan et al. (2023) proposed Surgical Vision Language Pre-training (SurgVLP) using 3k hours of surgical video lectures aligned with transcribed audio. While it outperformed non-surgical baselines, it underperformed compared to fully supervised methods on vision-specific surgical tasks, highlighting the difficulty of adapting general models to specialized tasks.

The only work exploring video use for surgical feedback tasks is Kocielnik et al. (2023), which developed

a method for automated feedback classification using the clinically validated system proposed by Wong et al. (2023). This approach utilized multi-modal inputs (text, audio, video) and fine-tuned a VideoMAE model pre-trained on human action data.

Although the tasks are similar, predicting feedback effectiveness is less visual compared to feedback classification, making it more challenging. Feedback categories such as "Visual Aid" and "Anatomic" involve distinct visual elements such as cues for directing the trainee's attention or the presence of anatomical landmarks (Wong et al., 2023). In contrast, understanding feedback effectiveness involves identifying unclear and variable visual cues (Kocielnik et al., 2023).

3. Methods

3.1. Data and Annotations

Data Acquisition We use a dataset containing real-time feedback delivered during live robot-assisted surgeries summarized in Table 1. The dataset comprises audio recordings captured using wireless microphones and video footage from an endoscopic camera, both of which were simultaneously recorded. Each surgery was conducted using the da Vinci Xi robotic surgical system (Freschi et al., 2013). This feedback data was previously collected and categorized by Wong et al. (2023).

A subset of the recorded OR conversations represents surgical feedback, defined by Wong et al. (2023) as trainer utterances intended to modify trainee behavior or thinking. Feedback instances were manually transcribed and timestamped by medical residents. The dataset encompasses 4204 distinct feedback instances across 33 surgeries.

For each feedback instance, the trainee's response was annotated into one of three categories: behavioral change, verbal acknowledgment, or request for clarification (Wong et al., 2023). Behavioral change was defined as an immediate adjustment made by the trainee in direct response to the feedback. Verbal acknowledgment involved the trainee confirming they had heard and understood the feedback. A request for clarification occurred when the trainee asked for the feedback to be restated due to a lack of understanding. Responses were annotated by three human raters using a shared codebook (Wong et al., 2023).

Given that the primary goal of surgical feedback is to influence trainee behavior (Wong et al., 2023),

Table 1: Statistics of video data and the annotations in our dataset per trainee response categories. Task-relevant SSL fine-tuning was performed on feedback video data, while domain-aware SSL fine-tuning was performed on all video data.

Category	Len/Count	% of Tot.
All Video	78.9 hr	100%
Feedback Video	11.7 hr	14.8%
Feedback Instances	4204	100%
Trainee Response		
Verbal Acknowledgment	1941	46.2%
Behavioral Change	1865	44.4%
Ask for Clarification	104	2.5%

we assess feedback effectiveness based on whether it results in trainee behavior change.

3.2. Model Pipeline

The model pipeline for predicting trainee behavior consists of four key steps, as shown in Figure 3. First, VideoMAE is pre-trained using self-supervised learning (SSL) on an unlabeled human action dataset (Kinetics-400). Next, we fine-tune VideoMAE with SSL on unlabeled surgical videos, using two strategies: task-relevant fine-tuning (using only feedback-related video) and domain-relevant fine-tuning (using all surgical video). After this, we perform supervised fine-tuning on labeled feedback videos to train the model specifically for predicting feedback effectiveness (trainee behavior change). Finally, we perform multi-modal supervised training which involves using both video features from VideoMAE and text features from SBERT to predict trainee behavior change.

3.3. Automated Speech Recognition

We employed Automated Speech Recognition (ASR) to generate automated transcriptions based on the starting timestamps of feedback instances (for a 10s duration). For this, we used the pre-trained Whisper medium model (Radford et al., 2022), which was trained on 680,000 hours of labeled English speech data. Given the conversational nature of feedback exchanges, we also applied speaker diarization using Pyannote (Bredin et al., 2020; Bredin and Laurent, 2021). Speaker diarization segments the audio

5

to identify different speakers, adding context to the dialogue around feedback delivery.

3.4. Self-Supervised Learning of VideoMAE

We used the VideoMAE base model, consisting of 86 million trainable parameters, pre-trained on Kinetics-400, a dataset comprising videos of 400 human action classes (Li et al., 2020). We investigated task-relevant and domain-relevant SSL fine-training strategies. Task-relevant SSL fine-tuning involved pre-training the model on video clips associated with feedback instances, which comprised approximately 12 hours of data. Domain-relevant SSL fine-tuning involved pre-training on all available surgical videos, which comprised approximately 80 hours of data.

3.5. Supervised Fine-tuning of VideoMAE

Following SSL fine-tuning, the VideoMAE model was fine-tuned using supervised learning to predict feedback effectiveness (trainee behavior change). For each feedback video, features were extracted from VideoMAE and averaged across the temporal and spatial dimension, producing a 768-dimensional vector. This vector was then passed through three fully connected linear layers, reducing the dimensionality to 512, 256, and finally 2 (logits) in a funnel-shaped architecture. ReLU activations were applied between layers to introduce non-linearity.

3.6. Supervised Training of Multi-modal Model

Finally, features were extracted from VideoMAE (for video) and SBERT (for text) and integrated through supervised multi-modal training to predict trainee behavior change. Features extracted from Video-MAE were averaged across the temporal and spatial dimension resulting in a 768-dimensional vector. Two fully connected layers with ReLU activations were applied, reducing the dimensions to 512 and then 256. Features extracted from SBERT, initially a 384-dimensional vector, were also processed through two fully connected layers with ReLU activations, reducing their dimensions to 128 and then 64. The resulting vectors were concatenated into a 320-dimensional vector, which was passed through three fully connected layers, reducing its dimensions to 256, 128, and finally 2 (the logits) in a funnel architecture. During multi-modal training, SBERT and VideoMAE weights remained frozen.

3.7. Data Processing

Following Kocielnik et al. (2023), we trimmed a 10-second video when human-annotated feedback appears. This includes 5 seconds before (to capture context) and 5 seconds after (to capture delivery) the feedback onset. We preprocessed each video by down-sampling the resolution to 320 x 250 and extracting 16 randomly uniformly sampled frames.

4. Experimental Results

4.1. Training and Implementation Details

Task-relevant SSL fine-tuning was conducted for 500 epochs, and domain-relevant for 1000 epochs, both using a batch size of 24 and an initial learning rate of 1.5×10^{-4}. A linear decay scheduler reduced the learning rate to zero over training, with a weight decay of 0.05 applied. Following Tong et al. (2022), which found that higher masking ratios result in better performance, an 85% masking ratio was used for the reconstruction task to improve performance.

Supervised fine-tuning of VideoMAE was conducted for 4 epochs with a batch size of 16, an initial learning rate of 1×10^{-4}, and a weight decay of 0.01. A ReduceLROnPlateau scheduler reduced the learning rate by 0.5 if validation accuracy didn't improve over two epochs. Multi-modal training was performed for another 4 epochs with a batch size of 16, a learning rate of 1×10^{-3}, and a weight decay of 0.01, with the learning rate managed similarly.

All training was executed on an NVIDIA GeForce RTX 4090 GPU.

4.2. Evaluation Protocols

The dataset was split 80% for training and 20% for testing. To address the class imbalance in trainee behavior change, the minority class was upsampled in both sets.

In task-relevant SSL fine-tuning, all feedback instances excluding those in the test set were used. In domain-relevant SSL fine-tuning, all available surgical video data, split into non-overlapping 10-second clips, was used.

For all methods, hyperparameters, including epochs, learning rate, weight decay, and layer dimensions, were fine-tuned using grid search. This tuning was performed on a validation set comprising 12.5% of the training data. Model performance was evaluated using Area Under the Receiver Operating

Method	AUROC	Precision	Recall
Text	$0.66_{\pm 0.004}$	$0.62_{\pm 0.02}$	$0.62_{\pm 0.13}$
Text + VideoMAE	$0.68_{\pm 0.01}$ ↑3.59%	$0.63_{\pm 0.01}$	$0.59_{\pm 0.10}$
Text + VideoMAE (Task-relevant)	$0.70_{\pm 0.02}$ ↑6.16%	$0.65_{\pm 0.03}$	$0.56_{\pm 0.15}$
Text + VideoMAE (Domain-relevant)	$0.70_{\pm 0.02}$ ↑6.55%	$0.63_{\pm 0.03}$	$0.66_{\pm 0.09}$

Table 2: Comparison of model performance for **predicting trainee behavior change** using auto-transcriptions of verbal feedback. The percentage improvement shown is relative to the text method. Values represent the mean and standard deviation across three splits with different seeds. The multi-modal methods, combining video and text features, outperformed the text-only method, showing the value of incorporating visual information for the prediction task.

Method	AUROC	Precision	Recall
SurgVLP	$0.52_{\pm 0.01}$	$0.55_{\pm 0.06}$	$0.42_{\pm 0.33}$
GSViT	$0.50_{\pm 0.00}$	$0.67_{\pm 0.29}$	$0.49_{\pm 0.44}$
VideoMAE	$0.58_{\pm 0.00}$	$0.55_{\pm 0.02}$	$0.58_{\pm 0.37}$
VideoMAE (Task-relevant)	$0.61_{\pm 0.01}$ ↑5.46%	$0.57_{\pm 0.04}$	$0.61_{\pm 0.11}$
VideoMAE (Domain-relevant)	$0.60_{\pm 0.01}$ ↑3.73%	$0.56_{\pm 0.02}$	$0.62_{\pm 0.22}$

Table 3: Comparison of video-only model performance for **predicting trainee behavior change** across SurgVLP, GSViT, and multiple VideoMAE pre-training methods. Values represent the mean and standard deviation across three splits with different seeds. SurgVLP and GSViT, baseline models pre-trained on large surgical datasets and fine-tuned for this task, were not predictive of behavior change. VideoMAE features were somewhat predictive, with SSL fine-tuning of VideoMAE improving performance by up to 5.5%.

Characteristic Curve (AUROC), with precision and recall metrics also reported. Methods were evaluated across three splits using distinct random seeds, with the mean and standard deviation reported.

4.3. Results

Table 2 shows the performance of text-only and multi-modal methods using auto-transcribed text for predicting trainee behavior change. The text-only method achieved an AUROC of 0.66. Incorporating video features (without SSL fine-tuning) helped improve performance to an AUROC of 0.68, a 3.59% improvement over the text baseline. With task-relevant and domain-relevant SSL pre-training, the AUROC improved to 0.70, reflecting gains of 6.16% and 6.55%, respectively. Both SSL fine-tuning approaches performed similarly, suggesting that using additional surgical video data did not yield significant benefits.

We further evaluated video-only methods to predict trainee behavior change as shown in Table 3. Methods compared included SurgVLP (Yuan et al., 2023), GSViT (Schmidgall et al., 2024), VideoMAE without SSL fine-tuning, and VideoMAE with task-relevant and domain-relevant SSL fine-tuning. SurgVLP and GSViT are existing surgery models pre-trained on large surgical datasets and serve as

baselines (implementation details are provided in Appendix C). Overall, video features are less predictive than text, with the best method achieving an AUROC of 0.61. SurgVLP and GSViT were not predictive, with AUROCs of 0.52 and 0.50, respectively. Task-relevant and domain-relevant SSL fine-tuning of VideoMAE improved AUROC by 5.46% and 3.73% over VideoMAE without fine-tuning.

4.4. Visual Inspection and Analysis

Appendices A and B provide examples of the multi-modal model's true positives, false positives, true negatives, and false negatives on the test data. Generally, in the cases where the model predicted a positive outcome, there was a noticeable change in the video during feedback delivery (at around 5 seconds) and after feedback delivery (at around 9 seconds). In contrast, in the cases where the model predicted a negative outcome, there was a minimal change.

False positive examples (Figure 4) displayed substantial movement in the video that did not correspond to the verbal feedback. For instance, in the case of "Careful, you're almost cutting it," the instruments were moving, but the movement was unrelated to the feedback content. This may have caused the model to make a false positive prediction.

False negative examples (Figure 5) displayed minimal movement, providing little visual evidence of a behavior change. For example, in the case of "You can clip that one," while the angle of the left instrument changed post-feedback, its position stayed the same. This lack of clear motion may have influenced the model to make a false negative prediction.

5. Discussion

Our findings reveal the value of using multi-modal learning to assess feedback effectiveness in real-world surgeries. Incorporating features from video alongside auto-transcriptions improved performance by up to 6.6%, resulting in an overall AUROC of 0.70.

Individual Predictiveness of Text and Video. We observe a significant difference in the individual predictiveness of the text and video modalities. The text-only model achieved an AUROC of 0.66, while the best video-only model achieved an AUROC of 0.61. This difference is likely because text is inherently more structured and less noisy compared to video data. The semantic meaning of the transcribed feedback is clearly relevant to the prediction task, whereas it is less clear which aspects of the video are relevant. Additionally, video data contains more noise (irrelevant information) due to its larger size and higher dimensionality, which makes it even more challenging to extract meaningful visual features for prediction with limited training data.

SSL Fine-tuning. We demonstrate that self-supervised fine-tuning of video models effectively addresses the challenges of extracting meaningful information from video data. SSL task-relevant and domain-relevant fine-tuning improved video-only performance by up to 5.5% and multi-modal performance by up to 6.6% for the prediction task. SSL fine-tuning also offers scalability due to its reliance on unlabeled data (Caron et al., 2020; Kaplan et al., 2020; Sorscher et al., 2022). As more surgical video data is collected, it can be incorporated into the pre-training process, further boosting prediction performance and improving model adaptability over time.

Interestingly, the task-relevant and domain-relevant fine-tuning methods performed comparably, suggesting that pre-training with additional video data beyond feedback instances was not significantly helpful. This may be because non-feedback video clips are (1) less likely to capture active surgical work that typically triggers feedback, and (2) less likely to show distinct visual changes seen in feedback clips, where trainee behavior often shifts visibly after feedback is delivered. So, the additional data might have been somewhat redundant or not as relevant (Sorscher et al., 2022). Although both methods achieve similar performance, they each have their advantages. Task-relevant SSL fine-tuning requires less data for the model to effectively learn task-relevant features (Sorscher et al., 2022). Domain-relevant SSL fine-tuning does not require knowledge of the task, eliminating the need to determine which video instances are task-relevant. It is also more flexible for training (Oquab et al., 2023) and the resulting model may easily generalize to other tasks (Pantazis et al., 2021; Tong et al., 2022).

Practical Value in Clinical Settings. The ability to automate prediction surgical feedback effectiveness has significant implications for improving the quality of surgical training in real-world surgeries. Compared to manual annotation, automated prediction is quick, standardized, and scalable (Wong et al., 2023). Trainers can review predictions post-surgery and use them to improve their feedback delivery and training quality (Ma et al., 2024). Accurate feedback effectiveness assessment is also useful for the future development of systems for automated feedback delivery during live surgery, where assessment of feedback options in real-time would be necessary (Kocielnik et al., 2023; Ma et al., 2024).

Limitations and Future Directions. Although we show that the visual features extracted from VideoMAE are useful for predicting feedback effectiveness, the incorporation of visual features alongside auto-transcribed text features provides limited additional benefit to prediction performance. Trainee behavior changes are directly observable through video, indicating that we have yet to fully leverage the richness and relevant information present in video data, which will be the focus of future work.

As a next step, while extracting general, abstracted visual features is valuable given the variability in visual cues, incorporating more structured visual information such as instrument kinematics and surgical gesture classification could further improve prediction performance. Another promising approach is to focus on identifying changes before and after feedback is delivered. For instance, contrastive learning could be applied to distinguish video segments pre- and post-feedback by creating positive pairs (related

to the same feedback) and negative pairs (unrelated segments).

6. Conclusion

From a clinical perspective, both the content of feedback and the visual context play a crucial role in determining the effectiveness of surgical feedback during live surgeries. In this work, we demonstrate the value of a multi-modal approach, combining transcribed feedback and surgical video data, for the novel task of predicting feedback effectiveness. By accurately predicting feedback effectiveness in a scalable and objective way, we can gain a deeper understanding of what constitutes effective feedback and the visual contexts in which feedback is most or least effective. Trainers can use this information to provide more effective, personalized guidance to trainees, enhancing surgical training. This understanding can also help pave the way for developing real-time systems that provide automated feedback during surgery. Ultimately, this work provides an important step towards advancing and automating surgical training in the operating room.

References

Riaz A Agha, Alexander J Fowler, and Nick Sevdalis. The role of non-technical skills in surgery. *Annals of medicine and surgery*, 4(4):422–427, 2015.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

Melissa Baralt. Coding qualitative data. *Research methods in second language acquisition: A practical guide*, pages 222–244, 2011.

Esther M Bonrath, Nicolas J Dedy, Lauren E Gordon, and Teodor P Grantcharov. Comprehensive surgical coaching enhances surgical skill in the operating room. *Annals of surgery*, 262(2):205–212, 2015.

Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Jenny Yun-Chen Chan and Michèle MM Mazzocco. Integrating qualitative and quantitative methods to develop a comprehensive coding manual: Measuring attention to mathematics in play contexts. *Methods in Psychology*, 4:100044, 2021.

Bonnie Chinh, Himanshu Zade, Abbas Ganji, and Cecilia Aragon. Ways of qualitative coding: A case study of four strategies for resolving disagreements. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–6, 2019.

Anne-Lise D D'Angelo, Andrew R Ruis, Wesley Collier, David Williamson Shaffer, and Carla M Pugh. Evaluating how residents talk and what it means for surgical performance in the simulation lab. *The American Journal of Surgery*, 220(1):37–43, 2020.

Cinzia Freschi, Vincenzo Ferrari, Franca Melfi, Mauro Ferrari, Franco Mosca, and Alfred Cuschieri. Technical review of the da vinci surgical telemanipulator. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 9(4):396–406, 2013.

Michael M Haglund, Andrew B Cutler, Alexander Suarez, Rajeev Dharmapurikar, Shivanand P Lad, and Katherine E McDaniel. The surgical autonomy program: a pilot study of social learning theory applied to competency-based neurosurgical education. *Neurosurgery*, 88(4):E345–E350, 2021.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Rafal Kocielnik, Elyssa Y Wong, Timothy N Chu, Lydia Lin, De-An Huang, Jiayun Wang, Anima Anandkumar, and Andrew J Hung. Deep multimodal fusion for surgical feedback classification. In *Machine Learning for Health (ML4H)*, pages 256–267. PMLR, 2023.

Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.

Runzhuo Ma, Ryan S Lee, Jessica H Nguyen, Andrew Cowan, Taseen F Haque, Jonathan You, Sidney I Roberts, Steven Cen, Anthony Jarc, Inderbir S Gill, et al. Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills—an unblinded pilot randomized controlled trial. *The Journal of Urology*, 208(2):414–424, 2022.

Runzhuo Ma, Dani Kiyasseh, Jasper A Laca, Rafal Kocielnik, Elyssa Y Wong, Timothy N Chu, Steven Cen, Cherine H Yang, Istabraq S Dalieh, Taseen F Haque, et al. Artificial intelligence-based video feedback to improve novice performance on robotic suturing skills: a pilot study. *Journal of Endourology*, 2024.

Syed Abdul Mateen, Niharika Malvia, Syed Abdul Khader, Danny Wang, Deepti Srinivasan, Chi-Fu Jeffrey Yang, Lana Schumacher, and Sandeep Manjanna. Thoracic surgery video analysis for surgical phase recognition. *arXiv preprint arXiv:2406.09185*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 10583–10592, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, et al. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL http://arxiv.org/abs/1908.10084.

Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023.

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.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pretraining. In *Advances in Neural Information Processing Systems*, 2022.

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*, 36(1):86–97, 2016.

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 431–441. Springer, 2022.

Elyssa Y Wong, Timothy N Chu, Runzhuo Ma, Istabraq S Dalieh, Cherine H Yang, Ashwin Ramaswamy, Luis G Medina, Rafal Kocielnik, Seyedeh-Sanam Ladi-Seyedian, Andrew Shtulman, et al. Development of a classification system for live surgical feedback. *JAMA Network Open*, 6(6): e2320702–e2320702, 2023.

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.

Appendix A. True and False Positives Visualizations
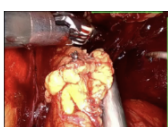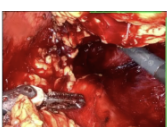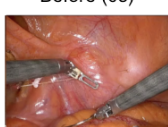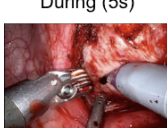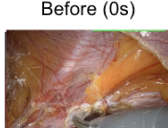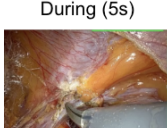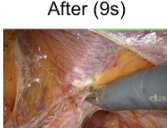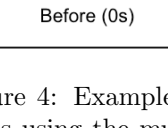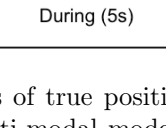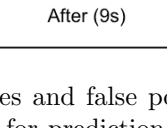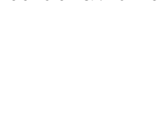
Video			Text (Auto Transcrip.)	Type
Before (0s)	During (5s)	After (9s)	"Here? No, no. All that's going to be hanging. You're not going to flip it down. Oh, the flap. Okay, I'm going to flip it down. All that is going to come down."	True Positive
Before (0s)	During (5s)	After (9s)	"Pull. Uh-huh, much better, thank you."	True Positive
Before (0s)	During (4s)	After (9s)	"So open up the easy space, uh-huh, good, just open all that up."	True Positive
Before (0s)	During (5s)	After (9s)	"Careful, you're almost cutting it, yep."	False Positive
Before (0s)	During (5s)	After (9s)	"Make sure it's free all around."	False Positive
Before (0s)	During (5s)	After (9s)	"Be careful, you have dp gas right over there, you saw it?"	False Positive

Figure 4: Examples of true positives and false positives using the multi-modal model for prediction of trainee behavior change.

Appendix B. True and False Negatives Visualizations
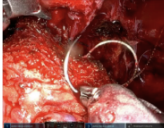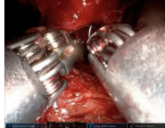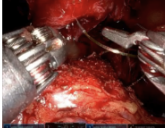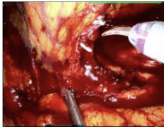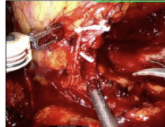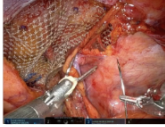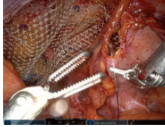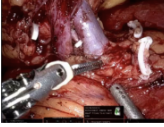
Video			Text (Auto Transcrip.)	Type
Before (0s)	During (5s)	After (9s)	"You need to be able to."	True Negative
Before (0s)	During (5s)	After (9s)	"I mean, we literally stabbed through the same vein in two different spots."	True Negative
Before (0s)	During (5s)	After (9s)	"That's good."	True Negative
Before (0s)	During (5s)	After (9s)	"I'm going to get the peritoneum first."	False Negative
Before (0s)	During (5s)	After (9s)	"You can clip that one."	False Negative
Before (0s)	During (5s)	After (9s)	"Okay. Robotic web, please"	False Negative

Figure 5: Examples of true negatives and false negatives using the multi-modal model for prediction of trainee behavior change.

Appendix C. Implementation Details of Baseline Video Models

SurgVLP and GSViT were both fine-tuned for the task of feedback effectiveness prediction. The details of their implementations and fine-tuning are discussed below.

SurgVLP: For each 10-second feedback video clip, 16 uniformly sampled frames were passed through the vision module of the SurgVLP model, producing a 16×768-dimensional output. This was averaged across the temporal dimension to generate a single 768-dimensional vector. A 2-layer multilayer perceptron (MLP) was then applied to obtain logits, and cross-entropy loss was minimized during training. The weights of the vision module were fine-tuned throughout the process. The model was trained for 5 epochs using the Adam optimizer with a learning rate of 1×10^{-4}.

GSViT: For each 10-second feedback video clip, 16 preprocessed uniformly sampled frames were passed through GSViT to obtain individual embeddings, which were flattened and concatenated. A 3-layer MLP was then applied to obtain logits, and cross-entropy loss was minimized during training. GSViT weights were fine-tuned throughout the process. The model was trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4}.

Appendix D. Segmented/Selective Manual Transcriptions

We investigated the performance of using manual transcriptions provided by the dataset for predicting feedback effectiveness. These manual transcriptions contain more information than simple verbatim transcriptions in two ways:

1. **Segmentation Based on Trainee Response:** Instead of transcribing a trainer's spoken feedback as a continuous, uninterrupted utterance, human annotators split the feedback into smaller segments. Each segment corresponds to a part of the feedback that individually resulted in a trainee response such as verbal acknowledgment (which is inferred through audio) or behavior change (which is inferred visually).

2. **Selective Transcription:** In some cases, human annotators only transcribe parts of the trainer's spoken utterance that they believe are directly associated with an observed trainee response.

Examples of manual transcriptions illustrating these factors are shown in Table 4. The extra judgment applied through segmentation and selective transcription makes manual transcriptions significantly more predictive of trainee behavior change, as shown in Table 5. Since approximately half of the transcribed feedback instances resulted in trainee behavior change (as shown in Table 1), the segmentation was partly informed by visual cues, meaning some video context is implicitly reflected in the manual transcriptions. This likely explains why the multi-modal approach offers only a modest performance improvement (up to 1.4%) over the text-only approach when using manual transcriptions. In contrast, the multi-modal approach yields a larger performance improvement of up to 6.6% when using auto-transcriptions.

Utterance	Manual Transcription
"That's not exactly it either. What I would suggest you do is I would suggest you grab this leading edge right there. Grab it. You're not doing right. No. No. This edge. The cut edge. Uh-huh. Grab it. Uh-huh. Good. Pull. Pull. Uh-huh. Open that up. But know where your vein is."	"that's not exactly it either."
	"what I would suggest you do is, is I would suggest you grab this leading edge right there, grab it."
	"this edge, the cut edge."
	"grab it."
	"good."
	"pull pull."
	"open that up."
	"wait, know where your vein is, know where your artery is."
"So take the... You're going to put only one clip to the artery. You're not going to cut the artery. Right. You want to be safe."	"only one clip to artery, wanna be safe."

Table 4: **Examples of manual transcriptions of feedback instances corresponding to continuous trainer utterances.** In the first example, the trainer's utterance is split into smaller segments, each resulting in a distinct trainee response (verbal acknowledgment or behavior change). In the second example, only the portions of the utterance directly relevant to the observed trainee response are transcribed (selective transcription).

Method	AUROC	Prec.	Recall
Text	$0.79_{0.04}$	$0.72_{0.03}$	$0.72_{0.11}$
Text + Video	$0.80_{0.03}$ ↑0.53%	$0.74_{0.03}$	$0.68_{0.06}$
Text + Video (Task-Aware)	$0.80_{0.03}$ ↑1.36%	$0.73_{0.04}$	$0.71_{0.02}$
Text + Video (Domain-Aware)	$0.80_{0.03}$ ↑1.22%	$0.73_{0.04}$	$0.74_{0.03}$

Table 5: Comparison of model performance for **predicting trainee behavior change** using **manually transcribed text**. Each model is evaluated based on AUROC, precision, and recall. The percentage improvement shown is relative to the text method. Values represent the mean and standard deviation (SD) across three splits with different seeds.