Best Paper

ML4H

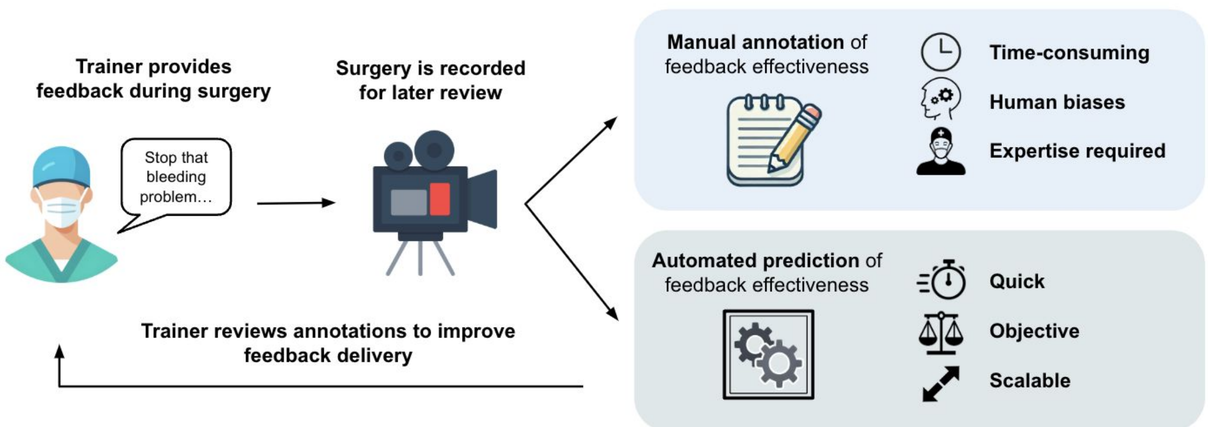
Multi-Modal Self-Supervised Learning for Surgical Feedback Effectiveness Assessment

Arushi Gupta*[1], Rafal Kocielnik*[1], Jiayun Wang[1], Firdavs Nasriddinov[1], Cherine Yang[2], Elyssa Wong[3], Anima Anandkumar[1], Andrew Hung[2]

[1]Caltech [2]Cedars-Sinai [3]USC

Motivation

- Real-time **surgical feedback** important for immediate **correction** and long-term **skill acquisition** [1]
- **Analyzing feedback effectiveness** crucial for improving surgical training but no automated approaches exist [2]
- Manual annotation **time** and **resource-demanding**

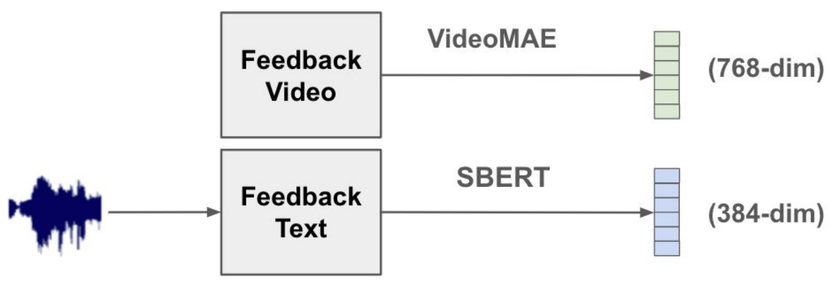


Dataset

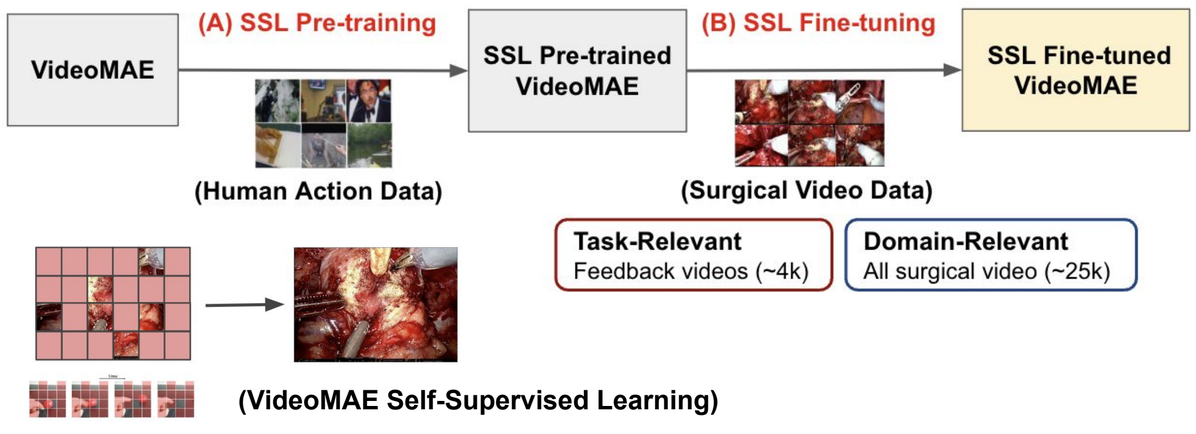
- Dataset of **4210** timestamped feedback instances from **live robot-assisted surgery**
  - **Trainer audio** and **surgical video**
- Manual annotations for whether feedback was effective (resulted in **trainee behavior change**)

Methods

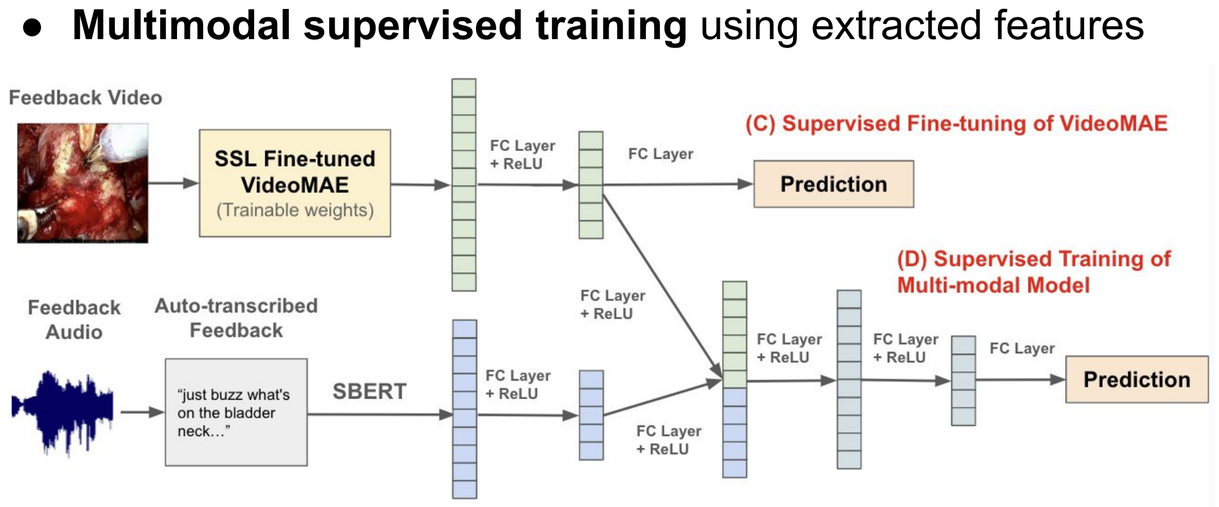
- Extract features using **VideoMAE** and **SBERT** [3, 4]



- Additional **fine-tuning of VideoMAE** on surgical video data



- **Multimodal supervised training** using extracted features



Results

Method	AUROC	Precision	Recall
VideoMAE	$0.58_{\pm 0.00}$	$0.55_{\pm 0.02}$	$0.58_{\pm 0.37}$
VideoMAE (Task-relevant)	$\mathbf{0.61}_{\pm 0.01}$ ↑5.46%	$0.57_{\pm 0.04}$	$0.61_{\pm 0.11}$
VideoMAE (Domain-relevant)	$0.60_{\pm 0.01}$ ↑3.73%	$0.56_{\pm 0.02}$	$0.62_{\pm 0.22}$

Performance of Video Models: Self-supervised fine-tuning of VideoMAE improves AUROC.

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)	$\mathbf{0.70}_{\pm 0.02}$ ↑6.55%	$0.63_{\pm 0.03}$	$0.66_{\pm 0.09}$

Performance of Text and Multimodal Models: Addition of video improves AUROC.

Conclusions

- Text and video **individually predictive** of feedback effectiveness; text is **more predictive** than video
- Adding video alongside text improves performance, but not significantly
- Task-relevant (using 14.8% of data) and domain-relevant fine-tuning of VideoMAE **perform similarly**
- Model has **practical use**: trainers can review predictions of model post-surgery to improve feedback delivery

Confidence	% of Instances	Accuracy
>90%	2.46%	87%
>85%	6.53%	80%
>80%	11.24%	76%
>75%	22.59%	72%
>70%	36.30%	70%

Model prediction accuracy at different **confidence score thresholds**

Future Work

- **Improve video component** of model
  - Extract specific, structured information
  - Leverage **contrastive learning** approaches directly comparing pre-/post-feedback
- Extract **general, interpretable insights** from model to improve practical use

[1] Wong et al., 2023. "Deconstructing and quantifying live surgical feedback in the operating room." *American Urological Association*.
[2] Agha et al., 2015. "The role of non-technical skills in surgery." *Annals of Medicine and Surgery*.
[3] Zhan et al., 2022. "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training." *NeurIPS*.
[4] Reimers and Gurevych, 2019. "Sentence-BERT: Sentence embeddings using siamese BERT-networks." *CoRR*.

Research reported in this publication was supported by the **National Cancer Institute of the National Institute of Health** under Award Numbers R01CA251579 and R01CA273031, and the **Northern California Associates SURF Fellowship**. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.