Automating Feedback Analysis in Surgical Training: Detection, Categorization, and Assessment

Firdavs Nasriddinov* **Rafal Kocielnik*** {FIRDAVS,RAFALKO}@CALTECH.EDU

Arushi Gupta AGUPTA5@CALTECH.EDU
California Institute of Technology, USA

Cherine Yang CHERINE.YANG@CSHS.ORG
Cedars-Sinai Medical Center, USA

Elyssa Wong 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

This work introduces the first framework for reconstructing surgical dialogue from unstructured real-world recordings, which is crucial for characterizing teaching tasks. In surgical training, the formative verbal feedback that trainers provide to trainees during live surgeries is crucial for ensuring safety, correcting behavior immediately, and facilitating long-term skill acquisition. However, analyzing and quantifying this feedback is challenging due to its unstructured and specialized nature. Automated systems are essential to manage these complexities at scale, allowing for the creation of structured datasets that enhance feedback analysis and improve surgical education. Our framework integrates voice activity detection, speaker diarization, and automated speech recognition, with a novel enhancement that 1) removes hallucinations (non-existent utterances generated during speech recognition fueled by noise in the operating room) and 2) separates speech from trainers and trainees using few-shot voice samples. These aspects are vital for reconstructing accurate surgical dialogues and understanding the roles of operating room participants. Using data from 33 real-world surgeries, we demonstrated the system's capability to reconstruct surgical teaching dialogues and detect feedback instances effectively (F1 score of 0.79±0.07). Moreover, our hallucination removal step improves feedback detection performance by ≈14%. Evaluation on downstream clinically relevant tasks of predicting Behavioral Adjustment of trainees and classifying Technical feedback, showed performances comparable to manual annotations with F1 scores of 0.82±0.03 and 0.81±0.03 respectively. These results highlight the effectiveness of our framework in supporting clinically relevant tasks and improving over manual methods.

Keywords: robot-assisted surgery, teaching, feedback, surgical assessment, automated speech recognition

Data and Code Availability Due to the nature of the data it will only be available on request. The code is publicly available on our github.

Institutional Review Board (IRB) The data used was collected under the IRB of the University of Southern California (HS-17-00113).

1. Introduction

Importance: Formative verbal feedback delivered by a trainer to a trainee during surgical procedures is crucial for immediate correction and guidance (Wong et al., 2023) as well as for fostering long-term skill acquisition (Agha et al., 2015). High-quality feedback has been shown to significantly enhance intraoperative performance (Bonrath et al., 2015), accelerate surgical skill development (Ma et al., 2022),

* These authors contributed equally

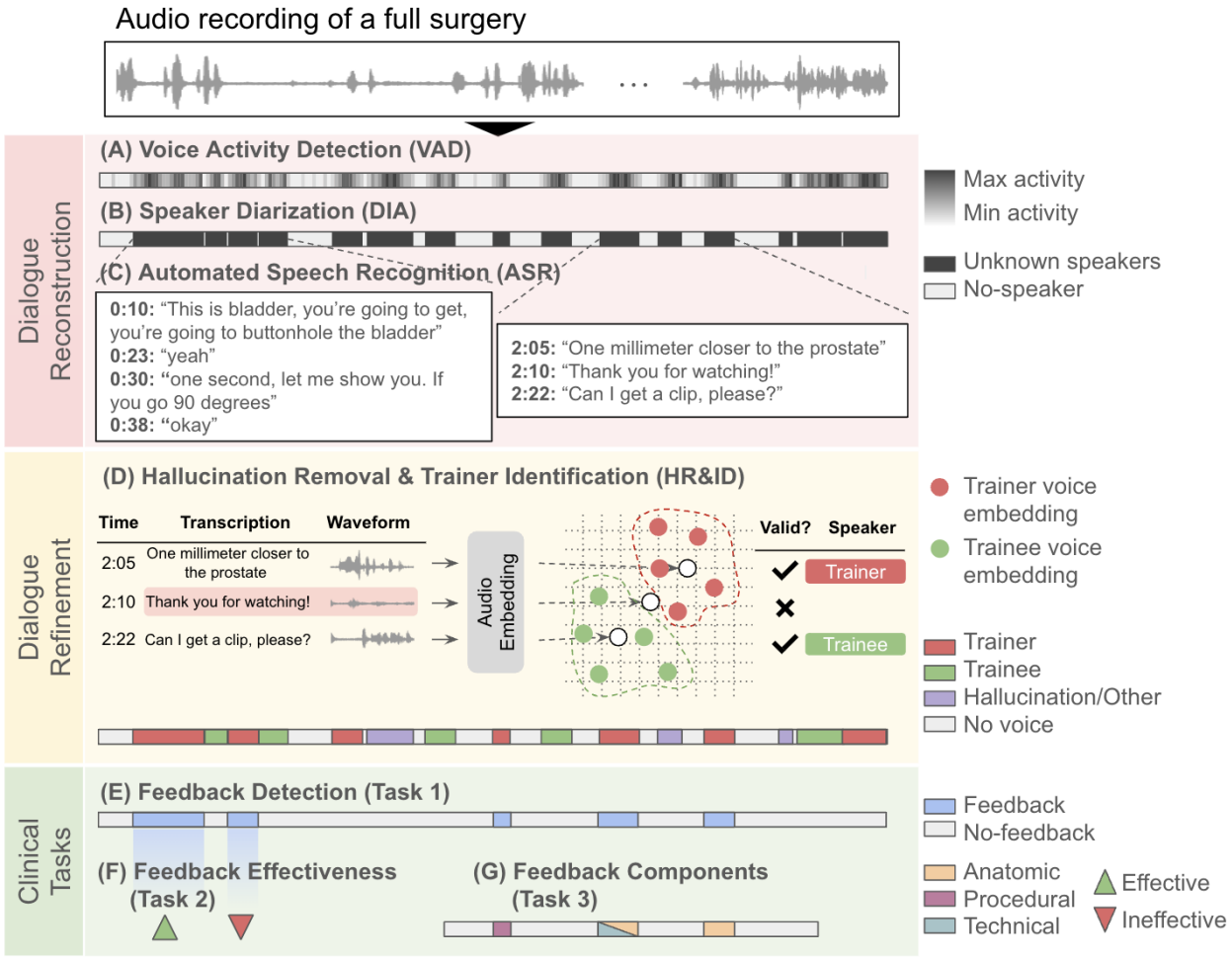

Figure 1: Overview of our automated surgical feedback detection and assessment framework, organized into: Dialogue Reconstruction, which integrates (A) Voice Activity Detection (VAD) to detect timespans of speech. (B) Speaker Diarization (DIA) to differentiate speech from different speakers. (C) Automated Speech Recognition (ASR) to transcribe the audio into text. Dialogue Refinement focuses on (D) Hallucination Removal & Trainer/Trainee Identification which ensures the correct identification of speakers and removes irrelevant audio snippets. Clinical Tasks step applied our framework to clinically relevant tasks from Wong et al. (2023), including (E) Feedback Detection (Task 1) which identifies when feedback is delivered. (F) Feedback Effectiveness (Task 2) which evaluates the impact and effectiveness of the feedback. (G) Feedback Components (Task 3) which categorizes the feedback into Anatomic, Procedural, and Technical.

and bolster trainees' sense of autonomy (Haglund et al., 2021). The quality of communication has also been directly linked to operative outcomes (D'Angelo et al., 2020). This underscores the importance of understanding current practices of feedback delivery on the quality of surgical education (El Boghdady and Alijani, 2017). Given the critical role of feedback in surgical training and patient outcomes, automating feedback analysis is essential. It can help streamline the understanding of current practices, enhance trainer-trainee communication, and reduce inconsistencies in manual analysis. By improving training methods and creating structured datasets, automation enables the development of standardized, cost-efficient guidance systems (Ma et al., 2024), ultimately elevating the quality of surgical education and patient care.

Challenges: Effectively quantifying and analyzing real-world surgical feedback at scale presents significant challenges. The contextual understanding of surgery, including its phases, specific procedures, and instructional tasks, necessitates deep domain expertise (Haque et al., 2024). Differentiating genuine feedback from mere comments or unrelated remarks requires a thorough familiarity with operating room interactions (Wong et al., 2023). Moreover, discerning which feedback is actionable and merits detailed annotation demands surgical teaching expertise. Additionally, the often extensive durations of surgeries coupled with the complex social dynamics among medical staff lead to many exchanges that do not qualify as clinically relevant feedback. Due to these challenges, previous studies have depended on labor-intensive manual identification and annotation of feedback by trained human raters (Wong et al., 2023; Hauge et al., 2001; Blom et al., 2007). Such manual annotation requires considerable time and effort from skilled annotators and is impractical to perform at scale, thereby hindering both the systematic quantification of feedback and the creation of structured, clinically-aligned datasets essential for developing automated feedback delivery systems (Laca et al., 2022).

Approach: We analyze full-length recordings of live surgical cases to automatically detect surgical feedback delivery moments based on the clinically validated definition of feedback from Wong et al. (2023). We further automatically categorize the content of this feedback and assess its effectiveness. We develop a *framework for reconstructing teaching interactions in the operating room* consisting of 3 consecutive steps as shown in Figure 1:

1. **Dialogue Reconstruction**: Combines Voice Activity Detection (VAD) to identify speech timespans in continuous audio, Speaker Diarization (DIA) to provide timespans for speech from different speakers, and Automated Speech Recognition (ASR) to transcribe audio within each timespan.
2. **Dialogue Refinement**: Removes ASR and DIA hallucinations and identifies trainer/trainee speaking turns by leveraging cosine similarity between embeddings of candidate speech segments and few-shot embedded voice samples from the trainer and trainee present in the operating room.
3. **Clinical Task Evaluation**: We demonstrate the crucial importance of each step on downstream clinically relevant tasks form Wong et al. (2023): *Feedback Detection*, *Feedback Effectiveness Assessment*, and *Feedback Components Identification*.

We perform experiments on recordings of 33 surgeries from prior work on feedback (Wong et al., 2023), which contains clinically validated human annotations of 4.2k feedback instances along with component and trainee behavior annotations (Table 1).

Findings:

- We show an ability to identify feedback utterances effectively with F1 of 0.79±0.07 (Table 2).
- Evaluation on the downstream clinically relevant tasks shows performance comparable to manual annotations: F1 of 0.82±0.03 for predicting *Behavioral Adjustment* of a trainee, F1 of 0.81±0.03 for classifying *Technical* feedback (Table 3).
- We show that our Hallucination Removal step improves performance on all tasks with the most performance gains in feedback detection with an increase of ≈14%.

Contributions:

- To the best of our knowledge, we are the first to attempt highly automated detection and assessment of surgical feedback from real-world teaching interactions in the operating room.
- We provide a robust evaluation of our framework on 33 real-world surgical cases involving unseen surgeries and clinically relevant downstream tasks.
- Aside from effectively combining existing voice tasks, we also introduce a novel step of Hallucination Removal & Speaker Identification (Figure 1), which leverages audio representations to identify trainers and filter out hallucinated text transcription fragments based on audio latent space.

2. Background and Related Work

Analysis of Feedback in Operating Room. Historical efforts in annotating surgical teaching interactions have primarily depended on manual observation. Hauge et al. (2001) and Blom et al. (2007) developed categorization schemes through manual annotation of recorded surgeries, involving hundreds of teaching behaviors. Ramprasad et al. (2024) manually transcribed 615 minutes of operating room interactions. Wong et al. (2023) analyzed 29 surgical videos and 3,711 interactions manually, establishing

Table 1: Statistics of the dataset obtained from Wong et al. (2023) containing full video and audio recordings of 33 surgeries along with available human annotations for different clinically relevant tasks which we support in our feedback detection framework.

Task	Dimension	Definition	Len/Count	% Pos
Dialogue	-	Verbal interactions in the operating room.	$78h\ 52m$	-
Feedback Detection	-	Any dialogue intended to modify trainee thinking or behavior.	4210	-
Component Classification	Anatomic	Familiarity with anatomic structures and landmarks.	1194	28.4%
	Procedural	Pertains to timing and sequence of surgical steps.	851	20.2%
	Technical	Performance of a discreet task with appropriate knowledge of exposure, instruments, and traction.	3489	82.9%
Feedback Effectiveness	Behavioral Adj.	Behavioral adjustment made by the trainee that corresponds directly with the preceding feedback	1866	44.3%
	Verbal Ack.	Verbal or audible confirmation by the trainee confirming that they have heard the feedback	1944	46.2%

a validated feedback classification and differentiating effective from ineffective feedback across various surgical stages and trainee experience levels. However, the scalability of these manual methods is limited, impacting the automation of teaching feedback systems and the generation of comprehensive, clinically relevant datasets needed for automation.

Recent efforts targeted the partial automation of feedback analysis in surgical training. Kocielnik et al. (2023) automated the feedback categorization scheme validated by Wong et al. (2023), using multimodal information. However, this automation was limited to the categorization phase, with the initial detection of feedback moments still dependent on manual annotations. Our work overcomes this limitation by introducing an automated system that detects feedback directly from raw recordings of surgical cases using validated definitions from Wong et al. (2023). We enhance the system's credibility by rigorously comparing its output to human annotations in clinically relevant tasks such as feedback categorization and trainee behavior change prediction.

Speech Reconstruction in Surgery. Research on speech recognition in healthcare has primarily targeted non-surgical settings (Schaaf et al., 2021; Corbisiero et al., 2023), leaving significant gaps in the surgical context. The complexity of the operating room (OR) poses unique challenges for automated speech analysis systems due to non-standardized communication, background noise (Hasfeldt et al., 2010), and the dynamic interplay of multiple roles including surgeons, nurses, and anesthesiologists (Blom et al., 2007; Gardezi et al., 2009).

In this setting, the hallucinations in transcribed speech, which can misrepresent critical verbal exchanges is a key issue (Koenecke et al., 2024; Kuhn et al., 2024; Hasfeldt et al., 2010). Magesh et al. (2024) reports that ASR hallucinations can affect 17% to 33% of content in specialized settings. Prior work tackled this through diverse methods: using multi-step verification processes (Taki et al., 2024), audio-visual alignment (Zhang et al., 2024), majority voting from multiple ASR runs (Koenecke et al., 2024), or directly predicting hallucinated outputs (Serai et al., 2022). In surgical settings, hallucinations manifest differently, best described as *"non-existent utterances generated during speech recognition fueled by noise in the operating room"* in line with Koenecke et al. (2024). These can occur across multiple steps of the dialogue reconstruction pipeline—voice activity detection, speaker diarization, and ASR transcription—often producing repetitive affirmations like "Yeah", "good", "thank you" rather than factual inaccuracies. Such errors are particularly problematic in the OR, where hallucinated critical verbal exchanges can be misrepresented as approvals from the trainer or trainee acknowledgments.

Our approach enhances dialogue reconstruction steps to address these challenges by filtering hallucinated responses using voice samples from the surgical team, improving both the accuracy and relevance of dialogue reconstruction in the OR. At the same

time, it provides critical identification of speaking turns from trainer and trainee, which is crucial for detecting clinically relevant feedback. Our method represents a targeted solution to overcome the specific difficulties of dialogue reconstruction in surgery.

3. Methods

3.1. Data Acquisition

Our study utilized a dataset of genuine feedback from trainers to trainees captured during real-world robot-assisted surgical procedures, detailed in Table 1 and obtained by Wong et al. (2023). The dataset covers multi-organ surgical contexts across 7 procedures, involving 4 trainers and 11 trainees. It has been rigorously annotated by 3 trained human raters. Feedback was recorded using wireless microphones on the surgeons and a video recorder capturing the surgeon's point-of-view through an endoscope camera, with all data synchronously captured using an external device and the da Vinci Xi surgical robot system (DiMaio et al., 2011). Each instance of feedback, totaling 4210, was timestamped and manually transcribed from the audio data.

3.2. Surgical Feedback Definition

Following the clinically validated definition by Wong et al. (2023), surgical feedback is *any dialogue intended to modify trainee thinking or behavior during a live surgery*. This feedback must be delivered in real-time by a trainer to a trainee who is actively operating the robotic console, allowing for immediate application and adjustment. The communication aims to influence the trainee's actions, decision-making, or understanding of the surgical task at hand and must be contextually relevant to the ongoing procedure. Social conversations and unrelated discussions within the operating room are not considered feedback.

3.3. Surgical Dialogue Reconstruction

The process of reconstructing surgical dialogues is outlined in Figure 1. Initially, the entire audio stream is divided into 3-minute chunks for individual processing. Each chunk is first processed using *Voice Activity Detection (VAD)* through python *py-webrtcvad* module (Wiseman, 2021), which assigns a value from 0 to 1 for every 10ms frame, where 0 indicates no activity and 1 maximum activity.

In addition to VAD, the raw audio undergoes *Speaker Diarization (DIA)* utilizing the speaker-diarization-3.1 model from *Pyannote* (Bredin, 2023; Plaquet and Bredin, 2023). This method identifies segments of audio containing speech and assigns random speaker IDs such as "SPEAKER 0." Each segment detected by DIA is cross-verified with the VAD output, and segments without significant VAD activity (below a threshold of 0.3) are discarded. This threshold was determined through empirical testing of various values (see Appendix B).

The remaining segments are then transcribed using *Automated Speech Recognition (ASR)* with the *Whisper-1* model (Radford et al., 2023), pre-trained on 680k hours of labeled English speech data for accurate transcription. This model was specifically fine-tuned for speech recognition tasks. Speech data was annotated using large-scale weak supervision.

3.4. Hallucination Removal & Trainer / Trainee Identification

The output of prior steps produces significant hallucinations due to background noise in the OR. Furthermore, the lack of knowledge about the role of the speaker hinders the detection of clinically meaningful guidance. To address both issues we introduced a custom dialogue refinement step.

We manually select and annotate a set of anchor audio segments for each trainer and trainee. These anchors are chosen to represent clear instances of each individual's voice. We identified at least 5 anchors per person, balancing the need for sufficient representation with practical constraints. The selection process involves visualizing speaking (or voice activity) times and choosing diverse segments across the surgery duration. We prioritize segments with minimal background noise and clear speech. The number of anchors (5+) was determined through empirical testing (Appendix L shows sufficient dissimilarity between separate speaker voices with the use of 5 anchors), to capture voice variations while remaining manageable in the applied context. These anchor segments are then embedded using *Pyannote*'s "embedding" model (Bredin et al., 2020; Coria et al., 2020) that is based on the x-vector TDNN-based architecture (Snyder et al., 2018) with 4.2M parameters.

Our method for hallucination removal and speaker role identification (Figure 1) processes each audio segment from the diarization step. We embed these segments using the same model as the anchors. For each segment, we compare its embedding to all anchor embeddings of the corresponding trainer and trainee us-

Technique	Data Processing	Classifier	F1-bin	Precision	Recall
Fixed-Window (baseline)	Voice Activity Detection (VAD)	-	$0.42_{\pm 0.20}$	$0.28_{\pm 0.17}$	$1.00_{\pm 0.00}$
	+ Audio	Wav2Vec2	$0.52_{\pm 0.11}$	$0.49_{\pm 0.19}$	$0.59_{\pm 0.11}$
	+ Text (ASR)	BERT	$0.59_{\pm 0.13}$	$0.55_{\pm 0.18}$	$0.66_{\pm 0.10}$
	+ Text (ASR)	GPT-4o	$0.60_{\pm 0.11}$	$0.60_{\pm 0.16}$	$0.62_{\pm 0.06}$
	+ Audio + Text (ASR)	Multimodal	$0.58_{\pm 0.13}$	$0.53_{\pm 0.19}$	$0.67_{\pm 0.07}$
Dialogue Reconstruction	Dialogue$^{\mp}$	GPT-4o	$0.58_{\pm 0.11}$	$0.64_{\pm 0.15}$	$0.55_{\pm 0.09}$
	+ Hallucination Rem. (baseline)	GPT-4o	$0.59_{\pm 0.07}$	$0.57_{\pm 0.07}$	$0.61_{\pm 0.08}$
	+ Hallucination Rem. **(our)**	GPT-4o	$0.66^{*}_{\pm 0.18}$	$0.65_{\pm 0.22}$	$0.71_{\pm 0.11}$
	+ Trainer/Trainee ID **(our)**	GPT-4o	$\mathbf{0.79}^{*}_{\pm 0.07}$	$\mathbf{0.76}_{\pm 0.12}$	$\mathbf{0.85}_{\pm 0.09}$

Table 2: Performance on **Feedback Detection**. Fixed-Window technique classifies on rolling 10-sec audio fragments through Temporal Event Detection. Acronyms: Voice Activity Detection (VAD) - used to detect audio containing speech. Automated Speech Recognition (ASR) - used to transcribe the audio of speech to text. Speaker Diarization (DIA) - used to separate the speech from different speakers as separate timespans. $^{\mp}$Dialogue is a combination of VAD, DIA, and ASR applied in sequence and grouped with the context of past utterances leading to feedback. Statistically significant gain compared to prior step in dialogue reconstruction at $^{*}p<0.05$, $^{\dagger}p<0.1$.

Data Processing	Feedback Effectiveness		Feedback Components		
	Beh. Adj.	Verb. Ack.	Anatomic	Procedural	Technical
Manual annotations	$0.78_{\pm 0.03}$	$0.63_{\pm 0.04}$	$0.64_{\pm 0.11}$	$0.46_{\pm 0.19}$	$0.78_{\pm 0.03}$
Dialogue	$0.80_{\pm 0.02}$	$0.61_{\pm 0.09}$	$\mathbf{0.69}_{\pm 0.09}$	$0.45_{\pm 0.18}$	$0.77_{\pm 0.03}$
$^{+}$ Hallucination Rem.	$\mathbf{0.82}^{\dagger}_{\pm 0.03}$	$\mathbf{0.66}^{*}_{\pm 0.06}$	$0.65_{\pm 0.09}$	$\mathbf{0.49}^{\dagger}_{\pm 0.17}$	$\mathbf{0.81}^{*}_{\pm 0.03}$
$^{+}$ Trainer/Trainee ID	$\mathbf{0.82}_{\pm 0.08}$	$0.64_{\pm 0.04}$	$0.66_{\pm 0.08}$	$0.46_{\pm 0.17}$	$\mathbf{0.81}_{\pm 0.03}$

Table 3: Performance on **Feedback Effectiveness Assessment** and **Feedback Component Classification** downstream clinically validated tasks (F1 binary). Manual annotation represents the human baseline provided in Wong et al. (2023). Human annotation involved ground truth labels for the tasks as well as limited transcription of the feedback itself, which was used as input for prediction. This manual transcription does not include dialogue leading to feedback and often only selected phrases in trainer feedback are transcribed. Automated transcription offers a more comprehensive transcript with dialogue context, which is also used as input for the tasks. Statistically significant gain compared to prior step in dialogue reconstruction at $^{*}p<0.05$, $^{\dagger}p<0.1$.

ing cosine similarity. We then calculate the average similarity for both trainer and trainee.

To identify hallucinations, we apply a threshold of 0.2 to both average similarities (see Appendix J for determining threshold). Segments falling below this threshold for both trainer and trainee are classified as hallucinations or other speakers and excluded. Otherwise, we assign the segment to either trainer or trainee based on the higher similarity score. This approach not only removes hallucinations, but also leverages domain knowledge to identify the critical roles of trainer and trainee, which are essential for understanding the surgical teaching context.

4. Experiments

We evaluate our method on several clinically relevant downstream tasks: *Feedback Detection (Task 1)*, *Feedback Effectiveness Assessment (Task 2)*, and *Feedback Component Classification (Task 3)*. We compare the performance of our approach to human annotator baselines as well as to other solutions.

We evaluate models on a test set including five unseen surgery cases covering feedback from different trainer-trainee pairs under different procedures. This test set is also representative of the whole dataset in terms of class distributions under all the tasks (see Table 5 in Appendix C). We compute binary recall, precision, and F1 score (Scikit, 2024) between true and predicted labels for each task.

We further statistically compare the impact of crucial *Hallucination Removal* and *Trainer/Trainee Identification* steps to the dialogue reconstruction without these steps using McNemar's non-parametric statistical test (Dietterich, 1998) further adapted to deep learning setups by Vanwinckelen and Blockeel (2012). We use a Python implementation of McNemar's test provided in Raschka (2018).

4.1. Task 1: Feedback Detection

This task involves identifying instances where the trainer provides feedback to the trainee operating the surgical console. Performance is evaluated as correctly labeling audio segments as feedback or not, compared to human annotations. Detecting feedback occurrences is crucial for assessing feedback quality. The annotation counts are in Table 1 under the "Feedback Detection" task. We introduce several baselines.

4.1.1. Baselines

Voice Activity Detection: This approach simply detects any speech and separates it from the non-speech sounds in the OR. We then use these speech times as predicted feedback instances and align them with the human expert annotations. The performance of this baseline indicates how much of the verbal interaction in the OR is related to clinically valid teaching feedback.

Fixed-Window Temporal Event Detection: We use a 10-sec moving window with 5-sec overlap, chosen based on average feedback length. VAD is applied as a first preprocessing step, which identifies speech with sub-second precision. For detected speech fragments, we apply *Automated Speech Recognition (ASR)* with *Whisper-1* (Radford et al., 2023) for text transcription. We then classify each window for feedback using Audio, Text, and Audio+Text late fusion models. We fine-tune Wav2Vec base (Baevski et al., 2020) for audio classification, and BERT base (Devlin et al., 2018) for text classification. For multimodal classification, we use late fusion, extracting 256-dimension vectors from audio and text, concatenating them, and passing through fully connected layers with ReLu activation and dropout.

For model training, we preprocess data using 10-second audio fragments (3 seconds before to account for any imperfections in annotating and 7 seconds after to capture feedback delivery). Audio is downsampled to 16kHz mono. We fine-tune each classifier (audio, text, audio+text) with 5 different IID dataset splits using a balanced set of feedback/no-feedback instances from each surgery case for training (due to class imbalance), which was further split into 80%/20% train/val sets, separate from the test set. All hyperparameters were selected based on performance on the validation split. Models are trained for 20 epochs with an initial learning rate of 5e-5, Adam optimizer, and linear LR reduction. The best model is determined by the highest binary F1 score. We use standard fine-tuning with all weights being trainable.

Hallucination Removal - Multiple ASR Runs: Several existing methods for ASR hallucination removal described in §2 address different types of hallucinations or rely on the presence of additional information not available in our context (e.g., aligned video modality, annotated hallucinations for direct prediction). One method we could compare to was proposed in Koenecke et al. (2024) and relies on running the ASR more than once and removing the transcriptions that differ as hallucinations.

4.1.2. Evaluation Setup

To assess the effectiveness of our dialogue reconstruction approach for feedback detection, we run our framework to obtain audio segments with transcriptions and speaker roles. GPT-4o classifies these for feedback presence (see prompt used in Appendix D), aligned with human annotations. We then apply a 5 second tolerance where we prompt GPT-4o to check if the phrase from the extracted dialogue and the human annotations have overlaps in transcriptions. This step is only needed for the alignment of ASR with human-transcribed instances for the purpose of evaluation. It is specifically relevant for situations where the beginning of the human-annotated transcription preceeds the ASR transcription segment, but part of the content of the human transcription is still within the ASR transcribed content.

4.2. Task 2: Feedback Effectiveness

This task evaluates how effectively the delivered feedback impacts subsequent trainee behavior. This is measured by human expert annotated observed behaviors of the trainee post feedback delivery in two categories of *Behavioral Adjustment* - *"Behavioral adjustment made by the trainee that corresponds directly with the preceding feedback"* and *Verbal Acknowledgment* - *"Verbal or audible confirmation from the trainee confirming that they heard the feedback"*.

4.2.1. Baselines

Human Selective Transcription: We leverage text transcription from human annotators available in our dataset with the same GPT-4o classifier using the same prompt but adapted for single phrases instead of a dialogue (see Appendix G). Human annotators provided transcriptions of only the trainer's feedback, without any conversational context. The details of the annotation and transcription process can be found in Wong et al. (2023).

4.2.2. Evaluation Setup

We prompt GPT-4o to predict Verbal Acknowledgement and Behavioral Change (see Appendix F). This evaluation is done on true positive feedback phrases obtained from Task 1. True labels come from aligned human annotations.

4.3. Task 3: Feedback Components

This task requires categorizing feedback into 3 clinically validated components based on Wong et al. (2023). These components represent feedback categorized as *Anatomic - "familiarity with anatomic structures and landmarks"*, *Procedural - "timing and sequence of surgical steps"*, and *Technical - "performance of a discrete task with appropriate knowledge of factors including exposure, instruments, and traction"*. This categorization is not mutually exclusive, as feedback instances can pertain to anatomic, procedural, and technical aspects simultaneously. The annotation counts for this task can be found in Table 1 under the component classification task.

4.3.1. Baselines

Human Selective Transcription: We leverage the same selective transcriptions as for Task 2 (See Appendix I for the GPT-4o prompt used in this task).

4.3.2. Evaluation Setup

We prompt GPT-4o to categorize feedback into Anatomic, Procedural, and/or Technical components in a multi-label fashion using the prompt in Appendix H. True labels come from aligned human annotations.

5. Results

We present results for Feedback Detection in Table 2 and for other downstream tasks of Feedback Effectiveness and Feedback Components in Table 3.

5.1. Task 1: Feedback Detection

Our results in Table 2 show the Precision, Recall, and F1 (Scikit, 2024) scores for detecting feedback using two techniques: Fixed-Window (baseline), and Dialogue Reconstruction. For the fixed-window, we see that the naive VAD-only classifier performed the worst with an F1 of $0.42_{\pm 0.20}$, indicating that not all speech in the operating room corresponds to feedback. Further, we see that the text and multimodal classifiers perform very similarly with F1's of $0.59_{\pm 0.13}$ and $0.58_{\pm 0.13}$, respectively, and outperform the audio classifier that achieves an F1 of $0.52_{\pm 0.11}$.

The results for the Dialogue row are obtained using initial dialogue reconstruction, and we see performance on par with best fixed-window classifiers. Our Hallucination Removal approach offers a big boost in performance, with an F1 of $0.66_{\pm 0.18}$, which is more effective than the Hallucination Removal approach from Koenecke et al. (2024) with F1 of $0.59_{\pm 0.07}$. Finally, identifying between Trainer and Trainee for each dialogue phrase achieves the highest F1 at $0.79_{\pm 0.07}$. Appendix K provides confusion matrices for various ablations of feedback detection.

High recall is more important than high precision, as it's more important to correctly detect all the feedback instances than to have the model be correct each time it determines a segment to be feedback. This can be corrected in a subsequent human verification step. Our method performed the best with a recall of $0.71_{\pm 0.11}$ after removing hallucinations and $0.85_{\pm 0.09}$ after identifying between Trainer and Trainee.

5.2. Task 2: Feedback Effectiveness

Table 3 shows results for predicting *Behavioral Adjustment* and *Verbal Acknowledgement* under Feedback Effectiveness. For this task, the baseline is the classifications done on the manual annotations that yield an F1 of $0.78_{\pm 0.03}$ and $0.63_{0.04}$ for *Beh. Adj.* and *Verb. Ack*, respectively. We see that the results for the initial dialogue reconstruction slightly improve for *Beh. Adj.* and slightly worsen for *Verb. Ack.*. Further, Hallucination Removal improves both metrics with F1 of $0.82_{\pm 0.03}$ for *Beh. Adj.* and $0.66_{\pm 0.06}$ for *Verb. Ack*. Finally, identifying Trainer/Trainee does not help further.

5.3. Task 3: Feedback Components

Table 3 also shows results for classifying feedback as *Anatomic*, *Procedural*, and/or *Technical* under Feedback Components. Similar to Feedback Effectiveness,

the baseline leverages manual annotations in classification. For classifying *Anatomic* feedback, the base Dialogue performs best at an F1 of $0.69_{\pm 0.09}$ while Hallucination Removal and Trainer/Trainee ID still outperform the manual annotations. For classifying *Procedural* feedback, Hallucination Removal yields the best results at $0.49_{0.17}$ with Trainer/Trainee ID performing on par with manual annotations. For classifying *Technical* feedback, we see that Hallucination Removal and Trainer/Trainee ID are equal and the highest with an F1 of $0.81_{\pm 0.03}$ for both.

6. Discussion

Our framework achieved F1 scores of $0.79_{\pm 0.07}$ in feedback detection and up to $0.82_{\pm 0.08}$ in analyzing effectiveness, demonstrating the high feasibility of automated surgical training analysis. We enable highly automated quantification and analysis of feedback in real surgeries, with implications for improving training practices and patient care. Given the wide coverage of procedures, tasks, anatomic settings, as well as trainer-trainee interactions in our dataset as in §3.1, our approach is likely to generalize to various educational clinical settings where guidance is performed via verbal feedback and dialogue. Our approach relies on data organization and preprocessing, rather than on a specific fine-tuned model, which should further reduce the reliance on a particular dataset.

Surpassing Human Annotation in Downstream Tasks Our automation surpassed human annotations in several downstream tasks, primarily because it captured more comprehensive contextual information. Unlike humans, who transcribed only feedback itself, our method transcribed the entire dialogue involving rounds of discussion leading to feedback, better reflecting the true nature of interactions in the operating room (Wong et al., 2023). This full transcription allowed for a more effective grouping of feedback instances, combining several human-labeled segments into a single, more informative one. Secondly, some of the human transcriptions used abbreviations and selectively transcribed parts of what actually has been said. This is due to the manual and cognitive effort associated with literal transcription, which further motivates the benefits of automation. Finally, the automated system applied consistent classification criteria across all samples, avoiding the biases and inconsistencies often introduced by human annotators (Kiyasseh et al., 2023).

Use of Few-shot Speech Samples Our dialogue refinement step using speech samples from known speakers enhances feedback detection but requires collecting clean speech before surgery and assumes consistent vocal traits during the actual procedure. This approach is effective for a fixed group of speakers. For unknown speakers or uncollectable samples, options like unsupervised speaker diarization (Xylogiannis et al., 2024) or role-based speaker recognition (Bellagha and Zrigui, 2020) offer alternatives, dependent on the surgical setting's constraints. Although our method improves accuracy, it demands thorough consideration of these implementation factors.

Limitations and Future Directions Our research identifies key areas for future exploration. First, enhancing feedback analysis through the integration of visual data from surgeries could link verbal feedback to specific surgical actions, enriching the context significantly (Kocielnik et al., 2023). Second, adapting our methods for real-time feedback during surgeries would not only improve teaching assessments but also help in documenting educational elements in a live setting (Akbari et al., 2023). Lastly, investigating how feedback patterns evolve over time and developing methodologies to track and improve feedback delivery could provide deeper insights into the effectiveness of surgical training and pedagogical evolution (Ma et al., 2021).

7. Conclusion

This work introduced a novel automated method for feedback detection in surgical education, utilizing dialogue reconstruction, hallucination removal, and speaker identification. Our method has shown robust performance, even surpassing human annotation. This scalable system promises to improve educational strategies and patient outcomes, marking a significant advancement in the automation of surgical education analysis. It sets the stage for future developments in real-time analysis and automated feedback delivery systems, with broad implications for healthcare training and patient care.

Acknowledgments

Research supported by NCI NIH under Award Numbers R01CA251579 and R01CA273031, and the Ron Sven Rat and Bfield SURF Fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

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.

Leila Akbari, Masoud Bahrami, and Akram Aarabi. Development and implementation of an intraoperative documentation protocol for enhancing patient safety in the operating room: A mixed methods protocol study. *Journal of Education and Health Promotion*, 12(1):279, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Mohamed Lazhar Bellagha and Mounir Zrigui. Speaker naming in tv programs based on speaker role recognition. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE, 2020.

EM Blom, EGG Verdaasdonk, LPS Stassen, HG Stassen, PA Wieringa, and J Dankelman. Analysis of verbal communication during teaching in the operating room and the potentials for surgical training. *Surgical endoscopy*, 21:1560–1566, 2007.

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, 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.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

Eugenio Corbisiero, Gennaro Costagliola, Mattia De Rosa, Vittorio Fuccella, Alfonso Piscitelli, and Parinaz Tabari. Speech recognition in healthcare: A comparison of different speech recognition input interactions. In *International KES Conference on Innovation in Medicine and Healthcare*, pages 142–152. Springer, 2023.

Juan M. Coria, Hervé Bredin, Sahar Ghannay, and Sophie Rosset. A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification. In Luis Espinosa-Anke, Carlos Martín-Vide, and Irena Spasić, editors, *Statistical Language and Speech Processing*, pages 137–148. Springer International Publishing, 2020. ISBN 978-3-030-59430-5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Simon DiMaio, Mike Hanuschik, and Usha Kreaden. The da vinci surgical system. *Surgical robotics: systems applications and visions*, pages 199–217, 2011.

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.

Michael El Boghdady and Afshin Alijani. Feedback in surgical education. *the surgeon*, 15(2):98–103, 2017.

Fauzia Gardezi, Lorelei Lingard, Sherry Espin, Sarah Whyte, Beverley Orser, and G Ross Baker. Silence, power and communication in the operating room. *Journal of advanced nursing*, 65(7):1390–1399, 2009.

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.

Taseen F Haque, J Everett Knudsen, Jonathan You, Alvin Hui, Hooman Djaladat, Runzhuo Ma, Steven Cen, Mitchell Goldenberg, and Andrew J Hung.

Competency in robotic surgery: Standard setting for robotic suturing using objective assessment and expert evaluation. *Journal of Surgical Education*, 81(3):422–430, 2024.

Dorthe Hasfeldt, Eva Laerkner, and Regner Birkelund. Noise in the operating room—what do we know? a review of the literature. *Journal of PeriAnesthesia Nursing*, 25(6):380–386, 2010.

Linnea S Hauge, Jeanne A Wanzek, and Constantine Godellas. The reliability of an instrument for identifying and quantifying surgeons' teaching in the operating room. *The American journal of surgery*, 181(4):333–337, 2001.

Dani Kiyasseh, Jasper Laca, Taseen F Haque, Brian J Miles, Christian Wagner, Daniel A Donoho, Animashree Anandkumar, and Andrew J Hung. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Communications Medicine*, 3(1):42, 2023.

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.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681, 2024.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23, 2024.

Jasper A Laca, Rafal Kocielnik, Jessica H Nguyen, Jonathan You, Ryan Tsang, Elyssa Y Wong, Andrew Shtulman, Anima Anandkumar, and Andrew J Hung. Using real-time feedback to improve surgical performance on a robotic tissue dissection task. *European Urology Open Science*, 46:15–21, 2022.

Runzhuo Ma, Sharath Reddy, Erik B Vanstrum, and Andrew J Hung. Innovations in urologic surgical training. *Current urology reports*, 22:1–13, 2021.

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.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*, 2024.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Aarya Ramprasad, Imaima Casubhoy, Austin Bachar, Melanie Meister, Brenda Bethman, and Gary Sutkin. Language in the teaching operating room: Expressing confidence versus community. *Journal of Surgical Education*, 81(4):556–563, 2024.

Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018. doi: 10.21105/joss.00638. URL https://joss.theoj.org/papers/10.21105/joss.00638.

Thomas Schaaf, Longxiang Zhang, Alireza Bayestehtashk, Mark Fuhs, Shahid Durrani, Susanne Burger, Monika Woszczyna, and Thomas Polzin. Are you dictating to me? detecting embedded dictations in doctor-patient conversations. In *2021*

IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 587–593. IEEE, 2021.

Scikit. 3.4. metrics and scoring: quantifying the quality of predictions — scikit-learn 1.5.1 documentation. https://scikit-learn.org/stable/modules/model_evaluation.html, 7 2024. (Accessed on 09/05/2024).

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900, 2022.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition, 2018.

SM Taki, Showmick Kar, Soumik Deb Niloy, Mazharul Islam Rakib, and Abdullah Al Nahid Biswas. *Mitigation of hallucination and interpretations of self attention of Mistral 7B AI to analyze and visualize context understanding ability of large language models*. PhD thesis, Brac University, 2024.

Gitte Vanwinckelen and Hendrik Blockeel. On estimating model accuracy with repeated cross-validation. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, pages 39–44, 2012.

John Wiseman. wiseman/py-webrtcvad: Python interface to the webrtc voice activity detector. https://github.com/wiseman/py-webrtcvad, Jan 2021. (Accessed on 09/02/2024).

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.

Paris Xylogiannis, Nikolaos Vryzas, Lazaros Vrysis, and Charalampos Dimoulas. Multisensory fusion for unsupervised spatiotemporal speaker diarization. *Sensors*, 24(13):4229, 2024.

Fang Zhang, Yongxin Zhu, Xiangxiang Wang, Huang Chen, Xing Sun, and Linli Xu. Visual hallucination elevates speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19542–19550, 2024.

. .

Appendix A. Automated and Manual Feedback Alignment Example

Context	Phrase	Pred Label	Aligned Human Annotations	True Label
00:37:58 Trainer: "That's good, I like that.", **00:38:12 Trainer**: "Make sure you stay in the same plane, more or less R.", **00:38:20 Trainer**: "And about there you can start doing a little mild smile towards the prostate.", **00:39:11 Trainer**: "So open up wide. This is what I mean by not digging yourself in the hole. Okay, I want all the lateral stuff opened up or dropped.", **00:39:25 Trainer**: "The reason for that is it gives you a sense of the prostate's contour as opposed to this distorted thing where you have...",	**00:39:44 Trainer**: "So even more than what you've actually done, I would have just buzzed that right there when you can see it, whatever the, yeah, butternut thing, uh-huh."	TRUE	**00:39:44**: "so even more than what you've actually done... I would've...", **00:39:48**: "just buzz that right there when you can see it... whatever the... yeah bladder neck thing"	TRUE
00:38:12 Trainer: "Make sure you stay in the same plane, more or less R.", **00:38:20 Trainer**: "And about there you can start doing a little mild smile towards the prostate.", **00:39:11 Trainer**: "So open up wide. This is what I mean by not digging yourself in the hole. Okay, I want all the lateral stuff opened up or dropped.", **00:39:25 Trainer**: "The reason for that is it gives you a sense of the prostate's contour as opposed to this distorted thing where you have...", **00:39:44 Trainer**: "So even more than what you've actually done, I would have just buzzed that right there when you can see it, whatever the, yeah, butternut thing, uh-huh.",	**00:40:14 Trainee**: "Yeah. Sorry, I wasn't aware of that. No problem."	FALSE		FALSE

Appendix B. VAD Threshold Experiment

VAD Thresh	Precision	Recall	F1-bin
0	$0.32_{\pm 0.17}$	$0.67_{\pm 0.09}$	$0.41_{\pm 0.16}$
0.1	$0.37_{\pm 0.18}$	$0.67_{\pm 0.09}$	$0.45_{\pm 0.15}$
0.3	$0.43_{\pm 0.17}$	$0.66_{\pm 0.10}$	$\mathbf{0.50}_{\pm 0.13}$
0.5	$0.39_{\pm 0.18}$	$0.65_{\pm 0.07}$	$0.49_{\pm 0.12}$

Table 4: Fixed-length temporal event detection metrics for varying VAD thresholds averaged across results on validation set using text binary classifier models.

Appendix C. Label Distributions for 5 Unseen Test Set Surgeries

Table 5 summarizes the class distribution for the test set of 5 unseen surgeries across all the tasks. The test set has been chosen to cover feedback from different trainer-trainee pairs under different procedures and represents 22.8% of the data (961 of 4210 total instances). The label distributions (%) are representative of the whole dataset (see Table 1).

Appendix D. GPT-4o Feedback Detection Prompts

System Prompt: You are a binary classifier that determines whether a given phrase contains delivery of feedback from a trainer to a trainee where the trainee is conducting urology surgery using the da Vinci robot. The dialogue is between two speakers, a trainer and a trainee. There are multiple turns in the dialogue where the same speaker can go back to back because a piece of dialogue from the other speaker might not have been picked up or because the other speaker didn't speak as much (usually the trainer speaks more than the trainee). There can be 6 types of feedback:

1. Anatomic: familiarity with anatomic structures and landmarks. i.e. 'Stay in the correct plane, between the 2 fascial layers.'

2. Procedure: pertains to timing and sequence of surgical steps. i.e. 'You can switch to the left side now.'

3. Technical: performance of a discrete task with appropriate knowledge of factors including exposure, instruments, and traction. i.e. 'Buzz it.'

4. Praise: a positive remark. i.e. 'Good job.'

5. Criticism: a negative remark. i.e. 'It should never be like this.'

User Prompt: Classify whether the following phrase contains the delivery of feedback considering the given context of the last couple turns in the dialogue where the phrase is the last entry in the context.

Format your response as follows. DO NOT DO ANY OTHER FORMATTING.:

{'feedback': 'yes'} if the dialogue contains feedback

{'feedback': 'no'} if the dialogue does not contain feedback

Context:

⟨ context string ⟩

Phrase:

⟨ phrase string ⟩

For example:

{'feedback': 'yes'}

Appendix E. GPT-4o Feedback Prediction Alignment Prompt

System Prompt: You are a binary classifier that determines whether two strings have any alignment or not. An alignment means that the two strings might have some common words or phrases that align with each other in terms of their order and/or meaning.

User Prompt: Classify whether the following two strings have any alignment or not.

Format your response as follows. DO NOT DO ANY OTHER FORMATTING.:

{'alignment': 'yes'} if the two strings have any alignment

{'alignment': 'no'} if the two strings do not have any alignment

For example: { 'alignment': 'yes' }

String 1:

⟨ phrase ⟩

String 2:

⟨ human annotation ⟩

Task	Dimension	Count	% Pos
Feedback Detection	Instances	961	
Component Classification	Anatomic	290	30.2%
	Procedural	194	20.2%
	Technical	734	76.4%
Feedback Effectiveness	Behavioral Adjustment	517	53.8%
	Verbal Acknowledgment	415	43.2%

Table 5: Summary of label distribution for the test set of 5 unseen surgeries across different tasks.

Appendix F. GPT-4o Feedback Effectiveness Prediction From Auto Dialogue Prompt

System Prompt: You are an AI assistant specializing in predicting trainee responses during urology surgery training using the da Vinci robot. Your task is to analyze dialogue between a trainer and a trainee, focusing on the trainee's reactions to feedback. The dialogue may contain multiple consecutive turns by the same speaker due to missed responses or varying speech patterns.

You will categorize potential trainee responses into two types:

1. Verbal Acknowledgement: This includes any verbal or audible confirmation from the trainee indicating they have heard and understood the feedback. Examples include: - "Okay, I see" - "Uh-huh, got it" - "Understood" - "Yes, I'll do that"

2. Behavioral Change: This refers to any physical or observable adjustment made by the trainee that directly corresponds to the feedback received. For example: - If the trainer suggests tightening a suture, the trainee immediately pulls the suture thread more tightly.

Your role is to predict which type(s) of response the trainee is likely to give based on the specific feedback provided by the trainer. Consider the context of the surgical procedure and the nature of the feedback when making your predictions.

User Prompt: Classify whether the following feedback phrase will lead to a trainee response, where a trainee response can be either 1) verbal acknowledgement, 2) behavioral change.

Context: ⟨ context string⟩

Phrase: ⟨ phrase string⟩

Format your response as follows. DO NOT DO ANY OTHER FORMATTING.:

'verbal acknowledgement': 'yes' if you predict the trainee to respond with a verbal acknowledgement otherwise 'no'

'behavioral change': 'yes' if you predict the trainee to respond with a behavioral change otherwise 'no'

Your output can be a combination of the two categories. For example:

{ 'verbal acknowledgement': 'yes', 'behavioral change': 'no' }

Appendix G. GPT-4o Feedback Effectiveness Prediction From Human Annotations Prompt

System Prompt: You are an AI assistant specializing in predicting trainee responses during urology surgery training using the da Vinci robot. Your task is to analyze feedback from a trainer surgeon to a trainee surgeon, focusing on the trainee's reactions to feedback.

You will categorize potential trainee responses into two types:

1. Verbal Acknowledgement: This includes any verbal or audible confirmation from the trainee indicating they have heard and understood the feedback. Examples include: - "Okay, I see" - "Uh-huh, got it" - "Understood" - "Yes, I'll do that"

2. Behavioral Change: This refers to any physical or observable adjustment made by the trainee that directly corresponds to the feedback received. For example: - If the trainer suggests tightening a suture, the trainee immediately pulls the suture thread more tightly.

Your role is to predict which type(s) of response the trainee is likely to give based on the specific feedback provided by the trainer. Consider the context of the surgical procedure and the nature of the feedback when making your predictions.

User Prompt: Classify whether the following feedback phrase will lead to a trainee response, where a trainee response can be either 1) verbal acknowledgement, 2) behavioral change.

Feedback: ⟨ human annotation string⟩

Format your response as follows. DO NOT DO ANY OTHER FORMATTING.:

'verbal acknowledgement': 'yes' if you predict the trainee to respond with a verbal acknowledgement, otherwise 'no'

'behavioral change': 'yes' if you predict the trainee to respond with a behavioral change otherwise 'no'

Your output can be a combination of the two categories. For example:

{ 'verbal acknowledgement': 'yes', 'behavioral change': 'no' }

Appendix H. GPT-4o Feedback Component Classification From Auto Dialogue Prompt

System Prompt: You are an AI assistant specializing in classifying feedback during urology surgery training using the da Vinci robot. Your task is to analyze dialogue between a trainer and a trainee, focusing on categorizing the feedback into anatomic, procedural, and/or technical. The dialogue may contain multiple consecutive turns by the same speaker due to missed responses or varying speech patterns.

You will categorize the feedback into three types:

1. Anatomic: Familiarity with anatomic structures and landmarks. Examples include: - "Stay in the correct plane, between the 2 fascial layers." - "Avoid the blood vessels here."

2. Procedural: Pertains to the timing and sequence of surgical steps. Examples include: - "You need to suture this area first." - "You can switch to the left side now."

3. Technical: Performance of a discrete task with appropriate knowledge of factors including exposure, instruments, and traction. Examples include: - "Adjust the tension on the suture." - "Buzz it."

Your role is to predict which type(s) of feedback the phrase contains based on the specific feedback provided by the trainer. Consider the context of the surgical procedure and the nature of the feedback when making your predictions.

User Prompt: Classify the feedback phrase into one or more of the following categories: 1) anatomic, 2) procedural, 3) technical. Do this while considering the given context of the last couple turns in the dialogue where the phrase is the last entry in the context.

Context: ⟨ context string ⟩

Phrase: ⟨ phrase string ⟩

Format your response as follows. DO NOT DO ANY OTHER FORMATTING.:

'anatomic': 'yes' if the feedback is anatomic otherwise 'no'

'procedural': 'yes' if the feedback is procedural otherwise 'no'

'technical': 'yes' if the feedback is technical otherwise 'no'

Your output can be a combination of the three categories. For example:

{ 'anatomic': 'yes', 'procedural': 'no', 'technical': 'yes', }

Appendix I. GPT-4o Feedback Component Classification From Human Annotations Prompt

System Prompt: You are an AI assistant specializing in classifying feedback during urology surgery training using the da Vinci robot. Your task is to analyze feedback from a trainer to a trainee, focusing on categorizing the feedback into anatomic, procedural, and/or technical.

You will categorize the feedback into three types:

1. Anatomic: Familiarity with anatomic structures and landmarks. Examples include: - "Stay in the correct plane, between the 2 fascial layers." - "Avoid the blood vessels here."

2. Procedural: Pertains to the timing and sequence of surgical steps. Examples include: - "You need to suture this area first." - "You can switch to the left side now."

3. Technical: Performance of a discrete task with appropriate knowledge of factors including exposure, instruments, and traction. Examples include: - "Adjust the tension on the suture." - "Buzz it."

Your role is to predict which type(s) of feedback the phrase contains based on the specific feedback provided by the trainer. Consider the context of the surgical procedure and the nature of the feedback when making your predictions.

User Prompt: Classify the feedback phrase into one or more of the following categories: 1) anatomic, 2) procedural, 3) technical. Do this while considering

the given context of the last couple of turns in the dialogue where the phrase is the last entry in the context.

Feedback: ⟨ human annotation string ⟩

Format your response as follows. DO NOT DO ANY OTHER FORMATTING.:

'anatomic': 'yes' if the feedback is anatomic otherwise 'no'

'procedural': 'yes' if the feedback is procedural otherwise 'no'

'technical': 'yes' if the feedback is technical otherwise 'no'

Your output can be a combination of the three categories. For example:

{ 'anatomic': 'yes', 'procedural': 'no', 'technical': 'yes', }

Appendix J. Cosine Similarity Threshold Experiment

Sim. Thresh	Prec.	Recall	Prec.-Leaning Mean
0	0.085	0.507	0.423
0.1	0.077	0.823	0.566
0.2	0.063	0.933	**0.593**
0.3	0.053	0.947	0.579
0.4	0.046	0.973	0.579
0.5	0.044	1.000	0.588

Table 6: Metrics for classifying segments as trivial hallucinations. Precision-Leaning Mean is calculated by $2 \times \text{Precision} + \frac{\text{Recall}}{2}$.

Trivial hallucinations are found using a method described in Koenecke et al. (2024) where ASR is run twice on the same phrase and if the outputs are different then it is considered as a trivial hallucination. Having this "true label" hallucinations dataset, we apply the different cosine-similarity thresholds in the Hallucination Removal step that filters out hallucinations by thresholding the cosine similarity between audio segment and trainer and trainee voice samples. Note that we prioritize precision more than recall because a high recall does imply we pick up all hallucinations but it also means that all of those phrases do not get considered to have feedback since they would be classified as hallucinations. Further, the identification process almost always identifies empty phrases as unknown and hallucinations

Appendix K. Dialogue Reconstruction Feedback Detection Confusion Matrices

	Pred: False	Pred: True
Label: False	2144	358
Label: True	474	613

Table 7: Dialogue (off-the-shelf dialogue reconstruction) feedback detection confusion matrix.

	Pred: False	Pred: True
Label: False	502	197
Label: True	180	500

Table 8: Hallucination Removal (our) dialogue refinement feedback detection confusion matrix.

	Pred: False	Pred: True
Label: False	656	115
Label: True	115	493

Table 9: Trainer/Trainee ID (our) dialogue refinement feedback detection confusion matrix.

Appendix L. Example Cosine Similarities of Embeddings

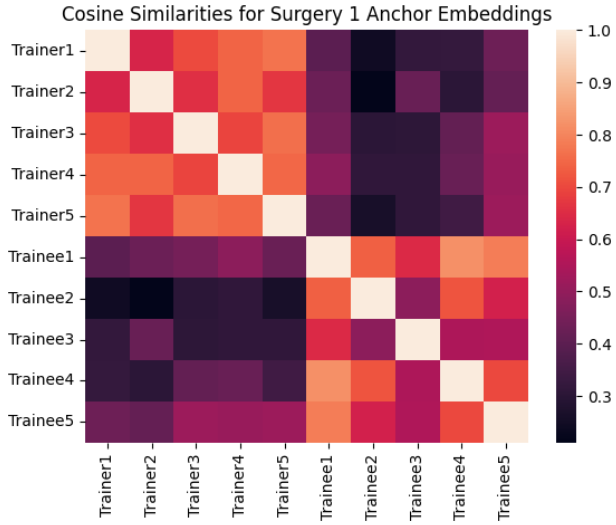

Figure 2: Cosine similarity between trainer and trainee anchor embeddings for surgery 1. Trainer1, ..., Trainer5 refer to the different audio examples for the same trainer and likewise for Trainee1, ..., Trainee 5.

Appendix M. Detailed Temporal Event Detection

This approach relies on a moving fixed-length window of 10 seconds length with 5 sec overlap between the windows. These settings have been chosen empirically based on the observed average length of feedback. Note that feedback annotations in our dataset only have the beginning times, but not the duration.

We classify each such moving window for the presence of feedback leveraging Audio, Text, and Audio+Text late fusion models. We apply *Automated Speech Recognition (ASR)* with *Whisper-1* (Radford et al., 2023) to obtain the text from a given timespan. For audio classification, we fine-tune Wav2Vec base model (Baevski et al., 2020) with 95M parameters.

For text classification, we fine-tune BERT base model with 110M parameters (Devlin et al., 2018). For multimodal classification, we apply a late fusion approach where we extract richer representations from the audio and text modalities as 256-dimension vectors. The representations are concatenated into one 512-dimension vector and passed via 2 fully-connected linear layers that reduce the dimensions to 64 and finally 2. This sequential architecture is augmented with ReLu activation and additional dropout in between. The additional steps can help the model calculate intermediate fusion features.