2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #51114

LLM Based Analysis of Using Videos in Teaching Robotics

Prof. hongbo zhang, Middle Tennnesee State University

I am Assistant Professor at Middle Tennnesee State University. I am currently working on the robotics engienering education research.

Benjamin Li, Middle Tennessee State University

LLM Based Analysis of Using Videos in Teaching Robotics

Abstract

Embodied learning involves the use of hands-on and physical interaction with the underlying subjects. This research explored the use of the embodied learning-centered videos for robotics instruction. The engagement of the users with the videos was studied through the video user comments with LLM models. The correlation between video user comments and the video frames were also analyzed with LLM models. The research identified the learnable elements in the videos for users to interact and engage through the analysis the video frame captions. The results showed the evident advantages of using embodied learning videos in engaging user interests in learning robotics technology. Overall, users have shown enhanced positive comments on the embodied learning videos. A user study was also performed involving both embodied learning and conventional learning videos. The results of the user study showed that the use of the embodied learning elements in the videos has reinforced the learning effectiveness with positive user comments and feedback. The research has practical implications for the enhancement of robotics instruction and workforce training.

1. Introduction

Learning robotics technology is challenging due to the multidisciplinary nature of the subject, which involves mathematics, electronics, software, mechanical engineering, and materials science. The use of videos has been shown effective in teaching robotics. Specifically, instructional videos is helpful to enhance the learning outcome in remote learning of robotics technology related classes[1,2]. The use of videos is also crucial for the flipped classroom teaching, where the use of videos plays critical roles in preparing students for the relevant lectures[3].

Conventional learning approaches mainly rely on lectures, slides, or textbooks for learning underlying subjects. While conventional learning has been adopted for STEM education, its disadvantages are evident where learners are frequently disengaged through the learning process[4]. In contrast, embodied learning paradigm emphasizes the integration of the cognitive, sensorial, affective and psychomotor functioning through the learning process[5]. Embodied learning represents the use embodied elements including physical objects and human gestures through learning[6,7,8]. Embodied learning emphasizes the use of experiential learning through the learning process thus to achieve better learning outcomes in contrast to the conventional learning approach.

Previous studies showed that embodied learning has positively impacted the learning

performance, which is shown able to promote the increased participation of the underlying learning process and motivates learners towards the learning subjects[9,10].Among them, Freeman et al. (2014) undertook a meta-analysis of 225 comparative research studies of conventional lecturing with studies of emphasis on embodied learning principles[4]. The study showed that the use of embodied learning has improved student performance by 6 percent for test scores and reduced failure rates by 55 percent, therefore making embodied learning more effective in improving overall class performance as compared to conventional learning approaches. Similarly, Kontra et al. (2015) showed that physical experience in science learning increases the understanding of challenging principles like angular momentum by thus engaging certain cerebral bodily systems[11]. It is shown that the embodiment of physical interactions allows students to better gain performance on quizzes and presented learning outcome showing the effectiveness of embodied learning in science. In the similar vein, Khodr et al. (2021) has leveraged the use of Cellulo robots for teaching the concept of virus propagation for high school students[12]. Specifically, the embodied learning technique was applied through the process, where the combination of a physical robot interaction and software simulation was used to improve the learning performance. It also shows that the use of embodied learning has led to significant appreciation of students with the enhanced level of interest of the subjects and improved engagement in the learning activities. The use of embodied learning is not only able to engage student interests but also enhance subject concept clarity and encourage the participation of students in classrooms.

While the benefits are clear, it is challenging to effectively execute embodied learning in practice. Embodied learning requires stronger teaching skills in execution the learning paradigm, which requires physical interaction, sensory engagement, kinesthetic learning, and mind-body connections [13,14,15]. Furthermore, embodied learning also requires the use of strong hands-on skills, effective visualization, and sensory engagement of students through the process [16,17]. There is also lack of dedicated study for the investigation of the effectiveness of the embodied learning paradigm for learning of the robotics technology. Until present, there is no research investigating the comparable effectiveness of the use of embodied learning and conventional videos for remote learning and flipped classrooms through the active learning process. It therefore becomes difficult for instructors to choose videos for robotics instruction and workforce training. This therefore has motivated our research to investigate the performance of embodied learning for robotics technology instruction using videos.

The project aims to understand the user engagement of robotics instruction videos featured with embodied and conventional learning methods for learning robotics and workforce training. By systematically analyzing user engagement and sentiment toward the video content with the large language model (LLM) methods, this study is able to provide a comprehensive understanding of how the videos featured with embodied learning method impact the learning effectiveness of the robotics technology. For this purpose, through the utilization of the LLM based data extraction, transcription, and sentiment analysis, our research is able to ensure accurate and reliable video analysis for the desirable research outcomes. We seek to assess whether videos featured with embodied learning technique is able to better engage with learners in comparison to conventional learning videos. Through the process, the video comments were analyzed for assessing the interaction and participation of the learning process. Subsequently, the sentiment analysis of video comments was also performed to understand the emotional (positive, negative, neutral) and

cognitive responses of both embodied learning and conventional learning to understand the differences of user engagement with the videos. We have also performed the user study for understanding of the differences of the embodied centered learning videos versus the conventional learning centered videos. Systematic comparison between different LLM methods was also compared for non-biased analysis for the assessment of the effectiveness of the embodied learning in robotics instruction. We have therefore concluded the effectiveness and reception of conventional and embodied robotics instruction videos. The research is expected to provide significant insights for robotics instructors to teach robotics technology effectively by selecting effective videos and also obtain instructional insights from the videos.

2. Methods

2.1 Overview

The research selected and categorized YouTube videos into conventional and embodied learning videos. The embodied learning videos represent the use of hands-on and physical interaction with objects. In contrast, conventional learning videos involve the demonstration and use of text, slides, graphics, maps across various topics related to robotics such as object manipulation, programming, and path planning for robots. We have made efforts in ensuring the representation of the videos. Diversified videos were studied in our research, which include programming, mathematics, AI, and hardware. Through the process, both formal and informal learning videos have been selected. Specifically, the video content range from introduction to expert level robotics technology so that the videos are more representative for learning robotics technology. The videos are also comprehensive for covering significant number of applications including agriculture, manufacturing, healthcare, logistics, construction, energy, and retails. These videos are representative so that it is able to be useful for robotics workforce training without suffering biased video selection.

We have also enforced the random selection of the videos for foundational robotics topics, robotics applications, and special robots. Given the list of the searched results of the videos on Youtube, we have selected the videos by following the definition of the conventional learning and embodied learning. It means that the embodied learning and conventional learning videos were selected randomly based on the learning content rather than the quality and experiences of the teaching and the level of user engagement with the videos.

The balanced selection of the videos in terms of user comments was also performed. Specifically, the videos with balanced comments such as very few numbers (1 - 10), somewhat (10+), and large number (100+) user comments were selected. Other balanced concern of video comments includes the selection of the sentiment of the videos, where both positive, neutral, and negative comments are all in need of presence in the conventional and embodied videos. Through the use of positive, neutral, and negative user comments for the videos, it is able to ensure the fair comparison between the conventional and embodied videos.

The descriptive statistics of the video length and total comments is shown in Table 1. Specifically, it shows the minium, maximum, and average video length and total comments. Specifically, the the minimum video length is 1 minute with maximum video length of 34 minutes, and average

Table 1: Data Collection of Robotics Instructional Videos (Conventional vs. Embodied)

Table I: Conventional Robotics Instructional Videos

Category	Video ID	Subcategory	Content Description	Comments
Applications	V1_con	Comp. Vision	AI-based virtual mouse using OpenCV and Mediapipe.	980
	V2_con	Deep Learning	Robotics advancements via training datasets and adaptation.	6
Foundation	V4_con	Electronics	Arduino basics: circuits, IDE setup, and beginner projects.	1408
	V6_con	ROS Prog.	ROS basics and Gazebo simulations for robotic systems.	112
	V8_con	Industrial	ABB robot programming fundamentals with virtual controllers.	15
	V7_con	ROS Setup	Installing/configuring ROS Noetic on Ubuntu 20.04.	62
Special Robots	V3_con	Grippers	Designing robotic grippers for drones and aerial navigation.	3
	V5_con	Combat	Beginner's guide to combat robots and RC electronics.	129
	V10_con	FRC	FRC roles/mechanics highlighting SolidWorks tools.	22
	V9_con	Quadcopter	Quadcopter mechanics: thrust, torque, and motion control.	29

Table II: Embodied Robotics Videos

Category	Video ID	Subcategory	Content Description	Comments
Foundation	V6_emb	Electronics	Guides on using breadboards, multimeters, and Arduino.	281
	V2_emb	ROS	AMR with enhanced odometry and ROS-based navigation.	271
	V1_emb	AMR	Waypoint guidance and 360-degree obstacle avoidance.	54
Special Robots	V3_emb	Combat	Guide to building combat robots and channel mixing.	128
	V4_emb	Combat	Electronics for beetleweight robots: ESCs and batteries.	24
	V5_emb	Robotic Arm	Mark 1 arm with hand gesture glove control via Bluetooth.	623
	V9_emb	Intel Robot	Jetson Nano/TurtleBot 3 for mapping and recognition.	248
Applications	V7_emb	Fire Fighting	Automated firefighting robot with flame sensors and pump.	xxx
	V8_emb	DIY Drone	Custom drone build with KK2.1.5 flight controller.	559
	V10_emb	Comp. Vision	Vision with Arduino: face tracking and object detection.	506

video length of 4 minutes. The number user comments per video has minimum value of 1 comment, maximum of 128 comments, and average of 20 comments.

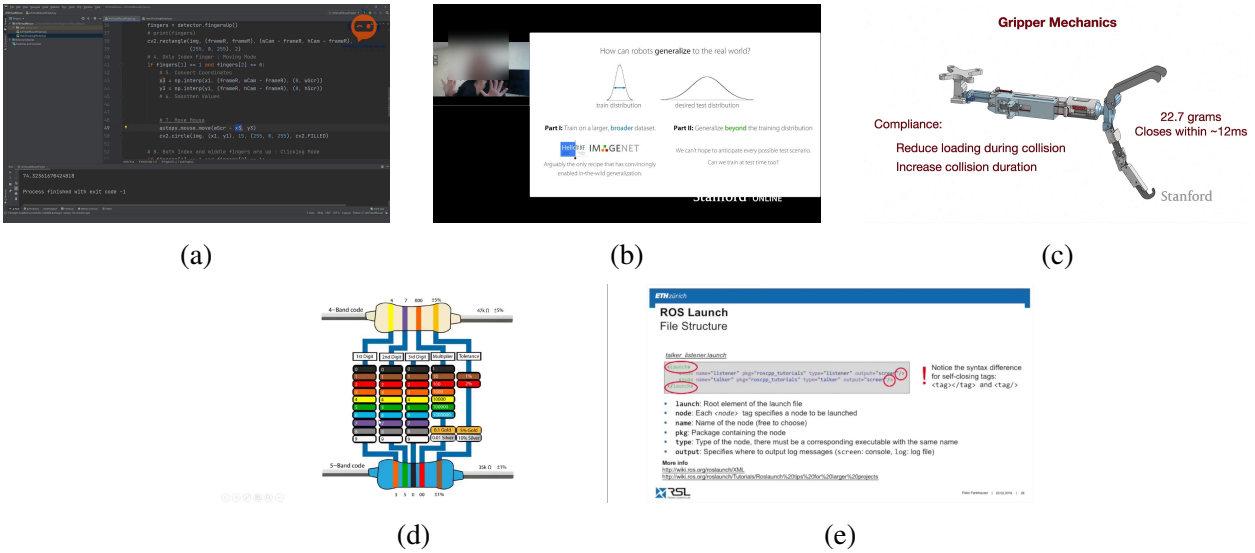
Following the video data collectioneach frame of the video was extracted for further analysis of the video content. Specifically, these frames were captioned with vision-language models, LLaVA (Large Language and Vision Assistant), and MiniCPM (Mini Chinese Pre-trained Language Model) to provide textual description of the video frames. The correlation analysis between frame captions and video user comments to understand user learning and engagement with videos, where the correlation analysis was performed with the LLaMA-2 LLM model.

Among the LLM models used for the analysis of the videos, LLaVA is the state-of-the-art multimodal model, combines natural language processing with computer vision to produce detailed and semantically accurate captions. On the other hand, MiniCPM, a lightweight yet robust pretrained model, focuses on generating concise and contextually relevant descriptions with high efficiency. For obtaining accurate LLM analysis results, both models were compared for their captioning performance. Additionally, user comments were extracted, and a correlation analysis was performed to evaluate the alignment between video content, generated captions, and user feedback. This integrated approach provided insights into the relationship between video content, descriptive accuracy, and user engagement. The following sections outline the methods and tools used in this study.

2.2 Data Collection

The identified Youtube videos were categorized into two distinct groups, videos featuring embodied learning methods, characterized by hands-on activities and physical interaction with robotic components, and videos featuring conventional learning methods, which primarily involve text-centered instruction along with visual demonstration. The selection of robotics videos in the two categories was based on topics such as robotic manipulation, kinematics of robots, robotic programming, deep reinforcement learning in robotics, autonomous navigation, and related subjects.

Both conventional and embodied learning videos are shown in Figure 1 and Figure 2, which show the examples of conventional and embodied videos selected respectively for the research. In Figure 1, videos demonstrated detailed explanations about electronics used in the building of compact robots starting from basic steps to more complicated configurations. The tutorial has explained the concepts verbally and visually while the tutorial provides the instructions, it doesn't involve the physical construction of robots through the video. The emphasis is however on understanding the principles of building the compact robot with text and images rather than hands-on activities. In contrast, Figure 2 shows the tutorial in a step-by-step fashion for how to build the robotic arm allowing viewers to follow along and actively participate in the creation process. Other hands-on and physical interaction driven robotics learning related videos were also included in the video dataset. The interactive nature of the embodied learning tutorial to control the robotic arm wirelessly and achieve automated functions showing the physical interaction nature of the learning process.

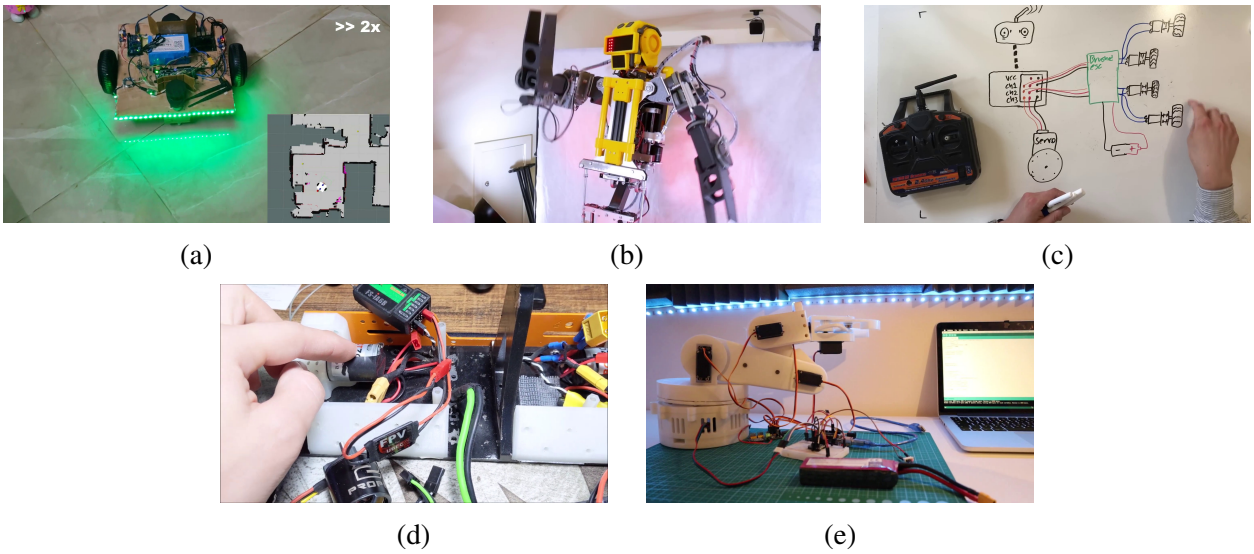


(a) (b) (c)

(d) (e)

Figure 1: **Conventional Learning Video Dataset:** (a) AI-based virtual mouse controller, (b) Deep reinforcement learning adaptation, (c) Aerial grasping and ReachBot grippers, (d) Arduino microcontroller fundamentals, (e) ROS nodes and architecture simulation.

Following the identification and categorization of the videos, the research utilized youtube-comment-downloader, which is a python tool to download the comments of the users systematically from YouTube videos. The tool effectively searches the comments by taking the



(a) (b) (c)

(d) (e)

Figure 2: **Embodied Learning Video Dataset:** (a) Humanoid robot dynamic tasks, (b) Autonomous navigation/mapping (c) Quadruped agility and terrain traversal, (d) Humanoid motion planning, (e) Arduino robotic arm with gesture control

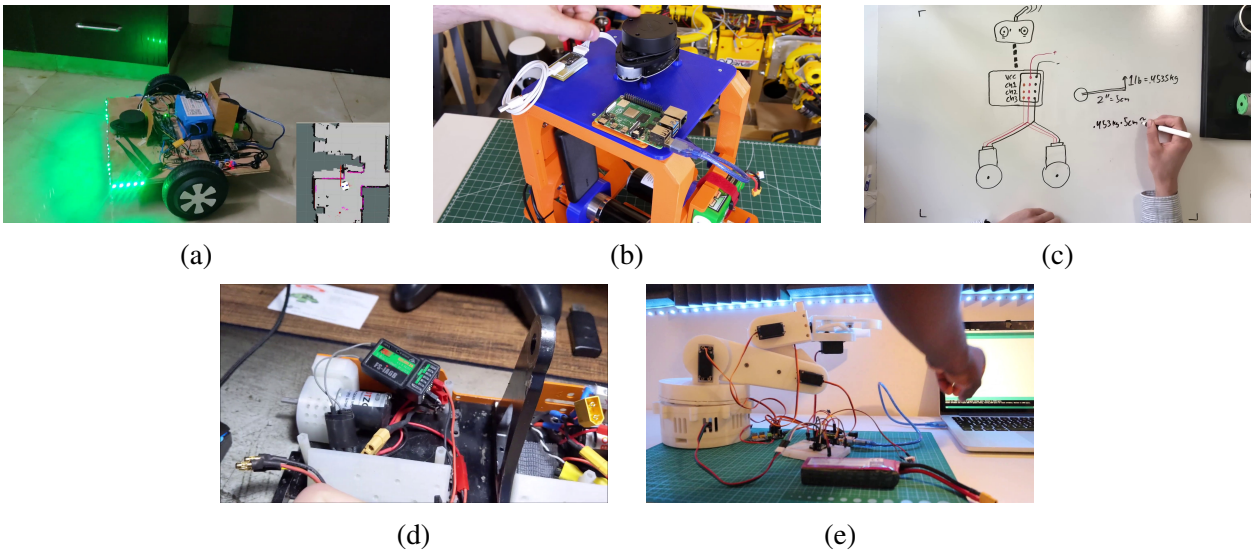
URL of the targeted video and the comments fetched from the mentioned video are stored into a JSON file. All of the comment data of the comments is included in the resulting JSON output file such as comment ID, comment text, comment timestamp, comment author, channel ID, comment votes, and reply status. It was also observed that with JSON format, the extracted data can easily be processed and analyzed by different data analysis tools so as to derive meaningful insights from the user generated content.The text comments were extracted from a JSON file using a Python script. Such approach was rather beneficial in terms of the fact that it allowed to extract and store the comments in a way that would be convenient for further analysis and finally the comments of the selected videos are stored in separate documents for further analysis.

2.3 Frame Extraction

Frames were systematically extracted from conventional and embodied robotics videos shown in Figure 3. A total of 100 frames were extracted from each video, providing a representative overview of the visual content. The frame extraction process used OpenCV, a robust library for computer vision, to process video files efficiently. The extraction relied on metadata such as frame count, resolution, and frame rate to ensure consistent sampling across all videos. The frames were stored in high-quality formats to preserve their visual details, ensuring suitability for subsequent processes such as caption generation and analysis. This standardized approach ensues that the extracted dataset will be consistent and comprehensive, supporting detailed comparisons between the hands-on and conventional videos.

2.4 Frame Captioning

Larger language model, LLaVA and MiniCPM were used to generate caption for the extracted frames shown in Figure 4. These models were chosen for their ability to generate accurate and



(a) (b) (c)

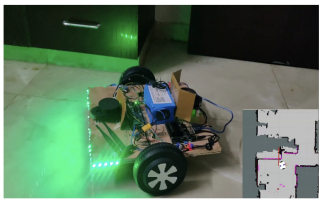
(d) (e)

Figure 3: **Video Frames Data:** (a) Video_1_emb (AMR mapping), (b) Video_2_emb (ROS hardware), (c) Video_3_emb (Combat robot wiring), (d) Video_4_emb (Beetleweight electronics), (e) Video_5_emb (Mark 1 robotic arm with gesture control).

contextually rich captions that give further insight into the visual elements of both embodied and conventional videos.

To generate captions for the extracted frames, the LLaVA model was applied to process each video frame to obtain the description of the frame benefiting from the model capability of analyzing and captioning a visual content. This involved loading the frames from video and parsing each image through the LLaVA model. The captions for all the frames were processed where the results were stored to enable subsequent analyses including content and user engagement correlation analysis.

For ensuring the non-biased LLM analysis, the frames extracted from videos were also captioned using the MiniCPM-V model, a vision-language model that is trained for fine-grained image description. MiniCPM-V also has strong multimodal analysis and can analyze the visual content and generate descriptive captions. Through the use of another LLM model, we can obtain the comparable analysis of the video frame caption dataset to ensure robustness of the analysis. In the process, for each frame, the model was used to generate captions that described the video frame scene in detail as well as provided a summary of the scene. The process involved loading and pre-processing each video frame, then used a structured dialogue-based approach to interact with the model and the prompt to generate video frame captions. The model employed sampling strategies with a regulated temperature to ensure the variety and applicability of the produced captions. The obtained captions reflected the fine-grained details of the visual content, providing a dense textual description for further analysis. This approach ensured the creation of accurate and high-quality captions for all video frames.



MiniCPM caption: The image features a small, green robot that is being controlled by a wooden board with a green light on top. The robot is equipped with a battery and has a black wheel on the front. It is positioned on a tiled floor in a room. The robot appears to be in the process of moving, possibly towards the right side of the image.
LLaVA caption: The image shows a small, homemade robot. It appears to be a self-driving vehicle with four wheels and some sort of chassis at the center. There is a visible electronics board or breadboard mounted on top, which could be part of the circuitry for controlling the robot's movements. There are also what look like LED strips glowing in various colors on its frame, suggesting that it might be used for demonstration purposes, perhaps to showcase its autonomous capabilities. The environment suggests this is a home setting, with a kitchen or dining area visible in the background.

Figure 4: The example of video frame along with the frame caption generated by MiniCPM and LLaVA model from video_1_emb

2.5 Sentiment Analysis of User Comments and Correlation of User Comments with Frame Caption

To reveal the user learning experiences of the robotics instructional videos. For the user comments of both conventional and embodied learning videos, the sentiment of the user comments were analyzed through the use of LLaMA-2 model to determine the positive, neutral, and negative user comments. The sentiment analysis of the user comments is expected to reveal the user perception of the video content and their sentimental reaction toward the video content and overall learning experiences through the learning process.

A correlation analysis was also conducted using the LLaMA-2 model to assess the relationship between the generated frame captions and user comments shown in Table 2. This analysis aimed to identify how effectively the captions, derived from video frames, aligned with the topics and technical details discussed in the user comments. By evaluating this correlation between user comments and video content, the study measured the user engagement of the videos through the learning process.

To obtain the video content reliably, the captions corresponding to three successive frames were combined. It therefore made it possible to have a continuous view of the video segment to correlate to the video comments. The concatenation enabled the model to capture sequential information of the video frames. For example, the video frames dedicated to the elaborations of robotics platform DIY steps and procedures that requires more than one frame of the video. Therefore, our approach will ensure summarized in a much better way.The process involved comparing each user comment with captions corresponding to three consecutive frames.

Table 2: Comprehensive LLaVA and MiniCPM caption correlation analysis for Embodied and Conventional robotics instructional video content.

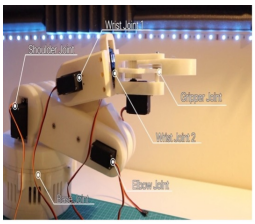
Embodied Video (LLaVA Analysis)

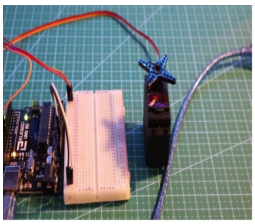
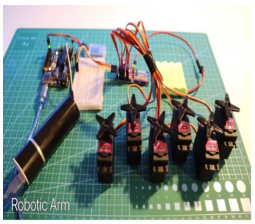


User Comment: "Man, I swear there is a button more brushless ESCs than brushed ones at 6 cells battery... might be worth considering if you want to build a brushed motor system for 60 lbs robots."
Correlated Captions: 1. Workspace featuring hand holding motor part, soldering iron. 2. DC motors with wiring harnesses and ESCs. 3. Detail of electronic device; text discusses peak power (420W).

Reason for Correlation: Correlation between mentioned **brushless ESCs/motors** and LLaVA identifying **DC motors/controllers**. Technical discussion aligns with visual assembly context.

Embodied Video (MiniCPM Analysis)

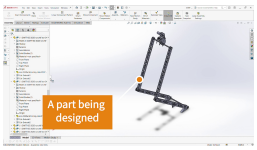


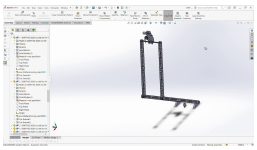
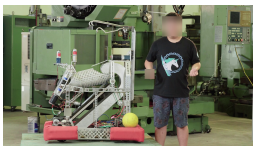



User Comment: "Can I use the MG 996R that rotates 180 degrees, or does it have to be 360?"
Correlated Captions: • Robotic arm wrist joint assembly. • DIY setup with blue servo motor and breadboard. • Network of wires connecting battery and circuit board to a robot arm.

Reason for Correlation: Query regarding **MG 996R servo** correlates with the "blue motor" (servo) identified in frames. Visual data of the joint assembly matches the technical query.

Conventional Video Analysis (LLaVA Model)

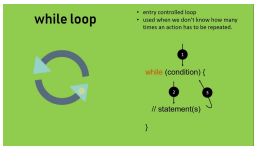


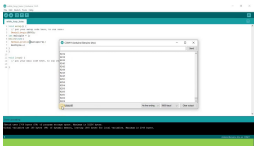
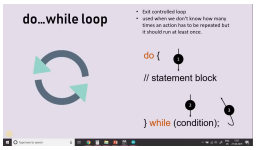



User Comment: "For CAD I would recommend: 1. Onshape, 2. Solidworks, 3. Autodesk Inventor, 4. Autodesk Fusion360."
Correlated Captions: • 3D modeling software featuring "PART BEING DESIGNED". • CAD interface with wireframe geometry and modeling functions. • Industrial workshop showing a CNC mill for manufacturing.

Reason for Correlation: The **CAD software recommendations** correlate with visual evidence of 3D modeling environments and the transition to industrial manufacturing (CNC milling).

Conventional Video Analysis (MiniCPM Model)



User Comment: "Great introduction to C programming. Not much Arduino though."
Correlated Captions: • Diagram illustrating `while` loop logic and flow. • Screenshot of text editor with "C++ Arduino code". • Flow diagram explaining iterative `do-while` loops and `if` statements.

Reason for Correlation: Connection is direct; frames provide instruction on core **C/C++ programming constructs** (loops and logic) as applied to Arduino.

The LLaMA-2 model was utilized to determine the correlation between user comments and video frame captions, where the user comment and frame caption were used as the input for the model. The LLaMA-2 model provided a confidence score (0 to 100), quantifying the strength of the correlation. High correlation scores indicated strong coherence between user comments and video content. The correlation was applied separately to the video frame captions generated by LLAVA and MiniCPM for ensuring the robustness and comparison of the results, providing insights into the comparative analysis of the video visual content and user engagement with videos.

2.6 User Study Data and Results Analysis

We performed the user study involving 20 participants for learning from 130 Youtube videos. The videos were categorized conventional and embodied videos. The 130 videos consisted of 74 embodied videos and 56 conventional videos. Videos were presented to the users, and user engagement and learning outcomes were assessed by and analyzing user feedback after watching the videos. With the user feedback and video frame captions, the correlation analysis was computed between these user comments and the video frame captions to quantify the user engagement of the videos.

We have performed the analysis of the videos characteristics as well as analyzing the relationship between video content and user engagement. For this, seven quantitative metrics were defined. Specifically, we have defined the following metrics for evaluating the videos and user engagement with the videos.

- **Per Video Caption Word Count (W_{cap}):** The total number of words in the video caption, serving as an indicator for the volume of verbal instructional content.
- **Per Video Duration (T):** The temporal length of the video in seconds.
- **Per Video Caption Density (ρ_{cap}):** The rate of verbal information delivery, defined as the number of per video caption words dividing by video length (words/seconds). This metric indicates the informational intensity of the video.

$$\rho_{cap} = \frac{W_{cap}}{T} \tag{1}$$

The results of video user comments are also shown in below.

- **Per Video Comments Count (N_{com}):** The total number of student comments collected for each video, quantifying the volume of learner feedback.
- **Engagement Rate (R_{eng}):** The engagement of users for video, measured by comments per minute.

$$R_{eng} = \frac{N_{com}}{T/60} \tag{2}$$

The results of correlation between video content and user comments are shown in below.

Correlation Score (S_{corr}): The average correlation between user comments and captions across user comments and video captions. It is calculated as:

$$S_{corr} = \frac{1}{N_c} \sum_{i=1}^{N_c} s(c_i, v) \tag{3}$$

where N_c is the total number of comments and $s(c_i, v)$ denotes the relevance score of the i-th comment c_i to video v.

- **Correlation Efficiency (η_{corr}):** The percentage of the correlation score greater than the set threshold score τ (set to 60%). This metric reflects the strength of the correlation between the user comment and the video frame caption.

$$\eta_{corr} = \frac{|\{s_i \mid s_i \geq \tau\}|}{N_{total}} \tag{4}$$

where N_{total} is the total number of candidate correlations evaluated and s_i represents the confidence score of the i-th candidate.

3. Results

Table 3 presents the correlation results for embodied videos using the LLaVA-generated captions. The results indicate a generally high level of alignment between user comments and the visual content analyzed by LLaVA. For instance, for the video of Video_8_emb, which focuses on a DIY drone, it shows a robust correlation with 899 out of 1332 comments being relevant to the video frame captions. Similarly, Video_5_emb (Robotic Arm) demonstrates strong engagement with 465 correlated comments out of 623. This suggests that LLaVA is effective at identifying the key visual elements in hands-on robotics tasks that users are actively discussing. However, there is some variability. As shown in Video_1_emb, where fewer comments are shown correlated, is possibly attributed to the specific nature of the discussions not relevant to video content. Table 4 summarizes the results for conventional videos using LLaVA captions. Here, the performance varies significantly depending on the video content. Video_1_con (Computer Vision) and Video_4_con (Electronics Basics) show decent correlation counts of 615 and 422 respectively, reflecting the popularity of these foundational topics. However, many other videos in this category, such as Video_2_con and Video_3_con, show almost negligible correlation (1 correlated comment). This highlights a potential limitation of LLaVA in understanding the abstract or theoretical concepts often presented in conventional lecture-style videos, or simply a lack of relevant user engagement with those visual elements.

Specifically, the overall ratio of the correlated comments versus total comment is more pronounced (Embodied: 266.4/395.9=0.67 versus Conventional: 127.3/275.3=0.46) for the embodied learning videos in comparison to the conventional learning videos. This has therefore suggests the overall effectiveness of the use of embodied learning videos for robotics instruction. The results were obtained through the use of videos with wide range of user comments from 3 - 1332, hence reinforcing that for both less engaging and engaging videos, they showed a similar patterns for the user engagement.

Table 3: Total and Correlated Comments for Embodied Learning Videos (LLaVA Captions)

Video Id	Total Comments	Correlated Comments
Video_1_emb_LLaVA	54	44
Video_2_emb_LLaVA	271	180
Video_3_emb_LLaVA	128	67
Video_4_emb_LLaVA	24	19
Video_5_emb_LLaVA	623	465
Video_6_emb_LLaVA	214	129
Video_7_emb_LLaVA	559	448
Video_8_emb_LLaVA	1332	899
Video_9_emb_LLaVA	248	147
Video_10_emb_LLaVA	506	266
Average	395.9	266.4

Table 4: Total and Correlated Comments for Conventional Learning Videos (LLaVA Generated Captions)

Video Id	Total Comments	Correlated Comments
Video_1_con_LLaVA	980	615
Video_2_con_LLaVA	6	1
Video_3_con_LLaVA	3	1
Video_4_con_LLaVA	1408	422
Video_5_con_LLaVA	129	75
Video_6_con_LLaVA	109	83
Video_7_con_LLaVA	54	47
Video_8_con_LLaVA	15	11
Video_9_con_LLaVA	29	13
Video_10_con_LLaVA	20	5
Average	275.3	127.3

Table 5 displays the correlation metrics for embodied videos when using MiniCPM for captioning. The results are comparable to LLaVA but with some notable differences. For Video_8_emb, MiniCPM achieved 829 correlated comments, slightly lower than LLaVA's 899. Conversely, for Video_2_emb, MiniCPM identified 182 correlated comments, marginally outperforming LLaVA. This indicates that MiniCPM produces captions that are competitive in describing physical robotics activities. The consistency across most embodied videos suggests that MiniCPM is a viable lightweight alternative LLM model for processing hands-on instructional content. Table 6 presents the performance of MiniCPM on conventional videos. Interestingly, MiniCPM demonstrates a distinct advantage in this category compared to LLaVA. For Video_1_con video, it achieved 809 correlated comments, significantly higher than LLaVA's 615. Similarly, for Video_4_con, the correlated count has increased to 700. This suggests that

MiniCPM's captioning style might be better suited for the static slides or specific visual formats common in conventional tutorials, allowing it to capture more of the text-heavy or schematic information in users comments of the videos. The contrasted differences (Embodied: 251.7/395.9, Conventional: 174.4/275.3) of the ratio of the correlated comments between embodied learning and conventional learning videos shows slightly stronger user engagement for the embodied learning videos.

Table 5: Total and Correlated Comments for Embodied Learning Videos (MiniCPM Generated Captions)

Video Id	**Total Comments**	**Correlated Comments**
Video_1_emb_MiniCPM	54	32
Video_2_emb_MiniCPM	271	182
Video_3_emb_MiniCPM	128	80
Video_4_emb_MiniCPM	24	21
Video_5_emb_MiniCPM	623	438
Video_6_emb_MiniCPM	214	100
Video_7_emb_MiniCPM	559	417
Video_8_emb_MiniCPM	1332	829
Video_9_emb_MiniCPM	248	128
Video_10_emb_MiniCPM	506	290
Average	395.9	251.7

Table 6: Total and Correlated Comments for Conventional Learning Videos (MiniCPM Generated Captions)

Video Id	**Total Comments**	**Correlated Comments**
Video_1_con_MiniCPM	980	809
Video_2_con_MiniCPM	6	3
Video_3_con_MiniCPM	3	1
Video_4_con_MiniCPM	1408	700
Video_5_con_MiniCPM	129	85
Video_6_con_MiniCPM	109	76
Video_7_con_MiniCPM	54	37
Video_8_con_MiniCPM	15	10
Video_9_con_MiniCPM	29	15
Video_10_con_MiniCPM	20	8
Average	275.3	174.4

Table 7 illustrates the results for conventional videos using the integrated approach, which combines insights from both LLaVA and MiniCPM. This method yields the most comprehensive correlation results. For Video_4_con, the number of correlated comments peaked at 911, and

Video_1_con saw an increase to 834. This improvement confirms that the integrated model successfully bridges the gaps left by individual models, showing a wider array of relevant content. The breakdown of the sentiment has revealed that a significant portion of these correlated comments are positive (e.g., 513 positive for Video_4_con), indicating that when the content is accurately identified, it aligns well with learner appreciation. Table 8 shows the results for embodied videos using the integrated captioning approach. This method consistently maximizes the matching between captions and user feedback. For example, for the video of video_8_emb, there are 1090 correlated comments, the highest specific value observed, with a strong positive sentiment ratio.

Table 7: Total Comments and Correlated Comments for Conventional Learning Videos (Integrated Captions)

Video ID	**Total**	**Correlated**	**Positive**	**Neutral**	**Negative**
Video_1_con	980	834	215	310	309
Video_2_con	6	4	3	1	0
Video_3_con	3	1	1	0	0
Video_4_con	1408	911	513	313	85
Video_5_con	129	82	35	47	10
Video_6_con	109	81	39	30	12
Video_7_con	54	46	17	20	9
Video_8_con	15	11	7	4	0
Video_9_con	29	16	8	7	1
Video_10_con	20	5	5	0	0
Average	275.3	199.1	84.3	73.2	42.6

Table 8: Sentiment Distribution for Correlated Comments in Embodied Learning Videos

Video ID	**Total**	**Corr.**	**Pos.**	**Neu.**	**Neg.**
Video_1_emb	54	49	33	16	0
Video_2_emb	271	240	124	89	27
Video_3_emb	128	96	41	42	13
Video_4_emb	24	23	16	7	0
Video_5_emb	623	551	318	204	29
Video_6_emb	214	166	113	39	14
Video_7_emb	559	486	141	273	72
Video_8_emb	1332	1090	515	514	61
Video_9_emb	248	199	139	44	16
Video_10_emb	506	398	224	121	53
Average	395.9	329.8	166.4	134.9	28.5

3.1 Statistical Analysis

To rigorously assess differences between embodied and conventional learning videos, statistical hypothesis tests were performed across the key metrics. The Shapiro-Wilk test was first applied to each group to assess the normality of the data distributions. For metrics satisfying the normality assumption ($p > 0.05$), an independent samples t-test (Welch's variant) was used. For non-normally distributed metrics, the non-parametric Mann-Whitney U test was employed. Cohen's d was computed as the effect size measure, where $|d| < 0.2$ is considered negligible, $0.2 \leq |d| < 0.5$ is small, $0.5 \leq |d| < 0.8$ is medium, and $|d| \geq 0.8$ is large[18]. The significance level was set at $\alpha = 0.05$. The results are summarized in Table 9.

For the user study involving 130 videos (74 embodied, 56 conventional), the engagement rate showed a small positive effect ($d = 0.275$, $p = 0.126$) favoring embodied videos, and the correlation efficiency metric also showed a small positive effect ($d = 0.230$, $p = 0.215$) in the same direction. Video duration showed a small negative effect ($d = -0.399$, $p = 0.132$), indicating that embodied videos tended to be shorter. While these differences did not reach statistical significance at $\alpha = 0.05$, the consistent directional trends across all engagement-related metrics support the descriptive observations.

For the YouTube comment analysis involving 20 curated videos (10 embodied, 10 conventional), the integrated correlation ratio demonstrated a statistically significant difference ($t = 3.254$, $p = 0.008$, $d = 1.455$, large effect), confirming that the integrated LLaVA-MiniCPM approach captured significantly higher content–comment alignment for embodied videos (M = 0.844) compared to conventional videos (M = 0.626). The LLaVA correlation ratio also approached significance ($p = 0.070$) with a large effect size ($d = 0.885$). These results provide strong statistical evidence that embodied learning videos generate higher content-aligned user engagement.

Table 9: Statistical Comparison of Embodied vs. Conventional Videos

Metric	**Emb. Mean±SD**	**Con. Mean±SD**	**Test**	**p-value**	**Cohen's d**	**Sig.**
User Study (130 Videos)						
Avg. Corr. Score	57.31 ± 19.76	53.51 ± 20.99	U	0.327	0.187 (Neg.)	n.s.
Caption Words	54510 ± 43491	73843 ± 80246	U	0.118	-0.312 (S)	n.s.
Comment Count	2.07 ± 1.53	1.91 ± 1.75	U	0.206	0.096 (Neg.)	n.s.
Duration (s)	699 ± 417	964 ± 891	U	0.132	-0.399 (S)	n.s.
Caption Density	74.06 ± 26.23	77.65 ± 29.80	U	0.080	-0.129 (Neg.)	n.s.
Engagement Rate	0.27 ± 0.34	0.19 ± 0.21	U	0.126	0.275 (S)	n.s.
Corr. Efficiency	0.62 ± 0.27	0.55 ± 0.28	U	0.215	0.230 (S)	n.s.
YouTube Comment Analysis (20 Videos)						
Corr. Ratio (LLaVA)	0.674 ± 0.111	0.507 ± 0.242	t	0.070	0.885 (L)	n.s.
Corr. Ratio (MiniCPM)	0.639 ± 0.117	0.578 ± 0.153	t	0.330	0.449 (S)	n.s.
Corr. Ratio (Integ.)	0.844 ± 0.067	0.626 ± 0.200	t	0.008	1.455 (L)	**
Pos. Sent. Ratio	0.560 ± 0.136	0.599 ± 0.252	t	0.672	-0.193 (Neg.)	n.s.
Neu. Sent. Ratio	0.360 ± 0.106	0.314 ± 0.185	t	0.508	0.304 (S)	n.s.
Neg. Sent. Ratio	0.080 ± 0.053	0.099 ± 0.119	U	1.000	-0.206 (S)	n.s.

Neg. = Negligible, S = Small, L = Large. * $p < 0.05$, ** $p < 0.01$, n.s. = not significant.

3.2 User Study Results

User study results showed the user engagement differences between the embodied versus conventional learning videos shown in Figure 5 and 6. Embodied robotics videos are associated with a higher level of user interaction, where users frequently discuss specific mechanical or programming details shown in the videos. For example, video Video_8_emb has engaged 1332 comments, indicating user interests in the DIY drone assembly process. In contrast, conventional videos showed higher variance in engagement, while video Video_4_con (Arduino basics) is associated with 1408 comments. Other conventional videos such as Video_2_con and Video_3_con have received negligible feedback (6 and 3 comments respectively).

There is a clear difference between the conventional learning videos and the embodied learning videos in terms of the user engagement. Such difference suggests that while foundational theory class of robotics can be used for robotics instruction, hands-on embodied demonstrations consistently foster more active learner's participation. The User study also yielded the descriptive statistics of the videos through the metrics of video frame caption word counts and video frame caption density.

The user study has also revealed the results of the user engagement with the videos. Specifically, Figure 5 shows a clear correlation between the user comments and video content, where the embodied learning videos are associated with stronger correlation between the user comments and video content, whereas the conventional learning videos shows less correlation between them. The results support the hypothesis that embodied learning is more supportive in engaging users through the video visual content. Specifically, embodied learning videos have provided rich visual context that generated detailed captions, further supporting the hypothesis that physical interaction has provided a more engaging learning experiences for users compared to sole theoretical-based robotics learning process.

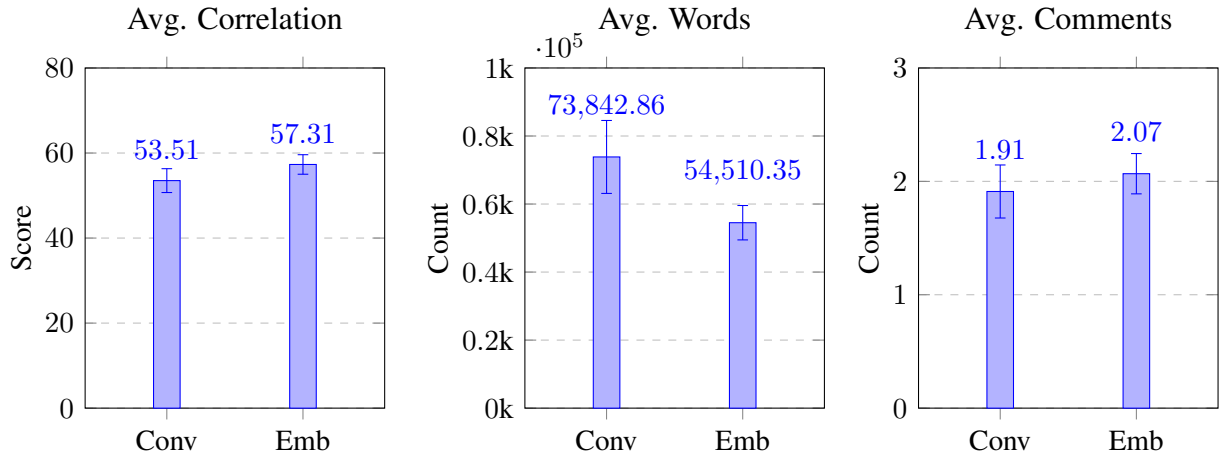


Figure 5: User study results for correlation between user comments and video content, caption density, user engagement, and efficiency in user engagement with video content.

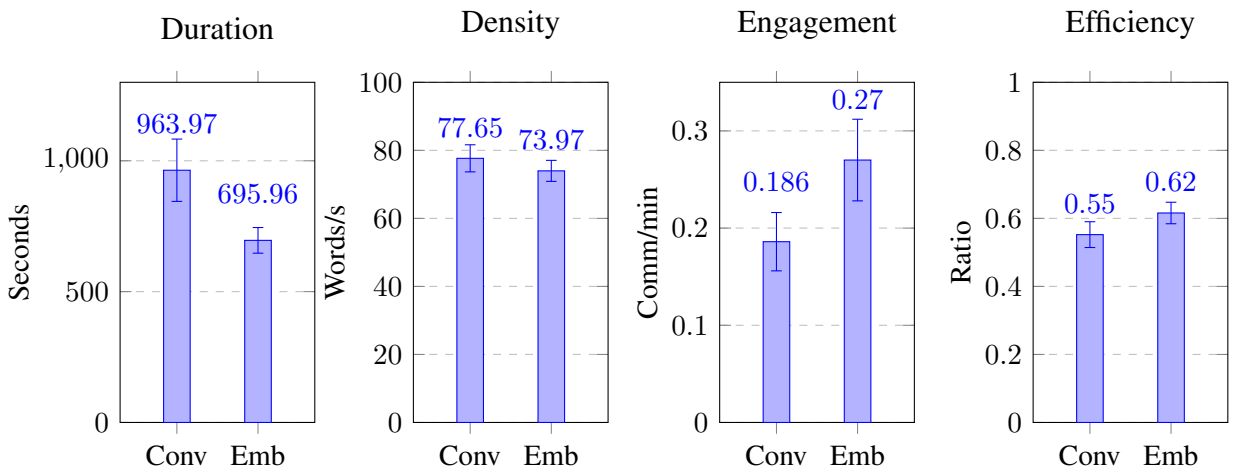


Figure 6: User study results for video duration, caption density, user engagement, and efficiency in user engagement with video content.

Conclusions

Our study has shown the difference of the user engagement for the conventional and embodied learning robotics instructional videos. The results have demonstrated the contrastable differences between the two videos in terms of the user engagement. The results are evident in showing the enhanced engagement among embodied learning videos.

Overall, in contrast to conventional videos, the embodied learning videos have more user comments showing that the embodied learning videos are more popular among learners. Furthermore, the ratio of the correlated user engagement versus the entire comments is slightly more pronounced among embodied learning videos in comparison with the conventional videos. Such results have been verified through both LLaVA and MiniCPM LLMs further reinforcing the results and conclusions.

The integration of both LLaVA and MiniCPM LLMs is also able to show promising results with higher confidence. The integration of the results have demonstrated the advantages and reliability of the results showing more correlated user comments with video content for both conventional and embodied learning videos. Notably, the integrated correlation ratio showed a statistically significant difference between embodied and conventional videos ($p = 0.008$, Cohen's $d = 1.455$, large effect), providing strong evidence that embodied learning videos achieve higher content–comment alignment. The LLaVA correlation ratio also exhibited a large effect size ($d = 0.885$), approaching significance ($p = 0.070$). It also shows that the embodied learning videos are associated with more positive reviews in comparison with the conventional videos. Fewer number of negative reviews are evident among the embodied learning videos. It thus shows the effectiveness of the use of embodied learning videos in comparison with the conventional videos.

The user study shows that there is increased user engagement with the embodied learning videos

in comparison with the conventional learning videos. Across the 130-video user study, the engagement rate ($d = 0.275$) and correlation efficiency ($d = 0.230$) both exhibited small positive effects favoring embodied videos. While these individual metrics did not reach statistical significance at $\alpha = 0.05$, the consistent directional trends across all engagement-related metrics collectively support the advantage of embodied learning videos. A stronger correlation between user comments and video content is evident. Furthermore, it is clear that there is increased user engagement efficiency among embodied learning videos, where the number of significant correlation (more than 60% correlation score) has been shown evident thus reinforcing the advantages in using the embodied learning videos for robotics instruction and workforce training.

The research shows that the use of embodied learning videos is insightful for engaging students in learning robotics. The statistical results show that there is significant difference between embodied learning and conventional learning videos thus indicating that the use of captions of the videos frame is able to correlated with the video content and potential student's interests and engagement with the videos. It therefore demonstrates that it is beneficial to engage students in learning robotics using videos which is known associated with strong visualization cues of the underlying topics. The study is limited in the use of relative small samples of videos thus motivating us for the use of larger-scale video dataset for future research.

Acknowledgment

This work was funded by National Science Foundation, Division Of Engineering Education and Centers, Grant/Award Number: 2306285.

Conflict of Interest

We do not declare the conflict of interests through the research.

References

[1] Amy Eguchi. Robotics as a learning tool for educational transformation. In *Proceeding of 4th international workshop teaching robotics, teaching with robotics & 5th international conference robotics in education Padova (Italy)*, volume 1, pages 1–6, 2014.

[2] Carlotta A Berry. Robotics education online flipping a traditional mobile robotics classroom. In *2017 IEEE Frontiers in Education Conference (FIE)*, pages 1–6. IEEE, 2017.

[3] J Berrío Perez. A "flipped classroom" for mobile robotics teaching. In *INTED2014 Proceedings*, pages 3076–3085. IATED, 2014.

[4] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, 111(23):8410–8415, 2014.

[5] Sofia Jusslin, Kaisa Korpinen, Niina Lilja, Rose Martin, Johanna Lehtinen-Schnabel, and Eeva Anttila. Embodied learning and teaching approaches in language education: A mixed studies review. *Educational Research Review*, 37:100480, 2022.

[6] Steven A Stolz. Embodied learning. *Educational philosophy and theory*, 47(5):474–487, 2015.

[7] Manuela Macedonia. Embodied learning: Why at school the mind needs the body. *Frontiers in psychology*, 10:2098, 2019.

[8] Marth Munro. Principles for embodied learning approaches. *SATJ: South African Theatre Journal*, 31(1):5–14, 2018.

[9] Carly Kontra, Susan Goldin-Meadow, and Sian L Beilock. Embodied learning across the life span. *Topics in cognitive science*, 4(4):731–739, 2012.

[10] Marianna Ioannou and Andri Ioannou. Technology-enhanced embodied learning. *Educational Technology & Society*, 23(3):81–94, 2020.

[11] Carly Kontra, Daniel J Lyons, Susan M Fischer, and Sian L Beilock. Physical experience enhances science learning. *Psychological science*, 26(6):737–749, 2015.

[12] Hala Khodr, Jerome Brender, and Aditi Kothiyal. How diseases spread: Embodied learning of emergence with cellulo robots. In *Proceedings of the 29th International Conference on Computers in Education (ICCE 2021)*, pages 51–60. Asia-Pacific Society for Computers in Education, 2021.

[13] Michael L Anderson. Embodied cognition: A field guide. *Artificial intelligence*, 149(1): 91–130, 2003.

[14] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645, 2008.

[15] Alexander Skulmowski and Günter Daniel Rey. Embodied learning: introducing a taxonomy based on bodily engagement and task integration. *Cognitive research: principles and implications*, 3:1–10, 2018.

[16] Chris Dede. The role of digital technologies in deeper learning. students at the center: Deeper learning research series. *Jobs for the Future*, 2014.

[17] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

[18] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.