Comparison of Text-Based Inputs for Human-in-the-Loop Feedback in Vision–Language Models

REZA SHAHRIARI, University of Florida, USA
AMAL HASHKY, University of Florida, USA
SHIVVRAT ARYA, New Jersey Institute of Technology, USA
TYLER AUDINO, University of Florida, USA
ERIC D. RAGAN, University of Florida, USA
VIBHAV GOGATE, University of Texas at Dallas, USA
JAIME RUIZ, University of Florida, USA

Human-in-the-loop methods leverage human feedback to enhance machine learning and artificial intelligence. Manual review of outputs can correct errors, identify model weaknesses, or expand labels to broaden model capabilities. Feedback collection methods range from simple flagging of outputs as correct or incorrect to more complex feature-level adjustments or natural language interpretations. This paper presents a user study evaluating changes in user performance over time and explores the trade-off between feedback quality and human effort. We compare four interactive input methods for reviewing and correcting outcomes in object detection and activity recognition in videos. Our findings indicate that while some complex input methods, such as free-text, require more time, the quality and impact of their feedback on model accuracy often surpass those of simpler methods that require less effort. However, more effort does not always lead to better-quality feedback, especially when aiming to improve the model. Our VLM experiments show that the most accurate models were trained using detailed natural language feedback or precise word-level corrections, while simple yes/no judgments also led to solid performance at a much lower annotation cost.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Human-in-the-Loop Feedback, User Study, Vision-Language Models

ACM Reference Format:
Reza Shahriari, Amal Hashky, Shivvrat Arya, Tyler Audino, Eric D. Ragan, Vibhav Gogate, and Jaime Ruiz. 2026. Comparison of Text-Based Inputs for Human-in-the-Loop Feedback in Vision–Language Models. *J. ACM* 1, 1, Article 1 (January 2026), 24 pages. https://doi.org/10.1145/3816700

1 INTRODUCTION

Human-in-the-loop (HITL) approaches combine human and machine intelligence to improve model accuracy and accelerate the achievement of desired performance levels [30]. Human feedback can be incorporated in various ways. One common method is active learning, where the model selects

Authors' addresses: Reza Shahriari, University of Florida, Gainesville, Florida, USA, rshahriari@ufl.edu; Amal Hashky, University of Florida, Gainesville, Florida, USA, ahashky@ufl.edu; Shivvrat Arya, New Jersey Institute of Technology, Richardson, Texas, USA, shivvrat.arya@njit.edu; Tyler Audino, University of Florida, Gainesville, Florida, USA, tyler.audino@ufl.edu; Eric D. Ragan, University of Florida, Gainesville, Florida, USA, eragan@ufl.edu; Vibhav Gogate, University of Texas at Dallas, Richardson, Texas, USA, vibhav.gogate@utdallas.edu; Jaime Ruiz, University of Florida, Gainesville, Florida, USA, jaime.ruiz@ufl.edu.

ACM 0004-5411/2026/1-ART1
https://doi.org/10.1145/3816700

its training inputs by querying specific domain areas. This feedback—whether binary or multi-class—allows the model to focus on the most uncertain instances, thereby reducing uncertainty [8]. HITL is also applicable to training methods such as few-shot learning (FSL), which requires the system to generalize from a small number of supervised examples from humans [44]. Additionally, in reinforcement Learning (RL), a training method based on rewarding desired behaviors and punishing undesired ones [19], human feedback has been shown to significantly improve algorithm efficiency [16].

Creating interactive machine learning (IML) systems that leverage the full range of human input and expertise requires understanding and implementing the best types of interaction. Previous research has highlighted the importance of designing user interfaces that bridge the gaps between human goals and the ML process [1, 36, 39, 42]. There are various methods to collect human input to enhance machine learning models [22], and each method may have varying degrees of influence on enhancing model accuracy. Each method offers distinct advantages and limitations, often balancing complexity with the required human effort or time against the depth and quality of feedback obtained. While it may be tempting to employ large-language models (LLMs) to support human effort by augmenting details of more simplistic human input, the introduction of new models for corrective feedback risks the introduction of new errors or hallucinations.

Different methods of feedback will require different amounts of human effort. Our research examines how observable response behaviors—such as response time and feedback quality—change as the effort required by different input methods varies. For example, more complex feedback methods may initially support more detailed or informative responses, but sustaining such effort over extended interaction may lead to changes in response behavior or performance consistency. Understanding these dynamics is essential to balancing feedback quality with user burden. When the cognitive cost of an input method outweighs its benefits, users may adapt their behavior in ways that affect the efficiency or reliability of feedback. Thus, optimizing input methods based on task complexity, duration, and feedback detail is crucial [9, 22, 23, 38].

To study these methods in a controlled and interpretable setting, we focus on a common human-in-the-loop task: text-based labeling and verification of AI-generated video descriptions [4, 45]. In this task, users review a video clip alongside a textual description produced by a vision-language model and provide feedback indicating whether the description is correct and, if not, how it should be revised. Although human feedback can be provided through many modalities (e.g., voice, sketches, demonstrations), this work intentionally restricts its scope to text-based input methods, which are widely used in large-scale annotation, auditing, and verification pipelines for vision-language systems [7, 32, 41]. Moreover, within this scope, we define an *input method* as the specific way in which users express feedback through text, ranging from low-effort evaluative judgments to more detailed corrective or explanatory input. Despite extensive work on human feedback in interactive machine learning, there is limited empirical understanding of how different text-based input methods compare in terms of human effort, feedback quality, and downstream model impact within repeated video description verification tasks.

Thus, we adopt a two-stage experimental design to capture both the human and model implications of different text-based input methods. In the first experiment, we analyze four input methods with varying levels of feedback detail to examine how user response time and feedback quality evolve over repeated interactions, reflecting differences in human effort and performance over time. In the second experiment, we build directly on these findings by evaluating how feedback collected via each input method influences downstream model learning. Importantly, more detailed or costly feedback for humans does not necessarily translate into better model performance. Together, this design allows us to compare input methods by jointly analyzing user behavior and model outcomes, clarifying how feedback detail, human effort, and learning effectiveness trade off in practice. Finally,

to explore opportunities for reducing annotation cost, we investigate whether structured feedback can be augmented using an LLM to generate natural language explanations, testing whether semantic richness can be synthesized post hoc to approximate the benefits of free-text input.

2 BACKGROUND

2.1 Human-in-the-Loop Machine Learning

Quality human annotation is essential for machine learning but can be challenging due to the risk of human errors in data labeling [30]. The impact of these errors varies significantly. For instance, critical applications like autonomous vehicles are highly sensitive to labeling mistakes, where even a minor error can lead to severe consequences, including accidents. Active learning is increasingly used to mitigate these issues, allowing the model to select its training inputs and reduce annotation costs by determining which data to label [8, 37]. Various methods have been developed to optimize this process.

Moreover, Interactive Machine Learning (IML) is a paradigm in which human users actively participate in the learning process, providing real-time feedback to enhance model accuracy [2, 5, 14, 25, 34, 43]. Early work in this field, such as that by Fails and Olsen [14], established foundational principles for involving end-users in the iterative training of machine learning models. They introduced the notion of "interactivity" in machine learning workflows, emphasizing the need for user-driven corrections and adjustments to fine-tune models based on real-world contexts and domain-specific knowledge. Building on this framework, Teso and Kersting [43] developed interactive mechanisms that allow users to intervene directly in the decision-making process by providing corrections or explanations. Their approach emphasized "explanatory interactions," where users can continuously refine the model's outputs and understand the reasoning behind its decisions. This interactive feedback loop is essential in systems that operate in dynamic environments, such as healthcare or finance, where human expertise is needed to navigate ambiguity and uncertainty.

Many studies have explored improving human-AI interaction [1, 12, 18, 26, 49]. A critical factor in this endeavor is studying how people interact and provide feedback in various scenarios. For instance, Dey et al. [10] found that asking specific and detailed questions in human-AI interactions resulted in better user responses than asking general questions. Combining this approach with elements such as showing uncertainty and requesting extra details can further improve response accuracy. However, since the human cognitive load is limited, it is important to take into account the potential impact on users' mental models and requirements during HITL interactions. For instance, interactions can affect user trust due to the awareness of system errors [17, 28]. In a study, Honeycutt et al. [17] developed a simulated object detection system that allowed participants to correct system errors by adjusting image regions for detected objects. The study found that allowing users to provide feedback can negatively impact their trust and perception of the system's accuracy, even if the system's performance improves.

In addition to user trust, user engagement can be impacted [20, 35]. For example, Dietvorst et al. [11] investigated the effect of allowing users to make slight modifications to algorithmic forecasts impacted their willingness to use the algorithms. Their findings revealed that allowing users to make minor adjustments to the forecasts significantly increased their willingness to use imperfect algorithms. The studies highlight the importance of incorporating user feedback mechanisms to effectively reduce algorithm aversion, where people prefer human judgment over imperfect algorithms and enhance human-algorithm interactions. Finally, while trust and engagement are considerable, it shows the importance of understanding user interaction when providing feedback.

Moreover, a human-in-the-loop topic modeling system was developed and evaluated to incorporate user-requested refinements and assess how human feedback affects user interaction [40]. The

authors found that participants preferred simple refinements, such as removing or adding words, which significantly improved the topic model's quality. However, challenges such as tracking complex changes and ensuring user confidence were identified. The study concluded that human feedback can enhance model quality and user engagement, emphasizing the need for user-centered feedback mechanisms that support user understanding and confidence in human-AI interactions [40].

2.2 Need for Diverse Feedback Collection Methods

Due to limitations in human memory and cognitive load, there is a trade-off between the human costs and the quality of feedback. For example, Cui et al. [9] investigated how various interaction types impact human performance, the quality of training data, and the overall learning outcomes of machine learning systems. By aligning feedback collection from human teachers with specific interaction types, the authors demonstrated how cognitive load and ease of use can influence data quality. The study results revealed that various types of interactions, such as *demonstrations*, *categorizing*, *sorting*, and *evaluating*, are consistently used to provide feedback on an agent's actions. These interactions can be effectively aligned with learning objectives to improve training, testing, and generalization performance in HITL systems. For example, *demonstrations*, where humans show the desired behavior and provide high-quality data, are very time-consuming and demanding. *categorizing*, which involves labeling or rating, is less demanding and quicker, offering structured data that are easier to collect.

Additionally, the study by Koppol et al. [23] focused on comparing different interaction methods in terms of usability and cognitive load. The findings revealed that high cognitive load and low usability have a negative impact on data quality. This aligns with the results of Cui et al. [9]. Both studies agree that *showing* and *categorizing* interactions are the least cognitively demanding and most usable, making them ideal for collecting high-quality data with minimal user burden. On the other hand, *evaluating* interactions are consistently found to be the most demanding and least user-friendly.

Although existing literature has demonstrated the effectiveness of HITL feedback and explored various methods for providing such feedback and their impact on cognitive load, there remains a gap in understanding how different levels of input method complexity perform over time. This gap can be addressed by evaluating user engagement metrics, such as response quality and time invested, at different stages of the HITL feedback process. This paper analyzes these metrics to assess the long-term sustainability and effectiveness of varying feedback inputs. This understanding helps design HITL systems that balance the quality of feedback with users' cognitive and temporal demands, ultimately leading to better-performing machine learning models.

3 EXPERIMENT MEASURING HUMAN FEEDBACK AND ENGAGEMENT

3.1 Research Goals

Different HITL feedback methods balance human effort and data quality. Some provide high-quality, rapid feedback initially, while others prioritize consistency of results over time. While detailed input methods may yield valuable responses early, they risk user boredom and declining quality. Understanding how these characteristics manifest over repeated feedback interactions is essential for designing input methods that are both efficient and sustainable for longer review tasks. Customizing interaction complexity to user needs ensures both short-term efficiency and long-term effectiveness, ultimately leading to more adaptable and durable input systems. To explore these challenges further, this study investigates the following research questions:

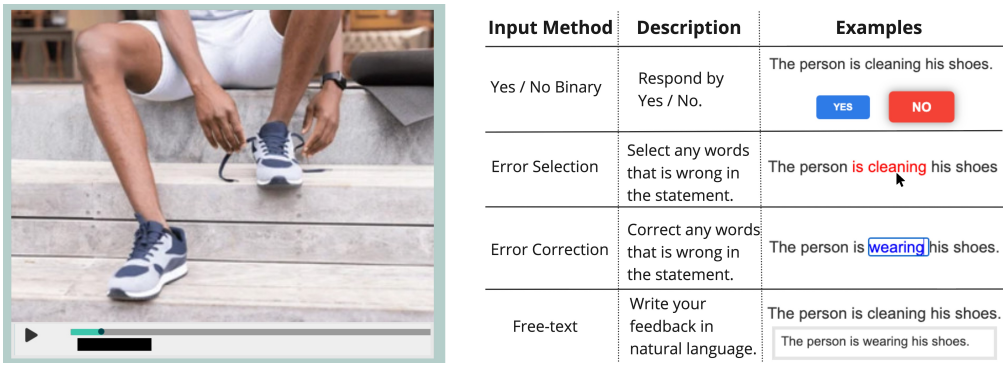

Fig. 1. The four input methods compared in the study involve different levels of feedback complexity.

- **RQ1**: How do response time and feedback quality as behavioral performance indicators reflect user effort and output consistency over repeated feedback inputs?
- **RQ2**: How does the trade-off between time and feedback quality vary across interaction types?

We hypothesize that user engagement and task accuracy decline over time in extended data review tasks, with steeper declines for more demanding input methods. The trade-off between human effort and feedback quality varies by interaction type, with some yielding higher-quality feedback at a greater cost. More complex input methods may reduce accuracy or increase response time, while simpler methods sustain performance by minimizing user effort. Alternatively, users might respond faster as they lose interest. Lastly, we expect more involved feedback to lead to greater improvements in model accuracy.

3.2 Experimental Design

To evaluate our hypothesis that input methods of varying complexity influence user behaviors and engagement, we selected a diverse set of methods, ranging from simple *yes/no* flagging to more complex *free-text* responses. This spectrum enables test scenarios that highlight differences in efficiency over time. The *input method* variable includes four variations (Figure 1): *Yes/No binary*, *error selection*, *error correction*, and *free-text*, differing in detail, ease of use, error identification accuracy, flexibility, user control, time required, and information richness. Prior work has identified several broad classes of human feedback, including showing (demonstration), categorizing or labeling, evaluating, correcting, and free-form explanation [9, 23]. Rather than exhaustively covering all possible interaction types, we focused on methods that (1) are directly applicable to text-based verification tasks, (2) vary primarily in input complexity rather than task structure, and (3) can be meaningfully compared along dimensions of human effort, feedback specificity, and downstream model utility.

The Yes/No binary method represents minimal evaluative feedback, commonly used in active learning and verification settings. Error selection extends this by requiring users to localize incorrect elements, corresponding to lightweight categorization and marking interactions. Error correction introduces explicit corrective supervision by allowing users to replace erroneous content, aligning with correction-based feedback studied in interactive learning. Finally, free-text input provides unconstrained natural language explanations, capturing the upper bound of feedback expressiveness and semantic richness. Together, these methods form an ordered progression from low-effort, low-information feedback to high-effort, high-information input.

Other interaction types discussed in prior work, such as demonstrations or ranking-based evaluations, were not included because they introduce fundamentally different task demands and interaction goals that are less compatible with sentence-level factual verification of AI-generated descriptions. Our goal was not to provide an exhaustive taxonomy of feedback methods, but to isolate how increasing feedback complexity within a consistent task context influences user behavior over time.

Moreover, Binary feedback simplifies the interaction process, allowing users to quickly provide input with minimal effort. This method is used in domains such as multiclass classification [29, 31], where humans can easily indicate whether a prediction is correct or incorrect, which results in reducing model uncertainty. With its simplicity, the *Yes/No binary* input method allows users to confirm whether a statement is accurate by answering yes or no. However, it does not provide any information about which parts of the statement are correct or incorrect.

Additionally, the *error selection* method requires users to identify specific incorrect elements or words in a statement. When users identify exact errors, the system can learn in a more targeted way. For example, in natural language processing tasks, users selecting incorrect words in generated text can help a model fine-tune its understanding of semantics or grammar [6]. Instead of general negative feedback, the system learns which elements are consistently causing mistakes and adapts accordingly. With this method, participants can select any number of words as incorrect or confirm the statement as correct without any selection. While this approach offers more precise error identification than binary feedback, it still lacks corrections for inaccurate information.

The *error correction* method enables participants to directly correct wrong information in the statement by selecting words in the description and typing new ones. It offers richer feedback than the above methods because it includes corrections. This method allows for correcting multiple words or validating the statement as correct without making modifications. Instead of leaving the system to infer corrections from binary or error selection feedback, it now has explicit correct answers to work with.

Finally, the *free-text* method offers the most flexibility, allowing users to rewrite descriptions, correct errors, add or remove information, or confirm accuracy. Unlike *error correction*, it enables adding new details beyond the original statement. Petrak et al. [33] showed that incorporating free-text feedback improves dialog system training by providing contextual insights. Additionally, analyzing the language and structure of human feedback can enhance model performance in processing natural language.

To analyze user engagement and feedback accuracy over time, we tracked each participant's behavioral patterns across multiple feedback instances. Each participant used the same assigned input method throughout the study, allowing us to observe interaction trends. To compare changes

Table 1. Independent variables for the 4x4 mixed-design user study

Independent Variables	**Levels**
Input Method (between-subjects)	1. Yes/No Binary 2. Error Selection 3. Error Correction 4. Free-text
Feedback Period (within-subjects)	1. First Period (Trials 1-25) 2. Second Period (Trials 26-50) 3. Third Period (Trials 51-75) 4. Fourth Period (Trials 76-100)

over time, we divided all trials into four equal intervals (*feedback periods*). As shown in Table 1, this segmentation was based on the number of trials rather than elapsed time to account for variability in review duration. Thus, the experiment followed a 4x4 mixed design with two independent variables: 1) *input method* (between subjects) and 2) *feedback period* (within subjects), each with four levels (see Table 1).

3.3 Task and Procedure

The study was Institutional Review Board (IRB) approved. Participants conducted the study online through an interactive web-based system. Participants first provided informed consent and completed a brief background questionnaire before being assigned to one of the four experimental conditions (Figure 1). Next, they received instructions and completed two example trials. To collect feedback over time, participants reviewed short video clips, each paired with an AI-generated textual statement (Figure 1).

This task was selected because it captures a common and practically relevant human-in-the-loop scenario in vision-language systems: verifying and correcting model-generated descriptions of visual content. Unlike classification or labeling tasks, this setting requires users to both evaluate correctness and, when necessary, provide corrective feedback, making it well-suited for comparing input methods with varying levels of expressiveness.

Additionally, video description verification introduces inherent ambiguity and multi-component structure (e.g., subject, action, object, and context), which allows us to systematically assess how different feedback modalities support error detection and correction at different levels of granularity. For instance, a description may be partially correct (e.g., correct action but incorrect object), requiring users to localize and fix specific components rather than reject the entire statement. This makes it a representative testbed for studying trade-offs between feedback effort and quality. This task closely reflects real-world workflows in dataset curation, model auditing, and alignment evaluation for vision-language models, where human reviewers are often required to validate and refine automatically generated outputs.

Each clip included a temporal indicator corresponding to the specific segment of the activity recognition statement, allowing precise mapping of participant feedback to exact moments in the clip. Also, for the provided textual statements, we adapted descriptions from the STAR dataset [47] for corresponding short video clips. This dataset was chosen for its rich detail, requiring sustained attention and cognitive effort, making them ideal for studying engagement variations.

Moreover, to simulate real-world variability, description accuracy was intentionally varied: 50% were entirely correct, 25% contained one error, 15% had two errors, and 10% were completely incorrect. This balanced distribution ensured all participants encountered a mix of description accuracies, allowing us to assess how fatigue and boredom over the 30-minute session influenced error detection and response.

Participants reviewed and provided feedback on up to 100 clip descriptions within a strict 30-minute window, balancing data collection needs with online study constraints. This time limit minimized dropout and extreme fatigue while allowing systematic tracking of feedback quality and response times. The high clip volume enabled us to capture fatigue and boredom effects. Interaction logs were also collected anonymously to ensure privacy. Participants completed the study in a single session via an online web app, engaging at their own pace without researcher intervention. Participation was voluntary, and extra credit was offered as compensation.

3.4 Measures

For each video clip reviewed, we recorded 1) *response time* and 2) the feedback provided. Response time was measured as the number of seconds it took a participant to review the video and description

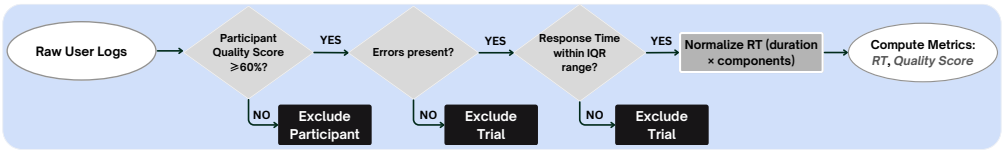

Fig. 2. Data processing pipeline applied prior to analysis (Details in Section 3.5).

and submit their feedback, serving as a direct indicator of the amount of human effort required to provide feedback. Other measures were calculated using the submitted feedback. The amount and type of recorded feedback varied according to the assigned feedback method: The *Yes/No binary* method captured simple affirmative or negative responses. The *error selection* method logged the specific words participants identified as incorrect. The *error correction* method included both the words selected as errors and the corresponding corrections made by participants. Lastly, the *free-text* method collected detailed free-text responses.

Because the study aimed to compare the accuracy and utility of the feedback collected from the different input methods, the study needed a quality score that could be calculated for each input method. To this end, we designed a *quality score* to assess the accuracy of responses as well as the units of information within each submission. The *quality score* was determined by comparing participants' feedback to a predefined standard for each video description. For the *Yes/No binary* case, possible scores were limited to 1 or 0, as the feedback lacked further information. For all other input methods, scoring was based on component-level correctness. Each statement was broken into fundamental components (e.g., "A person puts a pile of clothes on the floor" includes: Subject: "A person," Action: "puts," Object: "pile of clothes," and Place: "the floor"). Feedback accuracy was assessed based on participants' decisions to maintain, correct, remove, or add components. One point was awarded for each correct decision, including preserving valid components, correcting errors, or adding relevant details.

We calculated the *quality score* for each feedback by dividing the total number of correct decisions by the overall number of statement components, with scores typically ranging between 0 and 1. However, in the *free-text* and *error correction* conditions, scores could exceed one if participants added additional accurate components. Following a consensus among all authors on the scoring metrics, three coders performed the manual scoring. We maintained open communication to ensure consistency and revisited and discussed ambiguous cases to resolve discrepancies.

3.4.1 Quality Score Definition. To support consistent comparison across input methods, we formalize the quality score computation as follows. Let $S = \{c_1, c_2, \ldots, c_n\}$ denote the set of components in the original statement, where each component corresponds to a semantic unit (e.g., subject, action, object, location). For a given feedback instance, we define a correctness indicator for each component:

$$\delta(c_i) = \begin{cases} 1, & \text{if the participant correctly preserved, corrected, or removed } c_i \\ 0, & \text{otherwise} \end{cases}$$

The quality score Q for a feedback instance is computed as:

$$Q = \frac{\sum_{i=1}^{n} \delta(c_i)}{n}$$

3.5 Data Processing

As our goal is to compare how users provide feedback and how their performance changes over time, we applied several pre-processing (Figure 2) steps to improve comparability across trials and reduce noise unrelated to task behavior. First, trials with zero errors were excluded from the primary analysis. Error-free trials require little or no corrective interaction, and therefore reflect confirmation rather than active feedback generation. Including these trials would artificially lower average response times and obscure differences between input methods that specifically require user edits. Our analyses focus on feedback behavior; therefore, we restrict comparisons to trials that required user interaction.

Second, response times can be influenced by factors unrelated to cognitive effort (e.g., interruptions, pauses, or switching tasks), particularly in remote online studies. To reduce the influence of such artifacts, we removed extreme outliers using a standard interquartile range (IQR) criterion. Specifically, any trial with a response time greater than $Q_3 + 1.5 \times IQR$ was excluded. These outliers accounted for 7% of trials and exhibited substantially higher variance than typical responses.

Additionally, since the number of components in a statement and the duration of a video clip may affect response time, we normalized response times by the number of components and clip duration to allow fair comparisons across trials and participants. Also, to verify that these pre-processing choices did not meaningfully influence the findings, we repeated the analyses without removing outliers and including zero-error trials. The overall trends, relative differences between input methods, and statistical conclusions remained consistent. These steps, therefore, serve primarily to reduce noise rather than alter outcomes. Finally, we averaged quality scores and normalized response times across clips for each participant to compute two aggregate behavioral metrics: *Response Time* and *Quality Score*.

3.6 Participants

We conducted the experiment with 149 participants, setting a minimum average quality score of 60% to ensure data reliability. Thirteen participants were excluded, leaving a final sample of 136 after outlier removal. Participants were assigned to one of four interaction conditions, though exclusions led to the following distribution: *Yes/No Binary* (35), *Error Selection* (31), *Error Correction* (35), and *Free-text* (35). Ages ranged from 18 to 39, with a median of 20 years. The final group included 75 self-reported males, 56 females, 3 non-binary individuals, and 3 who chose not to disclose.

3.7 Results

The experiment studies patterns in user feedback over time based on speed and quality of feedback for the different input methods. The raw data did not meet the assumptions of normality and similar variances expected for parametric testing. We, therefore, utilized the ARTool [46] for nonparametric factorial analysis by the aligned rank transformation of the data, enabling the use of factorial parametric testing. We conducted two-way mixed ANOVA tests to account for effects due to (i) periods of time over the study duration and (ii) the four input methods for submitting input. For post-hoc analysis of significant main effects, we used paired-t tests with Tukey correction available within the ARTool [13]. We report test results along with partial eta squared (η_p^2) for effect sizes of ANOVA tests and Cohen's d for effect sizes of post-hoc tests.

3.7.1 Response Time. Time results are shown graphically in Figure 3, and Table 2 summarizes the results of the statistical analysis for *input method* and *feedback period*. As expected, input methods that require more feedback take significantly longer times. The post-hoc results show the *yes/no binary* condition was significantly faster than each of the other methods. *Error selection* was also

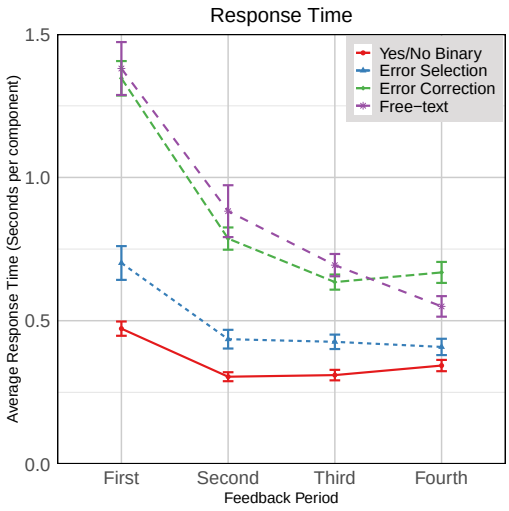

Fig. 3. Average human response time across feedback periods for different input methods (error bars show standard error).

Response Time	p-value	Effect Size
Main Effect of Input Method		
$F(3, 132.69) = 94.9$	$< 0.001^*$	$\eta_p^2 = 0.68$
Post hoc (Sig. Pairs)		
Free-text > Yes/No Binary	$< 0.001^*$	$d = 4.79$
Free-text > Error Selection	$< 0.001^*$	$d = 3.29$
Error Correction > Yes/No Binary	$< 0.001^*$	$d = 4.69$
Error Correction > Error Selection	$< 0.001^*$	$d = 3.19$
Error Selection > Yes/No Binary	$< 0.001^*$	$d = 1.49$
Main Effect of Feedback Period		
$F(3, 361.64) = 243.43$	$< 0.001^*$	$\eta_p^2 = 0.67$
Post hoc (Sig. Pairs)		
First > Second	$< 0.001^*$	$d = 2.24$
First > Third	$< 0.001^*$	$d = 2.86$
First > Fourth	$< 0.001^*$	$d = 2.99$
Second > Third	$< 0.001^*$	$d = 0.62$
Second > Fourth	$< 0.001^*$	$d = 0.75$
Interaction Effect of Input Method × Feedback Period		
$F(9, 361.48) = 32.3$	$< 0.001^*$	$\eta_p^2 = 0.45$

Table 2. ANOVA and Posthoc Tukey HSD Test Results for Input Method Differences on Response Time (* indicates $p < 0.05$).

significantly faster than both the *error correction* and *free-text* conditions. No statistically significant difference was detected for time results between *free-text* and *error correction* conditions.

The overall response time also significantly decreased over the study duration (Figure 3). In addition, a significant interaction effect indicates the decline over time may be more substantial for more-involved input methods requiring text entry (*error correction* and *free-text*). While response time drops over time for all methods, Figure 3 shows the rate levels out for most input methods. After the largest change between the first and second periods, the changes between subsequent periods are notably smaller. The behavior for the *free-text* shows an exception to this behavior for the final change (i.e., from third to fourth periods), as the response time continues to drop more than the other input methods. Figure 3 also shows average response times for *free-text* to be slightly above *error correction* for the first three periods; this changes in the fourth period, where posthoc testing for the significant interaction effect found *free-text* to have significantly faster times than *error correction* ($p < 0.001$, Cohen's d = 0.62). This crossover suggests that once users become familiar with the task, composing free-text explanations may be more efficient than performing structured corrections, which require locating and editing multiple specific components.

3.7.2 Quality Score. Figure 4 shows the quality score results. Table 3 shows statistical results showing a significant main effect on quality scores for both *input method* and *feedback period* as well as a significant interaction effect between the two factors. Figure 4 indicates that higher quality scores are generally associated with more demanding input methods, and scores decrease over time periods. Importantly, we interpret quality scores as behavioral performance indicators that reflect the amount and consistency of actionable feedback provided by users, rather than purely as measures of correctness. The *free-text* condition had a significantly higher quality score than all other methods, which shows that—as would be expected—more feedback information was collected when using input methods that allow greater flexibility.

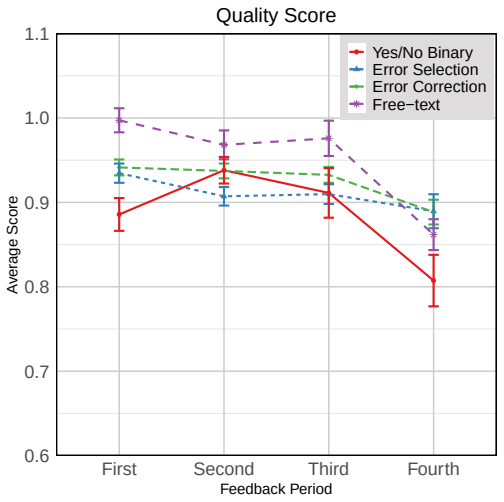

Fig. 4. Average score across feedback periods for different input methods (error bars show standard error). The minimum score of 0.6 was enforced for quality control.

Quality Score	p-value	Effect Size
Main Effect of Interaction Method		
$F(3, 132.77) = 7.06$	$< 0.001^*$	$\eta_p^2 = 0.14$
Post hoc (Sig. Pairs)		
Free-text > Yes/No Binary	$< 0.001^*$	$d = 0.83$
Free-text > Error Selection	$< 0.01^*$	$d = 0.79$
Free-text > Error Correction	$< 0.05^*$	$d = 0.59$
Main Effect of Feedback Period		
$F(3, 366.04) = 33.61$	$< 0.001^*$	$\eta_p^2 = 0.22$
Post hoc (Sig. Pairs)		
First > Fourth	$< 0.001^*$	$d = 1.12$
Second > Fourth	$< 0.001^*$	$d = 1.12$
Third > Fourth	$< 0.001^*$	$d = 1.11$
Interaction Effect of Input Method × Feedback Period		
$F(9, 365.56) = 7.07$	$< 0.001^*$	$\eta_p^2 = 0.15$

Table 3. ANOVA and Posthoc Tukey HSD Test Results for Input Method Differences on Score (* indicates $p < 0.05$).

The significant decrease in scores over time was primarily seen in the last period of the study session (further supported by a significant post-hoc difference between period four and all prior periods). However, the drop over time also varies across input methods, showing a significant interaction effect that corresponds with the previously noted significant interaction effect on response time. While the *free-text* method initially had higher scores in the first three periods, its score dropped significantly from the first to the last period ($p < 0.001$, Cohen's d = 2.03) and fell below *error selection* and *error correction*). This indicates the possibility of a larger drop in quality for more demanding input methods relative to a drop in other methods.

Furthermore, the variation in score changes in each period reveals that the *yes/no binary* method exhibits considerable variation, as overlooking a single detail can lead to the loss of an entire score point. On the other hand, *error selection* and *error correction*, which prompt users to pay closer attention to each word, resulted in fewer changes over time.

4 ANALYZING FEEDBACK SPECIFICITY IN HUMAN-IN-THE-LOOP CALIBRATION OF VISION-LANGUAGE MODELS

While the *human experiment* examined how different input methods affect human performance over repeated interactions, human effort alone does not determine the overall effectiveness of a human-in-the-loop system. In practice, the value of feedback also depends on how well it improves downstream model learning. Feedback that is faster or easier for users to provide may not necessarily produce better model updates, whereas more detailed input may offer stronger learning signals at a higher human cost. To complete this loop, we therefore evaluate in this experiment how feedback collected through each input method impacts the performance of vision-language models.

Specifically, this section investigates how varying levels of human feedback affect the calibration of vision-language models (VLMs) for verifying AI-generated video descriptions—a critical task for evaluating factual alignment between visual content and textual output. Building on the human-in-the-loop methods described earlier, we assess how the specificity of feedback influences VLM

accuracy. Our study explores the trade-off between the informational richness of different feedback modalities and their impact on model reliability, focusing on the integration of human guidance to improve alignment predictions. We detail the experimental setup, including VLM architectures and training protocols, and evaluate performance across multiple feedback types.

To support this investigation, we propose a post-hoc feedback conversion pipeline that employs a Large Language Model (LLM) to translate structured annotations—error corrections, span selections, and binary judgments—into natural language sentences. This approach aims to retain the expressiveness of free-text feedback while reducing annotation overhead. We assess the effectiveness of these LLM-generated textual proxies in VLM fine-tuning and compare them against models trained directly on the original structured feedback.

Our analysis is guided by the following research questions:

- **RQ3**: How do different types of feedback affect the accuracy of VLMs in the factual alignment task?
- **RQ4**: Can LLM-based conversion of low-performing structured feedback enhance the calibration accuracy of vision-language models?

4.1 Experimental Setup

We conducted an experiment to evaluate how different forms of human feedback, as characterized in our user study, affect VLM performance on the task of identifying inaccurate video descriptions.

4.1.1 Task Definition. The task requires a VLM to determine whether an AI-generated textual description accurately represents the content of a given video clip. Each instance consists of a video, a potentially inaccurate AI-generated description, and structured human feedback.

4.1.2 Data, Models, and Prompting. We adapted video clips and their corresponding initial descriptions from the dataset collected in Section 3, based on 100 clips sampled uniformly at random from the STAR dataset [47]. No explicit stratification was applied with respect to clip length or description complexity. This sampling strategy was chosen to preserve the natural distribution of clips encountered in the user study. Finally, the sampled videos were then randomly partitioned into three subsets: 80 clips for fine-tuning the VLMs, 10 for validation, and 10 for testing.

For evaluation, we selected two state-of-the-art vision-language models: LLaVA-NeXT-Video [51] and Video-LLaVA [27], both known for strong performance in multimodal understanding. Each VLM received input via a standardized prompt: "`Video Description: [description] User Feedback: [feedback] Question: Does the above information accurately describe the video? Answer Yes or No.`" This formulation casts the task as a binary classification problem, enabling consistent assessment of the impact of different feedback types. Figure 5 illustrates example interactions with the VLMs using different feedback types.

4.1.3 Feedback Modalities and Training Configurations. We evaluated six training configurations, each corresponding to a distinct feedback modality, to assess their impact on VLM performance and to establish upper and lower performance bounds:

- *True Model*: Trained exclusively on correctly matched video-description pairs, serving as an upper bound on achievable performance.
- *No Feedback Model*: Fine-tuned on AI-generated descriptions without any human feedback, some of which contain factual errors. This setting defines the lower bound.
- *Yes/No Binary*: Augmented training data with binary user judgments indicating whether the AI-generated description was correct or incorrect.
- *Error Selection*: Included user feedback identifying specific erroneous spans within the AI-generated text.

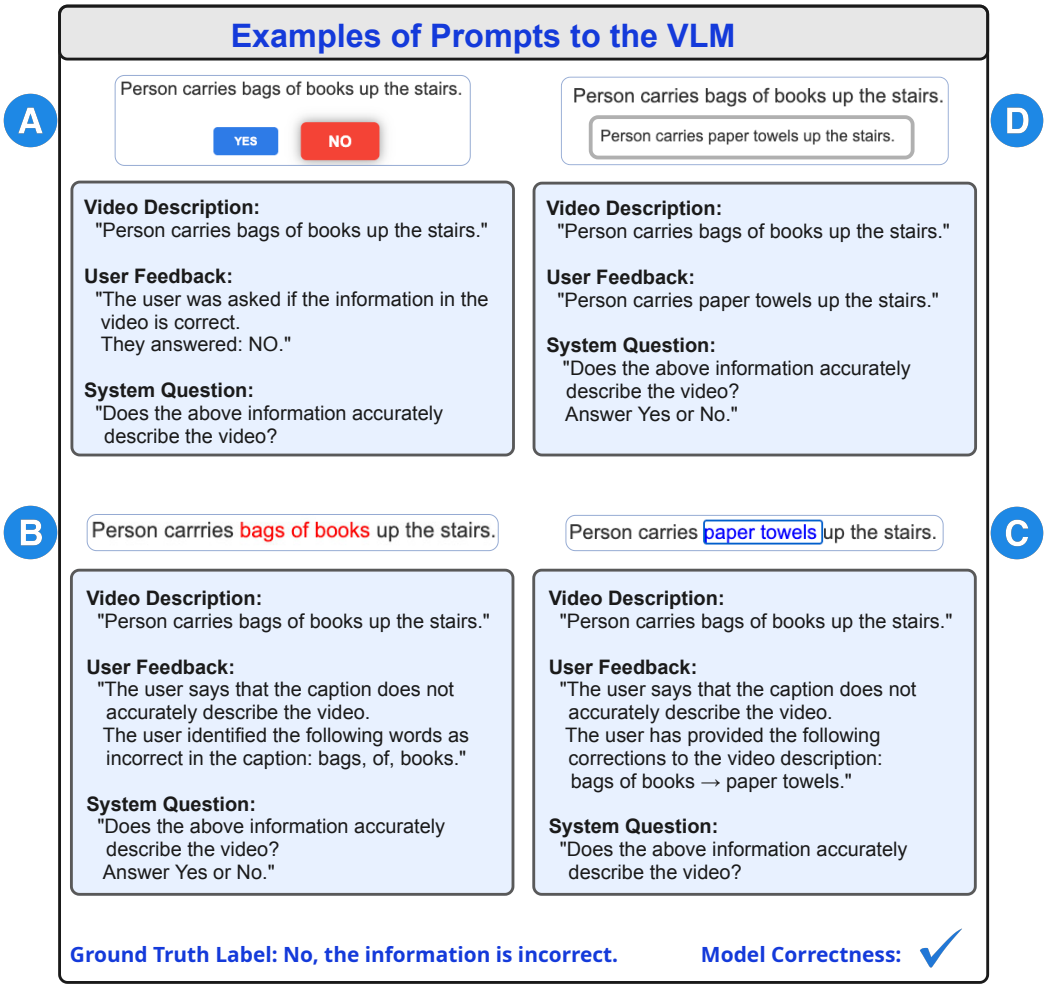

Fig. 5. Examples of prompts shown to the Vision-Language Model (VLM) across four feedback conditions: A Binary Yes/No, B Error selection, C Error correction, and D Free-text. Each prompt includes a video description, user feedback, and a system question. These examples illustrate how user-provided feedback was converted into structured prompts, clarifying both the input and the expected system action for each type of input method. In C, the user feedback shown uses the Unicode arrow symbol (→) to indicate a correction (e.g., "bags of books → paper towels"). This formatting was introduced during prompt construction for the VLM to enhance clarity and model performance. Importantly, participants in the user study did not submit feedback in this format; they provided corrections through a structured UI (Shown in Figure 1).

- *Error Correction*: Incorporated targeted, word-level corrections, mapping erroneous terms to their accurate counterparts as provided by users.
- *Free-Text*: Leveraged natural language feedback, allowing users to elaborate on inaccuracies or confirm correctness.

These configurations span a spectrum of feedback granularity, enabling systematic analysis of how the precision and richness of human input influence model calibration and verification

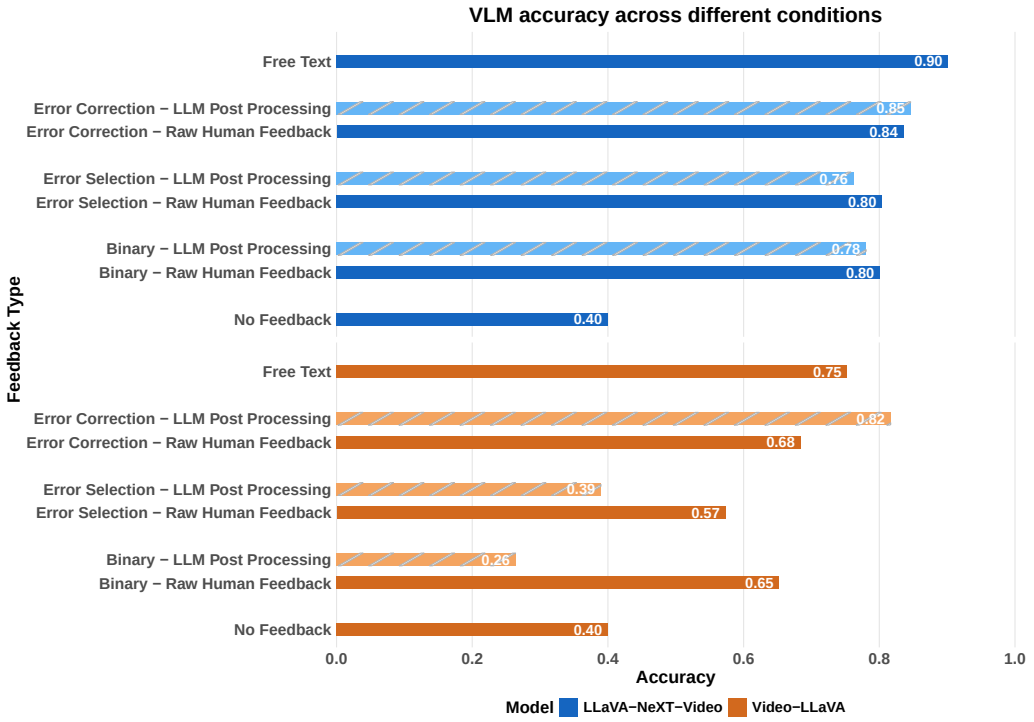

Fig. 6. Comparison of model accuracy across different types of human feedback in VLMs. Performance generally improves with more involved feedback. While LLM post-processing enhances accuracy for *error correction*, it leads to decreased performance for *error selection* and *yes/no binary feedback*.

accuracy. Together, they cover all feedback types evaluated in the user study and allow direct comparison against both the *No Feedback Model* (no feedback) and the *True Model* (trained on ground-truth descriptions).

4.1.4 Fine-tuning Protocol. We fine-tuned the VLMs using QLoRA with rank-8 LoRA adapters applied to all linear layers, excluding the language modeling head and multimodal projector. Training was performed with 4-bit quantization. We used the AdamW optimizer with a learning rate of 2×10^{-5}, and applied cosine learning rate scheduling with a 10% warmup phase. Models were trained for 100,000 steps with gradient accumulation over 16 steps.

4.2 Performance Analysis and Impact of Feedback Modality

Figure 6 presents the comparative accuracy of LLaVA-NeXT-Video and Video-LLaVA under the six training configurations.

Both VLMs display consistent performance trends across feedback modalities, indicating that the impact of feedback generalizes across architectures. The *True Model* achieves perfect accuracy, confirming that these models are capable of reliable video-text verification when trained with fully accurate supervision. In contrast, the *Baseline Model*—fine-tuned on AI-generated descriptions without human feedback—performs worst, underscoring the detrimental effect of training on noisy or incorrect data in the absence of corrective guidance.

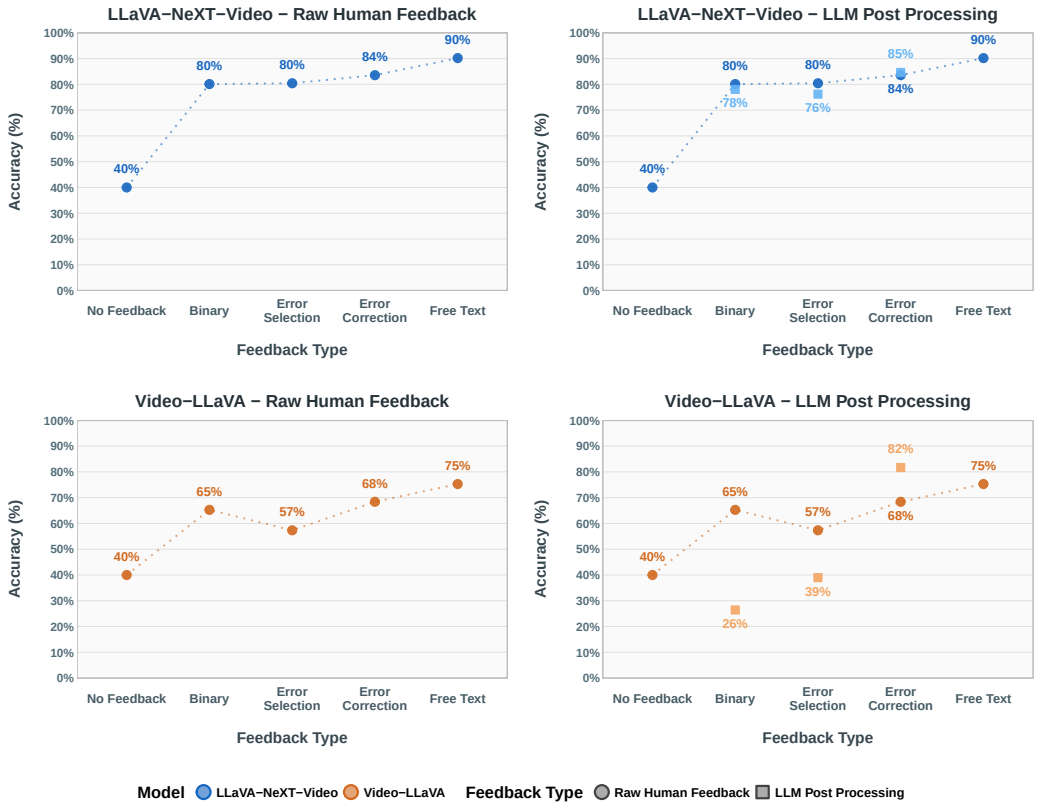

Fig. 7. Accuracy of two VLMs under four human feedback types. Performance improves with richer feedback. LLM post-processing helps *error correction* but degrades *binary* and *error selection.*

Incorporating human feedback consistently improved model accuracy over the baseline, with performance gains generally aligned with the specificity and informativeness of the feedback modality (Figure 6).

Free-text feedback led to the highest accuracy improvements across both VLMs. This modality offers unconstrained natural language input, providing the richest corrective signal. Although it required more annotator time over extended interactions (Figure 3), the resulting gains in model performance justify the additional cost.

Error Correction feedback, which offers direct word-level substitutions for erroneous tokens, ranked second in performance impact. Although *free-text* feedback offered the greatest boost in model accuracy (Figure 6), its difference in terms of time taken to get that boost was not statistically significant from *error correction* (Figure 3). Yet *free-text* feedback still earned a markedly higher accuracy than all other methods. This indicates that, while gathering free-text feedback demands more annotator time, it yields richer, more detailed inputs that can lead to stronger performance improvement.

The comparative effectiveness of *Error Selection* and *Yes/No Binary* feedback revealed a more model-dependent trend (Figures 6 and 7). *Yes/No Binary* feedback was more effective for Video-LLaVA, whereas *Error Selection* produced marginally better results for LLaVA-NeXT-Video. Notably, binary feedback required substantially less annotator time (Figure 3), underscoring its efficiency.

One contributing factor to its competitive performance is the clarity of the binary supervision signal, which can guide learning without introducing ambiguity. In contrast, feedback that highlights individual errors may lead to partial mislabeling or annotator disagreement about which segments are incorrect, thereby injecting noise into the training signal. Simple correct/incorrect judgments avoid this issue, reduce model uncertainty, and require minimal processing, as they do not involve correction content—enabling more stable and consistent performance gains.

4.3 Converting Structured Feedback to Free-Text via an LLM

Collecting free-text feedback from annotators is costly (Figure 3); yet, our analysis in Section 4.2 showed that this modality yields the highest accuracy. We therefore investigate whether *post-hoc* conversion of *structured* feedback—*error-correction* edits, *error-selection* spans, and *Yes/No binary* judgments—into text sentences using a Large Language Model (LLM) can recover some benefits of free-text feedback while maintaining low annotation overhead.

4.3.1 Conversion pipeline. For each feedback, we prompt the LLAMA 3 model [15] to rewrite the original description into a self-contained sentence, conditioned on the structured cues. In the *error-correction* setting, the LLM replaces each erroneous token with the corresponding user-provided correction. For *error-selection* and *binary* feedback, the model receives either the marked erroneous spans or a binary yes/no label and revises the AI-generated video description accordingly to reflect the feedback.

The LLM-rewritten outputs are treated as proxies for *free-text* feedback and format them using the prompt template shown in Figure 5. Using this format, we finetune the VLM following the procedure outlined in Section 4.1; all other hyper-parameters remain unchanged.

4.3.2 Results. Figures 6 and 7 compare models trained on original structured feedback (solid bars) with those trained on LLM post-processed *text* feedback (hatched bar). The impact of this conversion on model performance varies significantly across feedback types.

Converting *error correction* feedback to *text* improves accuracy for both VLMs—from 83.6% to 84.6% for LLaVA-NeXT-Video and more substantially from 68.4% to 81.7% for Video-LLaVA. This gain indicates that *error correction* feedback contains sufficient information to be effectively translated into richer *text* guidance, which the model can leverage more effectively during fine-tuning.

In contrast, converting *error selection* feedback into text feedback leads to performance degradation: -4.2 % for LLaVA-NeXT-Video and -18.3 % for Video-LLaVA. Qualitative analysis reveals that the LLM often generates syntactically fluent but semantically incoherent revisions, likely due to insufficient context, which introduces noise into the learning signal.

The sharpest decline occurs with *Yes/No Binary* feedback, where converting yes/no labels into text feedback results in drops of -2.1 and -38.9 percentage points for the two models, respectively. Lacking concrete semantic cues, the LLM resorts to generating plausible-sounding but often incorrect justifications, thereby increasing label noise instead of reducing it.

5 DISCUSSION

5.1 Interpretation of Results

5.1.1 Human Experiment. The study compared four different input methods, each varying in detail and feedback complexity, to examine how user accuracy and response times evolve over the course of the study. Our objective was to explore the balance between the human effort involved and the value of the feedback received, enabling us to choose input methods needed depending on the task effectively. While it is expected that methods allowing users to provide more information will allow

collection of a greater amount of potentially useful feedback data, our results show that the amount of user effort and the quality and detail of feedback may vary. As illustrated in Figure 3 and Table 2, the cost of data collection over time is an important consideration due to the significant drop in response time over extended periods of feedback collection. It can be observed that more involved methods—such as the *free-text* and *error correction* methods—took more time, but the increased effort and boredom led to a larger decrease in response time as users tend to spend less time over time.

Interestingly, although the *free-text* method exhibited longer response times than *error correction* during the first three periods, this relationship reversed in the final period, where free-text became faster (Figure 3). At the same time, its quality scores declined relative to both *error selection* and *error correction* (Figure 4). Together, these patterns suggest a shift in user behavior over repeated interaction: as participants became more familiar with the task, free-text input may have been produced more quickly but with less detail or precision, whereas structured correction required consistent, step-by-step edits that maintained higher quality but limited speed gains. This divergence indicates that increased expressiveness does not necessarily translate into sustained feedback value over time and highlights a practical trade-off between flexibility and consistency. From a design perspective, these results suggest that free-text input may be advantageous for early, exploratory feedback, while more structured correction mechanisms may better support stable, higher-quality input during prolonged use.

This was in a collection period capped at 30 minutes. These results indicate that the initial higher response time for *free-text* may have led to fatigue or decreased focus, ultimately impacting its overall quality. Considering the significant interaction effect, this drop in quality score for *free-text* suggests that if the feedback period had continued, it is even possible the *free-text* feedback quality may have decreased further. Therefore, while gathering more involved feedback can yield higher-quality responses, practical considerations for human effort are crucial for optimizing the trade-offs of potential high-quality information, cost of human time, and possible variation in human attention or effort with extended duration. A drop in quality over time could result in the collection of low-quality feedback, which would waste time for annotators and cause inconsistency for AI developers.

Based on the results from our study, although *free-text* inputs enabled users to provide much more detailed feedback, the level of consistency was lower, which we expect to be attributed to the increased effort due to the need to type out a sentence for each instance. For engaged participants, the method also involved deciding whether to provide simple corrections, add details, or use entirely new descriptive phrasing for the content. With the higher demands of open-ended input, users may have become disengaged over time. Conversely, *error selection* and *error correction* yielded more consistent quality, as they reduced the potential for components that could be selected or changed. Both methods supported greater focus on specific parts of the given description.

In contrast, the quality score results for *yes/no binary* input is more varied. While the response time for the *yes/no binary* method is significantly lower than for others, the quality score is not significantly lower, demonstrating the high value that binary responses can still provide. However, the considerable variation in changes to quality scores indicates that missing even a single detail can have a higher penalty for the value of the provided data. The impact of a mistake might be larger, and the lack of ability to provide explanations or additional information along with the feedback could result in the risk of less useful responses for ambiguous cases for the model.

5.1.2 VLM Experiments. Our VLM experiments moved our study of feedback methods beyond understanding of human behaviors and feedback quality to evaluate potential differences when attempting to improve models with the different forms of feedback. The experiments investigate

how feedback specificity modulates model calibration and accuracy across two architectures, LLaVA-NeXT-Video and Video-LLaVA, under six training configurations (Figure 6). The results reveal a direct correlation between the semantic richness of human feedback and the resulting model's verification accuracy, establishing a clear performance hierarchy among different human-in-the-loop strategies.

We first establish performance bounds. A *True Model*, trained on ground-truth data, achieves perfect accuracy, confirming the VLMs' capacity for the task. In contrast, a *No Feedback Model* trained on uncorrected AI-generated descriptions performs poorly, demonstrating the risk of error reinforcement without supervision. All forms of human feedback improved accuracy relative to this lower bound, validating the human-in-the-loop approach.

Across both VLM architectures, performance scales with feedback granularity. Consistent with its informational richness, *free-text* feedback yields the most significant accuracy gains, affirming that its high annotation cost translates directly to maximal model improvement. *Error correction* offers a structured yet highly effective alternative. Notably, the competitive performance of *Yes/No binary* feedback over *Error selection*, particularly for the Video-LLaVA architecture, indicates that a simple, unambiguous correctness signal is a highly efficient calibration method for specific models due to its minimal annotation cost.

We further analyze the nature of feedback value by attempting a *post-hoc* conversion to free-text format using an LLM. The successful conversion of semantically rich *error-correction* feedback demonstrates that structured data can be "up-leveled" to mimic the benefits of natural language, boosting model accuracy. Conversely, this process fails for semantically sparse *error-selection* and *binary* feedback. The LLM lacks sufficient context, hallucinates justifications, and introduces noise that degrades performance. This result underscores a core principle: the value of feedback lies in its semantic content, not its syntactic format. LLM-based conversion is only viable when the source feedback is already information-dense, as the initial quality of human input is paramount and cannot be retroactively engineered.

Collectively, these findings highlight a clear trade-off between feedback specificity, annotation cost, and model accuracy. For maximal performance, the higher cost of *free-text* or *error correction* is justified. For large-scale applications, *binary* feedback may offer the optimal cost-benefit ratio. The optimal strategy is therefore contingent on specific architectural properties and resource constraints. Finally, automated conversion pipelines can reduce effort only when upstream feedback is already information-dense; otherwise they risk propagating errors and should be avoided.

Taken together, our results reveal that the trade-off between human effort and feedback value is not a simple linear relationship, but instead depends on how feedback structure aligns with both human behavior and model learning. While more expressive input methods such as *free-text* and *error correction* provide richer and more informative signals, they also impose higher effort and time, leading to reduced consistency over time. In contrast, low-effort methods such as *yes/no binary* feedback offer stable and efficient interaction, and despite their simplicity, still provide strong signals that can effectively guide model learning. This suggests that feedback efficiency is not solely determined by the amount of information provided, but also by the clarity and reliability of that information under repeated use.

From the model perspective, however, a different pattern emerges. When high-quality feedback is sustained, more expressive inputs such as *free-text* and *error correction* lead to greater improvements in model performance, as they provide more explicit and informative signals about errors and their corrections. At the same time, simpler methods such as *yes/no binary* feedback remain effective, particularly when large volumes of consistent responses can be collected efficiently. For example, a high volume of reliable *binary* feedback may be more beneficial in practice than a smaller set of detailed but inconsistent explanations. These results suggest that there is no universally

optimal feedback method. Instead, the choice depends on the application context where high-effort, information-rich input is preferable when accuracy is critical, and data is limited, whereas low-effort input is more suitable for scalable or time-constrained settings. Overall, the trade-off is not solely between effort and quality, but between consistency, scalability, and the level of detail required for effective model improvement.

5.2 Implications for Feedback Collection

The tradeoffs we identified between different types of human feedback have direct implications for how HITL systems can be designed in practice. In various domains, HITL has been adopted to overcome the limitations of fully manual or fully automated approaches [3, 21, 24, 48, 50]. For example, in large-scale image dataset construction, HITL methods were used to reduce the burden of manual labeling by combining deep learning with human review, where models filtered easy cases and humans focused only on ambiguous images, significantly accelerating the process [50]. Also, in natural language processing, systems that generate paraphrases rely on humans not just to accept or reject outputs, but also to rank and rewrite them, helping models learn diverse and human-aligned language patterns [3]. These examples demonstrate how feedback design, ranging from minimal binary input to rich rewriting, plays a critical role depending on task complexity, resource constraints, and accuracy requirements. Our study helps make sense of these choices by showing how different types of feedback yield different benefits for both human effort and model performance.

Additionally, our study builds on findings from prior research, such as Koppol et al. [23], which identified that more detailed feedback methods, like free-text input or their *evaluating* approach, yield higher quality feedback initially. Their method requires users to critically assess and provide specific comments, which enhances the granularity of feedback. However, as they highlighted, this approach can lead to reduced consistency over time due to the increasing cognitive load. Our findings extend this understanding by explicitly focusing on the trade-off between user effort and feedback quality over time, with particular attention to how different input methods influence user accuracy and response time as tasks progress. While Koppol et al. [23] focused on cognitive load, usability, and performance across interaction types, our study examined the trade-off between human cost and feedback quality explicitly, focusing on how user accuracy and response time evolve over time with repeated use of various types of inputs and levels of detail. We found that while methods like *error selection* and *error correction* are initially effective, they can sustain feedback quality longer for more extended tasks.

This contrasts with Koppol et al. [23] recommendation for simpler methods like categorizing or showing, which, though less cognitively demanding, may not support long-term engagement as effectively when detailed feedback is required. This suggests that there is a threshold where balancing the duration of the task, input detail, and understanding these patterns can help maintain the quality of feedback. Moreover, Koppol et al. [23] focused on comparing interaction types in terms of cognitive load, usability, and user performance, they did not investigate how the different feedback types affect downstream model learning or perform any model update experiments. In contrast, our study explicitly examines the trade-off between user effort and feedback quality over time, including how these inputs influence model calibration and learning outcomes.

5.3 Limitations and Future Work

A major factor for consideration is the people who are reviewing data and providing feedback. The level of commitment, motivation, and engagement of the individuals is critical—especially for collecting high-quality, accurate data over extended periods of time. A limitation of our study is that it only evaluated feedback behaviors by participants who were not directly incentivized by

the performance of the hypothetical model. This is similar, however, to many practical situations where data annotators or data workers are simply doing a job without particular attachment to the project's goals. However, consideration of other cases is also important when interpreting the results from our study. For instance, for feedback collection cases where the data annotators are also the creators or the end-users of the technology, we might see a more consistent level of effort and feedback quality over time. Still, even heavily-invested people are still only people, and the limitations of human attention and fatigue will apply. Future work should also study trade-offs among different types of feedback collection while accounting for different levels of user engagement or incentives.

Another limitation concerns the scope of interaction modalities evaluated. This study intentionally focused on four text-based input methods that represent a spectrum of feedback granularity, from simple evaluative judgments to corrective and explanatory responses. While other interaction forms, such as demonstrations, ranking, sketches, or voice input, are also common in human-in-the-loop systems, these modalities introduce qualitatively different interaction paradigms and task demands that are not directly comparable within a sentence-level verification task. Our goal was therefore not to exhaustively cover all possible interaction types, but to isolate trade-offs within a consistent and widely used text-based feedback setting. Future work should extend this analysis to additional modalities to examine whether similar effort–quality trade-offs generalize beyond text-based interactions.

Also, we acknowledge a limitation related to the design of our quality metric, which was intentionally structured to reflect the informational granularity of each feedback modality. Specifically, we evaluated Yes/No responses at the full-statement level and allowed free-text and error-correction inputs to contribute additional components when they introduced new corrective details. While these choices better capture how downstream models would realistically use such feedback, they may create differences in the scoring scale across modalities and thus influence absolute quality values. Our analysis, therefore, emphasizes relative trends and comparisons over time rather than absolute score magnitudes. Future work could investigate alternative normalization strategies or component-matched scoring schemes to examine further the robustness of these findings under different evaluation assumptions.

We should also note that we focus on a single task: text-based verification and correction of AI-generated video descriptions. While this task is representative of many real-world human-in-the-loop workflows in vision-language systems, it does not capture the full diversity of feedback scenarios in interactive machine learning. Different tasks, such as image classification, ranking, summarization, or open-ended generation, may involve different behavioral patterns. As a result, the trade-offs observed in this study between feedback effort, quality, and model impact may not generalize uniformly across all domains. Future work should extend this analysis to a broader range of tasks and modalities to better understand how task characteristics interact with feedback design.

6 CONCLUSION

This paper presented an analysis of human-in-the-loop (HITL) feedback collection methods, evaluating how input method complexity affects both user engagement and machine learning outcomes. Through a large-scale user study (n=149), we compared four feedback modalities, ranging from simple binary inputs to unconstrained *free-text* to assess changes in response time and feedback quality over time. We found that while more detailed input methods like *free-text* and *error correction* initially yield higher-quality feedback, they also incur greater effort and demonstrate sharper performance drops due to user fatigue. Simpler input methods like *yes/no binary* responses maintain stable engagement and deliver surprisingly competitive feedback quality with minimal effort. Complementing the user study, we fine-tuned two vision-language models (VLMs) on feedback

collected from each input modality to evaluate their impact on model calibration. The results show that feedback specificity positively correlates with model accuracy: *free-text* and *error correction* inputs produced the most accurate models. However, we also found that the benefit of post-hoc feedback conversion using a Large Language Model (LLM) is highly dependent on the semantic richness of the original feedback, effective for detailed corrections but harmful when applied to sparse inputs like binary labels. Together, these findings underscore the trade-offs between feedback richness, human effort, and model performance. Effective HITL systems must carefully balance annotation cost with desired learning outcomes. In practice, the best feedback strategy may vary by task: low-effort methods are more scalable, while high-effort inputs are preferable when accuracy is critical.

ACKNOWLEDGMENTS

This work was supported by the DARPA ECOLE Program under award number HR00112390063.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Crystal Butler, Harriet Oster, and Julian Togelius. 2020. Human-in-the-loop ai for analysis of free response facial expression label sets. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*. 1–8.

[4] Lele Cao. 2025. A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content. *arXiv preprint arXiv:2504.02898* (2025).

[5] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–8.

[6] Po-Lin Chen, Shih-Hung Wu, Liang-Pu Chen, and Ping-Che Yang. 2016. Improving the selection error recognition in a Chinese grammar error detection system. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*. IEEE, 525–530.

[7] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14239–14250.

[8] David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning* 15 (1994), 201–221.

[9] Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca Fitzgerald. 2021. Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In *International Joint Conference on Artificial Intelligence*.

[10] Anind K Dey, Stephanie Rosenthal, and Manuela Veloso. 2009. Using interaction to improve intelligence: how intelligent systems should ask users for input. In *Workshop on Intelligence and Interaction: IJCAI*.

[11] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science* 64, 3 (2018), 1155–1170.

[12] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.

[13] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*. 754–768.

[14] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny

Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu,

Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[16] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).

[17] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.

[18] Liu Jiang, Shixia Liu, and Changjian Chen. 2019. Recent research advances on interactive machine learning. *Journal of Visualization* 22 (2019), 401–417.

[19] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[20] Hyunjin Kang and Chen Lou. 2022. AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement. *Journal of Computer-Mediated Communication* 27, 5 (2022), zmac014.

[21] Bongjun Kim and Bryan Pardo. 2018. A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–23.

[22] Pallavi Koppol, Henny Admoni, and Reid Simmons. 2020. Iterative interactive reward learning. In *Participatory Approaches to Machine Learning, International Conference on Machine Learning Workshop, Virtual.*

[23] Pallavi Koppol, Henny Admoni, and Reid G Simmons. 2021. Interaction Considerations in Learning from Humans.. In *IJCAI*. 283–291.

[24] Eric Krokos, Hsueh-Chen Cheng, Jessica Chang, Bohdan Nebesh, Celeste Lyn Paul, Kirsten Whitley, and Amitabh Varshney. 2019. Enhancing deep learning with visual interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 1 (2019), 1–27.

[25] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[26] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. 2023. Assessing human-ai interaction early through factorial surveys: A study on the guidelines for human-ai interaction. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–45.

[27] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *Conference on Empirical Methods in Natural Language Processing* (2023). https://doi.org/10.48550/arXiv.2311.10122

[28] Jie Liu, Kim Marriott, Tim Dwyer, and Guido Tack. 2023. Increasing user trust in optimisation through feedback and interaction. *ACM Transactions on Computer-Human Interaction* 29, 5 (2023), 1–34.

[29] Evan Lucas, Steven Whitaker, and Timothy C Havens. 2022. Online learning with binary feedback for multi-class problems. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 374–380.

[30] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Simon and Schuster.

[31] Hung Ngo, Matthew Luciw, Jawas Nagi, Alexander Forster, Jürgen Schmidhuber, and Ngo Anh Vien. 2014. Efficient interactive multiclass learning from binary feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 3 (2014), 1–25.

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[33] Dominic Petrak, Nafise Sadat Moosavi, Ye Tian, Nikolai Rozanov, and Iryna Gurevych. 2023. Learning From Free-Text Human Feedback–Collect New Datasets Or Extend Existing Ones? *arXiv preprint arXiv:2310.15758* (2023).

[34] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal*

of Human-Computer Studies (2024), 103301.

[35] Jeba Rezwana and Mary Lou Maher. 2022. Understanding user perceptions, collaborative experience and user engagement in different human-AI interaction designs for co-creative systems. In *Proceedings of the 14th Conference on Creativity and Cognition*. 38–48.

[36] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John A Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C North, and Daniel A Keim. 2017. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* 268 (2017), 164–175.

[37] Burr Settles. 2009. Active learning literature survey. (2009).

[38] Reza Shahriari, Eric D Ragan, and Jaime Ruiz. 2025. Natural Language Interaction for Editing Visual Knowledge Graphs. In *Proceedings of the 13th Knowledge Capture Conference 2025*. 26–34.

[39] Reza Shahriari, Yichi Yang, Danish Nisar Ahmed Tamboli, Michael Perez, Yuheng Zha, Jinyu Hou, Mingkai Deng, Eric D Ragan, Jaime Ruiz, Daisy Zhe Wang, et al. 2025. MuCHEx: A Multimodal Conversational Debugging Tool for Interactive Visual Exploration of Hierarchical Object Classification. *IEEE Computer Graphics and Applications* (2025).

[40] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*. 293–304.

[41] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems* 33 (2020), 3008–3021.

[42] Simone Stumpf, Erin Sullivan, Erin Fitzhenry, Ian Oberst, Weng-Keen Wong, and Margaret Burnett. 2008. Integrating rich user feedback into intelligent user interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 50–59.

[43] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.

[44] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

[45] Chloe Wittenberg, Ziv Epstein, Adam J Berinsky, and David G Rand. 2024. Labeling AI-generated content: promises, perils, and future directions. (2024).

[46] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[47] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*.

[48] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[49] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[50] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015).

[51] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/