# Transparent and Coherent Procedural Mistake Detection

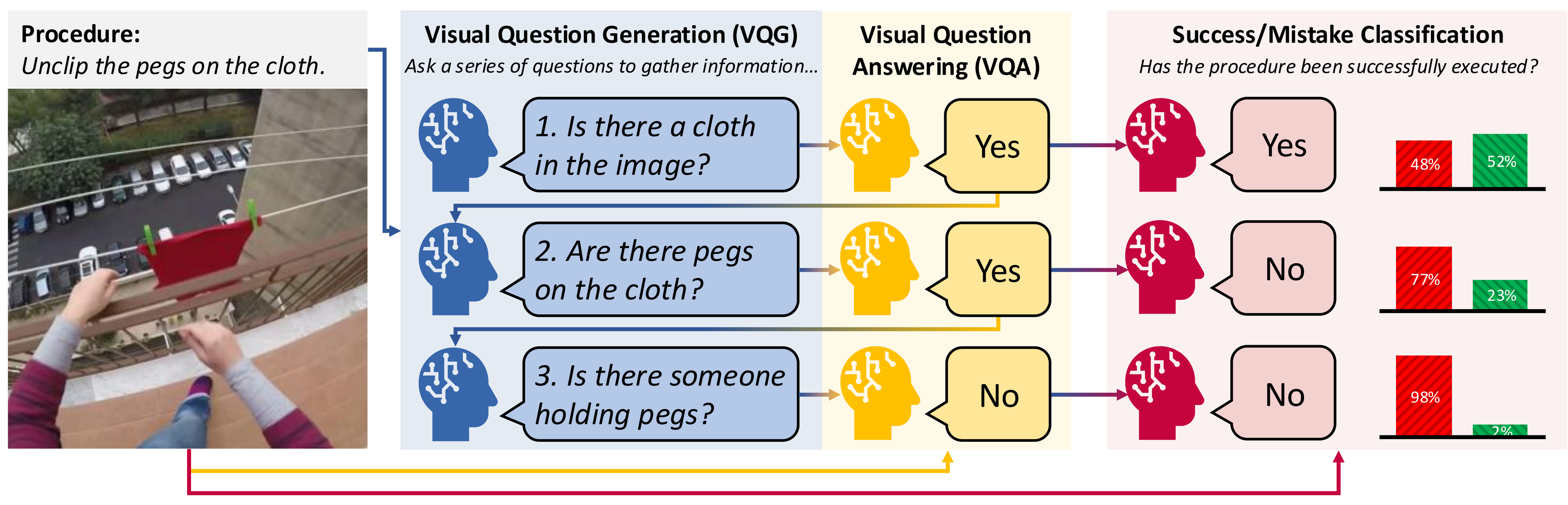
Shane Storks, Itamar Bar-Yossef, Yayuan Li, Zheyuan Zhang,
Jason J. Corso, & Joyce Chai (University of Michigan)

## Motivation

**Procedural mistake detection (PMD)** requires classifying whether a human (seen through egocentric video) has successfully executed a task (specified by a procedural text). Despite significant efforts, VLM performance in the wild is nonviable, and underlying knowledge and reasoning processes are opaque.

## Coherent Procedural Mistake Detection

We reformulate PMD to require a self-reflective dialog rationale from VLMs:

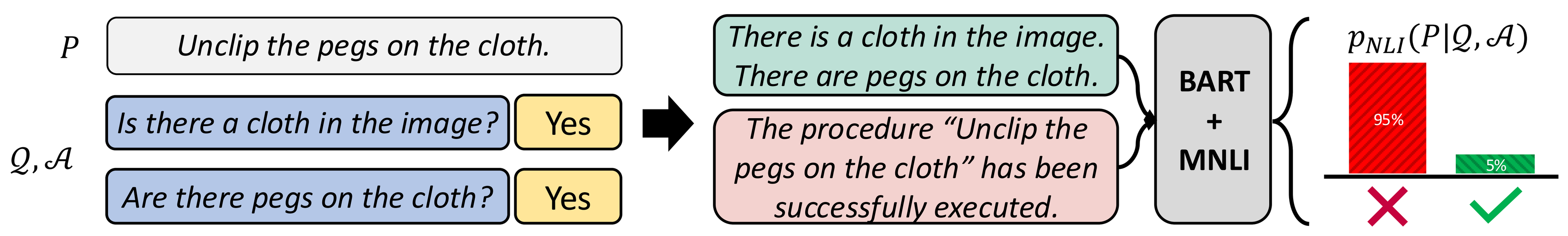



We generate diverse video frame mistake detection data from Ego4D [1]:


| Example Type | Train | (Sample) | Validation | (Sample) | Test | (Sample) |
|---|---|---|---|---|---|---|
| Success | 42,013 | 5,000 | 13,058 | 250 | 18,057 | 1000 |
| Mistake | 99,401 | 5,000 | 25,423 | 250 | 34,182 | 1000 |
| *Incomplete* | 15,057 | 755 | 4,908 | 51 | 6,545 | 194 |
| *Wrong V* | 11,780 | 604 | 2,694 | 31 | 3,747 | 108 |
| *Wrong N* | 36,434 | 1,853 | 8,914 | 87 | 11,843 | 344 |
| *Wrong V & N* | 36,130 | 1,788 | 8,907 | 81 | 12,047 | 354 |

## Rationale Coherence Metrics

To evaluate whether evidence collected from VLMs suggests success, we leverage fine-tuned natural language inference (NLI) models:




We use the NLI model $p_e$ to measure the **relevance** of a question $Q'$ to the success of a procedure $P$, given previous questions $\mathcal{Q}$ and answers $\mathcal{A}$:

$$\texttt{Rel}(Q'|T,\mathcal{Q},\mathcal{A}) = |p_e(T|\mathcal{Q}\cup Q',\mathcal{A}\cup No) - p_e(T|\mathcal{Q}\cup Q',\mathcal{A}\cup Yes)|$$

Relevance is summarized by example through a mean over questions:

$$\frac{1}{n}\sum_{i=1}^{n}\texttt{Rel}(Q_i|T,\{Q_j\colon j<i\},\{A_j\colon j<i\})$$

We also measure the **informativeness** of a predicted answer $A'$ for $Q'$:

$$\texttt{Inf}(A'|Q',T,\mathcal{Q},\mathcal{A}) = 1 - H(p_e(T|\mathcal{Q}\cup Q',\mathcal{A}\cup A'))$$

Reference-adjusted informativeness $\texttt{Inf}^*$ is negated if the most likely success label in $p_e$ disagrees with the ground truth label $y^*$. It is summarized by example through the maximum informativeness achieved:

$$\max_{1\leq i\leq n}\texttt{Inf}^*(A_i|Q_i,T,\{Q_j\colon j<i\},\{A_j\colon j<i\},y^*)$$

## References

[1] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[3] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

## Links

Paper | Code | Me | Lab

## Experimental Results

We apply 2 interventions to the selection of candidate questions generated by LLaVA-1.5 [2] through a greedy beam search:

1. **Coherence-based re-ranking** of candidate questions
2. **In-context learning** (ICL) from 20 sets of human-written questions

| Ranking | ICL | Acc. ↑ | Rel. ↑ | Inf. ↑ | # Iter. ↓ | Info. Gain ↑ |
|---|---|---|---|---|---|---|
| Likelihood | ✗ | 60.7 | 40.3 | .259 | 3.25 | .435 |
| Likelihood | ✓ | 61.8 | 36.5 | .272 | 3.34 | .429 |
| Coherence | ✗ | 61.4 | 66.5 | .321 | **3.06** | .540 |
| Coherence | ✓ | **67.8** | **75.5** | **.464** | 3.46 | **.663** |

We then **fine-tune LLaVA for question generation with our coherence metrics**, using DPO [3] over question pairs generated from training data:

| Ranking | ICL | Acc. ↑ | Rel. ↑ | Inf. ↑ | # Iter. ↓ | Info. Gain ↑ |
|---|---|---|---|---|---|---|
| Likelihood | ✗ | 62.2 | 75.7 | .318 | 2.33 | .617 |
| Likelihood | ✓ | 63.7 | 58.5 | .330 | 2.67 | .548 |
| Coherence | ✗ | 62.3 | 92.2 | **.340** | 2.06 | .719 |
| Coherence | ✓ | **64.2** | **95.0** | .304 | **1.81** | **.742** |

## Performance Analysis

Our metrics provide global and local insights into VLM performance: