ConfusionBench: An Expert-Validated Benchmark for Confusion Recognition and Localization in Educational Videos

Lu Dong*, Xiao Wang*, Mark Frank, Srirangaraj Setlur, Venu Govindaraju, Ifeoma Nwogu
* Equal contribution | University at Buffalo, SUNY, USA.

University at Buffalo

NATIONAL AI INSTITUTE for Exceptional Education

CVPR JUNE 3-7, 2026 DENVER COLORADO

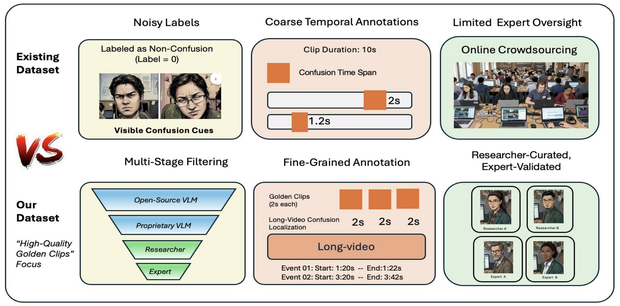
Introduction & Motivation

Confusion in Educational AI

Confusion is a cognitive-affective state during learning, yet it remains underexplored in educational AI. Reliable detection can support intelligent tutoring systems and expert intervention.

Limitations of Current Dataset

✗ Noisy labels — confusion cues appear in "non-confusion" labeled clips
✗ Coarse 10-sec temporal annotations — confusion cues last ~1–2 sec
✗ Large-scale crowdsourcing — limited expert oversight



Confusion Cues (FACS)

Primary: AU4 (brow lower), AU7 (eyelid tighten), AU4 + AU7
Secondary: AU10 (lip raise), AU24 (lip press)
Hand-to-face: chin touching, forehead pressing
Body posture: forward lean, restless movement
Exclusion: smiling (in most cases)

ConfusionBench Construction

① **Two-Second Clip Segmentation**
10-sec DAiSEE clips → 5 non-overlapping 2-sec clips
→ 45,000 candidate clips generated

② **Coarse VLM Screening (Qwen3-VL-4B-Instruct)**
Local inference on NVIDIA RTX 3090 GPU
4-level prediction: None / Low / Medium / High
→ 3045 source segments retained (Medium + High + Low)

③ **Fine-Grained VLM Screening (Gemini 3 Flash)**
Confidence-aware majority voting over 5 runs
High/Med/Low weighted: 3/2/1
→ 1,348 2-sec clips selected

④ **Researcher Curation**
Behavioral protocol: AU4, AU7, AU10, AU24, hand-to-face, posture cues. Two independent researchers, agreement required → 301 clips selected for expert annotation

⑤ **Expert Validation**
Two facial expression & human cognition experts
224 ✓ Yes · 46 ? Unsure · 31 ✗ No
→ 224 golden confusion clips confirmed

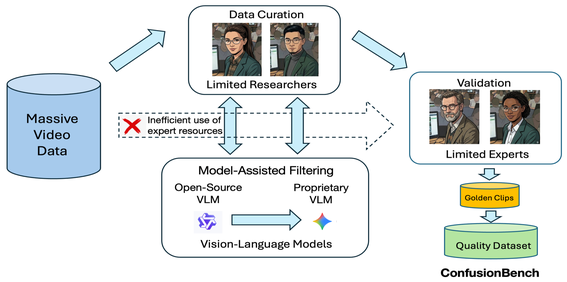


Figure 1 . ConfusionBench Multi-stage Construction Pipeline

Confusion Datasets

ConfusionBench — Final Dataset Statistics

➢ Balanced recognition set:
450 clips (224 Yes, 226 No)
➢ Long-video localization set:
10 five-minute videos for confusion localization
Both datasets are continuously being expanded.

Clip-Level Recognition Experiments

Model	Acc	Prec	Rec	F1
Qwen3-VL-4B	0.691	0.782	0.527	0.629
Gemini 3 Flash ★	0.798	0.751	0.888	0.814

Key Findings

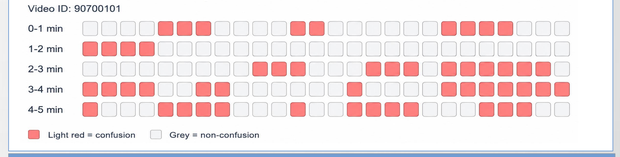
★ Gemini outperforms Qwen by large margin (F1: 0.814 vs. 0.629, +18.5%)
▶ Qwen: conservative predictions, low recall (0.527) → many missed confusions
▶ Gemini: higher sensitivity (recall 0.888), better detection of subtle cues
Both models show precision ≈ 0.75–0.78, suggesting reliable positive predictions when flagging confusion.

Video-Level Localization Experiments

Table 2. Zero-shot long-video confusion localization results using a proprietary VLM (Gemini 3 Flash Preview).

Source	Acc	Prec	Rec	F1	GT_ev	Pr_ev	tIoU	P@.1	R@.1	F1@.1	P@.3	R@.3	F1@.3
001	0.9067	0.8684	0.7857	0.8250	4	10	0.7021	0.4000	1.0000	0.5714	0.2000	0.5000	0.2857
002	0.7467	0.9111	0.5467	0.6833	11	19	0.5190	0.5263	0.9091	0.6667	0.3684	0.6364	0.4667
003	0.7200	0.7273	0.9072	0.8073	17	19	0.6769	0.6842	0.7647	0.7222	0.5263	0.5882	0.5556
004	0.6267	0.6116	0.8916	0.7255	17	23	0.5692	0.6087	0.8235	0.7000	0.4783	0.6471	0.5500
005	0.6800	0.5926	0.9412	0.7273	16	22	0.5714	0.3636	0.5000	0.4211	0.3182	0.4375	0.3684
006	0.9000	0.7903	0.9608	0.8673	11	20	0.7656	0.5000	0.9091	0.6452	0.5000	0.9091	0.6452
007	0.6867	0.6327	0.8493	0.7251	11	22	0.5688	0.4091	0.8182	0.5455	0.3636	0.7273	0.4848
008	0.7133	0.7407	0.3571	0.4819	7	19	0.3175	0.2632	0.7143	0.3846	0.1053	0.2857	0.1538
009	0.5933	0.8727	0.4706	0.6115	6	25	0.4404	0.1600	0.6667	0.2581	0.1200	0.5000	0.1935
010	0.3200	1.0000	0.3200	0.4848	1	22	0.3200	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[Micro-Avg]	**0.6893**	**0.7289**	**0.6612**	**0.6934**	**101**	**201**	**0.5307**	**0.3831**	**0.7624**	**0.5099**	**0.2985**	**0.5941**	**0.3974**
[Macro-Avg]	**0.6893**	**0.7747**	**0.7030**	**0.6939**	**10.1**	**20.1**	**0.5451**	**0.3915**	**0.7106**	**0.4915**	**0.2980**	**0.5231**	**0.3704**

Student Confusion Report Visualization

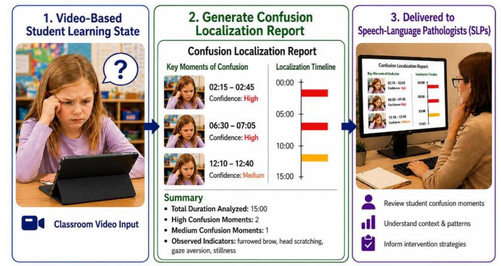


Conclusions & Future Work

▶ Expert-validated benchmark for confusion recognition and temporal localization in educational videos.
▶ Multi-stage pipeline balance annotation quality with limited expert effort.
▶ Gemini outperforms Qwen, but both remain far from reliable deployment.
▶ Confusion visualization report supports fast, interpretable expert intervention.

Future Work:
➢ Expand the dataset with more long videos and fine-grained confusion stage annotations.
➢ Develop an agentic confusion analysis system for educational intervention support.

Educational Application



Contact
LU DONG
Ph.D. University at Buffalo, SUNY, USA
Email: ludong@buffalo.edu
Website: https://dongludeeplearning.github.io/
Phone: 716-7300429

SCAN ME

Acknowledgement
This work is based upon work supported under the National AI Research Institutes program by the National Science Foundation (NSF) and the Institute of Education Sciences (IES), U.S. Department of Education, through Award #2229873 and NSF Award #2223507.

Project Page
https://dongludeeplearning.github.io/ConfusionBench.html

Project Page

Archival Paper: CV4Edu-20

CVPR JUNE 3-7, 2026 DENVER COLORADO CV4Edu