Delta-Gated Incremental Multi-Forward-Pass Modeling for Robust Multimodal Classroom Video Understanding

Chongyu He1; Peter Youngs2; Scott T. Acton1
1 ECE, University of Virginia 2 CISE, University of Virginia

CVPRW 2026 | CV4Edu-17 | Archival Paper

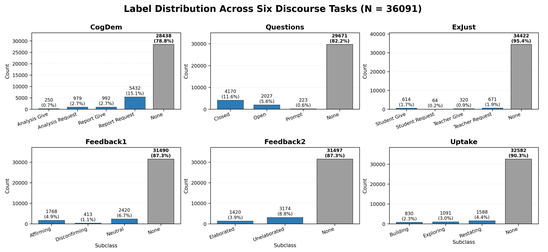
Takeaway: decompose multimodal fusion, trust reliable modality gains, and explain which cues changed the prediction.

Abstract

In noisy classrooms, do audio/video help or mislead the transcript model?

- Classroom recordings are noisy, missing, and weakly aligned.
- We estimate contribution and reliability, not just fusion.
- DG-MFP isolates audio/video gains and gates trusted evidence.

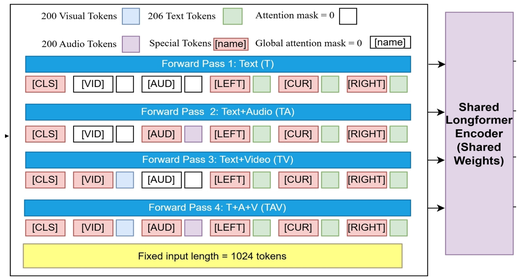
Data	**AIAIS; 107 videos; 36,091 samples**
Tasks	6 tasks, 25 labels
Signal	Text + HuBERT + CLIP
Metric	F1 + macro-F1 robustness



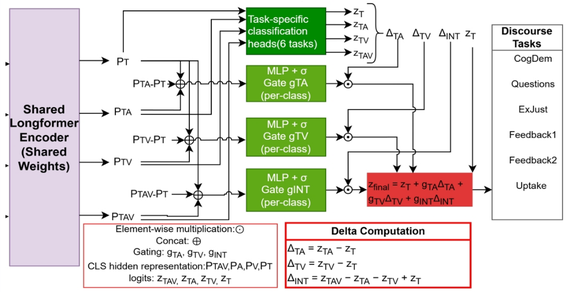
Long-tailed labels make rare-class robustness the real test.

Core idea
T = transcript text; A = audio; V = video.
Use T as the baseline prediction.
Compare TA, TV, and TAV passes to isolate added signal.
Logit deltas estimate contribution; gates estimate reliability.

Methods and Materials



1. Masked multi-pass inputs.



2. Delta-gated decomposition.

$$z = z_T + g_{TA} \odot \Delta_{TA} + g_{TV} \odot \Delta_{TV} + g_{\mathrm{INT}} \odot \Delta_{\mathrm{INT}}, \quad (1)$$

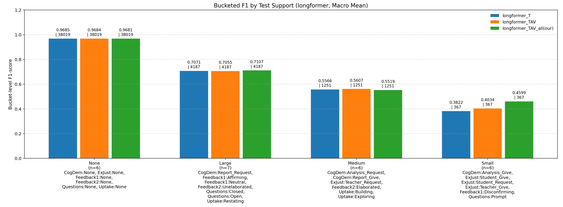
Notation: T=text, A=audio, V=video; TA/TV bimodal, TAV all modalities.
Question: do audio/video help, how much, or mislead?
Deltas measure gain over the transcript baseline.
Gates estimate reliability of each modality gain.
Stress tests verify trust under missing/misaligned cues.

Results

0.6777	+0.0565	-0.0012
best F1	**small-label gain**	**missing drop**

Overall F1 on full test set

Model	S42	S43	S44
Text	0.6588	0.6599	0.6636
Early	0.6711	0.6676	0.6678
DG-MFP	**0.6777**	**0.6712**	**0.6687**



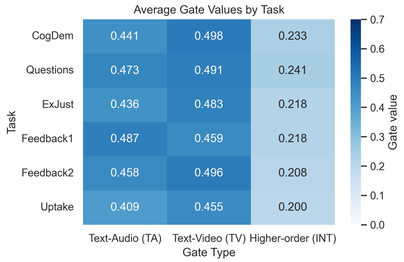
Bucketed F1 by label support (seed 42)

Bucket	#	Text	Early	Ours
None	6	0.9685	0.9684	0.9681
Large	7	0.7071	0.7055	**0.7107**
Medium	6	0.5566	**0.5607**	0.5519
Small	6	0.3822	0.4034	**0.4599**

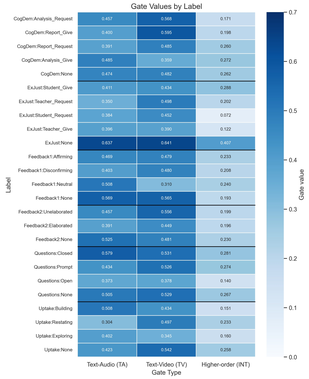
Stress robustness: macro-F1 (drop from clean)

Model	Clean	Misalign	Missing	NoPrompt
Early	0.6613	0.6392 (-0.0222)	0.6505 (-0.0109)	0.3522 (-0.3091)
DG-MFP	**0.6742**	**0.6736 (-0.0006)**	**0.6730 (-0.0012)**	**0.4304 (-0.2438)**

Discussion and Conclusions



Discussion: gate patterns explain model behavior.



Gate evidence: trusted increments vary by task and label.

Finding: audio-text and video-text carry most complementary signal.

Conclusion: rare discourse labels benefit most from gated multimodal cues.

DG-MFP turns fusion into measurable contribution and reliability.

Contact: Chongyu He | ekn8kz@virginia.edu | CV4Edu-17
Data reference: AIAIS dataset, AIAI Challenge 2025

Funding: This work is funded in part by the Gates Foundation. Views expressed here are those of the authors and do not necessarily reflect positions or policies of the foundation.