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

Anonymous CVPR submission

Paper ID *****

Abstract

*Multimodal transformers are attractive options for the analysis of human activity in the classroom, but real-world classroom data often contain missing or misaligned modalities, making robust multimodal learning challenging. In this work, we propose a **Delta-Gated Multi-Forward-Pass (DG-MFP) Longformer** for robust multimodal classroom discourse understanding. The model treats the transcript as the primary modality and models audio and video as incremental sources of information. A shared encoder performs multiple forward passes with different modality masks (text, text+audio, text+video, text+audio+video), allowing modality contributions to be isolated through differences between representations. These increments are fused through class-specific delta gates that modulate modality contributions relative to the text baseline. To evaluate robustness under realistic classroom conditions, we introduce controlled missing-modality and cross-modal misalignment tests. Experiments on the Artificial Intelligence for Advancing Instruction at Scale (AIAIS) dataset show that the proposed method consistently improves overall F1 over standard multimodal fusion baselines and exhibits substantially stronger robustness under missing or misaligned modalities. Further analysis of the learned gates reveals interpretable, task-specific patterns of modality activity, highlighting how multimodal signals provide complementary information when transcript evidence alone is insufficient.*

1. Introduction

Understanding classroom interactions requires reasoning over multiple modalities such as spoken conversation and visual behavior. Recent advances in transformer architectures have enabled powerful multimodal representation learning across vision, language, and audio signals [5, 12, 14]. These models are increasingly used for educational video analysis and classroom discourse understanding.

However, real classroom recordings often contain missing, noisy, or misaligned modalities due to recording conditions. In this context, misalignment refers to imperfect temporal alignment between the transcript segment and the associated audio or video streams. Audio tracks may be corrupted, video streams may be incomplete, and multimodal signals are often imperfectly synchronized. Despite these challenges, most multimodal learning systems are evaluated under ideal conditions where all modalities are available and perfectly aligned. As a result, it remains unclear how robust these models are when applied to realistic classroom environments where modalities may be unreliable.

A key challenge in multimodal learning is understanding how different modalities influence the final prediction. Standard multimodal fusion methods typically combine modalities within a single forward pass. Although this approach can improve predictive performance, it makes it difficult to isolate modality contributions or analyze model behavior when modalities are missing or misaligned.

To address this problem, we propose a **Delta-Gated Multi-Forward-Pass Longformer** for robust multimodal classroom discourse understanding, as illustrated in Figure 1. The central idea is to treat the transcript as the primary modality and to model audio and video as incremental sources of information.

A shared encoder performs multiple forward passes with different modality masks, producing representations for text-only, text+audio, text+video, and text+audio+video inputs. Because the encoder parameters are shared across passes, differences between these representations reflect the additional information contributed by each modality.

This formulation explicitly separates the baseline textual signal from modality-specific increments, allowing the model to estimate both the incremental contribution and the reliability of each modality. The resulting increments are fused through a delta-gated mechanism that models modality contributions relative to the text baseline while preserving interpretability through class-specific gating weights.

In addition to the model design, we introduce a diagnostic evaluation protocol for multimodal robustness under realistic classroom conditions. The protocol includes con-

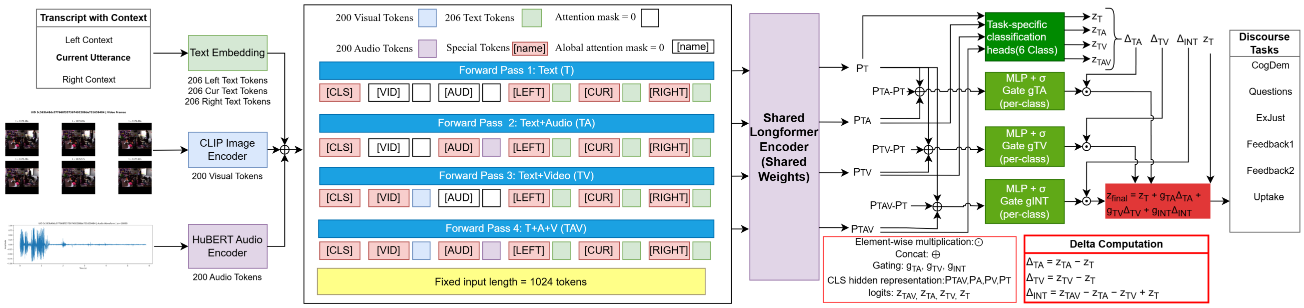

Figure 1. Overview of the proposed **DG-MFP** architecture. A unified multimodal token sequence is constructed from transcript context, video frames, and audio signals. A shared Longformer encoder is executed four times under different modality masks, producing representations for text-only (T), text+audio (TA), text+video (TV), and text+audio+video (TAV). Task-specific classification heads produce logits for each pass, while per-class delta gates modulate the incremental contributions of audio, video, and higher-order multimodal interaction terms.

trolled missing-modality scenarios and cross-modal misalignment tests that simulate common issues in classroom recordings. This design is particularly well suited to educational video analysis, where transcripts often provide the most reliable signal while audio and visual streams may be noisy, partially missing, or imperfectly aligned in real classroom recordings. Furthermore, the explicit modeling of modality contributions provides interpretable insights into how different signals support discourse prediction, which is important for educational research and learning analytics.

Our contributions are summarized as follows:

1. We propose a **Delta-Gated Multi-Forward-Pass Longformer** that treats the transcript as the primary modality and models audio and video as incremental contributions through structured logit increments and class-specific gates.
2. We introduce a **diagnostic evaluation protocol** for multimodal robustness that includes missing-modality and cross-modal misalignment stress tests.
3. Experiments with classroom interaction data demonstrate improved robustness and provide interpretable insights into modality usage in discourse classification tasks.

2. Related Work

2.1. Multimodal Transformers

Transformer architectures have become a dominant framework for modeling long-range dependencies across modalities. The transformer architecture introduced in [12] has been widely adopted for language, vision, and multimodal learning, while pretrained models such as BERT [5] demonstrated the effectiveness of large-scale transformer-based representation learning.

Recent work extends transformers to multimodal settings. The Multimodal Transformer [11] introduced cross-modal attention to jointly model language, visual, and acoustic signals in unaligned sequences. Other approaches such as VideoBERT [10] and vision–language models including CLIP [9] demonstrate the benefit of large-scale multimodal pretraining. More recent models such as Flamingo [1] and InternVideo [13] further explore large-scale vision–language and video representation learning.

Multimodal fusion strategies typically combine representations through early fusion, cross-modal attention, or gating mechanisms. Although these approaches improve predictive performance, they often make it difficult to explicitly analyze how individual modalities contribute to the final prediction.

For long-sequence modeling, the Longformer [3] introduces sparse attention that enables efficient processing of long contexts. In this work, we build upon Longformer as the shared encoder for multimodal classroom discourse analysis.

2.2. Robust Multimodal Learning with Missing Modalities

Many multimodal learning systems assume that all modalities are available during both training and inference. However, in practice, real-world data often contain missing, corrupted, or unreliable modalities. Early work such as ModDrop [8] improves robustness by randomly removing modalities during training.

More recent studies analyze the behavior of multimodal transformers when one modality is absent. For example, Ma et al. [7] investigated the robustness of transformer-based multimodal models under missing-modality conditions and showed that model performance can depend strongly on the fusion strategy.

Most prior work focuses primarily on improving predictive performance. In contrast, our work explicitly models modality contributions and evaluates model behavior under controlled missing-modality and cross-modal misalignment conditions.

2.3. Multimodal Analysis in Educational Settings

Multimodal signals play an important role in understanding classroom interactions and learning processes. Educational research increasingly uses multimodal learning analytics to analyze speech, gestures, and visual behaviors in instructional environments [4].

More broadly, multimodal machine learning integrates heterogeneous signals such as language, vision, and audio [2]. Transformer-based approaches further enable learning from partially aligned multimodal sequences by modeling cross-modal dependencies [11].

However, educational datasets often contain substantial noise, imperfect alignment, and incomplete signals due to real classroom recording conditions. These challenges make robustness particularly important for the practical implementation of multimodal models in educational settings.

Our work contributes to this area by evaluating multimodal transformers under realistic classroom conditions with missing modalities and temporal misalignment, while also analyzing how individual modalities contribute to discourse prediction tasks.

3. Dataset and Data Processing

3.1. Data Source

Our dataset is derived from the Artificial Intelligence for Advancing Instruction at Scale (AIAIS) dataset released as part of the AIAI Challenge [6], which contains classroom recordings for studying teacher–student interactions.

Rather than modeling entire videos, we focus on **transcript-centered discourse segments**, where each sample corresponds to a short unit of classroom conversation during elementary mathematics and language arts instruction.

We use a subset of **107 classroom videos**, producing **36,091 transcript-centered samples** aligned with the corresponding audio and video streams.

3.2. Multimodal Inputs and Context Construction

Each sample contains three modalities: classroom video frames, audio signals, and transcript text produced by automatic speech recognition (ASR). The transcript segment serves as the **primary unit of analysis**, while audio and video provide complementary contextual information.

Modalities are aligned using timestamps from the original ASR output. Although the text of the transcript was manually corrected, the timing information was not reannotated. Alignment is therefore approximate and occurs at the sentence level rather than precise frame synchronization, reflecting realistic classroom recording conditions.

To capture conversational context, each sample includes a context window consisting of left context, current transcript segment, and right context. The model processes these multimodal inputs within a maximum sequence length of **1024 tokens**.

Visual features are extracted using CLIP ViT-B/32 frame embeddings, while audio features are obtained using a HuBERT encoder. The transcript text is tokenized and processed by the transformer model.

3.3. Discourse Annotation Tasks and Dataset Statistics

The objective is to classify instructional discourse segments within classroom interactions. Each transcript-centered sample is annotated using a hierarchical scheme consisting of **six discourse tasks with 19 subclasses (25 labels in total including a *None* label for each task)**.

The six tasks capture different aspects of classroom instructional interaction. *CogDem* (Cognitive Demand) reflects the level of reasoning involved in the discourse, distinguishing between simple reporting of facts and higher-level analytical reasoning. *Questions* categorizes the types of teacher questions posed to students, such as open-ended or closed-ended questions. *ExJust* (Explanation and Justification) identifies when participants request or provide explanations and reasoning. *Feedback1* and *Feedback2* represent different types of teacher feedback and evaluation of student responses. Finally, *Uptake* captures how teachers incorporate or build on students' ideas during classroom interaction. Together, these tasks characterize key instructional behaviors such as questioning, reasoning, feedback, and the integration of student contributions.

The final dataset contains **36,091 samples** from **107 classroom videos**. We use an **80/20 split** with **28,787 training samples** and **7,304 test samples**. Each sample may contain labels for multiple discourse tasks, but within each task exactly one subclass is assigned (including *None*).

The dataset exhibits substantial class imbalance, with dominant *None* labels and long-tailed subclass distributions across tasks. This creates a challenging **multi-task long-tailed classification problem**. To better reflect performance on minority classes we adopt **macro-F1** as the primary evaluation metric. Figure 2 visualizes the label distribution across the six discourse tasks.

4. Method

4.1. Overview

Figure 1 illustrates the proposed **DG-MFP** framework for multimodal classroom discourse classification.

Given a transcript-centered classroom sample, we construct a unified multimodal token sequence containing transcript context, video tokens, and audio tokens. These tokens

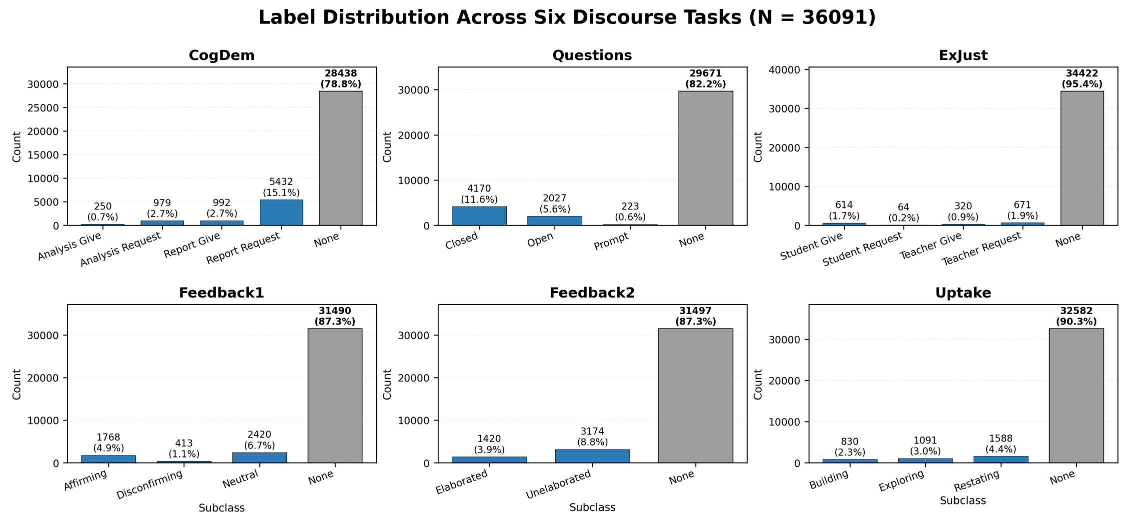

Figure 2. Label distribution across the six discourse tasks ($N = 36{,}091$ transcript-centered samples). Each task contains a dominant *None* class and several low-frequency subclasses, resulting in severe class imbalance and long-tailed label distributions. This motivates the use of macro-F1 for robustness experiments, as it better reflects performance on minority classes.

are processed by a **shared Longformer encoder** that is executed multiple times under different modality masks.

Unlike conventional multimodal models that fuse modalities in a single forward pass, our method performs four forward passes using the same encoder parameters: text-only (T), text+audio (TA), text+video (TV), and text+audio+video (TAV). Each pass always includes transcript tokens, while additional modalities are introduced incrementally through modality masks. As a result, differences between passes reflect only the additional information contributed by each modality.

Instead of directly predicting from the full multimodal representation, we model how additional modalities modify a text-only baseline through delta increments. The final logits are computed as

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

where $\odot$ denotes element-wise multiplication, Δ_{TA} and Δ_{TV} represent modality-specific increments relative to the text baseline, Δ_{INT} captures higher-order multimodal interactions, and the gates g_{TA}, g_{TV}, g_{INT} control their contributions. This formulation allows multimodal fusion to be interpreted as a set of incremental information gains relative to the baseline transcript prediction, with learned gates that estimate the confidence of each information source.

4.2. Multimodal Representation and Encoder

Each sample contains three modalities: transcript text, classroom video, and classroom audio. The text modality includes the target transcript together with its left and right discourse contexts (Section 3).

Visual and acoustic features are extracted using pretrained CLIP and HuBERT encoders and interpolated to fixed-length token sequences. Each sample contains 200 visual tokens, 200 audio tokens, and three text segments (left context, current transcript, and right context). The resulting multimodal sequence has a fixed length of **1024 tokens** with modality markers `[VID]`, `[AUD]`, `[LEFT]`, `[CUR]`, and `[RIGHT]`.

Let the multimodal sequence be denoted by x. A shared Longformer encoder $\mathrm{Enc}(\cdot)$ processes the input under different modality masks:

$$p_T = \mathrm{Enc}(x; M_T), \quad (2)$$
$$p_{TA} = \mathrm{Enc}(x; M_{TA}), \quad (3)$$
$$p_{TV} = \mathrm{Enc}(x; M_{TV}), \quad (4)$$
$$p_{TAV} = \mathrm{Enc}(x; M_{TAV}), \quad (5)$$

where M_T enables transcript tokens only, M_{TA} enables transcript and audio tokens, M_{TV} enables transcript and video tokens, and M_{TAV} enables all modalities.

The encoder parameters are shared across forward passes. The pooled representation is obtained from the final hidden state of the `[CLS]` token, producing a text baseline representation p_T, bimodal representations p_{TA} and p_{TV}, and the complete multimodal representation p_{TAV}.

4.3. Task Heads

The task is formulated as a **six-task discourse classification** problem. For each discourse task t, we attach a task-specific linear classifier $h^{(t)}(\cdot)$ to the pooled representation.

The same classifier is applied to all forward-pass representations, producing logits

$$z_T^{(t)} = h^{(t)}(p_T), \tag{6}$$

$$z_{TA}^{(t)} = h^{(t)}(p_{TA}), \tag{7}$$

$$z_{TV}^{(t)} = h^{(t)}(p_{TV}), \tag{8}$$

$$z_{TAV}^{(t)} = h^{(t)}(p_{TAV}). \tag{9}$$

4.4. Delta-Gated Multimodal Fusion

Rather than directly using multimodal logits $z_{TAV}^{(t)}$, we decompose the prediction into modality-specific increments relative to the text baseline.

Text–Audio Increment

$$\Delta_{TA}^{(t)} = z_{TA}^{(t)} - z_T^{(t)} \tag{10}$$

$$g_{TA}^{(t)} = \sigma\Big(W_{TA}^{(t)}[p_T,\ p_{TA},\ p_{TA} - p_T]\Big) \tag{11}$$

$$z^{(t)} = z_T^{(t)} + g_{TA}^{(t)} \odot \Delta_{TA}^{(t)} \tag{12}$$

Text–Video Increment

$$\Delta_{TV}^{(t)} = z_{TV}^{(t)} - z_T^{(t)} \tag{13}$$

$$g_{TV}^{(t)} = \sigma\Big(W_{TV}^{(t)}[p_T,\ p_{TV},\ p_{TV} - p_T]\Big) \tag{14}$$

$$z^{(t)} \leftarrow z^{(t)} + g_{TV}^{(t)} \odot \Delta_{TV}^{(t)} \tag{15}$$

Higher-Order Interaction

$$\Delta_{INT}^{(t)} = z_{TAV}^{(t)} - z_{TA}^{(t)} - z_{TV}^{(t)} + z_T^{(t)} \tag{16}$$

$$g_{INT}^{(t)} = \sigma\Big(W_{INT}^{(t)}[p_T,\ p_{TAV},\ p_{TAV} - p_T]\Big) \tag{17}$$

$$z^{(t)} \leftarrow z^{(t)} + g_{INT}^{(t)} \odot \Delta_{INT}^{(t)} \tag{18}$$

All gates are **task-specific and per-class**, allowing modality contributions to vary across discourse labels.

4.5. Training Objective

For each discourse task t, we apply cross-entropy loss

$$\mathcal{L} = \sum_{t \in \mathcal{T}} \mathrm{CE}(z^{(t)}, y^{(t)}) \tag{19}$$

where $\mathcal{T}$ denotes the set of six discourse tasks.

Table 1. Overall F1 on the full test set. Each result corresponds to the best checkpoint within 30 training epochs.

Model	S42	S43	S44
Longformer (T)	0.6588	0.6599	0.6636
EF Longformer (TAV)	0.6711	0.6676	0.6678
DG-MFP (Ours)	**0.6777**	**0.6712**	**0.6687**

5. Experiments

5.1. Baselines

We compare the proposed method with two baseline architectures designed to isolate the effect of the fusion strategy.

Text-only baseline. This baseline uses only the transcript together with the left and right context. It shares the same Longformer backbone and text formatting as our method but removes audio and video input entirely. This setting measures how much predictive signal is available from the transcript context alone.

Standard multimodal early-fusion baseline. This baseline performs multimodal fusion in a single forward pass. Visual and acoustic tokens are concatenated with the transcript sequence using the same prompt-token layout as our method, and the pooled Longformer representation is passed to the same task-specific classification heads. Unlike our method, this baseline does not perform multiple masked forward passes or delta-gated fusion.

5.2. Implementation Details

All models are trained for up to **30 epochs**, and we report the best checkpoint selected within this training schedule. To account for training variance, each main model is trained with **three random seeds (42, 43, 44)**.

For the main comparison, we report **overall F1 on the entire test set**. For robustness evaluation, we additionally report **macro-F1** because perturbations may disproportionately affect low-support classes.

5.3. Main Results

Table 1 reports overall F1 on the full test set.

The text-only baseline already achieves strong performance, indicating that the transcript context provides substantial predictive information for the classification of classroom discourse. Standard early fusion further improves performance, demonstrating that audio and visual signals provide useful complementary cues when integrated into the model.

The proposed DG-MFP model achieves the best overall F1 across all three random seeds used during training. For example, with the random seed set to 42, our model

Table 2. Bucket-level F1 comparison for seed 42.

Bucket	# Labels	Text-only	Early Fusion	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**

improves F1 from 0.6711 with early fusion to 0.6777. Although the absolute improvement is modest, the gain is consistent across seeds, indicating that explicitly modeling modality increments and higher-order interaction effects yields a more stable multimodal fusion strategy than single-pass early fusion.

Because the dataset is long-tailed and only approximately aligned across modalities, we next analyze performance by label frequency and under controlled robustness perturbations.

5.4. Performance Across Label-Frequency Buckets

As the dataset is strongly long-tailed, we further analyze which types of labels benefit the most from multimodal fusion. Using the seed-42 models, we group discourse labels into four frequency buckets: *None*, *Large*, *Medium*, and *Small*.

Figure 3 and Table 2 report the average F1 within each bucket. The *None* bucket contains the dominant background labels, whereas the *Small* bucket contains the rarest and most challenging discourse subclasses.

Several trends are apparent. First, all models perform similarly on the *None* bucket, indicating that dominant labels are already easy to predict from the transcript context.

Second, early fusion provides only limited and inconsistent gains across frequency buckets.

In contrast, the proposed DG-MFP model shows its clearest advantage on the *Small* bucket, which contains the rarest and most challenging discourse subclasses. Our method improves the average F1 from 0.3822 for the text-only baseline and 0.4034 for early fusion to 0.4599, corresponding to absolute gains of +0.0777 and +0.0565, respectively.

These results suggest that multimodal cues are particularly useful for low-support discourse labels, where the transcript context alone is often insufficient for reliable prediction.

5.5. Robustness Under Missing and Misaligned Modalities

To evaluate robustness beyond clean evaluation, we perform three controlled perturbation tests at inference time. These tests measure how multimodal fusion strategies behave when additional modalities are missing or semantically misaligned with the target transcript.

Cross-sample misalignment. We keep the transcript fixed, but replace the associated video and audio streams with mismatched signals drawn from other test samples. This produces a semantically incorrect multimodal context while preserving the text input. We denote this setting as **baseline_miss_alignment**.

Missing modality with prompt retention. We zero out all video and audio token values while retaining the `[VID]` and `[AUD]` prompt tokens. This tests whether the model remains stable when modality structure is present, but modality content is absent. We denote this setting as **baseline_missing**.

Missing modality with prompt removal. We remove both modality tokens and prompt tokens, leaving only the text portion of the input sequence. This tests robustness when the multimodal input structure itself changes. We denote this setting as **baseline_missing_removeprompt**.

All robustness tests are performed at evaluation time using the same trained checkpoints as in the main experiments, without additional training or finetuning. We report **macro-F1** under clean evaluation and the three perturbation settings. The text-only baseline is omitted because it does not use audio or video inputs.

Table 3 shows that the proposed model is substantially more robust than standard early fusion. Under cross-sample misalignment, early fusion drops by 0.0222, whereas our model drops only by 0.0006. Under the missing modality with prompt retention, the drop is 0.0109 for early fusion but only 0.0012 for our method. Even in the most severe setting, where both modality tokens and prompts are removed, our model retains a substantially higher macro-F1 (0.4304 vs. 0.3522).

These results indicate that modeling audio and video as incremental contributions relative to a baseline of transcripts makes the model less sensitive to misleading or absent multimodal signals.

5.6. Analysis of Learned Delta Gates

To understand how multimodal signals contribute to prediction, we analyze the learned gate values for the three modality pathways: text–audio (g_{TA}), text–video (g_{TV}), and higher-order interaction (g_{INT}). Each gate is a per-class scalar in the range $[0, 1]$ that modulates the contribution of the corresponding logit increment.

Relative importance of modality paths. Across all discourse tasks, the learned values g_{TA} and g_{TV} generally fall in the range of approximately 0.4–0.5, while the higher-order interaction gate g_{INT} remains substantially smaller, typically around 0.2. Figure 4 shows the average gate values

CVPR
#*****

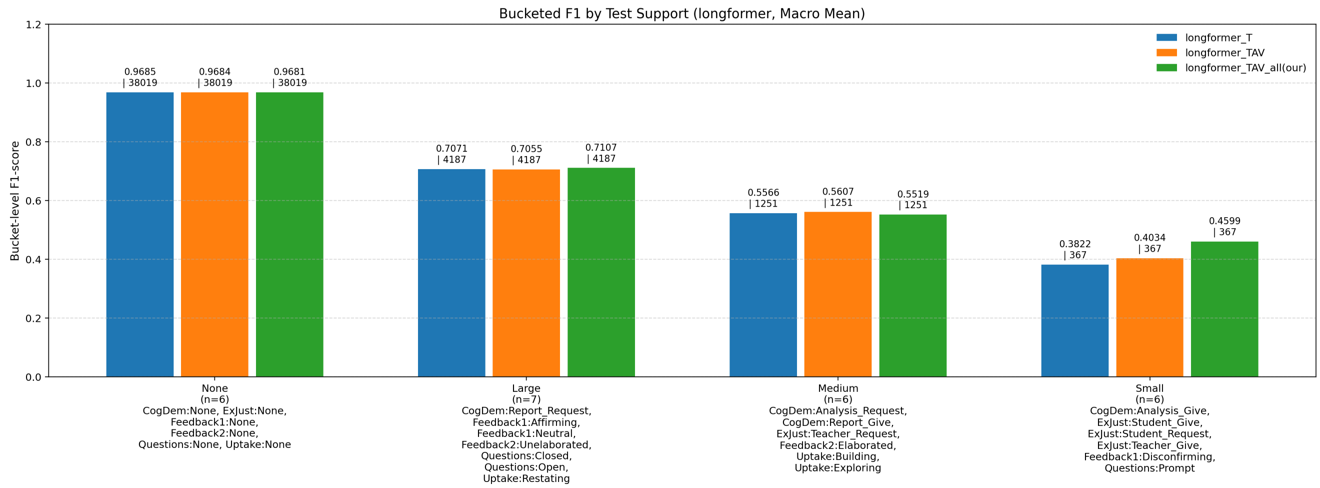

Figure 3. Bucketed F1 by test-set label frequency for the seed-42 models. Labels are grouped into *None*, *Large*, *Medium*, and *Small* buckets according to the number of training samples (support) for each discourse label. Label names follow the format *Task:Subclass* (e.g., *CogDem:Analysis_Give*, *Questions:Open*, *Feedback1:Affirming*), where the prefix denotes the discourse task and the suffix denotes the specific subclass. The proposed DG-MFP model achieves its clearest advantage on the *Small* bucket, indicating stronger benefits for low-support discourse subclasses.

Table 3. Robustness under missing and misaligned modalities. We report macro-F1, with the relative drop from the clean condition in parentheses.

Model	Clean	Misalignment	Missing	Missing w/o Prompt
Early-Fusion Longformer	0.6613	0.6392 (-0.0222)	0.6505 (-0.0109)	0.3522 (-0.3091)
DG-MFP (Ours)	**0.6742**	**0.6736** (-0.0006)	**0.6730** (-0.0012)	**0.4304** (-0.2438)

for each discourse task. For example, the *CogDem* task exhibits average gate values of $g_{TA} = 0.44$, $g_{TV} = 0.50$, and $g_{INT} = 0.23$, while the *Questions* task shows $g_{TA} = 0.47$, $g_{TV} = 0.49$, and $g_{INT} = 0.24$. This consistent pattern indicates that bimodal text–audio and text–video signals provide most of the complementary information beyond the text-only baseline, whereas higher-order three-way modality interactions appear less dominant.

Task-specific modality patterns. Although the overall trend is consistent, different discourse tasks exhibit subtle differences in modality usage. For example, the *Feedback1* task shows slightly stronger text–audio influence ($g_{TA} = 0.49$) than text–video ($g_{TV} = 0.46$), suggesting that acoustic cues may help distinguish different forms of teacher feedback. In contrast, the *CogDem* and *Feedback2* tasks exhibit slightly larger g_{TV} values (around 0.48–0.50), indicating that visual classroom cues can provide complementary signals for identifying cognitive demand and feedback structures.

Behavior on dominant *None* classes. Many *None* categories also exhibit relatively strong bimodal gates. This suggests that multimodal cues may help the model confirm the absence of specific discourse functions. However, the interaction gate g_{INT} remains consistently smaller for these dominant classes, indicating that detecting the absence of discourse signals typically does not require complex three-way cross-modal interactions.

Multimodal reliance for rare labels. A closer inspection of the label-level gate values (Figure 5) reveals that several low-frequency discourse labels rely more strongly on multimodal signals. For example, rare labels such as *Questions:Prompt* (test-set support = 44) and *CogDem:Analysis_Give* (test-set support = 43) exhibit relatively large text–audio or text–video gate values (g_{TA} or g_{TV} around 0.43–0.48). Similarly, *CogDem:Report_Give* shows a particularly strong visual contribution with $g_{TV} \approx 0.60$.

In contrast, dominant *None* categories, which have thousands of training examples, often show more moderate gate magnitudes. This suggests that the model can rely primarily on transcript context for frequent labels, whereas multimodal cues become more helpful when predicting less frequent discourse categories.

This observation is consistent with the bucket-level results in Table 2, where the proposed model achieves the

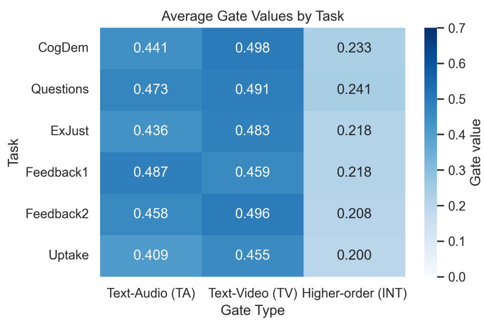

Figure 4. Average delta gate values across discourse tasks. Columns correspond (from left to right) to the text–audio (g_{TA}), text–video (g_{TV}), and higher-order interaction (g_{INT}) gates. The bimodal gates (g_{TA} and g_{TV}) consistently exhibit larger values than the interaction gate (g_{INT}), suggesting that bimodal modality signals contribute most of the complementary information beyond the text-only baseline, while three-way modality interactions are comparatively less dominant.

largest performance improvement on the small-label bucket. Together, these findings suggest that delta-gated fusion helps the model selectively leverage multimodal signals when transcript evidence alone is insufficient.

5.7. MoE Comparison as a Qualitative Reference

To further examine interpretability, we compare DG-MFP with a pooled-level Mixture-of-Experts (MoE) variant used as a qualitative reference. In this model, the pooled [CLS] representation is routed to multiple MLP experts through a soft router, and expert outputs are combined before classification. Unlike DG-MFP, this MoE variant does not explicitly separate text–audio, text–video, and higher-order interaction effects.

During training, routing statistics reveal limited expert specialization. The router often assigns most selections to a single expert in early epochs and gradually converges to nearly uniform expert usage ($\approx 25\%$ per expert, entropy $\approx \log 4$), suggesting weak semantic specialization.

In contrast, DG-MFP produces gates directly tied to structured modality contributions. These gates are task-specific and per-class, making them more interpretable than latent expert assignments in MoE.

6. Conclusion

In this work, we studied robust multimodal classroom discourse classification under realistic conditions where audio and video signals may be missing, noisy, or imperfectly aligned with the transcript. To address this problem, we proposed a **Delta-Gated Multi-Forward-Pass Longformer** that treats the transcript as the primary modality and models audio and video as incremental sources of information.

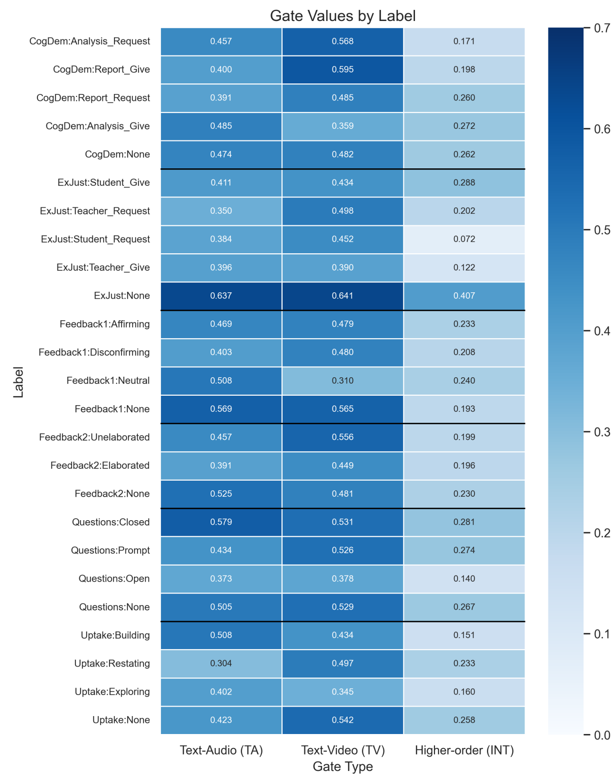

Figure 5. Delta gate values for each discourse label. Rows correspond to task-specific labels, and columns correspond (from left to right) to the text–audio (g_{TA}), text–video (g_{TV}), and higher-order interaction (g_{INT}) modality pathways. The visualization shows that multimodal contributions vary across labels, with several categories exhibiting relatively stronger bimodal modality gates.

The proposed framework uses a shared encoder in multiple masked forward passes and decomposes the prediction into a baseline transcript, modality-specific increments, and a higher-order interaction term. This formulation enables explicit modeling of modality contributions while maintaining a consistent representation space across forward passes.

Experiments on classroom discourse data show that the proposed method improves performance over standard early-fusion baselines, achieves strong gains on rare discourse labels, and demonstrates substantially stronger robustness under missing-modality and misalignment conditions. Analysis of the learned gates further reveals interpretable task-specific patterns of modality use, indicating that multimodal signals are most beneficial when transcript evidence alone is insufficient for reliable prediction.

In general, our results suggest that multimodal fusion can be effectively modeled as incremental contributions to a transcript-based prediction. We hope that this perspective will encourage further work on robust and interpretable multimodal learning in real-world educational environments.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 2

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2019. 3

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. arXiv:2004.05150. 2

[4] Paulo Blikstein. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 102–106, 2013. 3

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805. 1, 2

[6] DrivenData and University of Virginia. Artificial intelligence for advancing instruction (aiai) challenge. `https://github.com/drivendataorg/aiai-challenge`, 2025. Accessed: 2026. 3

[7] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18177–18186, 2022. 2

[8] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE TPAMI*, 38(8):1692–1706, 2016. 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. arXiv:2103.00020. 2

[10] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[11] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences, 2019. arXiv:1906.00295. 2, 3

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. arXiv:1706.03762. 1, 2

[13] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022. 2

[14] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE TPAMI*, 45(10):12113–12132, 2023. 1