A. Instructional Labels

The dataset contains 43 labels: 24 instructional activity labels (see Table 4) and 19 instructional discourse labels (see Table 5). For the discourse labels, the dataset includes information of labels that are usually paired together:

- **Analysis Request (Teacher)** and **Explanation/Justification Teacher Request**.
- **Analysis Give (Teacher)** and **Explanation/Justification Teacher Give**.
- **Feedback Neutral**, **Uptake Restating**, and **Feedback Unelaborated**.

B. Label Distribution

Table 6 reports the distribution of positive labels for both Instructional Activity and Instructional Discourse categories. For Activity labels, annotations are provided at the per-second level of the video; thus, the percentages reflect the proportion of seconds in which a given label is active. In contrast, Discourse labels are annotated at the utterance level. To enable comparison across modalities, we report both the proportion of utterances containing each label (% Utt) and a time-based measure showing the proportion of seconds in which the label occurs (% Sec).

The distributions highlight the strong imbalance across labels. For instance, highly frequent activity labels such as Teacher standing (52.4%) and Whole class activity (51.9%) dominate the dataset, while more fine-grained instructional practices like Teacher writing (4.1%) or On-task student talking with student (4.2%) are comparatively rare. A similar skew is evident in the discourse dimension: labels such as Report-Request appear in 20.5% of utterances, whereas others like Analysis-Give (0.5%) and Student-Request (0.2%) occur only sporadically. This imbalance underscores the difficulty of reliably detecting rare but pedagogically important practices, and motivates our exploration of imbalance-aware objectives and dynamic thresholding strategies.

C. Language Models Prompts

C.1. Multimodal Large Language Models

Figure 2 shows the prompt for zero-shot for MLLMs.

C.2. Large Language Models

Figure 3 shows the prompt for zero-shot and Figure 4 show the prompt for few-shot for LLMs.

D. F1 Score Per Label

In this section, we report per-label F1 scores. Table 7 presents results for Instructional Activity labels using MLLMs and fine-tuned models, while Table 8 shows results for Instructional Discourse labels using LLMs and fine-tuned models. Activity scores were computed on a per-second basis, and discourse scores on a per-utterance basis. Some labels are clearly more difficult to predict than others. For instructional activities, the best-performing model struggled with labels such as On-task student talking with student (0.080 macro-F1) and Teacher supporting multiple students without student interaction (0.169), while labels like Teacher standing (0.877) and Presentation with technology (0.870) achieved much higher scores. A similar pattern is seen in discourse recognition: Analysis-Give (0.115) and Teacher-Give (0.235) were among the lowest, while Feedback-Affirming (0.625) and Uptake-Restating (0.556) ranked highest. These discrepancies reflect both the rarity of certain labels (see Section B) and the inherent difficulty of modeling complex, fine-grained instructional practices.

Instructional Activity Label	Description
Whole class activity	All students are involved in one activity, with the teacher leading the learning (e.g., lecture, presentation, carpet time).
Individual activity	Students work independently (e.g., practice, reading), no interaction with peers.
Small group activity	Students working together with peers (e.g., think-pair-share, book club).
Transition	Teacher/students move between activities or locations; no meaningful instruction.
On-task student talking with student	Students conversing with each other without teacher support; can overlap with small group activity.
Student raising hand	Hand up for more than 1 second, clearly and purposefully.
Teacher sitting	Teacher seated (chair, stool, floor, crouching, kneeling, on desk).
Teacher standing	Teacher standing in the same spot/orientation to students.
Teacher walking	Teacher walking with purpose to change orientation.
Students sitting on carpet or floor	Students seated on the carpet/floor.
Students sitting at group tables	Students seated at tables.
Sitting at desks	Students seated at individual desks.
Students standing or walking	One or more students standing or moving around.
Teacher supporting one student	Teacher assists one student verbally or non-verbally.
Teacher supporting multiple students with interaction	Teacher assists multiple students who are also interacting with one another.
Teacher supporting multiple students without interaction	Teacher assists multiple students who are not interacting with one another.
Using or holding book	Book is used/held by teacher or student.
Using or holding worksheet	Worksheet is used/held by teacher or student.
Presentation with technology	Smartboard, Elmo, projector used to show content.
Using or holding instructional tool	Object (e.g., ruler, manipulative) used/held for instruction (not pen/pencil/-furniture).
Using or holding notebook	Notebook is used/held by teacher or student.
Individual technology	Laptop, tablet, or other personal tech used by student or teacher.
Teacher writing	Teacher writes/erases on paper, board, or document camera.
Student writing	Student writes/erases on paper or board.

Table 4. Instructional activity labels.

Instructional Discourse Labels	Description
Analysis Give (Teacher)	Teacher analyzes content (e.g., compare, justify, synthesize, connect).
Analysis Request (Teacher)	Teacher asks students to analyze content (e.g., why/how, connect ideas).
Report Give (Teacher)	Teacher states facts, definitions, or procedures.
Report Request (Teacher)	Teacher asks students to recall or report facts, definitions, or methods.
Questions Open-Ended	Teacher asks open content-related questions without a pre-scripted answer.
Questions Closed-Ended	Teacher asks content-related questions with a pre-scripted/fluency answer (e.g., yes/no).
Questions Task Related Prompt	Teacher reads/restates a task question/prompt from instructional materials.
Explanation/Justification Teacher Request	Teacher requests an explanation or justification.
Explanation/Justification Teacher Give	Teacher provides an explanation or justification, possibly a solution strategy.
Explanation/Justification Student Request	Student requests an explanation or justification.
Explanation/Justification Student Give	Student provides an explanation or justification.
Feedback Affirming	Teacher positively evaluates student's response (e.g., "Excellent").
Feedback Disconfirming	Teacher negatively evaluates correctness of a response (e.g., "No").
Feedback Neutral	Teacher neither confirms nor disconfirms, may repeat student's contribution.
Feedback Elaborated	Teacher expands on student's response or thinking.
Feedback Unelaborated	Teacher acknowledges, evaluates, or superficially uses a student's idea.
Uptake Restating	Teacher repeats/summarizes a student's idea without building on it.
Uptake Building	Teacher incorporates/clarifies or expands on a student's idea.
Uptake Exploring	Teacher probes further into a student's idea with follow-up questions.

Table 5. Instructional discourse labels.

Task Description
You are an expert classroom activity observer. You are given a one-second video clip from a classroom lesson. Carefully analyze the clip, think step-by-step, and identify all instructional activity labels that apply. You may select multiple labels. Each label is listed below with its definition to guide your judgment.

Instructional Activity Labels and Definitions
{*Names and definitions of 24 instructional activity labels*}

Instructions
Always return your answer as a JSON list containing the label names that exactly match those provided above (verbatim). If no labels apply, return an empty list. Do not include descriptions, explanations, commentary, or any modified text.

Figure 2. Prompt for multimodal large language models. {*Names and definitions of 24 instructional activity labels*} can be found in Section A.

Activity Labels	% Sec	Discourse Labels	% Utt	% Sec
Teacher standing	52.4	Report-Request	20.5	13.4
Whole class activity	51.9	Closed-Ended	13.4	7.8
Presentation with technology	44.5	Feedback-Neutral	10.0	6.0
Sitting at desks	43.2	Feedback-Unelaborated	10.8	6.0
Using or holding worksheet	40.6	Open-Ended	9.1	6.0
Students sitting at group tables	33.1	Feedback-Elaborated	4.9	4.0
Students standing or walking	30.3	Uptake-Restating	5.6	3.5
Students sitting on carpet or floor	29.0	Feedback-Affirming	4.9	3.3
Teacher sitting	26.8	Uptake-Building	3.1	2.6
Small group activity	22.1	Report-Give	2.3	2.3
Student writing	19.2	Analysis-Request	3.3	2.1
Using or holding notebook	18.7	Teacher-Request	2.6	1.4
Individual activity	17.6	Uptake-Exploring	3.9	1.4
Using or holding instructional tool	17.3	Task Related Prompt	0.7	1.4
Individual technology	14.0	Teacher-Give	0.9	1.3
Teacher supporting multiple students with student interaction	13.9	Analysis-Give	0.5	0.8
Using or holding book	13.3	Feedback-Disconfirming	0.8	0.7
Teacher walking	8.9	Student-Give	2.3	0.7
Student raising hand	6.5	Student-Request	0.2	0.1
Teacher supporting one student	6.5			
Transition	5.7			
Teacher supporting multiple students without student interaction	4.8			
On task student talking with student	4.2			
Teacher writing	4.1			

Table 6. Percentages of positive and negative labels for Activity and Discourse. For Discourse labels, we report both the proportion of utterances containing each label (% Utt) and the proportion of seconds in which the label occurs (% Sec), consistent with the time-based reporting used for Activity labels.

Task Description: For each of the following 19 categories, output 1 if that teaching activity appears in the given transcript excerpt, otherwise 0. Take into account who is the speaker and context of the transcript excerpt. Output exactly one JSON object with all 19 keys.

Categories:
{*Names and definitions of 19 instructional discourse labels*}

Further Instructions:
If you're unsure, assign 0 rather than inventing a 1.
Cognitive-Demand_Analysis-Request and Expl-Just_Teacher-Request are usually paired together.
Cognitive-Demand_Analysis-Give and Expl-Just_Teacher-Give are usually paired together.
Feedback_Neutral, Feedback_Unelaborated and Uptake_Restating are usually paired together.
Feedback_Affirming, Feedback_Disconfirming and Feedback_Neutral are mutually exclusive.
Feedback_Elaborated and Feedback_Unelaborated are mutually exclusive.
Uptake_Building, Uptake_Exploring and Uptake_Restating are mutually exclusive.
Questions_Closed-Ended and Questions_Open-Ended are mutually exclusive.
Output exactly one JSON object, no extra keys, no explanations.
Ensure valid JSON: double-quoted keys, commas between pairs, no trailing comma.

Task:
Context:
{*Context sentence t-2*}
{*Context sentence t-1*}
Speaker:
{*Current speaker*}
Transcript excerpt:
{*Current sentence*}
Result:

Figure 3. Prompts used for zero-shot prompting of *Llama3.1-Instruct* (8B and 70B).{*Names and definitions of 19 instructional discourse labels*} can be found in Section A, {*Context sentence t-2*} and {*Context sentence t-1*} are the two previous sentences, {*Current speaker*} is the speaker of the current sentence, and {*Current sentence*} is the sentence to be classified.

Task Description: For each of the following 19 categories, output 1 if that teaching activity appears in the given transcript excerpt, otherwise 0. Take into account who is the speaker and context of the transcript excerpt. Output exactly one JSON object with all 19 keys.

Categories:
{*Names and definitions of 19 instructional discourse labels*}

Further Instructions:
If you're unsure, assign 0 rather than inventing a 1.
Cognitive-Demand_Analysis-Request and Expl-Just_Teacher-Request are usually paired together.
Cognitive-Demand_Analysis-Give and Expl-Just_Teacher-Give are usually paired together.
Feedback_Neutral, Feedback_Unelaborated and Uptake_Restating are usually paired together.
Feedback_Affirming, Feedback_Disconfirming and Feedback_Neutral are mutually exclusive.
Feedback_Elaborated and Feedback_Unelaborated are mutually exclusive.
Uptake_Building, Uptake_Exploring and Uptake_Restating are mutually exclusive.
Questions_Closed-Ended and Questions_Open-Ended are mutually exclusive.
Output exactly one JSON object, no extra keys, no explanations.
Ensure valid JSON: double-quoted keys, commas between pairs, no trailing comma.

Example 1:
Context:
Teacher 1 (14:35): What about the boy? If you had to describe the boy, how would you describe him? [Student 8]'s hand shot up. [Student 8]. How would you describe him?
Student 8 (14:41): Greedy.
Speaker:
Teacher 1
Transcript excerpt:
Teacher 1 (14:42): Greedy? Why do you say greedy?
Result:
{*Valid JSON*}

Example 2:
Context:
Student 14 (16:07): And I know that four plus five is nine.
Teacher (16:08): Okay, so you knew that was nine. Student 5 turn around.
Speaker:
Student 14
Transcript excerpt:
Student 14 (16:16): I know that 10 plus seven is 17, so I put five and four together and it made nine. And then six and then [00:16:30] that would make 16.
Result:
{*Valid JSON*}

Task:
Context:
{*Context sentence t-2*}
{*Context sentence t-1*}
Speaker:
{*Current speaker*}
Transcript excerpt:
{*Current sentence*}
Result:

Figure 4. Prompts used for few-shot prompting of *Llama3.1-Instruct* (8B and 70B). {*Names and definitions of 19 instructional discourse labels*} can be found in Section A, {*Valid JSON*} is a valid JSON output for the example, here shortened for space, {*Context sentence t-2*} and {*Context sentence t-1*} are the two previous sentences, {*Current speaker*} is the speaker of the current sentence, and {*Current sentence*} is the sentence to be classified.

Label	Qwen2.5-VL	X-CLIP		V-JEPA 2		Foster et al. (2024)
	32B-Instruct	ViT-B/16	ViT-L/14	ViT-L/16	ViT-G/16	
On task student talking with student	**0.163**	0.088	0.080	0.114	0.080	0.12
Student raising hand	0.640	0.549	0.668	0.670	**0.705**	0.36
Individual technology	0.698	0.520	**0.774**	0.553	0.612	0.35
Presentation with technology	0.851	0.865	**0.870**	0.798	0.840	0.63
Student writing	0.144	0.501	**0.592**	0.502	0.563	0.43
Teacher writing	0.423	0.284	**0.501**	0.431	0.454	0.23
Using or holding book	**0.663**	0.234	0.521	0.347	0.502	0.46
Using or holding instructional tool	0.098	0.406	0.437	0.314	**0.443**	0.40
Using or holding notebook	0.384	0.374	**0.390**	0.315	0.315	0.24
Using or holding worksheet	0.388	0.599	**0.603**	0.602	0.646	0.56
Sitting at desks	0.363	0.713	**0.763**	0.722	0.762	0.74
Students sitting at group tables	0.572	0.617	0.575	0.590	0.524	**0.72**
Students sitting on carpet or floor	0.429	0.776	**0.845**	0.798	0.760	0.64
Students standing or walking	0.373	0.719	0.762	0.776	**0.788**	0.62
Teacher sitting	0.638	0.747	0.759	0.592	0.755	**0.78**
Teacher standing	0.820	0.844	0.877	0.875	**0.894**	0.64
Teacher walking	0.014	0.465	0.538	0.606	**0.624**	0.34
Teacher supporting multiple students with student interaction	0.472	0.522	0.426	0.461	0.446	**0.54**
Teacher supporting multiple students without student interaction	0.202	0.157	0.169	0.175	0.091	**0.40**
Teacher supporting one student	0.067	0.197	**0.286**	0.190	0.269	0.27
Individual activity	0.165	0.299	**0.491**	0.299	0.299	0.38
Small group activity	0.627	**0.652**	0.633	0.610	0.528	0.55
Transition	0.106	0.371	**0.426**	0.423	0.386	0.38
Whole class activity	0.854	0.817	**0.864**	0.834	0.832	0.37
Macro F1	0.423	0.513	**0.577**	0.525	0.547	0.469

Table 7. Per-label F1 scores for instructional activity recognition in videos using different models.

Label	**Llama 8B**		**Llama 70B**		**BCE**		**Focal**		**Asym**	
	Zero	Few	Zero	Few	-	PW	-	PW	-	PW
Feedback-Affirming	0.093	0.192	0.096	0.405	0.604	0.571	0.594	0.576	0.516	**0.625**
Feedback-Neutral	0.161	0.161	0.095	0.348	**0.659**	0.634	0.628	**0.659**	0.612	0.654
Feedback-Disconfirming	0.042	0.100	0.018	0.234	0.294	0.220	0.277	0.299	0.084	**0.322**
Feedback-Elaborated	0.089	0.101	0.049	0.275	**0.463**	0.420	0.444	0.451	0.387	0.459
Feedback-Unelaborated	0.190	0.142	0.088	0.344	0.581	0.603	0.588	**0.618**	0.586	0.586
Uptake-Building	0.000	0.000	0.061	0.213	0.383	0.356	0.384	0.369	0.338	**0.387**
Uptake-Exploring	0.052	0.124	0.065	0.394	0.406	0.375	0.370	0.364	0.296	**0.453**
Uptake-Restating	0.100	0.173	0.075	0.329	0.543	0.554	0.514	0.534	0.511	**0.556**
Report-Give	0.000	0.037	0.064	0.163	0.242	0.348	0.296	**0.376**	0.333	0.340
Report-Request	0.000	0.000	0.031	0.176	**0.765**	0.763	0.742	0.763	0.760	0.740
Analysis-Give	0.000	0.093	0.000	0.171	0.065	0.140	0.176	0.143	**0.214**	0.115
Analysis-Request	0.049	0.086	0.084	0.256	0.511	0.511	0.439	0.480	**0.517**	0.480
Teacher-Give	0.007	0.000	0.022	0.223	0.071	0.214	0.225	0.154	**0.294**	0.235
Teacher-Request	0.050	0.108	0.071	0.244	0.502	0.500	0.464	0.493	0.482	**0.505**
Student-Give	0.064	0.074	0.063	0.337	**0.571**	0.523	0.506	0.495	0.473	0.526
Student-Request	0.000	0.038	0.029	0.087	0.125	**0.200**	**0.200**	0.000	0.033	**0.200**
Closed-Ended	0.077	0.095	0.118	0.262	**0.657**	0.629	0.644	0.650	0.623	0.635
Open-Ended	0.161	0.195	0.167	0.456	**0.575**	0.600	0.562	0.572	0.563	0.572
Task Related Prompt	0.012	0.006	0.065	0.117	0.000	0.222	0.194	0.136	0.040	**0.344**
Macro F1	0.060	0.091	0.066	0.265	0.422	0.441	0.434	0.428	0.403	**0.460**

Table 8. Per-label F1 scores of Instructional Discourse Labels for different large language models, prompted with zero-shot and few-shot prompts and fine-tuned model with different methods to overcome class imbalance: Binary Cross-Entropy loss (BCE), Focal loss (Focal), and Asymmetric loss (Asymmetric) in their normal variants (-), as well as their variants with positive label weighting (PW).