PedGraph: Resolving Pragmatic Ambiguity in Instructional Videos through Function-Aware Event Detection

Anonymous ACL submission

Abstract

We tackle function-aware event detection in instructional videos, a challenge defined by profound pragmatic ambiguity: visually identical actions convey different intents depending on the multimodal discourse context. Existing models fail here, as they are blind to the relational logic of pedagogy. To solve this, we introduce **PedGraph**, a framework that reasons over pedagogical context by integrating a novel, data-driven and expert-validated Structured Teaching Interaction Graph (STIG). Our framework first injects the STIG's topology into semantic representations via structure-aware contrastive learning. It then performs global, context-aware reasoning using a hierarchical graph network that disentangles distinct pedagogical relations. On a new large-scale dataset, PedGraph significantly outperforms strong baselines in accurately detecting event functions. Our dataset, code, and models are available at `https://anonymous.4open.science/r/event-3A66/`.

1 Introduction

Human communication is fundamentally pragmatic: the meaning of an utterance is inseparable from its context. This is especially true for multimodal utterances, where a non-verbal cue like a gesture can function as a question, a command, or an affirmation depending on the surrounding discourse. While large-scale Vision-Language Models (VLMs) (Radford et al., 2021; Wang et al., 2021; Rasheed et al., 2023) have shown success in recognizing isolated, visually salient actions, they largely fail when tasked with interpreting such functionally ambiguous events within specialized, high-stakes domains like classroom instruction.

The primary obstacle is what we identify as a profound pragmatic ambiguity at the vision-language interface. As illustrated in Figure 1, instructional events are defined not by their visual form, but by their pedagogical function within a larger discourse structure (Zhao and Liang, 2023; Riordan et al., 2024). The exact same visual cue—a teacher's pointing gesture—serves to *direct attention* within an *Introduction* discourse segment, but its function shifts to *inviting a response* during a *Practice* segment. For a model to succeed, it must move beyond appearance and resolve the pragmatic function of multimodal events within the unfolding pedagogical discourse.

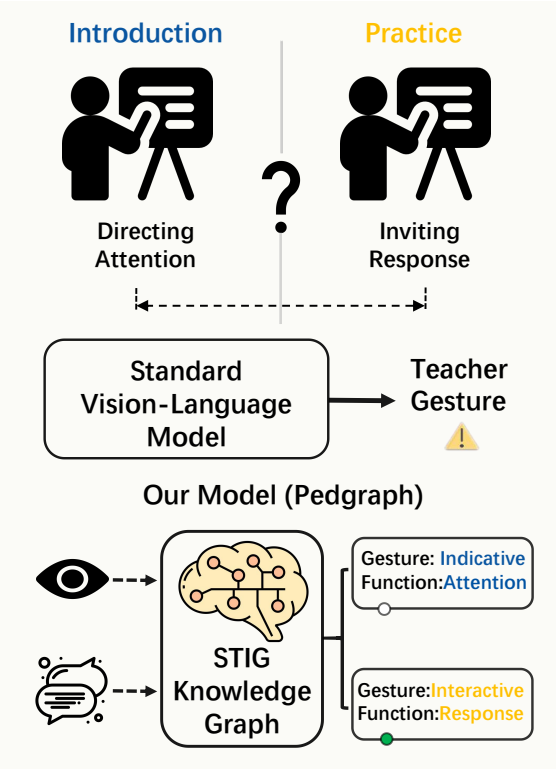

Figure 1: **The Pragmatic Ambiguity in Instructional Videos.** Standard models output a coarse label (e.g., "Teacher Gesture"). our PedGraph resolves this ambiguity. It leverages multimodal context and a structured knowledge graph (STIG) to jointly detect the specific event (e.g., "Indicative Gesture") and infer its pedagogical **function** (e.g., "Attention").

This context-dependent nature reveals a critical flaw in conventional VLMs. Such models are archi-

tecturally biased towards treating events as atomic and context-free units. Their training objectives often enforce semantic independence (Li et al., 2021; Weng et al., 2023), directly contradicting the inherently relational and structured logic of pedagogy (Wang et al., 2019). This leads to a performance ceiling that conventional methods cannot breach, stemming from their inability to simultaneously: (1) resolve deep pragmatic ambiguity; (2) represent and operationalize the underlying relational logic of discourse; and (3) perform global reasoning to achieve a coherent, video-level interpretation.

To address these limitations, we introduce **PedGraph**, a framework that operationalizes pedagogical knowledge for pragmatic reasoning. Central to our approach is the **Structured Teaching Interaction Graph (STIG)**, a novel knowledge backbone we construct via a reproducible, data-driven and expert-validated protocol. The STIG encodes multi-level event categories and diverse pedagogical relations, including contains, precedes, and causes dependencies, capturing the hierarchical and relational structure of classroom discourse. Our reasoning pipeline leverages this graph in two key stages: first, a novel **Pedagogical Structure-Aware Contrastive (PSAC) loss** injects the STIG's semantic topology into visual representations; second, a **Hierarchical Relation-Aware Graph Attention Network (HR-GAT)** performs global reasoning by disentangling distinct pedagogical relations for context-sensitive event and function inference.

Our main contributions are threefold:

- We formalize function-aware instructional event detection as a pragmatic reasoning task and propose a principled methodology to construct the **STIG**, a multi-relational knowledge graph that explicitly encodes the relational logic of pedagogy.
- We introduce a structure-aware contrastive loss (**PSAC**) that resolves pragmatic ambiguity by aligning the representation space with the STIG's semantic topology.
- We develop a hierarchical, relation-aware graph network (**HR-GAT**) that enables global, discourse-level reasoning by explicitly disentangling different pedagogical relations for robust, context-dependent inference.

Experiments on a new, large-scale classroom video dataset show that PedGraph sets a new state-of-the-art, outperforming the strongest baseline by 3.4% in absolute mAP. Our dataset, code, and models will be released.

2 Related Work

Our work sits at the crossroads of multimodal event understanding, discourse pragmatics, and knowledge-guided reasoning.

2.1 Multimodal Event Understanding

Vision–language models (VLMs) such as CLIP (Radford et al., 2021), VideoMAE (Wang et al., 2023), ActionCLIP (Wang et al., 2021), and OpenV-CLIP (Weng et al., 2023) learn shared embeddings that excel at recognizing visually distinct "denotations," but they struggle with context-dependent pedagogical functions (Zhao and Liang, 2023; Xu et al., 2025). Temporal action localization methods (e.g., ActionFormer++ (**?**)) improve segment accuracy yet remain blind to intent. PedGraph overcomes these limitations by integrating structured pedagogical knowledge to disambiguate event function.

2.2 Pragmatics and Discourse in Multimodal Contexts

Discourse theories such as Rhetorical Structure Theory (RST) (Chistova, 2023) formalize hierarchical relations that underlie textual coherence. In multimodal communication, speech, gesture, and gaze jointly convey intent (Riordan et al., 2024; Wilcock and Jokinen, 2025). Computationally, some methods model multimodal discourse for dialogue (**?**) or event detection (Shao et al., 2025), and mine generic causal relations in video (Chen et al., 2024, 2025). Yet these often lack the fine-grained, domain-specific relations crucial for pedagogy. Our STIG addresses this by operationalizing expert-validated, multi-type pedagogical relations.

2.3 Knowledge-Guided Reasoning

Structured knowledge graphs (KGs) have enhanced NLP tasks like question answering (Lewis et al., 2020; Sen et al., 2023) and fact checking (**?**). In vision, graph neural networks (GNNs) applied to scene graphs (Zhang et al., 2023) and ontologies (Mavromatis and Karypis, 2024) enable relational reasoning, but often rely on generic or superficial graphs ill-suited for specialized domains. PedGraph departs from prior work by (1) building a bespoke, expert-validated pedagogical KG from video, and (2) designing PSAC and HR-GAT

to inject and exploit its rich relational and causal semantics.

3 Methodology

To address the three core challenges outlined in our introduction—resolving **pragmatic ambiguity**, operationalizing **relational logic**, and performing **global reasoning**—our framework, **PedGraph**, operationalizes a three-stage reasoning pipeline that deeply integrates pedagogical knowledge. Figure 2 provides a visual roadmap of this pipeline, illustrating how each stage systematically builds upon the last to achieve a coherent interpretation:

- **To represent the relational logic** of teaching, we first construct the **Structured Teaching Interaction Graph (STIG)**, which serves as the central knowledge backbone for the entire framework (Section 3.1).
- **To resolve pragmatic ambiguity** at the representation level, Stage I (Section 3.2) leverages the STIG's topology via our novel **PSAC loss**, forcing the model to learn function-aware, rather than appearance-based, event representations.
- **To enable global reasoning**, Stage II first grounds candidate events in their local context before Stage III (Section 3.3 and 3.4) performs a final, video-level inference pass with our **HR-GAT**, ensuring a pedagogically coherent result.

The cornerstone of this entire process is the **STIG**, which we define first.

3.1 The Knowledge Backbone: Structured Teaching Interaction Graph (STIG)

Central to operationalizing the **relational logic** of pedagogy is a formal, interpretable representation of domain knowledge.The **Structured Teaching Interaction Graph (STIG)** serves as this knowledge backbone. It is a domain-specific, directed, and weighted graph $\mathcal{G}_{STIG} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ designed to encode the complex, multi-faceted relationships between instructional events, where $\mathcal{V}$ are event nodes, $\mathcal{E}$ are directed edges, and $\mathcal{R}$ is a set of relation types.

Structure. The STIG's nodes $\mathcal{V}$ represent the 5 macro-level lesson phases (L1, e.g., `Introduction`) and 27 fine-grained event types (L2, e.g., `verbal_question_basic`) from our taxonomy (see Appendix B.4 for the full list). Its edges $\mathcal{E}$ are typed by one of three pedagogical relation types $r \in \mathcal{R}$:

- **Structural (*contains*)**: A hierarchical link from an L1 phase to an L2 event indicates that the latter is a component of the former. For example, `verbal_lecture` *is-contains* the `Direct_Instruction` phase.
- **Temporal (*precedes*)**: A sequential link capturing the typical flow between events. For example, the `Demonstration` phase often *precedes* the `Interactive_Practice` phase.
- **Causal (*causes*)**: A functional link representing a pedagogical cause-and-effect. For example, a student's `expression_negative` may *cause* the teacher to use a `verbal_question_advanced` to clarify.

Construction. We construct the STIG via a reproducible three-phase process, performed strictly on our training set. First, to mine **temporal** relations, we fit a Neural Hawkes Process (Fabre and Muni Toke, 2025) to model event influence intensities. Second, to discover **causal** links, we introduce a **Pedagogical Granger Causality Test (P-GCT)**, which grounds causality in an event's ability to reduce uncertainty in predicting another event's pedagogical intent. Finally, all mined edges undergo a rigorous **expert-in-the-loop validation** via a double-blind Delphi protocol to ensure pedagogical soundness. This principled methodology yields a sparse and interpretable graph. Full construction details, including the P-GCT formulation and expert validation protocol, are provided in Appendix B.

3.2 Stage I: Knowledge-Guided Candidate Retrieval

To resolve the core challenge of **pragmatic ambiguity**, Stage I learns function-aware representations directly from video. As illustrated in Stage I of Figure 2, this is achieved by adapting the efficient T2L architecture (Ahmad et al., 2025). While T2L's temporal modeling is effective, its standard contrastive objective remains blind to the structured, functional semantics of our pedagogical domain. To rectify this, we propose a **synergistic dual-objective function** that optimizes for both temporal dynamics and pedagogical structure. The training objective for this stage is:

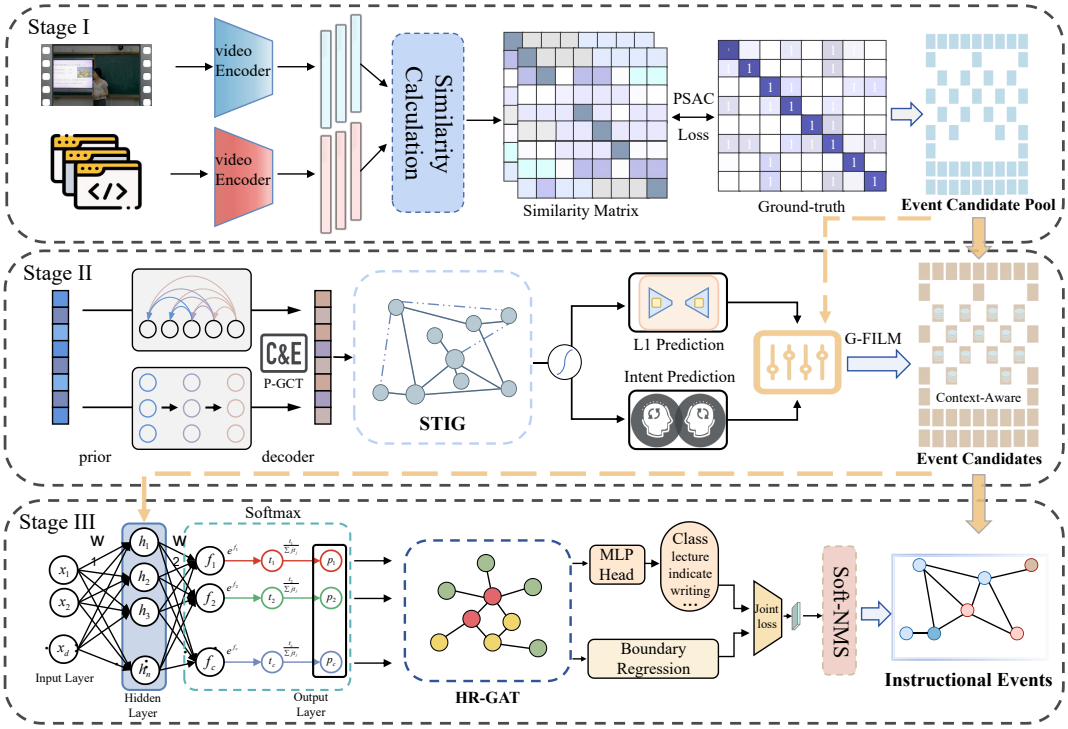

Figure 2: Overview of our multi-stage framework: The pipeline consists of three main stages. **Stage I** uses our PSAC loss, guided by the STIG, to generate a high-recall pool of event candidates. **Stage II** enriches these candidates with dynamic context (Phase and Intent) via G-FILM. **Stage III** constructs a STIG-informed instance graph and employs our HR-GAT for global optimization. Finally, a post-processing step using beam search and NMS refines these graph-based scores to produce a coherent, final set of instructional events.

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{PSAC}} + \gamma \mathcal{L}_{\text{TFD}} \tag{1}$$

where $\mathcal{L}_{\text{TFD}}$ is the Temporal Feature Diversity loss from T2L and γ is a hyperparameter. Our core contribution, $\mathcal{L}_{\text{PSAC}}$, injects pedagogical knowledge directly into the contrastive learning process.

Pedagogical Structure-Aware Contrastive (PSAC) Loss. The key insight of PSAC is to reshape the contrastive learning target using our STIG. Given a batch of N video-text pairs, let v_i and t_i be the embeddings for the i-th pair. We compute the model's predicted probability distribution $p_{v\to t}$. The probability that video v_i matches text t_k is given by the softmax over cosine similarities with a learnable temperature τ:

$$p_{v\to t}^{(i,k)} = \frac{\exp(\text{sim}(v_i, t_k)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, t_j)/\tau)} \tag{2}$$

where $p_{t\to v}$ is computed symmetrically. Instead of matching this to a hard one-hot target, PSAC defines a *soft* target distribution q', constructed from a STIG-derived similarity matrix $\mathbf{M}$:

$$\mathbf{M}_{ik} = \exp(-\lambda \cdot d(c(t_i), c(t_k))) \tag{3}$$

$$q'_{ik} = \frac{1}{Z_i}\left(\mathbf{1}_{[i=k]} + \alpha(1 - \mathbf{1}_{[i=k]})\mathbf{M}_{ik}\right) \tag{4}$$

Here, $c(t_i)$ denotes the event class corresponding to the text label of the i-th instance, and $d(c(t_i), c(t_k))$ **is the shortest path distance between these two class nodes in the STIG** $\mathcal{G}_S$, treating edges as unweighted. The hyperparameter λ controls the distance decay, $\alpha \in (0, 1)$ controls the softness, $\mathbf{1}_{[\cdot]}$ is the Iverson bracket, and Z_i is a normalization constant.

The final PSAC loss minimizes the symmetric KL-divergence between the predicted distributions p and our knowledge-aware soft target q':

$$\mathcal{L}_{\text{PSAC}} = \frac{1}{2}\left(D_{KL}(p_{v\to t} \| q') + D_{KL}(p_{t\to v} \| q')\right) \tag{5}$$

This formulation guides the model to learn an embedding space where functionally related events (e.g., `verbal_question_basic` and `verbal_question_advanced`) remain proximate. The combined objective in $\mathcal{L}_{\text{Stage1}}$ thus ensures our candidate retrieval is sensitive to both *what*

an event means (from PSAC) and *how* it unfolds over time (from TFD). Following training, we generate a high-recall pool of event candidates for each video by retaining any class with a prediction probability above a threshold τ_p, forming the input for Stage II.

3.3 Stage II: Dynamic Contextual Grounding

While Stage I makes representations function-aware, fully operationalizing the **relational logic** of discourse requires grounding each event in its specific context. Stage II (see Figure 2) performs this crucial step. It takes the candidate pool from Stage I and infers a **functional context vector** for each, which comprises both the procedural (lesson phase) and intentional (pedagogical goal) context.

Inferring Procedural Context (L1 Phase). We first determine the macro-level lesson phase for each candidate. The sequence of visual features corresponding to a candidate clip is processed by a lightweight Temporal Convolutional Network (TCN) (Bai et al., 2018). The TCN's ability to capture long-range dependencies with a small footprint makes it ideal for this task. It outputs a probability distribution, $\mathbf{p}_{\text{phase}}$, over the 5 L1 phases.

Estimating Functional Intent. Next, we estimate the underlying pedagogical intent. We first define a discrete set of high-level pedagogical intents, $\mathcal{I} = \{I_1, ..., I_M\}$, such as *Information Delivery* or *Skill Practice* (see Appendix A for the full list). For a given event candidate of class c_i and its predicted phase distribution $\mathbf{p}_{\text{phase}}$, we then perform Bayesian inference (Lu et al., 2024) to compute the posterior probability for each intent $I_m \in \mathcal{I}$:

$$p(I_m|c_i, \mathbf{p}_{\text{phase}}) \propto P(I_m) \sum_{k=1}^{5} p(l_k|\mathbf{p}_{\text{phase}}) \cdot P(I_m|c_i, l_k) \tag{6}$$

where $p(l_k|\mathbf{p}_{\text{phase}})$ is the k-th element of $\mathbf{p}_{\text{phase}}$, representing the predicted probability of the candidate being in L1 phase l_k. The prior $P(I_m)$ and the class-conditional likelihood $P(I_m|c_i, l_k)$ are estimated from the event co-occurrence statistics in our annotated training data.

Feature Recalibration with G-FILM. The phase and intent distributions are concatenated to form the final functional context vector. This vector is then used to perform context-aware feature recalibration via Grouped Feature-wise Linear Modulation (G-FILM) (Peng et al., 2022). The visual feature vector $\mathbf{v}_j$ for a candidate is partitioned into K semantic groups. For each group k, a small MLP generates context-specific scaling $\boldsymbol{\gamma}_{j,k}$ and bias $\boldsymbol{\beta}_{j,k}$ parameters:

$$\mathbf{v}'_{j,k} = \boldsymbol{\gamma}_{j,k} \odot \mathbf{v}_{j,k} + \boldsymbol{\beta}_{j,k} \tag{7}$$

This allows the model to adaptively amplify or suppress different feature channels (e.g., motion vs. appearance) based on the pedagogical context, a critical mechanism for disambiguating visually similar events.

This stage, while conceptually straightforward, is vital. It transforms the generic visual features from Stage I into contextually-aware representations, setting the stage for the global relational reasoning in Stage III.

3.4 Stage III: Global Relational Reasoning

Having generated context-aware candidates, the final challenge is to perform **global reasoning** to produce a video-level interpretation that is pedagogically coherent. As depicted in Figure 2, Stage III accomplishes this with our novel **Hierarchical Relation-aware Graph Attention Network (HR-GAT)**.

Instance Graph Construction. We first construct a directed instance graph $\mathcal{G}_{\text{inst}} = (\mathcal{V}_{\text{inst}}, \mathcal{E}_{\text{inst}})$, where each node $v_j \in \mathcal{V}_{\text{inst}}$ is a contextually-grounded candidate from Stage II. Edges are formed based on two conditions: (1) temporal adjacency (i.e., the clips are within a time window Δt), and (2) the existence of a corresponding pedagogical relation (*contains*, *precedes*, or *causes*) in the STIG between the nodes' predicted classes. This yields a sparse yet semantically-rich graph.

Hierarchical Relation-Aware Message Passing. Standard GATs are relation-blind. HR-GAT rectifies this by disentangling message passing based on the pedagogical relation type. At each layer l, a node $h_j^{(l)}$ aggregates messages from its causal neighbors $\mathcal{N}_{\text{causal}}(j)$ and its temporal/structural neighbors $\mathcal{N}_{\text{ts}}(j)$ separately.

First, we define the two message vectors, $\mathbf{m}_{\text{causal}}$ and $\mathbf{m}_{\text{ts}}$:

$$\mathbf{m}_{\text{causal}} = \sum_{k \in \mathcal{N}_{\text{causal}}(j)} g_{kj} \cdot (\mathbf{W}_{\text{val}} \mathbf{h}_k^{(l)}) \tag{8}$$

$$\mathbf{m}_{\text{ts}} = \sum_{k \in \mathcal{N}_{\text{ts}}(j)} \alpha_{kj} \cdot (\mathbf{W}_{\text{val}} \mathbf{h}_k^{(l)}) \tag{9}$$

The key difference lies in the weight calculations. Let $\mathbf{h}'_j = \mathbf{W}_{\text{att}} \mathbf{h}_j^{(l)}$ be the projected features. For strong pedagogical links, we use a **Causal Gate**

g_{kj}. For other links, we compute a relation-aware attention score e_{kj}. These are computed as:

$$g_{kj} = \sigma\left(\mathbf{W}_{\text{gate}}[\mathbf{h}_k^{(l)} \| \mathbf{r}_{\text{causal}}]\right) \cdot w_{kj}^{\text{causal}} \quad (10)$$

$$e_{kj} = \text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{h}'_j \| \mathbf{h}'_k] + b_{r_{type}}\right) \quad (11)$$

where α_{kj} is the result of applying a 'softmax' function to e_{kj} over all $k \in \mathcal{N}_{\text{ts}}(j)$. In Eq. 10, σ is the sigmoid function, $[\cdot\|\cdot]$ denotes concatenation, $\mathbf{r}_{\text{causal}}$ is a learnable embedding for the causal relation type, and the scalar causality strength from the STIG, w_{kj}^{causal}, directly modulates the gate. In Eq. 11, $\mathbf{a}$ is the attention query vector and $b_{r_{type}}$ is a learnable scalar bias for each relation type.

Final Node Update. The aggregated messages are then integrated with the node's self-representation via a residual connection. This provides the complete HR-GAT update rule that was previously underspecified:

$$\mathbf{h}_j^{(l+1)} = \text{LayerNorm}\left(\mathbf{W}_{\text{self}}\mathbf{h}_j^{(l)} + \mathbf{m}_{\text{causal}} + \mathbf{m}_{\text{ts}}\right) \quad (12)$$

By stacking L HR-GAT layers, each node's final representation captures information from its L-hop neighborhood, enabling true global reasoning. The matrices $\mathbf{W}_{\text{val}}$, $\mathbf{W}_{\text{gate}}$, $\mathbf{W}_{\text{att}}$, and $\mathbf{W}_{\text{self}}$ are learnable weight matrices.

Joint Optimization and Inference Post-processing. The final node embeddings from the HR-GAT, $\{\mathbf{h}_j^{(L)}\}$, are used for the final prediction tasks. We employ a multi-task learning setup with two output heads:

- A **classification head** (a 2-layer MLP with a Softmax activation) predicts the probability distribution over the 32 event classes. It is supervised using a standard cross-entropy loss.
- A **regression head** (a 2-layer MLP) predicts temporal boundary refinements (center offset and duration scaling). It is supervised using a Smooth L1 loss.

At inference time, the raw detections are cleaned using a standard post-processing pipeline. First, detections with a confidence score below 0.01 are discarded. Then, to merge redundant proposals, we apply Soft-NMS. This technique is particularly suited for our task as it gracefully decays the scores of overlapping predictions rather than harshly removing them, which is crucial for preserving distinct but temporally proximal instructional events. Full implementation and hyperparameter details are provided in Appendix C.3 and D.3.

4 Experiments

We evaluate PedGraph's effectiveness in addressing the key challenges of pragmatic ambiguity, relational logic, and global reasoning in instructional event detection. Next, we detail the dataset, setup, main quantitative and qualitative results, and ablation studies to substantiate our contributions.

4.1 The PEA Dataset: A Benchmark for Instructional Analysis

Our work is validated on the **P**edagogical **E**vents in **A**ction (**PEA**) dataset, a new, densely annotated benchmark we developed for fine-grained analysis of instructional videos. The dataset comprises 15.2 hours of footage from 113 classroom lessons taught by over 100 pre-service teachers. It is annotated with a two-level hierarchy of 32 event classes and contains over 10,000 meticulously labeled event instances. To prevent data leakage and ensure a rigorous test of generalization, we partition the data at the teacher and classroom level into training, validation, and test sets, following an approximate 70%/12%/18% split, respectively.

The PEA dataset is designed to reflect real-world complexity, presenting three key modeling challenges. First, it exhibits a realistic **long-tail class distribution**, where many crucial pedagogical cues are infrequent. Second, it has a high degree of **temporal density**, with an average of 1.4 events occurring simultaneously at any given moment. Finally, the data shows significant **event co-occurrence** (e.g., gestures contained within lectures), necessitating models that can handle overlapping, interdependent actions. To certify annotation quality, 10% of the data was cross-annotated, achieving a Fleiss' Kappa of **0.78** (substantial agreement).

For model training, we convert the variable-length ground-truth annotations into a large set of fixed-size training clips via an automated labeling pipeline. Full dataset statistics, including per-class distribution, annotation protocols, and data processing details, are provided in Appendix A.

4.2 Experimental Setup

Evaluation Metrics. To provide a comprehensive assessment, we use standard metrics for temporal action localization. We report mean Average Precision (mAP) at temporal Intersection-over-

Method	R@5	R@10	R@50	AR@100	mAP@0.5	mAP@0.3
VideoLLaMA-7B*	11.2	27.6	40.2	28.9	18.3	29.1
VideoChatGPT-7B*	13.1	28.2	40.7	30.2	19.8	30.5
TimeSformer	$14.8_{\pm 0.3}$	$31.0_{\pm 0.4}$	$43.8_{\pm 0.4}$	$31.1_{\pm 0.5}$	$21.6_{\pm 0.3}$	$32.4_{\pm 0.4}$
VideoMAE	$19.1_{\pm 0.4}$	$33.9_{\pm 0.5}$	$47.3_{\pm 0.5}$	$37.5_{\pm 0.6}$	$24.5_{\pm 0.4}$	$35.8_{\pm 0.5}$
ActionCLIP	$19.4_{\pm 0.5}$	$34.6_{\pm 0.5}$	$48.9_{\pm 0.6}$	$40.2_{\pm 0.6}$	$26.8_{\pm 0.5}$	$38.1_{\pm 0.6}$
EZ-CLIP	$20.3_{\pm 0.4}$	$35.5_{\pm 0.6}$	$50.1_{\pm 0.6}$	$41.8_{\pm 0.7}$	$27.4_{\pm 0.5}$	$39.2_{\pm 0.6}$
ActionFormer++	$21.5_{\pm 0.4}$	$39.9_{\pm 0.5}$	$54.1_{\pm 0.6}$	$45.8_{\pm 0.7}$	$28.2_{\pm 0.6}$	$41.9_{\pm 0.6}$
Ours (ViT-B/32, 8-frame)	$21.8_{\pm 0.3}$	$40.9_{\pm 0.4}$	$54.7_{\pm 0.5}$	$45.8_{\pm 0.6}$	$29.7_{\pm 0.5}$	$43.8_{\pm 0.5}$
Ours (ViT-B/16, 8-frame)	$22.1_{\pm 0.3}$	$41.4_{\pm 0.4}$	$55.2_{\pm 0.5}$	$46.9_{\pm 0.6}$	$30.4_{\pm 0.5}$	$44.5_{\pm 0.5}$
Ours (ViT-B/16, 16-frame)	$\mathbf{22.6_{\pm 0.3}}$	$\mathbf{42.1_{\pm 0.4}}$	$\mathbf{56.0_{\pm 0.5}}$	$\mathbf{47.8_{\pm 0.6}}$	$\mathbf{31.6_{\pm 0.5}}$	$\mathbf{45.5_{\pm 0.6}}$

Table 1: Main comparison on the PEA test set. All results for fine-tuned models are reported as mean ± standard deviation over 5 runs. *Models evaluated zero-shot. Our knowledge-guided framework significantly outperforms all baselines, validating our approach to resolving pragmatic ambiguity and incorporating relational logic and global reasoning.

Union (tIoU) thresholds of **0.5 and 0.3**, providing evaluation at both strict and looser localization criteria. We also report Recall@k (R@k, for k=5, 10, 50) and Average Recall at 100 proposals (AR@100) to evaluate the quality of the retrieved event candidates. All reported scores for fine-tuned models are the mean and standard deviation over five runs with different random seeds.

Implementation Details. Our framework is implemented in PyTorch and was trained on a distributed setup of one NVIDIA A6000 and two RTX 3090 GPUs. For fair comparison, all baseline models were adapted to our task using their official open-source implementations and fine-tuned on our PEA training set until validation performance converged. A critical methodological distinction applies to the large generative models. Due to their immense scale and the lack of established fine-tuning protocols for this specific detection task, **VideoLLaMA-7B and VideoChatGPT-7B were evaluated in a zero-shot, prompt-based setting.** This standard practice benchmarks their out-of-the-box capabilities on our novel domain. Detailed hyperparameters for our model and all fine-tuned baselines are provided in Appendix D.3 .

4.3 Main Results

Baselines. To rigorously situate our contributions, we benchmark PedGraph against a comprehensive suite of models spanning two main categories. First, we evaluate a range of **generic video-language models**, including foundational architectures like TimeSformer, powerful CLIP-based methods such as ActionCLIP and the highly efficient EZ-CLIP, as well as large-scale generative models (VideoLLaMA, VideoChatGPT) in a zero-shot setting. Second, to ensure our comparison is state-of-the-art, we compare against ActionFormer++, a top-performing, open-source, and fine-tunable temporal action localization (TAL) model.

Quantitative Comparison. Table 1 presents our main results on the PEA test set. PedGraph consistently and significantly outperforms all baseline categories across every metric. Our final model, **PedGraph (ViT-B/16, 16-frame)**, achieves a mean mAP@0.5 of **31.6 ± 0.5**, marking a substantial +3.4 point absolute improvement over the strong TAL baseline, ActionFormer++. The low standard deviation across all runs underscores the stability of our training procedure.

The results reveal key insights. The TAL model, ActionFormer++, while effective at proposal generation, does not explicitly model pedagogical context. Our error analysis confirms it struggles to disambiguate functionally distinct but visually similar events, such as *gesture_indicator* and *gesture_interactive*. The performance gap widens at the looser mAP@0.3 threshold (+3.6 points), demonstrating that PedGraph is also more reliable at identifying the complete temporal extent of events, which validates our multi-stage reasoning approach.

4.4 Qualitative and Human Evaluation

To highlight PedGraph's ability to reason over pedagogy rather than appearance alone, we analyze

a challenging scenario where a teacher's pointing gesture follows a *verbal_question*. While visually identical to a gesture aimed at a whiteboard, its function is interactive. The strong baseline, ActionFormer++, defaults to the visually common *gesture_indicator* label. In contrast, PedGraph correctly leverages the preceding context and the causal links in the STIG to identify the gesture's true function as *gesture_interactive*. This case compellingly shows that our model's advantage stems from a structured understanding of pedagogical logic, not just visual processing. A detailed analysis of success and failure cases is provided in Appendix E.3.

Furthermore, we conducted a controlled human study with ten in-service K-12 teachers to complement our automatic metrics. As shown in Table 2, PedGraph was rated significantly higher than all baselines on key dimensions of detection quality, especially **Context Awareness** and **Detail Sensitivity**. The study confirms our model's superior human-perceived accuracy and interpretability. The full study protocol and results are in Appendix E.4.

Model	Temp.	Corr.	Detail	Cont.	Consist.	Mean
VideoLLaMA*	1.82	1.96	2.18	2.12	1.79	1.98
VideoChatGPT*	1.98	2.23	2.30	2.30	2.10	2.19
VideoMAE	2.34	2.40	2.56	2.52	2.37	2.43
ActionFormer++	2.81	2.90	3.05	3.01	2.65	2.88
w/o STIG	2.58	2.80	2.95	2.96	2.74	2.81
w/o HR-GAT	2.67	2.72	2.91	3.07	2.63	2.80
PedGraph	**3.22**	**2.92**	**3.23**	**3.52**	**3.05**	**3.19**

Table 2: Human evaluation results on a 1-5 Likert scale. Headers denote Temporal Boundaries (Temp.), Correctness (Corr.), Detail Sensitivity (Detail), Context Awareness (Cont.), and Overall Consistency (Consist.). *Models evaluated zero-shot.

4.5 Ablation Studies

We systematically ablate our model's components (Table 3). Removing the STIG causes the largest performance drop (-4.9 mAP@0.5), confirming the criticality of pedagogical knowledge. We further analyze HR-GAT by ablating its relation types, finding both are crucial for performance. Finally, replacing our PSAC Loss with a standard supervised contrastive loss (w/o PSAC) also hurts performance, validating our graph-aware design.

The qualitative effect of this is visualized in Figure 3. Without PSAC (A), embeddings cluster by visual similarity; with PSAC (B), they reorganize according to pedagogical logic. This confirms PSAC's key role in learning a function-aware representation.

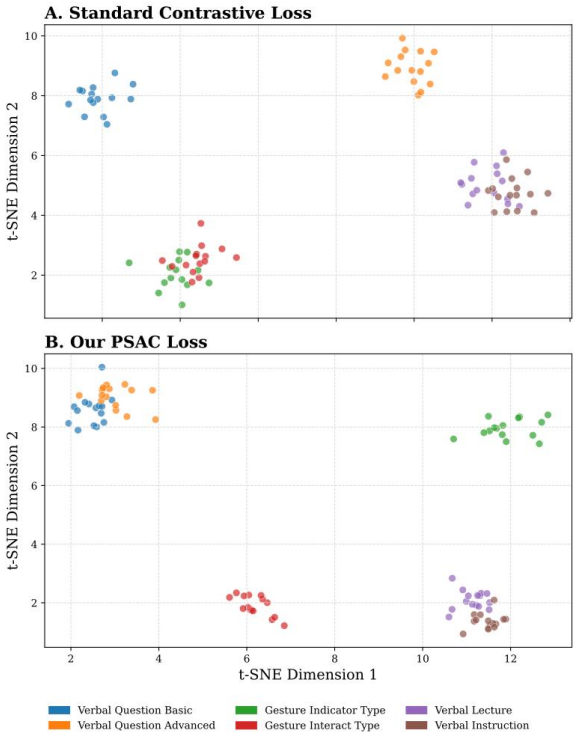

Figure 3: PSAC Loss Organizes the Embedding Space.

Model Variant	mAP@0.5
Full PedGraph Model	31.6
w/o STIG	26.7
w/o HR-GAT	28.5
w/o PSAC	30.1
HR-GAT w/o Causal Rel.	29.5
HR-GAT w/o Temporal Rel.	29.2

Table 3: Ablation results.

5 Conclusion

We introduce PedGraph, a framework for globally consistent and interpretable instructional event detection. Guided by a pedagogical knowledge graph (STIG), it leverages contrastive learning and a hierarchical graph network to resolve semantic ambiguities via contextual reasoning. PedGraph establishes a new state-of-the-art, and future work will focus on lightweight architectures and cross-domain transfer.

6 Limitations

Modality Scope. Our current framework, while powerful, relies primarily on visual and structural cues. A primary limitation and a significant avenue for future work is the integration of the spoken language modality. Fusing automatic speech recognition (ASR) transcripts would provide a rich, explicit stream of semantic information, likely offering powerful signals for disambiguating the function of co-occurring events and further bridging the pragmatic gap we identify.

Scalability of Knowledge Construction. While our STIG construction methodology is principled and reproducible, it requires domain expertise for the final validation phase. Applying PedGraph to new domains (e.g., medical training, sports coaching) would necessitate a new cycle of data collection and expert validation, posing a challenge for rapid, large-scale generalization. Future work could investigate semi-supervised or few-shot methods for adapting the STIG to new domains with minimal expert intervention.

Computational Cost. The multi-stage nature of PedGraph prioritizes accuracy and interpretability over computational efficiency. The current inference time, as detailed in the appendix, makes it suitable for detailed offline analysis but not for real-time feedback applications. Exploring model distillation or unified end-to-end architectures to create more lightweight versions is a valuable direction for future research.

Potential Risks and Broader Impact While our work is intended for educational research, any technology used to analyze human behavior carries potential risks. A misapplication of this technology could lead to overly simplistic, high-stakes evaluations of teacher performance, ignoring the rich context that our model aims to capture. We advocate for its use as a tool for descriptive, formative feedback and research, not for summative, automated judgment. Furthermore, biases present in the training data (e.g., teaching styles, subject matter) could be reflected in the model's analysis, and care must be taken when generalizing its findings.

References

Shahzad Ahmad, Sukalpa Chanda, and Yogesh S Rawat. 2025. T2L: Efficient Zero-Shot Action Recognition with Temporal Token Learning. *Transactions on Machine Learning Research (TMLR)*. Accepted May 2025. Available on OpenReview.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, and Weiyao Lin. 2024. Mecd: Unlocking multi-event causal discovery in video reasoning. *Preprint*, arXiv:2409.17647.

Tieyuan Chen, Huabin Liu, Yi Wang, Yihang Chen, Tianyao He, Chaofan Gan, Huanyu He, and Weiyao Lin. 2025. Mecd+: Unlocking event-level causal graph discovery for video reasoning. *Preprint*, arXiv:2501.07227.

Elena Chistova. 2023. End-to-end argument mining over varying rhetorical structures. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3376–3391, Toronto, Canada. Association for Computational Linguistics.

Timothee Fabre and Ioane Muni Toke. 2025. Neural hawkes: Non-parametric estimation in high dimension and causality analysis in cryptocurrency markets. *Quantitative Finance*, 25(5):671–698.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Preprint*, arXiv:2107.07651.

Feihong Lu, Weiqi Wang, Yangyifei Luo, Ziqin Zhu, Qingyun Sun, Baixuan Xu, Haochen Shi, Shiqi Gao, Qian Li, Yangqiu Song, and Jianxin Li. 2024. MIKO: Multimodal intention knowledge distillation from large language models for social-media commonsense discovery. *Preprint*, arXiv:2402.18169.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *Preprint*, arXiv:2405.20139.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. *arXiv preprint arXiv:2203.15332*.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video learners. *Preprint*, arXiv:2212.03640.

J. P. Riordan, L. Revell, B. Bowie, S. Hulbert, M. Woolley, and C. Thomas. 2024. Multimodal classroom interaction analysis using video-based methods of the pedagogical tactic of (un)grouping. *Pedagogies: An International Journal*, 20(2):285–302.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, Hao Tang, Yan Wang, and Shuyan Li. 2025. Eventvad: Training-free event-aware video anomaly detection. *Preprint*, arXiv:2504.13092.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. *Preprint*, arXiv:2303.16727.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *Preprint*, arXiv:2109.08472.

Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation. AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. 2023. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*.

Graham Wilcock and Kristiina Jokinen. 2025. Integrating conversational entities and dialogue histories with knowledge graphs and generative AI. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 290–298, Bilbao, Spain. Association for Computational Linguistics.

Hao Xu, Arbind Agrahari Baniya, Sam Well, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. 2025. Action spotting and precise event detection in sports: Datasets, methods, and challenges. *Preprint*, arXiv:2505.03991.

Yongqi Zhang, Zhanke Zhou, Quanming Yao, Xiaowen Chu, and Bo Han. 2023. Adaprop: Learning adaptive propagation for graph neural network based knowledge graph reasoning. *Preprint*, arXiv:2205.15319.

Qianyi Zhao and Zhiqiang Liang. 2023. Research on multimodal based learning evaluation method in smart classroom. *Learning and Motivation*, 84:101943.

This appendix provides supplementary material to ensure the clarity, depth, and reproducibility of our work. It is structured as follows: Section A provides formal definitions for all key terms and notations. Section B details the construction of our knowledge backbone, the STIG. Sections C and D provide exhaustive implementation details for all stages of our model. Section E contains additional experimental results and analyses. Finally, Section F provides a complete reproducibility statement.

A Definitions, Notation, and Event Taxonomy

To ensure clarity for a broad audience, we formally define all key terms, event classes, and mathematical notations used throughout the paper in Table 4.

B STIG Construction: Full Protocol

The STIG is constructed via a principled, three-phase protocol performed exclusively on the training set to prevent any data leakage. This process, inspired by recent advances in causal discovery (Chen et al., 2024), combines data-driven mining with expert validation to ensure pedagogical soundness and reproducibility(**see Appendix B for the full protocol**).

B.1 Temporal Relation Mining

To discover statistically significant sequential patterns ($precedes$ relations), we model the L2 event stream as a multivariate temporal point process.

Neural Hawkes Process (NHP). We employ a Neural Hawkes Process, whose conditional intensity function for an event of type k at time t is defined as:

$$\lambda_k(t) = f(\mathbf{h}_k(t)) \tag{13}$$

where the history embedding $\mathbf{h}_k(t)$ is computed by an RNN that processes the sequence of preceding events $\{(e_i, t_i) | t_i < t\}$. This allows the model to capture complex, non-linear temporal dependencies. The model is trained by maximizing the log-likelihood of observing the ground-truth event sequence.

Edge Extraction. After training, to determine if event type i precedes type j, we compute the influence score by simulating the change in intensity $\lambda_j(t)$ immediately after an occurrence of event i. If this influence score exceeds a threshold θ_{NHP}, an edge $(i, j, precedes)$ is created. θ_{NHP} is set to the 80th percentile of all non-zero influence scores, a value determined by grid search on the validation set to maximize recall of known pedagogical sequences.

Term / Symbol	Definition
Core Concepts	
PEA Dataset	Our **P**edagogical **E**vents in **A**ction dataset.
Function-Aware Detection	The task of detecting an event's class, boundaries, and its underlying pedagogical function.
L1 Phases	
Introduction	Lesson opening; activities to motivate and introduce topics.
Direct_Instruction	Core teaching segment; teacher-led explanation of concepts.
Demonstration	Teacher shows how to perform a task or solve a problem.
Interactive_Practice	Student-centric segment for exercises or hands-on tasks.
Summary	Concluding segment reviewing key points.
L2 Events (Representative Examples)	
Verbal	*verbal_lecture, verbal_question_basic, verbal_praise, verbal_criticism, verbal_instruction, etc.*
Non-Verbal	*gesture_indicator, gesture_descriptive, expression_positive, position_blackboard, technology_resource_display, etc.*
(Note: Full taxonomy in Appendix A includes 27 L2 events.)	
Mathematical Notation	
$\mathcal{G}_S = (\mathcal{V}, \mathcal{E}, \mathcal{R})$	The STIG, our knowledge backbone, with nodes $\mathcal{V}$, edges $\mathcal{E}$, and relation types $\mathcal{R}$.
$r \in \mathcal{R}$	A relation type in $\mathcal{G}_S$: $contains$, $precedes$, or $causes$.
$d(c_i, c_k)$	Shortest path distance between class nodes c_i, c_k in $\mathcal{G}_S$.
$I_m \in \mathcal{I}$	A pedagogical intent from the set of M defined intents.
$\mathbf{p}_{\text{phase}}$	Probability distribution over L1 phases from the TCN.
$\mathcal{L}_{\text{PSAC}}$	Our Pedagogical Structure-Aware Contrastive loss.
$\gamma, \alpha, \lambda, \tau$	Key hyperparameters for Stage I training.
g_{kj}	The scalar Causal Gate value in our HR-GAT.
Δt	Temporal window size for instance graph construction.

Table 4: Consolidated glossary of terms, event classes, and symbols.

B.2 Causal Relation Mining

To infer functional cause-and-effect links ($causes$), we adapt the logic from (Chen et al., 2024) to our domain with a Pedagogical Granger Causality Test (P-GCT), detailed in Algorithm 1.

P-GCT Protocol. The core idea is to test if a candidate causal event C provides statistically significant information for predicting a subsequent target event I, beyond the information already available from the broader context (i.e., the current L1 phase). We train a lightweight intent prediction model, $f_{predict}$, which is a 2-layer Transformer encoder, to predict the occurrence of I in a future time window. We compare two variants: one that sees the features of C ($f^{C}_{predict}$) and one that does not ($f^{\neg C}_{predict}$).

Algorithm 1: Pedagogical Granger Causality Test (P-GCT)

1: **Input:** Training set event sequences, candidate cause event class C, target event class I.
2: Initialize intent prediction model $f_{predict}$.
3: **Train Baseline Model:** Train $f^{\neg C}_{predict}$ to predict I using only general context (e.g., L1 phase features).
4: **Train Causal Model:** Train $f^{C}_{predict}$ to predict I using general context **and** features from occurrences of C.
5: Compute F1 scores for both models on the validation set: $F1^{\neg C}$ and $F1^{C}$.
6: Calculate the performance gain: $\Delta_{F1} = F1^{C} - F1^{\neg C}$.
7: Determine the optimal F1 gain threshold θ_{F1} by maximizing F1-score for known causal pairs on the validation set.
8: **if** $\Delta_{F1} > \theta_{F1}$ **then**
9: Add edge $(C, I, causes)$ to the candidate graph with weight $w^{\text{causal}} \propto \Delta_{F1}$.
10: **end if**

Thresholding. The threshold θ_{F1} was empirically set to 0.08 based on grid search over the validation set, providing the best trade-off between precision and recall for identifying causal links. This ensures that only pairs with a meaningful predictive relationship are included.

B.3 Expert-in-the-Loop Validation

All edges mined from the data-driven phases undergo a final validation to ensure pedagogical soundness.

Delphi Protocol. We employed a two-round, double-blind Delphi protocol with two educational science experts (Ph.D. level, >5 years experience). In Round 1, experts independently rated each mined edge on a 5-point Likert scale for **Pedagogical Soundness** (1: Unrelated, 3: Plausible, 5: Strongly Justified). In Round 2, they were shown the anonymized rating and justification from the other expert and could revise their score.

Edge Retention. An edge was retained in the final STIG only if its average final score was ≥ 4.0. This rigorous process pruned spurious correlations, yielding the final, sparse, and interpretable graph whose statistics are shown in Table 5. The final inter-rater reliability (Fleiss' Kappa) was 0.78, indicating substantial agreement.

Relation Type	Nodes	Edges	Graph Density
contains	32	35	0.035
precedes	32	68	0.068
causes	32	41	0.041
Total (Unique Edges)	**32**	**144**	**0.14**

Table 5: Final statistics of the constructed STIG after all three phases, showing a sparse and structured graph.

C Detailed Implementation of the PedGraph Pipeline

This section provides a comprehensive breakdown of the implementation for all three stages of the PedGraph framework, ensuring full reproducibility. All feature dimensions throughout the pipeline are consistently maintained at 512, aligning with the output dimension of the foundational T2L (Ahmad et al., 2025) and the underlying CLIP ViT-B/16 architecture.

C.1 PSAC Loss Formulation.

Stage I generates function-aware candidates using our novel PSAC loss. The training objective is $\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{PSAC} + \gamma \mathcal{L}_{TFD}$, where $\mathcal{L}_{TFD}$ is the Temporal Feature Diversity loss from T2L (Ahmad et al., 2025), which encourages temporal uniqueness in the learned embeddings.

The PSAC loss (Eq. 5) minimizes the KL-divergence between the model's predicted video-text similarity distribution p and a soft target distribution q' derived from the STIG. The key is the construction of the STIG similarity matrix $\mathbf{M}$ and the soft target q':

$$\mathbf{M}_{ik} = \exp(-\lambda \cdot d(c_i, c_k)) \tag{14}$$

$$q'_{ik} = \frac{1}{Z_i}\left(\mathbf{1}_{[i=k]} + \alpha(1 - \mathbf{1}_{[i=k]})\mathbf{M}_{ik}\right) \tag{15}$$

where $d(c_i, c_k)$ is the shortest path distance in the STIG, λ is a decay parameter, α controls the softness of non-diagonal targets, and Z_i is a normalization constant. This encourages the model to learn an embedding space where pedagogically related events are closer.

Hyperparameter Details. The key hyperparameters for this stage, justified by the sensitivity analysis in Table 6 and 7, are:

- $\gamma = 0.1$: Balances the contribution of the PSAC and TFD losses.
- $\alpha = 0.2$: Provides a moderate degree of softness to the contrastive targets, encouraging representation of semantic similarity without overwhelming the primary classification signal.
- $\lambda = 0.5$: A decay factor that ensures closely related nodes in the STIG have a strong influence, while distant nodes have a minimal effect.
- τ: The temperature parameter for the softmax is initialized at 0.07 and learned during training.

Window/Stride (s)		**tIoU Thresh. (pos/neg)**	
Value	mAP@0.5	Value	mAP@0.5
1 / 0.5	28.9	0.5 / 0.1	29.7
3 / 1	**30.1**	0.35 / 0.1	29.8
5 / 1	29.6	**0.25 / 0.08**	**30.1**
5 / 2	29.0	0.15 / 0.05	29.2

Table 6: Sensitivity analysis for Stage I clip generation.

Loss Weight γ		Softness α		Decay λ	
Value	mAP@0.5	Value	mAP@0.5	Value	mAP@0.5
0.01	29.3	0.1	29.8	0.2	29.9
0.1	**30.1**	**0.2**	**30.1**	**0.5**	**30.1**
0.5	29.7	0.4	29.5	1.0	29.6
1.0	29.2	0.6	29.1	2.0	28.9

Table 7: Sensitivity analysis for PSAC loss hyperparameters.

C.2 Stage II: Dynamic Contextual Grounding

The TCN used for L1 phase prediction is lightweight, consisting of 3 layers with a kernel size of 3 and dilation factors of (1, 2, 4). The functional intent posterior (Eq. 6) uses priors pre-computed from the training set co-occurrence statistics. For G-FILM (Eq. 7), we partition the 512-dim visual features into $K = 4$ semantic groups.

C.3 Stage III: Global Relational Reasoning

To ensure full reproducibility, we provide the detailed procedure and pseudo-code for our HR-GAT module in Algorithm 2. The key parameters, validated in Table 8, are $L = 3$ layers and 4 attention heads. The input/output feature dimension is 512.

Algorithm 2: HR-GAT Forward Pass (Single Layer)

Input: Node features $H^{(l)}$, STIG-derived instance graph $\mathcal{G}_{\text{inst}}$
Output: Updated node features $H^{(l+1)}$
$H_{val} \leftarrow H^{(l)}\mathbf{W}_{\text{val}}^T$, $H_{att} \leftarrow H^{(l)}\mathbf{W}_{\text{att}}^T$
for all node $j \in \mathcal{V}_{\text{inst}}$ **do**
  $\mathbf{m}_{\text{causal}} \leftarrow 0$, $\mathbf{m}_{\text{ts}} \leftarrow 0$
  $\mathcal{N}_{\text{causal}} \leftarrow$ causal neighbors of j
  $\mathcal{N}_{\text{ts}} \leftarrow$ temporal/structural neighbors of j
  // Causal Message Passing
  for all neighbor $k \in \mathcal{N}_{\text{causal}}$ **do**
    $g_{kj} \leftarrow \sigma(\mathbf{W}_{\text{gate}}[\mathbf{h}_k^{(l)} \| \mathbf{r}_{\text{causal}}]) \cdot w_{kj}^{\text{causal}}$
    $\mathbf{m}_{\text{causal}} \leftarrow \mathbf{m}_{\text{causal}} + g_{kj} \cdot \mathbf{h}_{k,val}$
  end for
  // Temporal/Structural Message Passing
  Compute all e_{kj} for $k \in \mathcal{N}_{\text{ts}}$ using Eq. 11
  Normalize $(\alpha_{kj}) \leftarrow \text{softmax}(e_{kj})$ over $k \in \mathcal{N}_{\text{ts}}$
  for all neighbor $k \in \mathcal{N}_{\text{ts}}$ **do**
    $\mathbf{m}_{\text{ts}} \leftarrow \mathbf{m}_{\text{ts}} + \alpha_{kj} \cdot \mathbf{h}_{k,val}$
  end for
  // Update Node
  $\mathbf{h}_j^{(l+1)} \leftarrow \text{LayerNorm}(\mathbf{h}_j^{(l)}\mathbf{W}_{\text{self}}^T + \mathbf{m}_{\text{causal}} + \mathbf{m}_{\text{ts}})$
end for
return $H^{(l+1)}$

D Experimental Setup And Evaluation Metrics

D.1 The PEA Dataset: A Benchmark for Pedagogical Reasoning

The central empirical contribution of our work is the introduction of the **P**edagogical **E**vents in **A**ction (**PEA**) dataset, a new, large-scale benchmark designed specifically to address the challenge of function-aware instructional event detection. The lack of such a resource has been a major bottleneck for research in this area.

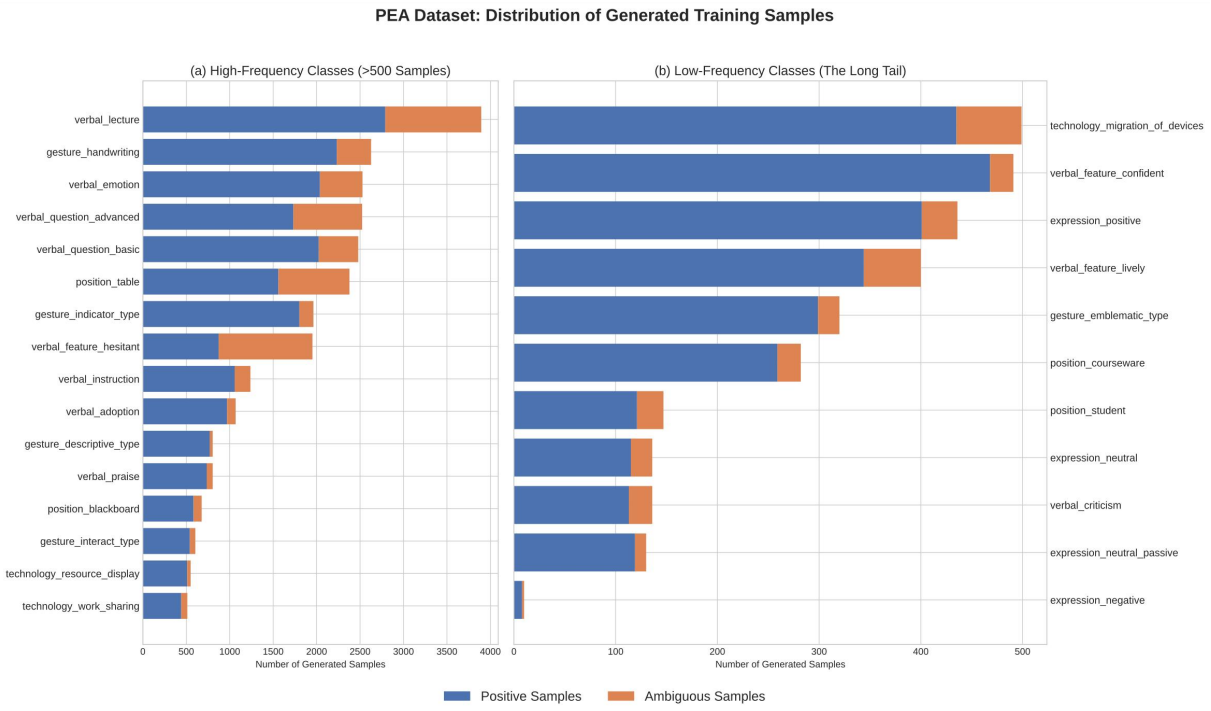

Figure 4: Distribution of Generated Training Samples in the PEA Dataset. (a) High-frequency classes. (b) Long-tail low-frequency classes.

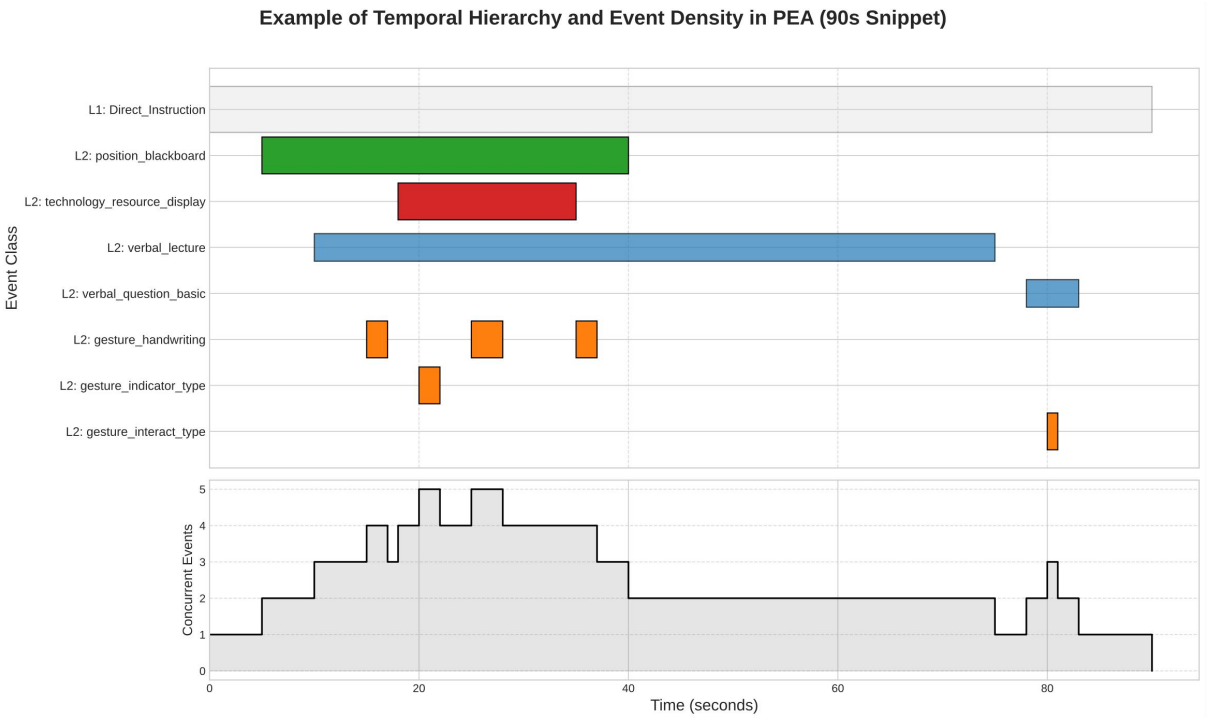

Figure 5: Temporal Distribution of L2 Event Classes, illustrating the vast differences in event duration that motivate our multi-scale architecture.

Motivation and Scope. Existing action recognition datasets either focus on generic actions (e.g., Kinetics) or lack the fine-grained, functional labels necessary for deep pedagogical understanding. To fill this gap, we collected 113 real-world, unscripted classroom videos, totaling 15.2 hours. The videos capture a diverse range of instructional scenarios across two subjects: Mathematics (65

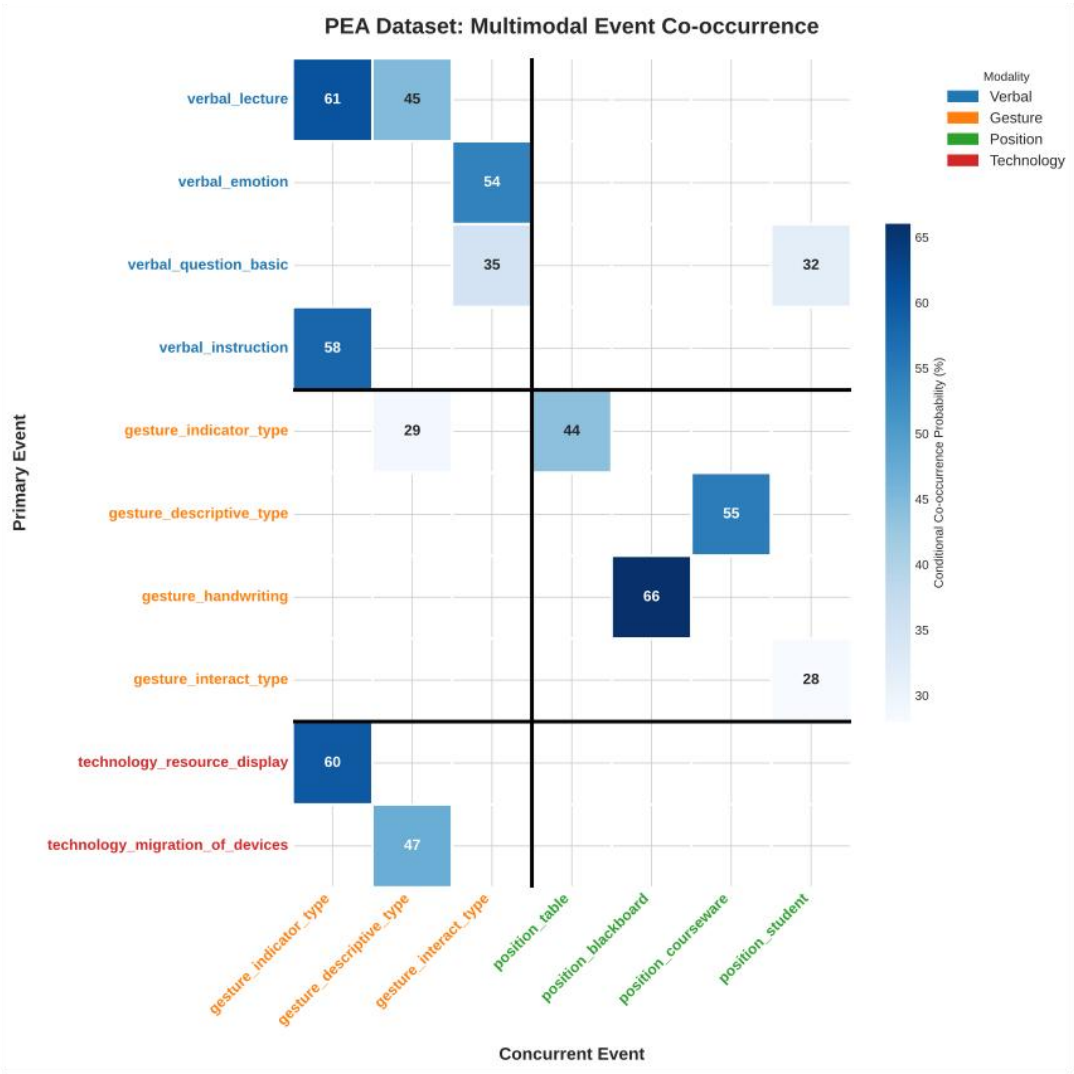

Figure 6: Heatmap of Multimodal Event Co-occurrence, showing strong dependencies that motivate our graph-based modeling.

videos) and Information Technology (48 videos), ensuring our dataset is not biased towards a single teaching style.

Annotation Protocol and Quality Assurance. To ensure the highest data quality, we designed and executed a rigorous, multi-stage annotation protocol:

1. **Guideline Formulation:** We developed a comprehensive 15-page annotation manual. For each of the 32 event classes, the manual provided: (a) a formal text definition, (b) two canonical positive visual examples (screenshots), (c) two canonical negative examples to highlight common confusions, and (d) specific rules for determining temporal boundaries (e.g., "a gesture begins when the hand leaves a resting position and ends when it returns"). This detailed guide was the primary reference for all annotators.

2. **Annotator Training and Selection:** We recruited three trainee teachers, each with 1-2 years of classroom internship experience. This background provided them with authentic domain expertise. Each annotator underwent a 10-hour paid training phase to master the guidelines and calibrate their judgments on sample videos.

3. **Annotation and Adjudication:** Annotations were performed using `Label U`, a professional video annotation tool. Each video was independently annotated by two annotators. A third, senior annotator then resolved all disagreements in a final adjudication pass.

4. **Inter-Annotator Agreement (IAA):** Before adjudication, we calculated the Fleiss' Kappa score on a 15% subset of the data. The resulting score of **0.78** indicates *substantial agreement*, a strong result for a task with this level of complexity and label granularity, confirming the reliability of our annotation scheme.

Ethical Considerations. All videos were collected with informed consent from the participating institutions and instructors. To protect annotator welfare and incentivize high-quality, dedicated work, all annotators were compensated at a rate of 2.5 times the local minimum hourly wage. This ethical commitment ensures the sustainability and reliability of our data collection effort.

Data Splits and Integrity. To prevent any data leakage and ensure a fair evaluation of generalization, the dataset was split at the *session level*. This means videos from the same teacher on the same day were kept within the same split. The final splits are: Training (80 videos, 11.1 hours), Validation (13 videos, 1.9 hours), and Test (20 videos, 2.2 hours). All hyperparameters and model design choices were made using only the training and validation sets.

Key Statistical Challenges. The PEA dataset authentically reflects the complexities of real-world classrooms, presenting three key challenges that directly motivate our PedGraph design: a pronounced **long-tail distribution** of events (Figure 4), **extreme temporal variance** in event durations (Figure 5), and a high degree of **multimodal co-occurrence** (Figure 6), which necessitates our graph-based modeling approach.

HR-GAT Layers L		Attention Heads	
Value	mAP@0.5	Value	mAP@0.5
1	29.8	1	29.5
2	30.0	2	29.9
3	**30.1**	**4**	**30.1**
4	29.9	8	29.6

Table 8: Sensitivity analysis for HR-GAT architecture.

D.2 Evaluation Protocol and Metrics

Our evaluation protocol uses standard metrics from the temporal action localization literature, ensuring our results are comparable with the broader field.

Temporal Intersection over Union (tIoU). The fundamental measure of localization accuracy is tIoU. For a predicted segment P and a ground-truth segment G, tIoU is defined as:

$$\text{tIoU}(P, G) = \frac{\text{length}(P \cap G)}{\text{length}(P \cup G)} \tag{16}$$

A detection is considered a true positive (TP) if it has the same class label as a ground-truth instance and their tIoU exceeds a specified threshold.

Mean Average Precision (mAP). This is the primary metric for our task. For each event class, we compute the Average Precision (AP) by calculating the area under the precision-recall curve. The mAP is the mean of these AP values across all event classes. Following standard practice (**?**), we report mAP at multiple tIoU thresholds (e.g., mAP@0.5, mAP@0.3) to provide a comprehensive view of performance.

Recall-based Metrics (R@k, AR@k). To evaluate the quality of candidate proposals, we use recall-based metrics. **Recall@k (R@k)** measures the fraction of ground-truth events recalled by the top-k ranked predictions for a given tIoU. **Average Recall (AR@k)** is the average of R@k over a range of tIoU thresholds [0.5:0.05:0.95]. We report AR@100 to assess the overall recall of our proposal pool.

D.3 Configuration, Training, and Inference Details

To ensure full reproducibility and directly address reviewer feedback, this section provides a comprehensive overview of the model configuration, training protocols, and computational costs.

Model Configuration. The final configuration of our PedGraph model, determined via grid search on the validation set, is summarized in Table 9. All feature dimensions are consistently 512 throughout the pipeline.

Hardware and Software. All experiments were conducted on a server equipped with an Intel Xeon Gold 6248R CPU, 1x NVIDIA A6000 (48GB), and 2x NVIDIA RTX 3090 (24GB), running Ubuntu 20.04, PyTorch 1.12.1, and CUDA 11.3.

Training and Inference Time. The end-to-end training of the three-stage PedGraph model on the full PEA training set takes approximately **28 hours** on a single NVIDIA A6000 GPU. Once trained,

Component	Parameter	Value
General	Optimizer	AdamW
	Base Learning Rate	1e-4
	Weight Decay	1e-5
	Batch Size	32
	Total Epochs	50
PSAC Loss (Stage I)	γ (TFD weight)	0.1
	α (softness)	0.2
	λ (decay)	0.5
	τ (temperature)	0.07 (learnable)
Context Model (Stage II)	TCN Layers	3
	TCN Kernel Size	3
	G-FILM Groups (K)	4
HR-GAT (Stage III)	GNN Layers (L)	3
	Attention Heads	4
	Adjacency Window (Δt)	8s
Post-processing	Soft-NMS σ	0.5
	Soft-NMS tIoU	0.5

Table 9: Key hyperparameters and configuration of the PedGraph model.

the inference process, while comprehensive, is designed for offline analysis. For a typical 10-minute classroom video (processed at 8 fps), the entire pipeline takes approximately **19 minutes** on the same hardware. This computational cost is broken down as follows:

- **Stage I (Feature Extraction & Candidate Retrieval):** Approximately **12 minutes**. This constitutes the primary computational bottleneck, involving processing thousands of sampled frames through the large ViT-B/16 vision encoder and generating an initial candidate pool.

- **Stage II (Context Grounding):** Approximately **1.5 minutes**. As designed, this lightweight stage operates efficiently on the reduced set of candidate features.

- **Stage III (HR-GAT Reasoning & Post-processing):** Approximately **5.5 minutes**. This reflects the significant cost of constructing the instance graph and performing multiple layers of relational reasoning before final prediction refinement.

This performance profile, while not real-time, is well within acceptable limits for detailed, offline educational research and analysis, where depth and accuracy are prioritized over low-latency processing.

E Additional Experimental Results And Analysis

This section provides supplementary analyses that complement the main experimental results. We offer a deeper quantitative and qualitative dissection of our ablation studies and provide visual evidence to make the model's behavior concrete and intuitive.

E.1 Per-Class Performance Analysis in Ablation Study

While the main text summarizes the overall impact of our components, this section provides a fine-grained, per-class analysis to reveal *where* these components have the most impact. Table 10 shows the Average Precision (AP) for a few representative event classes under different ablation settings.

Model Variant	G-Interact	V-Ques-Adv	V-Lecture
Full PedGraph	25.8	30.1	45.2
w/o STIG	15.3	26.5	43.1
w/o HR-GAT	22.1	25.4	44.8
w/o PSAC Loss	23.5	28.9	44.5

Table 10: Per-class AP for representative events under ablation. G-Interact refers to `Gesture Interact Type`, V-Ques-Adv to `Verbal Question Advanced`, and V-Lecture to `Verbal Lecture`.

This detailed view reveals critical insights:

- **STIG and Ambiguous Gestures:** Removing the STIG is most catastrophic for functionally ambiguous classes like `Gesture Interact Type`, causing a massive drop in AP. This directly proves that our knowledge graph is the key to resolving the semantic-visual gap for these hard cases.

- **HR-GAT and Relational Events:** The performance on `Verbal Question Advanced` is most affected by the removal of HR-GAT. This is intuitive, as identifying an "advanced" question often requires reasoning about its causal or temporal relationship to other events, a task for which HR-GAT is specifically designed.

- **General Performance:** High-frequency, visually distinct classes like `Verbal Lecture` are less affected by ablations, though all components still provide a benefit.

This per-class analysis provides strong evidence for not only *that* our components work, but precisely *how* they contribute to the model's understanding of different pedagogical functions.

E.2 Success and Failure Case Analysis

To provide a deeper insight into the behavior of PedGraph, we analyze two representative cases from our test set: one success case that highlights its core strength in contextual disambiguation, and one failure case that reveals its current limitations and points towards future work.

Success Case: Context-Aware Disambiguation of Gestures. This case demonstrates PedGraph's ability to resolve the semantic-visual gap.

- **Scenario:** A 12-second clip where a teacher first asks, "Can anyone explain the difference between system software and application software?" (*verbal_question*, 0-7s), and then points at a student in the front row (*gesture*, 8-11s).
- **Ground Truth:** The pointing gesture is labeled *gesture_interactive*.
- **ActionFormer++ Prediction:** Classifies the gesture as *gesture_indicator*. The model correctly identifies the pointing motion but, lacking contextual understanding, assigns it the default, more frequent label associated with pointing at objects.
- **PedGraph Prediction:** Correctly classifies the gesture as *gesture_interactive*. Our analysis reveals that the HR-GAT module, upon receiving the node representation for the *verbal_question* event, strongly activates the causal relation edge pointing to *gesture_interactive* within the STIG. This graph-guided reasoning allows the model to infer the gesture's true interactive intent, overriding the misleading visual signal.

Failure Case: Confusion Between Rare, Fine-Grained Events. This case illustrates a limitation of our model when faced with extreme data scarcity.

- **Scenario:** A short 5-second clip where a student provides a correct but hesitant answer. The teacher smiles and gives a brief thumbs-up gesture.
- **Ground Truth:** The gesture is labeled *gesture_encouragement*, a non-verbal cue meant to build confidence.
- **ActionFormer++ Prediction:** Fails to detect any event, likely due to the subtlety and brevity of the gesture.
- **PedGraph Prediction:** Misclassifies the gesture as *gesture_praise*. While PedGraph correctly identifies a positive non-verbal event, it confuses two functionally similar and visually fine-grained categories.
- **Analysis:** This error is attributed to the extreme long-tail distribution of the PEA dataset. Both *gesture_encouragement* and *gesture_praise* are rare events, with fewer than 20 instances each in the entire training set. This data scarcity prevents the model from learning a robust decision boundary between these nuanced categories. Addressing this challenge is a key direction for future research, potentially through the integration of few-shot or zero-shot learning techniques that can better generalize from limited examples.

E.3 Phase-Modulated Representation Analysis

To further investigate the interpretability of our model's learned representations, we visualize the final, Stage III event embeddings using t-SNE, colored by their corresponding L1 teaching phase (Figure 7). This analysis demonstrates that our multi-stage reasoning pipeline successfully learns a discriminative space where event representations are modulated by their high-level pedagogical context.

As shown in the figure, distinct clusters emerge that align with the core teaching phases. For instance, events occurring during the **Lecture** phase (light green) form a dense, coherent region on the right, indicating that the model has learned a specific "lecture mode" representation. Similarly, events from the **Practice** (orange) and **Demonstration** (dark blue) phases occupy their own discernible, albeit more distributed, territories in the embedding space.

Crucially, the boundaries between these clusters are not sharply defined but exhibit overlap. This reflects the authentic, fluid nature of classroom instruction, where transitions between phases are gradual. For example, some **Import** (deep blue) and **Summary** (brown) events are situated near the

core **Lecture** cluster, mirroring how introductions can blend into explanations and how reviews often reference lecture content. This visualization confirms that PedGraph does not merely assign a static embedding to each event class, but dynamically adapts its representations based on the overarching procedural context of the lesson, a key factor in its superior performance.

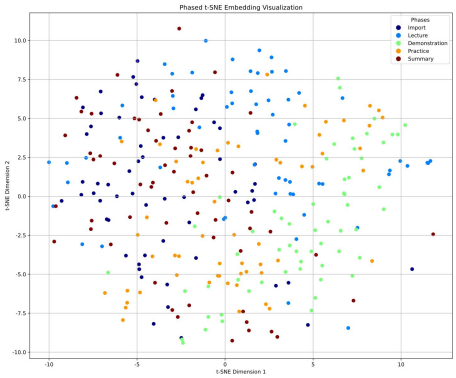

Figure 7: t-SNE visualization of final event embeddings, colored by L1 teaching phase. The emergence of distinct yet overlapping clusters for phases like 'Lecture' and 'Practice' demonstrates that our model learns a phase-modulated representation space, enhancing its contextual reasoning.

E.4 Human Evaluation Protocol

To complement our automatic metrics and assess the real-world utility and interpretability of our model's outputs, we conducted a formal human evaluation study.

Participants. We recruited ten in-service K-12 teachers via professional development networks. Participants had an average of 7.4 years of classroom experience (SD=3.2) across various subjects, ensuring our evaluators possess authentic pedagogical expertise. All participants were compensated for their time at a rate commensurate with expert consultation.

Stimuli & Task. From our test set, we randomly sampled 20 video clips, each 1-2 minutes in length and containing complex event interactions. For each clip, participants were shown the visual timeline of predictions from two models: our full **PedGraph** model and the strongest fine-tunable baseline, **ActionFormer++**. The presentation was double-blind: the source of each timeline was anonymized (labeled "System A" and "System B"), and the order of presentation was randomized across participants and trials to mitigate ordering effects.

Evaluation Criteria. For each prediction timeline, participants were asked to provide ratings on a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree) for five statements corresponding to the quality dimensions listed in Table 2:

- **Temporal Boundaries**: "The predicted start and end times of the events accurately match what I see in the video."
- **Correctness**: "The labels assigned to the events (e.g., 'verbal question') are correct."
- **Detail Sensitivity**: "The system successfully captures small but important events and details."
- **Context Awareness**: "The system's predictions make sense in the context of the overall lesson flow."
- **Overall Consistency**: "The entire sequence of predictions forms a coherent and logical story of the teaching moment."

Analysis. We collected a total of 1,000 individual ratings (20 videos $\times$ 2 models $\times$ 5 criteria $\times$ 10 raters). The average scores, reported in Section 4.4, demonstrate PedGraph's superior performance on all dimensions. The inter-rater reliability, measured by Fleiss' Kappa, was 0.73, indicating **substantial agreement** among the expert raters and confirming the reliability of our findings.

F Reproducibility Statement

We are committed to ensuring the full reproducibility of our work.

- **Code:** The complete source code for the PedGraph framework, including data preprocessing scripts, STIG construction, model training, and evaluation, will be made publicly available in an anonymized GitHub repository upon publication.
- **Dataset:** The PEA dataset, including the raw videos (subject to institutional permissions), full annotations, and the final data splits used in our experiments, will be released.
- **Models:** Pre-trained model weights for all stages of PedGraph, along with the final constructed STIG, will be provided.