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

Anonymous ACL submission

Abstract

Understanding classroom instruction requires not only localizing what happens in a video, but also inferring *why* it happens: visually similar behaviors (e.g., pointing) can serve different pedagogical functions under different discourse phases, creating profound pragmatic ambiguity. Yet existing video and vision-language models excel mainly at appearance-driven recognition and often lack an explicit representation of the relational logic that governs instructional interaction. To address this gap, we introduce **PedGraph**, a knowledge-guided framework that integrates a data-driven and expert-validated **S**tructured **T**eaching **I**nteraction **G**raph (**STIG**) to represent hierarchical, multi-relational pedagogical context; PedGraph injects STIG topology into representation learning via a structure-aware contrastive objective and performs global inference with a hierarchical relation-aware graph network to disambiguate event functions. We evaluate on **PEA**, a densely annotated instructional video benchmark (15.2 hours, 113 lessons, 32 event classes), where PedGraph outperforms strong baselines by 3.4 mAP@0.5 on function-aware event detection. Code, models, and the dataset will be released at `https://anonymous.4open.science/r/event-3A66/`.

1 Introduction

Human communication is fundamentally pragmatic: the meaning of a communicative act is inseparable from its context. This is especially true in multimodal communication, 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 often struggle to *localize and label* functionally ambiguous instructional events. In such cases, the correct interpretation depends on structured discourse context (e.g., whether the moment occurs in *Introduction* vs. *Practice*, and whether it follows a teacher question or a student answer).

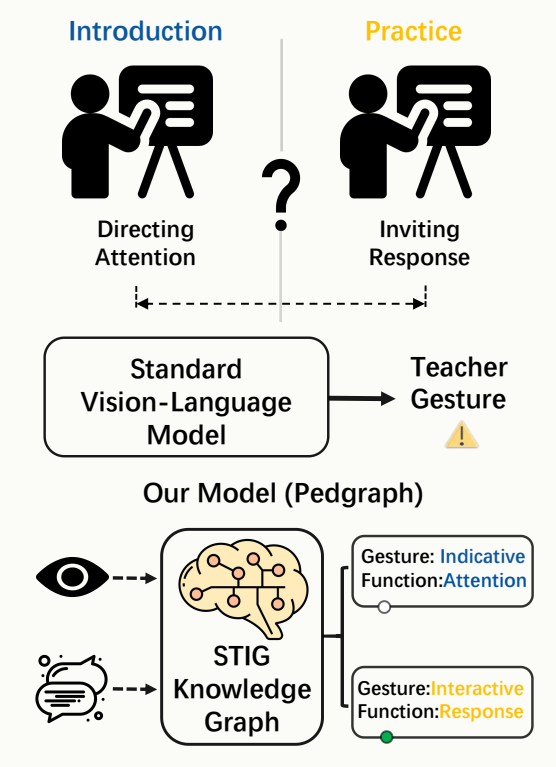

Figure 1: **The Pragmatic Ambiguity in Instructional Videos.** Standard models output a coarse label (e.g., "Teacher Gesture"). Our PedGraph resolves this ambiguity. It leverages multimodal context and a structured knowledge graph (STIG) to jointly detect the specific event (e.g., "Indicative Gesture")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

This difficulty stems from what we identify as a profound pragmatic ambiguity in instructional video understanding. 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). For example, the same vi-

sual cue (a teacher's pointing gesture) may *direct attention* during an *Introduction* segment, but *invite a response* when it follows a teacher question (or a student answer) in a *Practice* segment. To correctly interpret such events, a model must move beyond appearance and resolve the pragmatic function of multimodal events within unfolding pedagogical discourse. In this work, we focus on *visual pragmatics* and operationalize discourse context as *structured interaction context* (phases and relations); we intentionally treat spoken language as a controlled factor in the current benchmark to isolate the effect of pedagogical structure.

This context-dependent nature exposes a critical limitation of conventional VLMs. These models are architecturally biased towards treating events as atomic and context-free units, and their training objectives often encourage instance-level semantic independence (Li et al., 2021; Weng et al., 2023), which can be misaligned with the inherently relational and structured logic of pedagogy (Wang et al., 2019). Without an explicit relational scaffold, the model is forced to rely on local appearance cues, making it difficult to enforce cross-event consistency at the discourse level. This creates a performance ceiling that conventional methods struggle to breach, because they cannot simultaneously (1) resolve pragmatic ambiguity, (2) represent the relational logic of discourse, and (3) perform global reasoning to produce a coherent video-level interpretation.

To address these limitations, we introduce **PedGraph**, a knowledge-guided framework for pragmatic reasoning in instructional videos. At its core is the **S**tructured **T**eaching **I**nteraction **G**raph (**STIG**), a domain knowledge backbone constructed via reproducible data-driven mining followed by expert validation. STIG serves as an explicit relational scaffold: it encodes multi-level event categories and pedagogical relations (*contains*, *precedes*, *causes*) that capture the hierarchical and temporal/causal structure of classroom discourse. Importantly, we contribute an adaptive STIG construction pipeline rather than a rigid template, allowing the graph topology to vary across instructional ecologies. We leverage STIG in two key stages that directly address the above gaps: (i) **PSAC** injects STIG topology into representation learning to better separate functionally ambiguous events; and (ii) **HR-GAT** performs STIG-constrained global inference by disentangling relation types for discourse-consistent prediction.

Our main contributions are threefold:

- We formalize function-aware instructional event detection as a pragmatic reasoning task, and propose **STIG** together with a principled, reproducible construction-and-validation protocol.
- We propose **PSAC**, a structure-aware contrastive objective that aligns visual representations with STIG topology to reduce pragmatic ambiguity.
- We propose **HR-GAT**, a hierarchical relation-aware graph network that supports global, discourse-level inference by explicitly disentangling pedagogical relations.

Experiments on PEA (15.2 hours, 113 lessons, 32 event classes) show that PedGraph achieves the best performance among evaluated baselines, improving mAP@0.5 by 3.4 points on function-aware event detection. Beyond in-domain evaluation, we further probe *initial* generalization in two settings: cross-subject transfer and a pilot structural analysis on unseen instructional domains.

2 Related Work

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 (Zhang et al., 2022)) 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 (Chen et al., 2023) 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 (Kim et al., 2023). 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

Problem Setup. Given a video x, we aim to predict a set of instructional events $\mathcal{Y} = \{(s_n, e_n, c_n)\}_{n=1}^{N}$, where (s_n, e_n) are temporal boundaries and c_n is an event label whose interpretation is context-dependent in pedagogy; events may temporally overlap due to dense classroom interactions. Our framework produces $\mathcal{Y}$ via three stages and uses an explicit pedagogical knowledge graph as the shared scaffold across stages.

Roadmap. As illustrated in Figure 2, PedGraph closes the loop of our three challenges by: (i) encoding pedagogical relational logic in a Structured Teaching Interaction Graph (STIG; Section 3.1), (ii) injecting this structure into representation learning for high-recall candidate retrieval (Stage I; Section 3.2), and (iii) grounding candidates in local phase/intent context (Stage II; Section 3.3) and performing STIG-constrained global optimization to produce a coherent video-level interpretation (Stage III; Section 3.4).

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

We introduce the **Structured Teaching Interaction Graph (STIG)**, a domain-specific directed graph $\mathcal{G}_{STIG} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ that makes pedagogical structure explicit: nodes $\mathcal{V}$ are event types, edges $\mathcal{E}$ are typed relations, and $\mathcal{R}$ denotes relation categories. STIG serves as a reusable scaffold that shapes representation learning in Stage I and constrains relational reasoning in Stage III.

Structure. Nodes $\mathcal{V}$ include 5 macro-level lesson phases (L1; e.g., `Introduction`) and 27 fine-grained event types (L2; e.g., `verbal_question_basic`) in our taxonomy (Appendix A). Each edge $(u, v, r) \in \mathcal{E}$ is typed by one of three pedagogical relations $r \in \mathcal{R}$:

- **Structural (*contains*)**: a hierarchical link from an L1 phase to an L2 event (e.g., `Direct_Instruction` *contains* `verbal_lecture`).
- **Temporal (*precedes*)**: a sequential link capturing typical instructional flow (e.g., `Demonstration` *precedes* `Interactive_Practice`).
- **Causal (*causes*)**: a functional cause-effect link (e.g., `expression_negative` *causes* `verbal_question_advanced`).

Construction (reproducible and leakage-free). We construct STIG using *training data only*; all thresholds and hyperparameters are selected on the validation split, and the test split is never accessed in graph mining or calibration. The protocol has three phases with explicit outputs: (i) a Neural Hawkes Process mines statistically stable *precedes* candidates from event streams, (ii) our Pedagogical Granger Causality Test (P-GCT) adds *causes* edges only when the candidate event yields a measurable intent-prediction gain, and (iii) a double-blind Delphi validation filters spurious correlations and retains only pedagogically sound relations. Full details and implementation are provided in Appendix B.

3.2 Stage I: Knowledge-Guided Candidate Retrieval

Goal and output. Stage I is a *high-recall* retrieval stage: given a video, we score densely sampled sliding-window clips and output a candidate set $\mathcal{C} = \{(s_m, e_m, c_m, \hat{p}_m)\}_{m=1}^{M}$ for downstream contextual grounding and global reasoning. We therefore evaluate Stage I primarily by proposal quality (Recall@K and AR@100 under standard tIoU), rather than final mAP, since later stages are designed to trade recall for precision via contextual grounding and global reasoning.

Backbone and training objective. We adopt the efficient T2L architecture (Ahmad et al., 2025) for

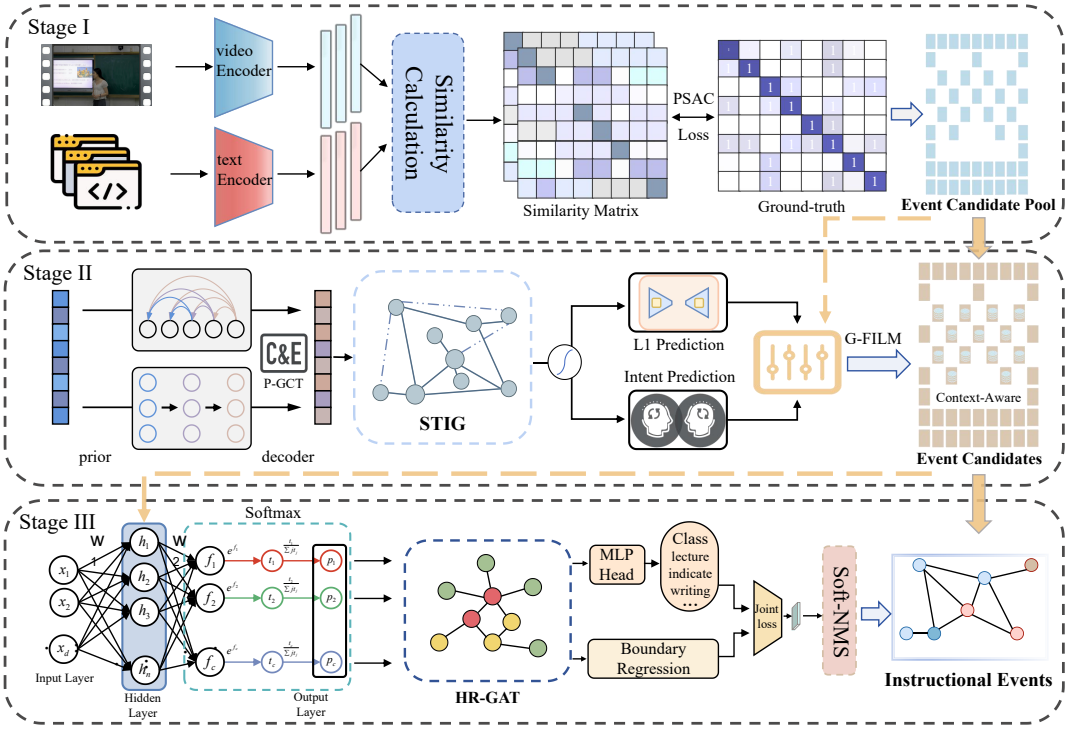

Figure 2: Overview of our multi-stage framework: The pipeline consists of three main stages. **Stage I** uses our PSAC loss, guided by the STIG, to generate a high-recall pool of event candidates. **Stage II** enriches these candidates with dynamic context (phase and intent) via G-FILM. **Stage III**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

temporal modeling, and train it with a dual objective:

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

where $\mathcal{L}_{\text{TFD}}$ is the Temporal Feature Diversity loss in T2L and γ balances the two terms.

Pedagogical Structure-Aware Contrastive (PSAC) loss. PSAC reshapes contrastive supervision using the STIG topology. Here, each "text" input is a templated label prompt for an event class (not spoken-language transcripts). Given a batch of N clip–label pairs, let v_i be the clip embedding and t_k be the label-text embedding. We compute the clip-to-text matching distribution:

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

and define a STIG-induced similarity matrix

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

where $d(\cdot,\cdot)$ is the (unweighted) shortest-path distance on the STIG. We then form a soft target distribution:

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

Finally, PSAC minimizes the KL divergence from the target to the prediction (symmetric for both directions):

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

We use the unweighted shortest-path distance on STIG as a simple and interpretable relatedness prior, and keep all hypotheses above a threshold $p_{\min}$ to form a high-recall pool for Stages II–III.

3.3 Stage II: Dynamic Contextual Grounding

Goal and I/O. Stage II grounds each retrieved candidate in its procedural and intentional context. Specifically, each candidate $j \in \mathcal{C}$ corresponds to a clip segment (s_j, e_j) with a Stage I predicted class c_j and a clip embedding $\mathbf{v}_j$. Stage II outputs a functional context vector $\mathbf{z}_j$ and a context-conditioned representation $\tilde{\mathbf{v}}_j$, which will be used as node features in the Stage III instance graph.

Inferring procedural context (L1 phase). We predict the macro lesson phase using a lightweight Temporal Convolutional Network (TCN) (Bai et al., 2018), producing $\mathbf{p}_{\text{phase}} \in \mathbb{R}^5$ over the 5 L1 phases.

Estimating functional intent (statistical cali-

brator). We define a discrete intent set $\mathcal{I} = \{I_1, \ldots, I_M\}$ (Appendix A). Given $(c_j, \mathbf{p}_{\text{phase}})$, we compute an intent posterior by mixing phase-conditioned intent statistics estimated on the training split only. Specifically, let $\boldsymbol{\pi}_{c_j,m} \in \mathbb{R}^5$ denote the phase-conditioned intent table, where $\boldsymbol{\pi}_{c_j,m}[k] = P(I_m \mid c_j, l_k)$.

$$\tilde{p}(I_m \mid c_j, \mathbf{p}_{\text{phase}}) = P(I_m) \cdot \mathbf{p}_{\text{phase}}^{\top} \boldsymbol{\pi}_{c_j,m}, \quad (6)$$

$$\mathbf{p}_{\text{intent}}[m] = \frac{\tilde{p}(I_m \mid c_j, \mathbf{p}_{\text{phase}})}{\sum_{m'=1}^{M} \tilde{p}(I_{m'} \mid c_j, \mathbf{p}_{\text{phase}})}. \quad (7)$$

We then form the functional context vector $\mathbf{z}_j = [\mathbf{p}_{\text{phase}} \| \mathbf{p}_{\text{intent}}]$. Because $\mathbf{p}_{\text{phase}}$ and $\mathbf{p}_{\text{intent}}$ capture function-level pedagogy rather than subject-specific content, this calibration signal tends to be more stable under cross-subject shifts, which we further analyze in our transfer experiments.

Feature recalibration with G-FILM. We use $\mathbf{z}_j$ to generate feature-wise modulation parameters and recalibrate $\mathbf{v}_j$ via G-FILM (Peng et al., 2022). We partition $\mathbf{v}_j$ into K groups; for each group k, an MLP $g_k(\cdot)$ maps $\mathbf{z}_j$ to $(\boldsymbol{\gamma}_{j,k}, \boldsymbol{\beta}_{j,k})$:

$$\tilde{\mathbf{v}}_{j,k} = \boldsymbol{\gamma}_{j,k} \odot \mathbf{v}_{j,k} + \boldsymbol{\beta}_{j,k}. \quad (8)$$

Stage II is lightweight by design: it preserves Stage I recall while injecting a minimal, interpretable notion of phase and intent, preparing candidates for STIG-constrained global reasoning in Stage III.

3.4 Stage III: Global Relational Reasoning

Goal and output. Stage III produces the final event set $\mathcal{Y}$ by enforcing **global pedagogical coherence** over the context-conditioned candidates from Stage II. Given candidate segments (s_j, e_j) with features $\tilde{\mathbf{v}}_j$ (optionally concatenated with $\mathbf{z}_j$), we jointly refine class scores and temporal boundaries, and this final stage is evaluated by standard detection metrics (mAP at tIoU thresholds).

STIG-informed instance graph. We construct a directed instance graph $\mathcal{G}_{\text{inst}} = (\mathcal{V}_{\text{inst}}, \mathcal{E}_{\text{inst}})$, where each node corresponds to a Stage II candidate. An edge $(k \to j)$ is added if (i) the two segments are temporally adjacent within Δt, and (ii) their predicted event classes are connected by a STIG relation type (*contains*, *precedes*, or *causes*), yielding a sparse but semantically grounded reasoning structure.

HR-GAT message passing. Standard GATs are relation-blind. HR-GAT disentangles message passing by relation type: causal edges act as strong, directed functional constraints (via a gated message), while temporal/structural edges provide softer discourse context (via relation-aware attention). This design operationalizes the STIG's relational logic and prevents dense temporal adjacency from overwhelming sparse but informative causal dependencies.

Prediction and inference. We apply a multi-task head on $\mathbf{h}_j^{(L)}$ to predict (i) class probabilities and (ii) boundary refinements, trained with cross-entropy and Smooth L1 losses, respectively. At inference time, we filter low-confidence detections and apply Soft-NMS to merge redundant proposals while preserving temporally proximal events.

Connection to the three challenges. Together, Stage I resolves *pragmatic ambiguity* at the representation level, Stage II grounds candidates in local procedural/intentional context, and Stage III performs *global reasoning* by explicitly operationalizing the STIG's *relational logic* to produce a coherent video-level interpretation.

4 Experiments

Goal. We evaluate PedGraph on PEA with standard TAL metrics, reporting both proposal quality (R@k, AR@100; Stage I) and final detection accuracy (mAP; Stages II–III).

4.1 PEA Dataset and Task Challenges

Scope and controlled variables. PEA contains 113 classroom lessons (15.2 hours) with over 10,000 densely annotated event instances under a two-level taxonomy of 32 event classes. We split the data by teacher and classroom into train/-val/test (70%/12%/18%) to prevent leakage. Following a controlled-variable design, the language of instruction is fixed to Mandarin Chinese (native speakers), allowing us to isolate pedagogical pragmatics from linguistic variation. Subject coverage includes Mathematics (65 lessons, 57.5%) and Information Technology (48 lessons, 42.5%).

Why PEA is challenging. PEA reflects real classroom complexity with three properties that directly motivate our design: (1) **temporal density and overlap**: events frequently overlap (1.4 concurrent events on average), making multi-event reasoning essential; (2) a **long-tail** distribution, where rare but function-critical events often require context from frequent head events; and (3) **co-occurrence**

Item	PEA
Language	Mandarin Chinese (controlled variable)
Subjects	Math 65 (57.5%), InfoTech 48 (42.5%)
Scale	113 lessons, 15.2 hours, >10k instances
Label space	32 classes (L1+L2 hierarchy)
Split	Teacher/classroom split: 70/12/18
Overlap	1.4 concurrent events on average
Agreement	Fleiss' Kappa 0.78 (10% double-coded)

Table 1: PEA scope and key properties.

structure (e.g., gestures contained within lectures), violating the independence assumptions of flat per-clip classifiers. Annotation reliability is ensured via 10% cross-annotation with Fleiss' Kappa of 0.78.

4.2 Experimental Setup and Fair Comparisons

Metrics. We report mAP at tIoU 0.5/0.3 for final detection quality, and Recall@k (k=5,10,50) plus AR@100 for proposal quality.

Baselines. We evaluate three baseline groups: (1) *Zero-shot generative VLMs* (VideoLLaMA-7B, VideoChatGPT-7B; marked with *), (2) *Fine-tuned recognition/TAL baselines* (TimeSformer, VideoMAE, ActionCLIP, EZ-CLIP, InternVL, ActionFormer), and (3) our variants. InternVL is fine-tuned on PEA to strengthen baseline coverage.

Implementation. All fine-tuned results are reported as mean±std over 5 runs. To preempt concerns that VLMs underperform only due to imprecise boundaries, we additionally evaluate them under very loose tIoU thresholds (Section 4.6). Hyperparameters and prompts are provided in Appendix D.3.

4.3 Main Results

Quantitative results. PedGraph achieves the best overall performance, improving both proposal quality and final mAP. Compared to a strong TAL baseline (ActionFormer), PedGraph yields higher mAP@0.5 while keeping AR@100 competitive, suggesting that gains stem from resolving context-dependent ambiguity rather than aggressive post-processing.

4.4 Generalization: Cross-subject Transfer and Structural Pilot

Cross-subject transfer. To test whether STIG-induced structure overfits subject-specific visual artifacts, we conduct leave-one-subject-out transfer: training on one subject and evaluating on the other unseen subject.

Pilot structural generalization. To probe broader subject regimes beyond the two STEM classes in PEA, we run a pilot study on newly collected History (discussion-heavy) and Physics (lab-demo) videos. We extract their event transition matrices and compare them with the PEA-induced structure, obtaining a cosine similarity of 0.81, suggesting that the pedagogical backbone is largely reusable while allowing topology to adapt to different classroom ecologies. (For this pilot, we collected approximately 3 hours of videos and extracted event sequences to build a row-normalized transition matrix over event types; due to limited pilot annotations, we report structural similarity as a lightweight proxy and leave full cross-domain detection evaluation to future work; see Appendix E.5.)

4.5 Ablations: Why Hierarchy and Graph Matter

Component and hierarchy ablations. We ablate STIG, HR-GAT, and PSAC, and include a hierarchy-removal variant (PedGraph-Flat) to isolate the necessity of the L1/L2 taxonomy.

Interpretation. Removing STIG causes the largest drop, confirming that explicit pedagogical relations are critical for this task. PedGraph-Flat performs substantially worse than the full model, indicating that hierarchical decomposition is not cosmetic but essential for resolving pragmatic ambiguity in dense, overlapping instructional interactions.

4.6 LLM Fairness: Loose tIoU Evaluation

Are LLMs only bad at boundaries? A frequent concern is that generative VLMs may capture the correct event semantics but fail to localize precise temporal boundaries. To factor out boundary sensitivity, we re-evaluate zero-shot VLM baselines under very loose thresholds (tIoU 0.2/0.1), where coarse overlap is sufficient for a true positive.

Implication. Even at tIoU 0.1, PedGraph retains a large absolute margin over the best zero-shot VLM baseline. This result indicates that the performance gap cannot be explained solely by boundary regression errors, but is also consistent with missing or incorrect pragmatic intent inference in the absence of explicit pedagogical structure.

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±0.3	31.0±0.4	43.8±0.4	31.1±0.5	21.6±0.3	32.4±0.4
VideoMAE	19.1±0.4	33.9±0.5	47.3±0.5	37.5±0.6	24.5±0.4	35.8±0.5
ActionCLIP	19.4±0.5	34.6±0.5	48.9±0.6	40.2±0.6	26.8±0.5	38.1±0.6
EZ-CLIP	20.3±0.4	35.5±0.6	50.1±0.6	41.8±0.7	27.4±0.5	39.2±0.6
InternVL	21.1±0.4	38.6±0.5	54.4±0.6	43.6±0.7	27.9±0.6	42.2±0.6
ActionFormer	21.5±0.4	39.9±0.5	54.1±0.6	45.8±0.7	28.2±0.6	41.9±0.6
Ours (ViT-B/32, 8-frame)	21.8±0.3	40.9±0.4	54.7±0.5	45.8±0.6	29.7±0.5	43.8±0.5
Ours (ViT-B/16, 8-frame)	22.1±0.3	41.4±0.4	55.2±0.5	46.9±0.6	30.4±0.5	44.5±0.5
Ours (ViT-B/16, 16-frame)	**22.6**±0.3	**42.1**±0.4	**56.0**±0.5	**47.8**±0.6	**31.6**±0.5	**45.5**±0.6

Table 2: Main comparison on PEA. We report proposal quality (R@k, AR@100; Stage I) and final detection accuracy (mAP; Stages II–III). *Zero-shot evaluation.

Train	Test	mAP@0.5	mAP@0.3	R@5	Ret.
Mixed	All	31.6	45.5	22.6	100
Math only	InfoTech	27.2	39.2	19.5	86.1
InfoTech only	Math	25.4	36.6	18.2	80.4

Table 3: Leave-one-subject-out transfer on PEA. Ret.:retention w.r.t. Mixed→All at mAP@0.5.

Model	0.5	0.3	0.2	0.1
VideoLLaMA-7B*	18.3	29.1	33.5	37.2
VideoChatGPT-7B*	19.8	30.5	35.2	38.6
PedGraph (Ours)	31.6	45.5	52.2	56.5

Table 5: mAP under loose tIoU thresholds (*zero-shot).

Variant	mAP@0.5
Full PedGraph	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 Temp./Struct. Rel.	29.2
PedGraph-Flat (remove hierarchy)	24.7

Table 4: Ablation results (single-column).

4.7 Qualitative and Human Evaluation

Qualitative analysis: resolving pragmatic ambiguity. We examine a representative ambiguity case where the teacher first asks a question and then produces a visually identical pointing gesture. Although the gesture resembles a common indicative motion (e.g., pointing at the board), its pedagogical function is interactive (inviting a student response). The strong flat TAL baseline (ActionFormer) defaults to the frequent label *gesture_indicator*, while PedGraph correctly predicts *gesture_interactive* by leveraging the preceding *verbal_question* context and the corresponding STIG causal links during HR-GAT reasoning. Additional success and failure cases are provided in Appendix E.2.

Human study: perceived coherence and usefulness. We conducted a controlled human evaluation with ten in-service K–12 teachers, who rated prediction timelines on five criteria (Temporal Boundaries, Correctness, Detail Sensitivity, Context Awareness, and Overall Consistency) using a 1–5 Likert scale in a double-blind, randomized setup. As shown in Table 2, PedGraph receives higher ratings than all baselines, with particularly strong gains on Context Awareness and Detail Sensitivity, suggesting that its structured reasoning better aligns with human pedagogical interpretation. The full protocol (sampling, blinding, and reliability analysis), along with a typical failure mode under extreme long-tail scarcity, is provided in Appendix E.4.

5 Conclusion

We first study function-aware instructional event detection and define it as resolving pragmatic ambiguity under pedagogical discourse context. We then examine why appearance-centric MLLMs and flat temporal models struggle to enforce globally coherent, interpretable predictions without an explicit relational scaffold. To mitigate, we propose PedGraph, which integrates a pedagogical knowledge graph (STIG) with structure-aware contrastive learning and a hierarchical graph network, and empirically validate its effectiveness across the evaluated baselines while leaving lightweight architectures and cross-domain transfer as future work.

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.

Ethical Considerations

Human subjects and privacy. Our work analyzes real-world classroom videos, which may contain identifiable individuals and sensitive contextual information. We emphasize that any data collection and release must follow institutional and legal requirements, and that the dataset should be distributed under appropriate access controls and usage agreements when full public release is not feasible.

Risk of misuse in high-stakes evaluation. Although our goal is to support educational research and formative feedback, automated analysis of teaching behaviors could be misapplied for summative, high-stakes evaluation (e.g., ranking teachers) without sufficient context. We therefore position PedGraph as a descriptive tool rather than an automated decision-making system, and we recommend that any deployment include human oversight, clear documentation of error modes, and safeguards against over-interpretation.

Bias and representational limitations. Model outputs may reflect biases present in the training data, such as subject coverage, teaching styles, classroom settings, or demographic factors not controlled in the benchmark. Users should be cautious when generalizing findings to new populations or instructional contexts, and future work should expand coverage and report subgroup-level analyses when metadata permits.

Transparency and accountability. To support responsible use, we provide detailed protocols for data splitting and leakage prevention, and we document the STIG construction procedure and evaluation settings so that limitations and assumptions are explicit. We also encourage reporting uncertainty and failure cases when presenting model predictions in downstream analyses.

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*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *Preprint*, arXiv:2306.15195.

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.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. *Preprint*, arXiv:2305.06590.

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.

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.

Chenlin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. *Preprint*, arXiv:2202.07925.

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 6.

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.

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:

$$\lambda_k(t) = f(\mathbf{h}_k(t)) \tag{9}$$

where the history embedding $\mathbf{h}_k(t)$ is computed by an RNN that processes the sequence of preceding events $\{(e_i, t_i) \mid t_i < t\}$. 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 $(i, j, \textit{precedes})$ is created. θ_{NHP} is set to the 80th percentile of all non-zero influence scores, determined on the validation set.

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 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 $\mathcal{G}_S$: $\textit{contains}$, $\textit{precedes}$, or $\textit{causes}$.
$d(c_i, c_k)$	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 6: Consolidated glossary of terms, event classes, and symbols.

B.2 Causal Relation Mining

To infer functional cause-and-effect links ($\textit{causes}$), 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 whether a candidate cause event C provides additional predictive information for 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} (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 ($f^{\neg C}_{\text{predict}}$).

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

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

Thresholding. In our implementation, θ_{F1} is set to 0.08 based on validation tuning, to retain only pairs with a meaningful predictive relationship.

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 process prunes spurious correlations, yielding the final, sparse, and interpretable graph whose statistics are shown in Table 7.

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 7: 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 details the implementation of all three stages in PedGraph for reproducibility. Following the main paper, we keep a unified feature dimension of 512 throughout the pipeline, matching the CLIP ViT-B/16 embedding size and the T2L feature interface used in Stage I.

C.1 Stage I: PSAC Loss and Candidate Retrieval

Stage I trains a temporal model with the objective $\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{PSAC}} + \gamma\mathcal{L}_{\text{TFD}}$, where $\mathcal{L}_{\text{TFD}}$ is the Temporal Feature Diversity loss from T2L (Ahmad et al., 2025).

PSAC construction. PSAC minimizes the KL-divergence between the predicted clip-to-text distribution and a STIG-induced soft target. We construct the STIG similarity matrix and the soft target as:

$$\mathbf{M}_{ik} = \exp(-\lambda\, d(c_i, c_k)) \tag{10}$$

$$q'_{ik} = \frac{\mathbf{1}_{[i=k]} + \alpha(1 - \mathbf{1}_{[i=k]})\mathbf{M}_{ik}}{Z_i} \tag{11}$$

$$Z_i = 1 + \alpha \sum_{k \neq i} \mathbf{M}_{ik}. \tag{12}$$

where $d(c_i, c_k)$ is the (unweighted) shortest-path distance on G_{STIG}, λ is a decay parameter, and α controls target softness.

Hyperparameters. Unless otherwise specified, we use $\gamma = 0.1$, $\alpha = 0.2$, $\lambda = 0.5$, and a learnable temperature τ initialized as 0.07. We tune Stage I clip generation parameters (window/stride and positive/negative tIoU thresholds) on the validation split.

Sensitivity analysis protocol. To avoid confounding factors, Tables 8 and 9 report validation performance under a fixed backbone and fixed downstream settings, and we select hyperparameters by end-to-end mAP@0.5 (Stages I–III) while

keeping the main paper's proposal-quality metrics (Recall@K, AR@100) for Stage I analysis.

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 8: Validation sensitivity analysis for Stage I clip generation hyperparameters, reported by end-to-end mAP@0.5 under fixed settings.

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 9: Validation sensitivity analysis for PSAC hyperparameters, reported by end-to-end mAP@0.5 under fixed settings.

C.2 Stage II: Dynamic Contextual Grounding

Phase predictor (TCN). The TCN for L1 phase prediction has 3 layers with kernel size 3 and dilation factors (1, 2, 4), outputting $\mathbf{p}_{\text{phase}} \in \mathbb{R}^5$.

Intent posterior (statistical calibrator). Let $\pi_{c,m} \in \mathbb{R}^5$ denote the phase-conditioned intent table estimated on the training split only, where $\pi_{c,m}[k] = P(I_m \mid c, l_k)$. Given a Stage I class prediction c_j and phase posterior $\mathbf{p}_{\text{phase}}$, we compute:

$$\tilde{p}(I_m \mid c_j, \mathbf{p}_{\text{phase}}) = P(I_m) \cdot \mathbf{p}_{\text{phase}}^{\top} \pi_{c_j,m} \quad (13)$$

$$p_{\text{intent}}[m] = \frac{\tilde{p}(I_m \mid c_j, \mathbf{p}_{\text{phase}})}{\sum_{m'} \tilde{p}(I_{m'} \mid c_j, \mathbf{p}_{\text{phase}})} \quad (14)$$

We then form the context vector $\mathbf{z}_j = [\mathbf{p}_{\text{phase}} \| \mathbf{p}_{\text{intent}}]$.

G-FILM. We partition the 512-dim visual feature $\mathbf{v}_j$ into $K = 4$ groups. For each group k, an MLP $g_k(\cdot)$ maps $\mathbf{z}_j$ to modulation parameters $(\gamma_{j,k}, \beta_{j,k})$, and applies:

$$\tilde{\mathbf{v}}_{j,k} = \gamma_{j,k} \odot \mathbf{v}_{j,k} + \beta_{j,k}. \quad (15)$$

C.3 Stage III: Global Relational Reasoning (HR-GAT)

We use an HR-GAT with $L = 3$ layers and 4 attention heads. The input/output node feature dimension is 512.

Relation-aware attention for temporal/structural edges. For temporal/structural neighbors, we compute attention scores with a standard GAT-style compatibility function:

$$e_{kj} = \text{LeakyReLU}\left(\mathbf{a}^{\top}\left[\mathbf{h}_{k,\text{att}}^{(l)} \| \mathbf{h}_{j,\text{att}}^{(l)} \| \mathbf{r}_{\text{ts}}\right]\right), \quad (16)$$

$$\alpha_{kj} = \text{softmax}_{k \in \mathcal{N}_{\text{ts}}}(e_{kj}), \quad (17)$$

where $\mathbf{r}_{\text{ts}}$ is a learned embedding shared by temporal/structural relations, and $\mathbf{h}_{\cdot,\text{att}}$ is a linear projection of $\mathbf{h}^{(l)}$.

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 $H^{(l+1)}$
$H_{\text{val}} \leftarrow H^{(l)}\mathbf{W}_{\text{val}}^{T}$, $H_{\text{att}} \leftarrow H^{(l)}\mathbf{W}_{\text{att}}^{T}$
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
  for all neighbor $k \in \mathcal{N}_{\text{causal}}$ **do**
    $g_{kj} \leftarrow \sigma(\mathbf{W}_{\text{gate}}[\mathbf{h}_k^{(l)} \| \mathbf{r}_{\text{causal}}]) \cdot w_{kj}^{\text{causal}}$
    $\mathbf{m}_{\text{causal}} \leftarrow \mathbf{m}_{\text{causal}} + g_{kj} \cdot \mathbf{h}_{k,\text{val}}$
  end for
  Compute e_{kj} for $k \in \mathcal{N}_{\text{ts}}$ using Eq. 17
  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**
    $\mathbf{m}_{\text{ts}} \leftarrow \mathbf{m}_{\text{ts}} + \alpha_{kj} \cdot \mathbf{h}_{k,\text{val}}$
  end for
  $\mathbf{h}_j^{(l+1)} \leftarrow \text{LayerNorm}(\mathbf{h}_j^{(l)}\mathbf{W}_{\text{self}}^{T} + \mathbf{m}_{\text{causal}} + \mathbf{m}_{\text{ts}})$
end for
return $H^{(l+1)}$

D Experimental Setup and Evaluation Metrics

D.1 The PEA Dataset

Scope and controlled variables. PEA contains 113 real-world classroom lessons (15.2 hours) under a two-level taxonomy of 32 classes, spanning

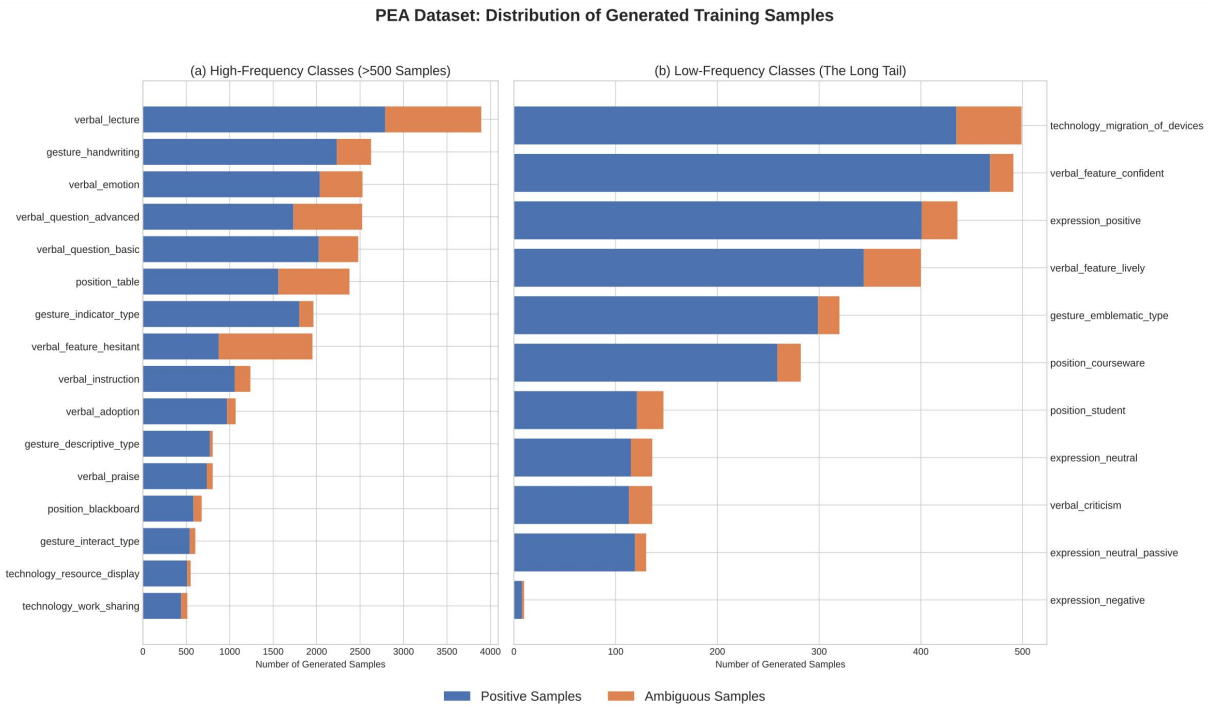

Figure 3: Distribution of training instances in PEA. (a) High-frequency classes. (b) Long-tail low-frequency classes.

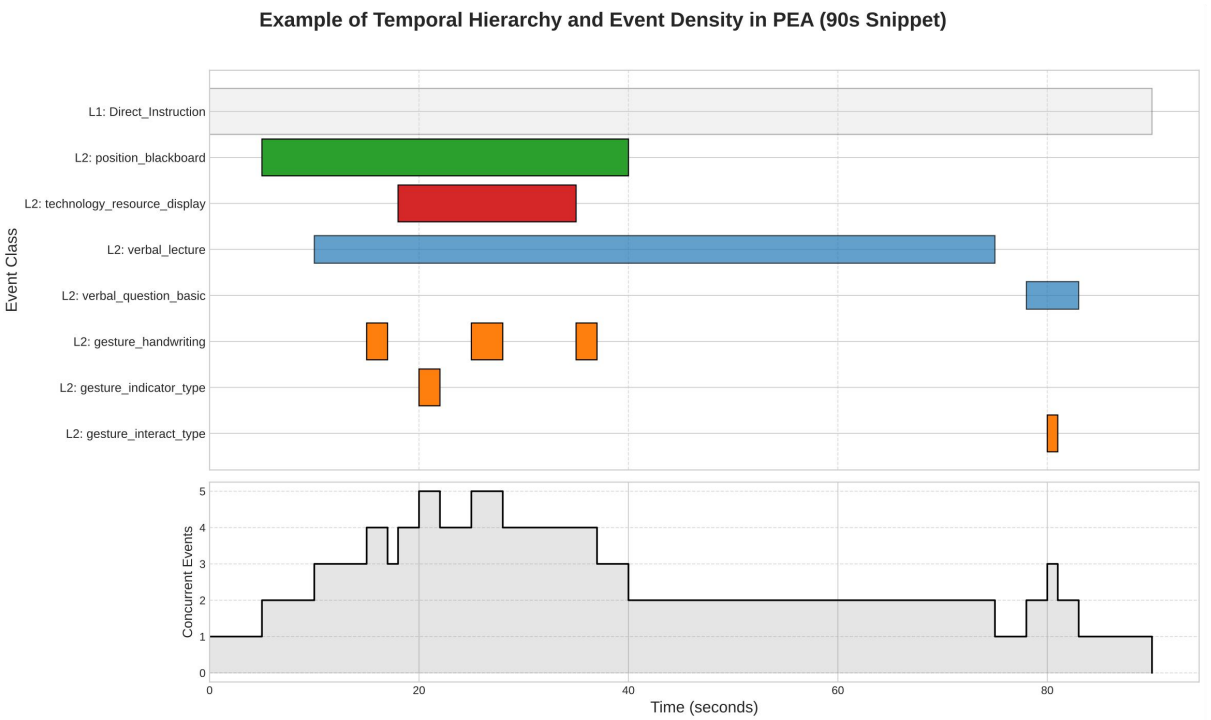

Figure 4: Temporal distribution of L2 event durations in PEA.

Mathematics (65 lessons) and Information Technology (48 lessons). Following the controlled-variable design in the main paper, the language of instruction is fixed to Mandarin Chinese.

Annotation protocol and quality assurance. We use a structured annotation manual and annotator training procedure to ensure label consistency. To measure inter-annotator agreement (IAA), we double-annotate a 10% subset of the data before adjudication and compute Fleiss' Kappa; the score is 0.78, indicating substantial agreement. Disagreements in the double-annotated subset are resolved

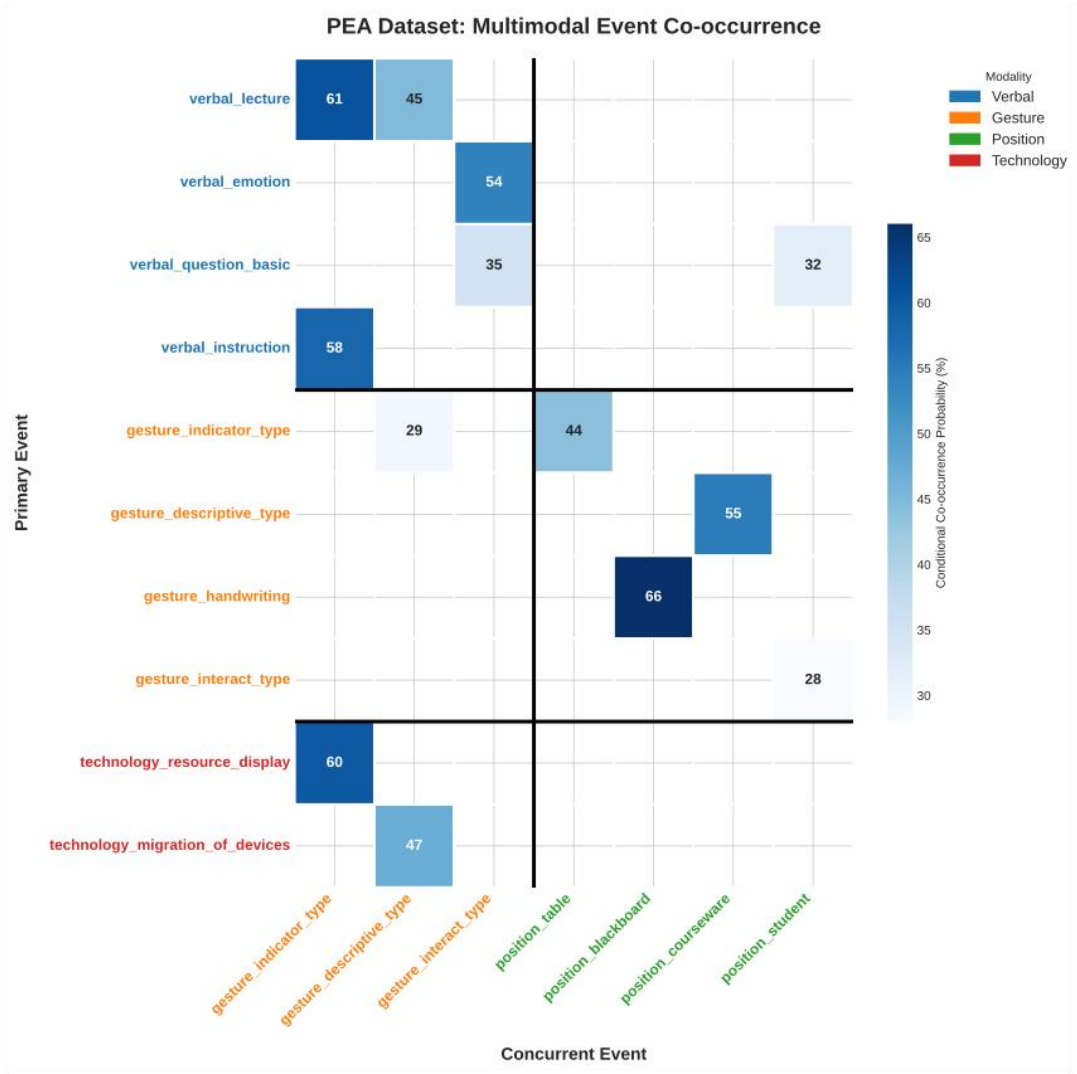

Figure 5: Heatmap of event co-occurrence in PEA.

by a senior adjudicator, and we perform periodic spot-checks during the annotation process to maintain quality.

Data splits and leakage prevention. We split the dataset by teacher and classroom into train/val/test = 70%/12%/18% to avoid leakage across splits. For reference, this corresponds to approximately 80/13/20 lessons (rounded to integers).

Key statistical challenges. PEA exhibits a long-tail class distribution (Figure 3), large temporal variance of event durations (Figure 4), and strong event co-occurrence dependencies (Figure 5), which motivate our multi-stage, graph-based modeling.

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 10: Validation sensitivity analysis for HR-GAT architecture (end-to-end mAP@0.5 under fixed settings).

D.2 Evaluation Protocol and Metrics

Temporal Intersection over Union (tIoU). For a predicted segment P and a ground-truth segment G, we define:

$$\text{tIoU}(P, G) = \frac{\text{length}(P \cap G)}{\text{length}(P \cup G)}. \quad (18)$$

A prediction is a true positive if it has the correct class label and its tIoU with a matched ground-truth instance exceeds the specified threshold.

Mean Average Precision (mAP). We report mAP at tIoU thresholds 0.5 and 0.3 as the primary detection metrics, consistent with the main paper.

Proposal-quality metrics (R@k, AR@100). To evaluate candidate proposals from Stage I, we report Recall@k (k=5,10,50) and AR@100 under standard tIoU thresholds, consistent with the main paper.

D.3 Configuration, Training, and Inference Details

Model configuration. Table 11 summarizes the hyperparameters used in our final model configuration. All feature dimensions are set to 512 throughout the pipeline.

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 11: Key hyperparameters and configuration of the PedGraph model.

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. End-to-end training on the PEA training set takes approximately 28 hours on a single NVIDIA A6000 GPU. For a typical 10-minute classroom video (8 fps), inference takes approximately 19 minutes on the same hardware, broken down into Stage I (12 min), Stage II (1.5 min), and Stage III (5.5 min).

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 12 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 12: Per-class AP for representative events under ablation. G-Interact refers to `gesture_interactive`, 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_interactive`, causing a large drop in AP. This supports that the knowledge backbone is crucial for resolving pragmatic ambiguity when visual cues alone are insufficient.
- **HR-GAT and relational events:** The performance on `verbal_question_advanced` is most affected by removing HR-GAT. This aligns with the need to reason over temporal/causal dependencies to identify advanced questioning patterns.
- **Head classes:** High-frequency, visually distinctive classes like `verbal_lecture` are less affected by ablations, though all components still contribute measurable gains.

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*. We find that HR-GAT leverages the STIG connectivity from the preceding *verbal_question* to support an interactive interpretation, overriding the misleading visual prior.

Failure Case: Confusion Between Rare, Fine-Grained Events. This case illustrates a limitation of our model under 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*.
- **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 detects a positive non-verbal event, it confuses two functionally similar and visually subtle categories.
- **Analysis:** This error is consistent with the long-tail property of PEA, where rare fine-grained classes have limited supervision. A promising direction is to incorporate few-shot adaptation or stronger text supervision to improve separation among rare, semantically close labels.

E.3 Phase-Modulated Representation Analysis

To further investigate interpretability, we visualize final Stage III event embeddings using t-SNE, colored by the corresponding L1 teaching phase (Figure 6). This analysis suggests that PedGraph learns a representation space modulated by high-level procedural context.

As shown in the figure, embeddings tend to form phase-associated regions. For example, events in **Direct_Instruction** often cluster more tightly than those in **Interactive_Practice**, reflecting differences in temporal structure and interaction patterns.

Notably, cluster boundaries overlap, consistent with the fluid nature of classroom instruction. For instance, some **Introduction** and **Summary** events appear near **Direct_Instruction**, mirroring how teachers gradually transition into core explanations and frequently reference lecture content during wrap-up.

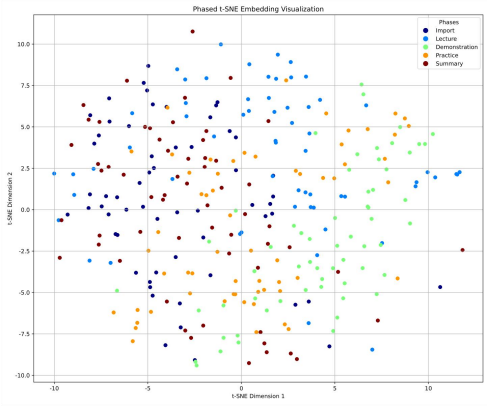

Figure 6: t-SNE visualization of final event embeddings colored by L1 phase (Introduction, Direct_Instruction, Demonstration, Interactive_Practice, Summary). Distinct yet overlapping regions indicate phase-modulated representations.

E.4 Human Evaluation Protocol

To complement automatic metrics and assess the utility and interpretability of model outputs, we conducted a 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 subjects. All participants were compensated for their time at an expert consultation rate.

Stimuli & Task. From the 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 ("System A" and "System B"), and the order was randomized across participants and trials.

Evaluation Criteria. For each timeline, participants rated five statements on a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree), corresponding to the quality dimensions in the main paper:

- **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 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 sequence of predictions forms a coherent and logical story of the teaching moment."

Analysis. We collected 1,000 ratings (20 clips $\times$ 2 systems $\times$ 5 criteria $\times$ 10 raters). The averaged results are reported in the main paper (human study section), and we compute inter-rater reliability using Fleiss' Kappa (0.73), indicating substantial agreement.

E.5 Topology Adaptation Visualization

This section provides supplementary evidence for the topology adaptation claim discussed in Sec. 4.4 of the main paper. We visualize the filtered STIG edges induced from the STEM training domain (PEA) and a pilot History domain using a consistent node set and a fixed layout, and we further report simple topology statistics computed on the same filtered graphs.

Visualization protocol. To ensure comparability across domains, we apply the following rules:

- **Consistent node set:** We use the same 32 nodes (L1 phases + L2 event types) defined in our taxonomy.
- **Top-edge filtering:** For *precedes* and *causes*, we only visualize edges whose weights fall into the top percentile within each relation type (equivalently, keep top-K per relation), to prevent dense low-confidence edges from dominating the plot.
- **Fixed layout:** We use the same 2D layout (same initialization seed and node ordering) for both subplots so that structural differences correspond to edge changes rather than arbitrary placement.
- **Node coloring:** Nodes are colored by event group (L1 phase vs. L2 verbal vs. L2 non-verbal) to make shifts in cross-group connectivity easier to interpret.

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.

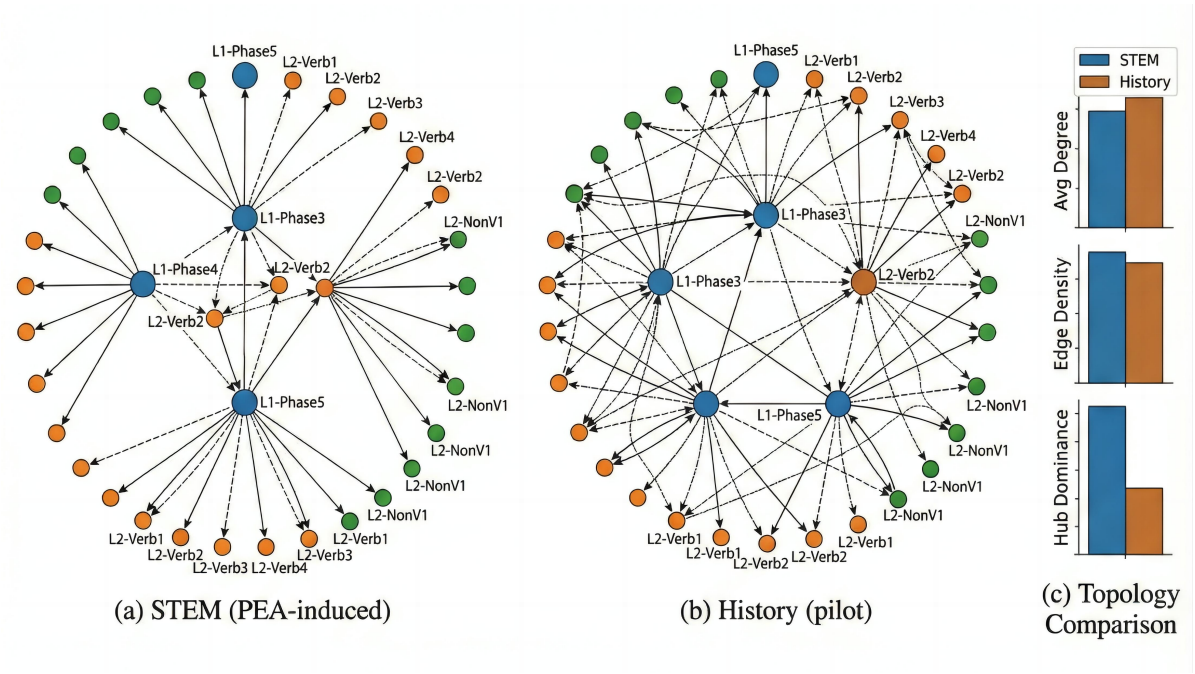

Figure 7: Topology adaptation visualization under a consistent node set and fixed layout. (a) STEM (PEA-induced) filtered topology. (b) History (pilot) filtered topology. In both graphs, we keep only top edges (same filtering rule) to highlight dominant dependencies. (c) Topology statistics computed on the same filtered graphs (Avg Degree, Edge Density, and Hub Dominance), showing a shift from a more hub-centric pattern to a more distributed "mesh-like" connectivity in History.