PROPOSAL: ĐÁNH GIÁ SỰ THAM GIA DỰA TRÊN PHÂN TÍCH HÀNH VI CƠ THỂ

Tên đề tài (Tiếng Việt): Nghiên cứu mô hình Mạng Nơ-ron Đồ thị Không gian - Thời gian (Spatio-Temporal Graph Convolutional Networks) để nhận diện hành vi và định lượng mức độ tham gia của người học trong môi trường bị che khuất một phần.

Tên đề tài (Tiếng Anh - Suggested): *Student Engagement Assessment in Partially Occluded Environments: A Skeleton-based Action Recognition Approach using Adaptive Spatio-Temporal Graph Convolutional Networks (ST-GCN).*

1. Đặt vấn đề (Problem Statement)

Các hệ thống EdTech hiện nay phụ thuộc quá nhiều vào khuôn mặt (Facial Analysis). Tuy nhiên, trong thực tế:

- Sinh viên có thể đeo khẩu trang hoặc cúi đầu viết bài (mất khuôn mặt).
- Hành vi học tập được thể hiện qua *chuỗi hành động* toàn thân: Giơ tay phát biểu, ghi chép, gục xuống bàn, hay quay ngang nói chuyện.
- **Hạn chế kỹ thuật:** Các phương pháp Pose Estimation truyền thống (như OpenPose) thường thất bại khi mất thông tin chân (do bàn che) hoặc khi các khớp (keypoints) bị chồng chéo.
- **Câu hỏi nghiên cứu (Research Question):** Làm thế nào để nhận diện chính xác hành vi học tập chỉ dựa trên thông tin **nửa thân trên (Upper-body)** và xây dựng một chỉ số Engagement định lượng từ chuỗi hành vi đó theo thời gian thực?

2. Mục tiêu nghiên cứu (Research Objectives)

1. **Tối ưu hóa Pose Estimation cho lớp học:** Phát triển mô hình phát hiện khung xương (Skeleton) nhẹ, chấp nhận đầu vào bị che khuất (Occlusion-robust), tập trung vào độ chính xác của các khớp cổ, vai, khuỷu tay và cổ tay.
2. **Mô hình hóa hành vi bằng đồ thị (Graph Modeling):** Biểu diễn cơ thể người dưới dạng đồ thị (Nodes = Khớp, Edges = Xương) và sử dụng mạng **ST-GCN** để học mối quan hệ không gian - thời gian. Ví dụ: Hành động "Giơ tay" là sự thay đổi tọa độ của khớp cổ tay so với vai theo thời gian.
3. **Xây dựng "Engagement Index":** Đề xuất công thức toán học chuyển đổi từ nhãn hành vi (Action Label) sang điểm số tham gia (Continuous Score).

3. Tổng quan tài liệu & Khoảng trống nghiên cứu (Literature Review)

3.1. Ước lượng tư thế (Pose Estimation):

- **OpenPose (2017):** Kinh điển nhưng quá chậm (Bottom-up approach).
- **YOLO-Pose / RTMPose (2023):** Đây là hướng đi mới (Top-down). RTMPose hiện là SOTA về tốc độ và độ chính xác, rất phù hợp cho real-time [1].

3.2. Nhận diện hành động (Action Recognition):

- **CNN-based (2D/3D):** Coi chuỗi pose như một "ảnh nhiệt" (Heatmap) để đưa vào CNN. Nhược điểm: Mất thông tin về cấu trúc hình học tự nhiên của cơ thể.
- **Skeleton-based (GCN):** Đột phá từ **ST-GCN** (AAAI 2018) đã thay đổi cuộc chơi. Nó xử lý trực tiếp tọa độ (x, y, c) của các khớp.
- **CTR-GCN (ICCV 2021) & InfoGCN (CVPR 2022):** Các cải tiến mới nhất giúp mạng tự học được cấu trúc đồ thị (Topology) thay vì dùng cấu trúc cố định, giúp nhận diện tốt hơn các hành vi tinh tế (như ngồi nghe vs. ngồi ngủ) [2].

3.3. Khoảng trống nghiên cứu (The Gap):

- Hầu hết các nghiên cứu ST-GCN tập trung vào dữ liệu chuẩn (NTU-RGB+D) nơi diễn viên đứng trọn vẹn trong khung hình.
- **Thiếu vắng nghiên cứu về "Seated & Occluded Pose":** Rất ít công trình tối ưu hóa GCN cho trường hợp sinh viên ngồi sau bàn (mất thông tin chân). Đây là điểm đề tài cần tập trung khai thác để tạo ra đóng góp mới (Novelty).

4. Tài liệu tham khảo minh chứng (References)

Các bài báo uy tín (2022-2024) cần trích dẫn:

1. **Về Pose Estimation hiện đại:**
   - *Paper:* "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose" (arXiv 2023). *Minh chứng cho việc chọn model nhẹ.*
   - *Paper:* "YOLOv8-Pose: Unified Framework for Object Detection and Pose Estimation" (2023).
2. **Về GCN & Behavior Analysis:**
   - *Paper:* "Student Class Behavior Recognition Based on Improved ST-GCN" (IEEE Access, 2023). *Bài báo trực tiếp liên quan đến đề tài.*
   - *Paper:* "InfoGCN: Representation Learning for Human Skeleton-based Action Recognition" (CVPR 2022).
3. **Về Engagement Assessment:**
   - *Paper:* "Automated Student Engagement Monitoring System using Computer Vision and Pose Estimation" (Education and Information Technologies, Springer, 2024).

5. Phương pháp nghiên cứu & Kiến trúc đề xuất (Methodology)

Đề tài cần thiết kế một **Two-Stage Framework** (Khung làm việc 2 giai đoạn):

Giai đoạn 1: Trích xuất khung xương mạnh mẽ (Robust Skeleton Extraction)

- **Input:** Video stream từ camera lớp học.
- **Model:** Sử dụng **RTMPose** (Real-time Multi-Person Pose Estimation).
    - *Lý do:* RTMPose sử dụng kiến trúc CSPNeXt, cân bằng cực tốt giữa tốc độ và độ chính xác.
- **Cải tiến xử lý che khuất (Occlusion Handling Strategy):**
    - Đối với các điểm bị che (như chân dưới gầm bàn), gán trọng số tin cậy (confidence score) $C = 0$.
    - Sử dụng kỹ thuật **Data Augmentation**: Trong quá trình train, chủ động "mask" (xóa) phần thân dưới của dữ liệu huấn luyện để mạng học cách nhận diện hành vi chỉ dựa vào thân trên.

Giai đoạn 2: Phân loại hành vi bằng ST-GCN (Skeleton-based Classification)

Dữ liệu đầu vào không phải là ảnh, mà là chuỗi tọa độ (N, C, T, V, M) tương ứng với (Batch, Channel, Frames, Vertices/Joints, Person).

- **Kiến trúc:** Sử dụng **Adaptive Graph Convolutional Network (AGCN)**.
    - Khác với ST-GCN gốc (cấu trúc xương cố định), AGCN có khả năng học các mối quan hệ phi tự nhiên. *Ví dụ:* Mối quan hệ giữa "Bàn tay" và "Đầu" không có xương nối trực tiếp, nhưng rất quan trọng để nhận diện hành vi "Gãi đầu" hoặc "Chống cằm".
- **Định nghĩa tập nhãn hành vi (Class Labels):**
    1. *High Engagement:* Giơ tay (Raising Hand), Ghi chép (Writing), Đứng lên (Standing up).
    2. *Medium Engagement:* Ngồi thẳng hướng lên bảng (Listening), Đọc sách (Reading).
    3. *Low Engagement:* Gục đầu (Sleeping), Quay ngang/sau (Looking around), Sử dụng điện thoại (Phone using).

Giai đoạn 3: Tính toán điểm Engagement (Scoring Function)

Đây là bước biến đổi từ AI sang EdTech.

- Mỗi hành vi A_i được gán một trọng số w_i (dựa trên thang đo tâm lý học Bloom hoặc khảo sát giáo viên).
- Điểm số tại thời điểm t :

$$S_t = \sum_{k=0}^{T} \gamma^k \cdot w(Action_{t-k})$$

- γ : Hệ số suy giảm (decay factor), giúp làm mượt điểm số, tránh việc điểm nhảy loạn xạ khi model nhận diện sai trong 1 frame.

6. Kế hoạch thực nghiệm (Implementation Plan)

6.1. Dữ liệu (Datasets)

Dữ liệu cho bài toán này hiếm hơn Face Recognition. Đề tài cần chiến lược:

1. **NTU RGB+D 120:** Bộ dữ liệu lớn nhất thế giới về hành động. Dùng để **Pre-train** mô hình GCN. Đề tài cần lọc ra các class liên quan đến lớp học (đọc, viết, ngồi, đứng…).
2. **Mimetics (Behavioral Dataset):** Tìm kiếm các dataset chuyên về hành vi lớp học (như dataset của bài báo "Student Class Behavior…" đã trích dẫn).
3. **Tự thu thập & Giả lập (Simulation - Điểm cộng):**
   - Dùng phần mềm 3D (Unity/Blender) để tạo ra các nhân vật ngồi bàn học, sau đó trích xuất khung xương. Cách này giúp tạo ra hàng ngàn mẫu dữ liệu bị che khuất (occluded samples) mà không tốn công gán nhãn thủ công. Đây là hướng đi rất phù hợp.

6.2. Metrics

- **Pose Accuracy:** OKS (Object Keypoint Similarity).
- **Action Recognition:** Top-1 Accuracy & Confusion Matrix.
- **Engagement Correlation:** Tính hệ số tương quan Pearson (r) giữa điểm số máy chấm (S_t) và điểm số do giáo viên chấm thủ công qua video.

7. Tài nguyên & Công cụ (Resources)

Đây là các thư viện "gối đầu giường" cho mảng Action Recognition:

A. GitHub Repositories (Frameworks):

1. **MMPose & MMAction2 (OpenMMLab):** github.com/open-mmlab/mmpose
   - Đây là kho vũ khí tối thượng. Nó chứa sẵn RTMPose, ST-GCN, CTR-GCN. Code được viết theo dạng module, cực kỳ chuẩn mực để phát triển luận văn.
   - *Lời khuyên:* Yêu cầu sinh viên học cách config file .py trong thư viện này thay vì code lại từ đầu.
2. **Awesome-Skeleton-Action-Recognition:** github.com/cagbal/Skeleton-Based-Action-Recognition-Papers-and-Notes
   - Danh sách tổng hợp tất cả các bài báo và code mới nhất.

B. Sách tham khảo:

- *Graph Neural Networks: Foundations, Frontiers, and Applications* - Lingfei Wu et al. (Để hiểu sâu về lý thuyết đồ thị).
- *Deep Learning for Human Activity Recognition* - Springer (Các chương về Skeleton-based methods).

Lời khuyên:

1. **Đừng quá tham lam về số lượng khớp (Keypoints):** Các mô hình chuẩn dùng 17 hoặc 25 khớp (COCO format). Với bài toán này, các khớp ở bàn chân (Ankle) là vô nghĩa và gây nhiễu.
    - *Tip:* Hãy định nghĩa lại cấu trúc đồ thị (Custom Graph Strategy), loại bỏ các node chân, chỉ tập trung vào thân trên. Điều này giúp model nhẹ hơn và chính xác hơn.
2. **Xử lý "Temporal Action Localization":** Sinh viên A có thể "Giơ tay" trong 3 giây, sau đó "Ghi chép" trong 10 giây.
    - Mô hình không nên chỉ phân loại cho 1 clip đã cắt sẵn. Nó cần chạy trên cửa sổ trượt (Sliding Window) để phát hiện *khi nào hành động bắt đầu và kết thúc*.