ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN THỊ CẨM VÂN

CÁC MÔ HÌNH HỌC SÂU KẾT HỢP ĐA PHƯƠNG THỨC TIÊN TIẾN NHẬN DIỆNCẢM XÚC TRONG HỘI THOẠI VỚI THÔNG TIN KHÔNG ĐẦY ĐỦ VÀ MẤT CÂN BẰNG

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Hà Nội, 2026

Luận án được thực hiện tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Hướng dẫn chính:	PGS.TS Hà Quang Thụy	Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội
Đồng hướng dẫn:	TS. Lê Đức Trọng	Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội
Reviewer 1:	PGS. TS Bùi Thu Lâm	Học viện Kỹ thuật Mật mã
Reviewer 2:	PGS.TS Lê Hồng Phương	Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội
Reviewer 3:	PGS.TS Nguyễn Đức Dũng	Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Đại học Quốc gia tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội (VNU-UET) vào …...giờ…..., thứ…, ngày…tháng..… năm 2026.

Luận án có thể được tìm thấy tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin – Thư viện, Đại học Quốc gia Hà Nội.

Tóm tắt

Một thách thức trung tâm trong bài toán nhận diện cảm xúc đa phương thức trong hội thoại (MERC) là thiết kế các mô hình kết hợp (fusion) có khả năng nắm bắt được các tương tác phức tạp giữa các phương thức văn bản, âm thanh và thị giác, đồng thời vẫn tôn trọng được động lực hội thoại và ngữ cảnh đặc trưng theo từng người nói. Các tiếp cận truyền thống thường dựa trên các chiến lược kết hợp sớm hoặc muộn, coi các phương thức như nhau và bỏ qua cấu trúc thời gian, quan hệ giữa các người nói cũng như các phụ thuộc ngữ cảnh bậc cao trong cuộc thoại. Trong các điều kiện thực tế, dữ liệu hội thoại đa phương thức còn bị suy giảm bởi hiện tượng thiếu phương thức và sự đóng góp mất cân bằng giữa các phương thức, làm suy yếu đáng kể độ vững và khả năng khái quát của các hệ thống MERC hiện có.

Luận án này giải quyết các thách thức trên bằng cách phát triển các mô hình học sâu kết hợp đa phương thức tiên tiến, được thiết kế chuyên biệt cho bài toán nhận diện cảm xúc trong hội thoại trong bối cảnh dữ liệu thiếu và mất cân bằng phương thức. Phần thứ nhất tập trung vào mô hình hóa hội thoại đa phương thức có cấu trúc trong điều kiện đầy đủ phương thức, với mục tiêu học được các biểu diễn giàu thông tin và nhận thức ngữ cảnh cho MERC. Chúng tôi đề xuất CORECT, một khung mô hình đồ thị quan hệ–thời gian tích hợp Relational Temporal Graph Convolutional Network (RT-GCN) với mô-đun Pairwise Cross-modal Feature Interaction (P-CM) để đồng thời mô hình hóa phụ thuộc thời gian ở mức phát ngôn, các tương tác đa phương thức và quan hệ hội thoại có xét đến người nói. Tiếp đó, chúng tôi giới thiệu MultiDAG+CL, kết hợp suy luận ngữ cảnh dựa trên đồ thị có hướng không chu trình (DAG) với chiến lược curriculum learning nhằm xử lý dần dần các chuyển đổi cảm xúc và mức độ khó của mẫu, từ đó cải thiện khả năng nhận diện cảm xúc trong các hội thoại đa người nói phức tạp.

Phần thứ hai của luận án tập trung vào bài toán kết hợp đa phương thức vừa vững vừa cân bằng trong điều kiện dữ liệu chất lượng thấp, nơi không thể giả định tính đầy đủ và cân bằng của các phương thức. Để đối phó với bài toán thiếu phương thức, chúng tôi đề xuất Mi-CGA, một khung mô hình đồ thị hai giai đoạn: trước hết xây dựng các

biểu diễn đa phương thức không đầy đủ, sau đó sử dụng Cross-modal Graph Attention Network (CGA-Net) với các cơ chế ước lượng đặc trưng phương thức, graph attention và cross-modal attention để tái dựng các tín hiệu bị thiếu và truyền tải thông tin bổ sung giữa các phương thức. Để xử lý hiện tượng mất cân bằng phương thức, chúng tôi phát triển hai chiến lược bổ trợ cho MERC. Ada2I đưa vào các cơ chế Adaptive Feature Weighting (AFW) và Adaptive Modality Weighting (AMW) để tái cân bằng động đóng góp ở cả mức đặc trưng và mức phương thức trong quá trình kết hợp, trong khi Self-Paced Curriculum Learning (SPCL) đóng vai trò một sơ đồ huấn luyện "plug-and-play", dần dần sắp lịch các mẫu theo thước đo độ khó hai tầng, giúp ổn định việc học đa phương thức trong bối cảnh chất lượng phương thức không đồng nhất.

Các thực nghiệm quy mô lớn trên những bộ dữ liệu MERC được sử dụng rộng rãi, bao gồm IEMOCAP, CMU-MOSI và CMU-MOSEI, cùng với các đánh giá bổ sung trong các thiết lập dữ liệu thiếu và mất cân bằng phương thức, cho thấy các mô hình được đề xuất liên tục vượt trội so với các đường cơ sở mạnh về chất lượng kết hợp, khả năng mô hình hóa ngữ cảnh, độ vững trước thiếu phương thức và động lực học cân bằng trong quá trình học. Nhìn chung, luận án đóng góp một góc nhìn thống nhất về kết hợp đa phương thức cho MERC: (1) kết hợp hội thoại có cấu trúc với mô hình hóa quan hệ và thời gian (CORECT, MultiDAG+CL); (2) kết hợp vẫn duy trì độ vững khi thiếu phương thức (Mi-CGA); và (3) các chiến lược huấn luyện nhận thức kết hợp giúp giảm thiểu hiện tượng chi phối phương thức (Ada2I, SPCL), từ đó cho phép nhận diện cảm xúc đáng tin cậy hơn trong các hội thoại đa phương thức thực tế.

Mở đầu

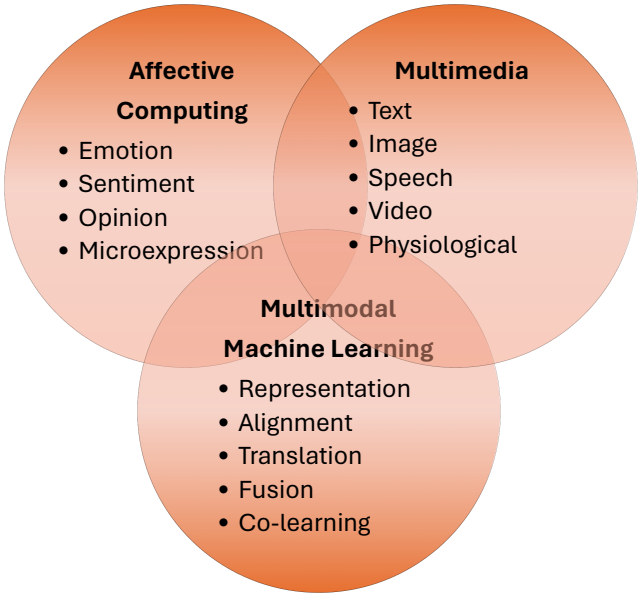
Động lực

Sự phát triển nhanh chóng của các nền tảng giao tiếp trực tuyến đã làm thay đổi cách con người tương tác, chia sẻ thông tin và bộc lộ cảm xúc, khiến hội thoại số trở thành môi trường trung tâm của trải nghiệm cảm xúc hằng ngày. Trong bối cảnh này, cảm xúc của con người được truyền tải thông qua sự phối hợp giữa ngôn ngữ, giọng nói và biểu cảm khuôn mặt, từ đó thúc đẩy bài toán Nhận diện cảm xúc đa phương thức trong hội thoại (Multimodal Emotion Recognition in Conversation – MERC) trở thành một nhiệm vụ quan trọng trong lĩnh vực tính toán cảm xúc. MERC hướng tới việc tự động suy luận cảm xúc của từng phát ngôn trong một đoạn hội thoại bằng cách mô hình hóa đồng thời các tín hiệu đa phương thức cùng với ngữ cảnh hội thoại và tương tác giữa các người nói.

Kết hợp đa phương thức hiệu quả là trọng tâm của MERC, bởi các tín hiệu văn bản, âm thanh và hình ảnh vừa khác biệt về cấu trúc vừa mang tính bổ sung trong cách chúng mã hóa các dấu hiệu cảm xúc. Các mô hình học sâu hiện đại đã giúp cải thiện việc kết hợp cho nhận diện cảm xúc trong hội thoại, nhưng nhiều mô hình vẫn được thiết kế dưới các giả định lý tưởng rằng mọi phương thức đều được quan sát đầy đủ, có độ tin cậy như nhau và cấu trúc hội thoại chỉ được mô hình hóa một phần. Trong các kịch bản thực tế, dữ liệu hội thoại đa phương thức thường có chất lượng thấp: một số phương thức có thể nhiễu, bị che khuất hoặc lệch pha theo thời gian, trong khi các phương thức khác có thể bị thiếu một phần, thiếu hoàn toàn hoặc yếu hơn hẳn so với phần còn lại. Những khiếm khuyết này khiến các mô hình kết hợp dễ bị phụ thuộc quá mức vào phương thức chiếm ưu thế, học theo các tương quan nhiễu và suy giảm mạnh hiệu năng khi tính sẵn có hoặc chất lượng của phương thức thay đổi.

Các quan sát trên dẫn tới định hướng trọng tâm của luận án về kết hợp đa phương thức sâu tiên tiến cho MERC trong bối cảnh phương thức thiếu và mất cân bằng. Một mặt, MERC đòi hỏi cơ chế kết hợp hội thoại có cấu trúc, có khả năng tích hợp các phương

thức khác nhau đồng thời mô hình hóa tường minh các phụ thuộc theo thời gian và quan hệ giữa các người nói trong hội thoại. Mặt khác, các hệ thống MERC thực tế phải duy trì được độ vững khi phương thức bị thiếu và ổn định khi mức đóng góp giữa các phương thức bị mất cân bằng, tránh hiện tượng một kênh chi phối và vẫn khai thác được các tín hiệu hữu ích từ những phương thức yếu hơn.



Hình 1: Tổng quan về khung phân tích cảm xúc đa phương thức

Phạm vi nghiên cứu, Mục tiêu và Câu hỏi nghiên cứu

Các phương pháp hiện có trong bài toán Nhận diện cảm xúc đa phương thức trong hội thoại (MERC) thường gặp khó khăn trong việc (1) căn chỉnh và kết hợp hiệu quả các đặc trưng dị thể từ văn bản, âm thanh và hình ảnh, (2) nắm bắt bối cảnh hội thoại theo thời gian và phụ thuộc vào người nói, và (3) duy trì độ vững trong các điều kiện dữ liệu đa phương thức chất lượng thấp như thiếu phương thức hoặc mất cân bằng giữa các phương thức. Để giải quyết các khoảng trống này, luận án tập trung vào hướng kết hợp đa phương thức sâu tiên tiến cho MERC trong bối cảnh dữ liệu thực tế có nhiều sai lệch và khiếm khuyết.

Phạm vi của luận án được xác định theo hai khía cạnh bổ trợ: phạm vi nhiệm vụ và phạm vi phương pháp luận. Về **phạm vi nhiệm vụ**, luận án tập trung vào MERC, trong đó mục tiêu là suy luận cảm xúc ở mức phát ngôn trong các đoạn hội thoại đa người nói bằng cách mô hình hóa đồng thời các tín hiệu đa phương thức và cấu trúc hội thoại. Về **phạm vi phương pháp luận**, luận án phát triển các mô hình học sâu cho kết hợp đa

phương thức nhằm (i) học được các biểu diễn đa phương thức giàu thông tin, phù hợp với dữ liệu hội thoại, và (ii) đảm bảo độ vững và tính cân bằng khi các phương thức bị thiếu hoặc có mức độ thông tin không đồng đều.

Trong các phạm vi đó, luận án được tổ chức xoay quanh hai mục tiêu nghiên cứu (O) cốt lõi và các câu hỏi nghiên cứu (RQ) tương ứng:

- **O1**: Kết hợp đa phương thức và mô hình hóa ngữ cảnh cho MERC. Mục tiêu này hướng tới việc thiết kế các kiến trúc kết hợp đa phương thức được điều chỉnh chuyên biệt cho bài toán nhận diện cảm xúc trong hội thoại, tập trung vào cách tích hợp các tín hiệu đặc trưng của từng phương thức trong khi vẫn mô hình hóa động lực hội thoại.
  - **RQ1**: Làm thế nào để học và kết hợp hiệu quả các biểu diễn đa phương thức ở mức phát ngôn, đồng thời nắm bắt được cả đặc trưng nội tại của từng phương thức và các tương tác bổ trợ giữa các phương thức cho bài toán nhận diện cảm xúc trong hội thoại?
  - **RQ2**: Làm thế nào để mô hình hóa có cấu trúc và tiến triển các phụ thuộc theo thời gian và tương tác giữa người nói trong hội thoại, nhằm hỗ trợ suy luận ngữ cảnh hiệu quả cho MERC?

  Các câu hỏi nghiên cứu này được giải quyết trong **Chương 2**.

- **O2**: Kết hợp đa phương thức vững và cân bằng cho MERC. Mục tiêu này tập trung nâng cao độ vững và độ tin cậy của các mô hình MERC khi dữ liệu đa phương thức không hoàn hảo, đặc biệt trong các điều kiện thiếu phương thức và mất cân bằng phương thức.
  - **RQ3**: Trong các kịch bản MERC có phương thức bị thiếu, làm thế nào để bù đắp hiệu quả các đặc trưng đa phương thức bị thiếu và truyền tải thông tin bổ sung giữa các phương thức để duy trì các biểu diễn đa phương thức vững chắc? *Câu hỏi nghiên cứu này được giải quyết trong* **Chương 3**.
  - **RQ4**: Làm thế nào để thiết kế các chiến lược huấn luyện và tối ưu hóa nhằm giảm thiểu mất cân bằng ở các mức đặc trưng, phương thức, cảm xúc và hội thoại, từ đó cải thiện tính ổn định và khả năng khái quát của mô hình MERC? *Câu hỏi nghiên cứu này được giải quyết trong* **Chương 4**.

Ý nghĩa và Đóng góp

Các đóng góp chính của luận án xuất phát từ một hướng nghiên cứu thống nhất về kết hợp đa phương thức sâu tiên tiến cho MERC, với nhấn mạnh vào mô hình hóa hội thoại có cấu trúc và độ vững trong bối cảnh phương thức thiếu và mất cân bằng. Cụ thể:

1. **Kết hợp hội thoại đa phương thức có cấu trúc**: Chúng tôi đề xuất **CORECT** [**VanNTC 1**] và **MultiDAG+CL** [**VanNTC 2**] như các mô hình MERC hướng kết hợp, đồng thời xét đến cả đặc trưng đa phương thức và cấu trúc hội thoại. CORECT đưa ra mô hình đồ thị quan hệ–thời gian kết hợp với tương tác đa phương thức bổ trợ để nắm bắt các phụ thuộc theo thời gian ở mức phát ngôn và các quan hệ đa phương thức ở mức hội thoại, trong khi MultiDAG+CL tận dụng suy luận ngữ cảnh dựa trên đồ thị có hướng không chu trình (DAG) kết hợp với curriculum learning để xử lý tốt hơn các chuyển dịch cảm xúc và mức độ khó đa dạng của mẫu trong hội thoại đa người nói.

2. **Kết hợp vững dưới điều kiện thiếu phương thức**: Chúng tôi phát triển **Mi-CGA** [**VanNTC 3**], một khung mô hình đồ thị được thiết kế cho MERC trong bối cảnh thiếu phương thức. Bằng cách kết hợp giai đoạn Incomplete Multimodal Representation (IMR) với Cross-modal Graph Attention Network (CGA-Net), Mi-CGA tái dựng các đặc trưng phương thức bị thiếu và truyền tải thông tin bổ sung giữa các phương thức, cho phép kết hợp ổn định ngay cả khi một hoặc nhiều phương thức chỉ có mặt một phần hoặc vắng mặt hoàn toàn.

3. **Kết hợp cân bằng khi phương thức bị chi phối**: Chúng tôi xử lý bài toán mất cân bằng phương thức trong MERC thông qua hai chiến lược học có nhận thức về kết hợp. **Ada2I** [**VanNTC 4**] đưa vào các cơ chế Adaptive Feature Weighting (AFW) và Adaptive Modality Weighting (AMW) được dẫn dắt bởi một tỉ lệ sai khác (discrepancy ratio) để tái cân bằng động đóng góp ở cả mức đặc trưng và mức phương thức trong quá trình kết hợp, trong khi **Self-Paced Curriculum Learning (SPCL)** [**VanNTC 5**] đóng vai trò một mô-đun plug-and-play gọn nhẹ, dần dần sắp lịch các mẫu huấn luyện dựa trên thước đo độ khó hai tầng, giúp ổn định quá trình kết hợp đa phương thức trên nhiều kiến trúc nền khác nhau.

Chương 1

Học sâu nâng cao cho nhận diện cảm xúc đa phương thức

Cảm xúc và nhận diện cảm xúc đa phương thức. Cảm xúc là các trạng thái cảm xúc rời rạc, được kích hoạt bởi những sự kiện cụ thể và thể hiện qua những thay đổi phối hợp trong trải nghiệm, hành vi và sinh lý. Trong bối cảnh tính toán, nhận diện cảm xúc nhằm gán các trạng thái này cho những đoạn dữ liệu dựa trên các tín hiệu quan sát được. Luận án này tập trung vào bài toán Nhận diện cảm xúc đa phương thức trong hội thoại, trong đó mục tiêu là dự đoán nhãn cảm xúc cho từng phát ngôn trong một hội thoại đa người nói dựa trên đầu vào văn bản, âm thanh và hình ảnh. Một cuộc hội thoại được mô hình hoá như một dãy $C = \{u_1, \ldots, u_N\}$ các phát ngôn, mỗi phát ngôn gắn với người nói và các đặc trưng theo từng phương thức, và MERC tìm một hàm ánh xạ các đầu vào hội thoại đa phương thức này sang nhãn cảm xúc ở mức phát ngôn, đồng thời khai thác cả nội dung cục bộ lẫn ngữ cảnh hội thoại.

Học máy đa phương thức. Học máy đa phương thức nghiên cứu cách biểu diễn, căn chỉnh và kết hợp các nguồn dữ liệu dị thể như ngôn ngữ, âm thanh và hình ảnh trong một khung học thống nhất. Những thách thức trung tâm bao gồm việc học các biểu diễn chung hoặc phối hợp, căn chỉnh các tín hiệu không đồng bộ theo thời gian và hỗ trợ suy luận, dịch chuyển thông tin giữa các phương thức. Đối với tính toán cảm xúc, đặc biệt là MERC, các thách thức này càng nổi bật do ý nghĩa cảm xúc phụ thuộc mạnh vào ngữ cảnh, danh tính người nói và lịch sử tương tác. Do đó, các phương pháp đa phương thức hiệu quả cần đồng thời xử lý được cấu trúc riêng của từng phương thức và các động lực hội thoại bậc cao.

Kết hợp dữ liệu đa phương thức. Kết hợp dữ liệu đa phương thức quan tâm đến cách tích hợp thông tin từ các phương thức khác nhau để phục vụ các tác vụ phía sau. Các chiến lược kinh điển trải dài từ kết hợp sớm, nơi các đặc trưng thô hoặc mức thấp được nối lại, đến kết hợp trung gian, nơi các không gian tiềm ẩn chung và tương tác đa

phương thức được học, và kết hợp muộn, nơi các dự đoán riêng theo từng phương thức được kết hợp. Trên thực tế, cơ chế kết hợp cần cân bằng giữa việc giữ lại các đặc tính phân biệt của từng phương thức và khai thác thế mạnh bổ sung giữa chúng. Trong bối cảnh hội thoại, điều này càng phức tạp khi dữ liệu có chất lượng thấp: một số phương thức có thể nhiễu, thiếu một phần hoặc yếu hơn hẳn so với các phương thức khác. Do vậy, các cơ chế kết hợp hiện đại cho MERC ngày càng có xu hướng tích hợp mô hình hoá thời gian, tương tác chéo dựa trên attention và độ vững trước các điều kiện phương thức không đầy đủ hoặc mất cân bằng.

Nhận diện cảm xúc đa phương thức trong hội thoại. MERC đặc tả bài toán dự đoán cảm xúc ở mức phát ngôn đồng thời xét đến cấu trúc hội thoại. Một cuộc hội thoại thường được mô hình hoá như một dãy $C = \{u_1, \ldots, u_N\}$, trong đó mỗi phát ngôn u_i gắn với một người nói p_i, một tập đặc trưng theo phương thức $\{x_i^m\}_{m \in M}$ (ví dụ văn bản, âm thanh, hình ảnh) và một nhãn cảm xúc $y_i \in Y$. Bài toán MERC là học một hàm

$$f : (C, \{x_i^m\}_{i=1..N,\, m \in M}) \rightarrow \{y_1, y_2, \ldots, y_N\},$$

ánh xạ các đầu vào hội thoại đa phương thức sang nhãn cảm xúc ở mức phát ngôn, đồng thời khai thác cả nội dung cục bộ và ngữ cảnh ở mức hội thoại.

Việc phát biểu bài toán thường giả định rằng mỗi phát ngôn được gán nhãn cảm xúc từ một tập cố định, và mô hình phải tận dụng cả nội dung đa phương thức của phát ngôn lẫn các phụ thuộc giữa các lượt thoại lân cận và giữa các người nói. Những tiếp cận ban đầu chủ yếu dựa trên các kiến trúc hồi quy hoặc Transformer để mô hình hoá ngữ cảnh, trong khi các nghiên cứu gần đây sử dụng cấu trúc đồ thị để biểu diễn quan hệ giữa các phát ngôn và giữa các người nói, thường kết hợp với cơ chế attention để làm nổi bật các tín hiệu quan trọng. Xuyên suốt các hướng tiếp cận, chủ đề cốt lõi là kết hợp hiệu quả giữa kết hợp đa phương thức giàu biểu đạt với mô hình hoá hội thoại có cấu trúc, đồng thời duy trì hiệu năng khi các phương thức nhiễu, thiếu hoặc mất cân bằng.

Bộ dữ liệu và tiêu chí đánh giá. Luận án đánh giá các mô hình MERC trên các bộ dữ liệu đa phương thức như IEMOCAP, CMU-MOSI và CMU-MOSEI. Các kết quả được báo cáo ở mức phát ngôn theo các tiêu chí Accuracy và F1 có trọng số/không trọng số, nhằm phản ánh cả hiệu năng tổng thể và ảnh hưởng của sự mất cân bằng nhãn.

Chương 2

Kết hợp đa phương thức cho nhận diện cảm xúc trong hội thoại

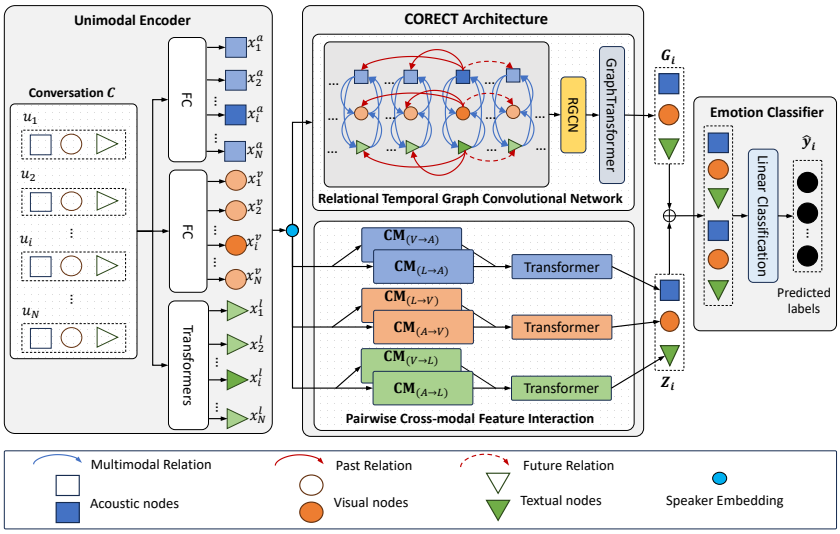
Chương này tập trung vào việc nâng cao cơ chế kết hợp đa phương thức cho MERC, trong đó mô hình cần suy luận hiệu quả từ các tín hiệu văn bản, âm thanh và hình ảnh trong bối cảnh hội thoại có cấu trúc. Cảm xúc trong hội thoại biến đổi theo các lượt thoại và phụ thuộc vào dòng chảy ngữ cảnh cũng như tương tác giữa các người nói, khiến MERC khó hơn nhiều so với bài toán phân loại cảm xúc từng phát ngôn độc lập. Các phương pháp dựa trên mạng hồi quy, đồ thị và kết hợp đa phương thức hiện có mới chỉ nắm bắt được một phần cấu trúc này, và thường gặp khó khăn với phụ thuộc dài hạn, mất cân bằng giữa các phương thức.

Để khắc phục những hạn chế đó trong điều kiện dữ liệu đầy đủ phương thức, chúng tôi đề xuất **CORECT** [VanNTC 1], một khung mô hình đồ thị quan hệ–thời gian mô hình hoá đồng thời ngữ cảnh hội thoại và tương tác đa phương thức. CORECT tích hợp Relational Temporal Graph Convolutional Network (RT-GCN) nhằm nắm bắt các quan hệ theo thời gian và phụ thuộc vào người nói ở mức phát ngôn, cùng với mô-đun Pairwise Cross-modal Feature Interaction (P-CM) mô hình hoá tường minh các tương tác giữa các phương thức ở mức hội thoại.

Trên nền tảng đó, chúng tôi tiếp tục giới thiệu **MultiDAG+CL** [VanNTC 2], mở rộng suy luận dựa trên đồ thị có hướng không chu trình (Directed Acyclic Graph – DAG) sang đầu vào đa phương thức và tích hợp chiến lược curriculum learning. MultiDAG mô hình hoá dòng chảy thông tin dọc theo cấu trúc hội thoại có hướng để nắm bắt tốt hơn các phụ thuộc dài hạn và các chuyển dịch cảm xúc, trong khi curriculum learning dần dần cho mô hình tiếp xúc với các cuộc hội thoại khó hơn, giúp cải thiện độ ổn định và khả năng khái quát của mô hình.

2.1 Kết hợp đa phương thức với mô hình hoá quan hệ và thời gian trong hội thoại

CORECT tích hợp một Mạng Tích chập Đồ thị Quan hệ–Thời gian (Relational Temporal Graph Convolution Network – RT-GCN) với một mô-đun Tương tác Đặc trưng Liên phương thức theo cặp (Pairwise Cross-modal Feature Interaction – P-CM) để thực hiện nhận diện cảm xúc đa phương thức trong hội thoại.



Hình 2.2: Minh họa khung mô hình CORECT

Cho một cuộc hội thoại đa người nói C gồm N lượt lời $[u_1, u_2, \ldots, u_N]$, ký hiệu S là tập người nói tương ứng. Mỗi lượt lời u_i được gắn với ba phương thức gồm âm thanh (a), hình ảnh (v) và ngôn ngữ (l), lần lượt ký hiệu là u_i^a, u_i^v, u_i^l. Dựa trên biểu diễn ngữ cảnh cục bộ và toàn cục, bài toán ERC nhằm dự đoán nhãn cảm xúc cho mỗi lượt lời $u_i \in C$ từ tập nhãn cảm xúc được xác định trước với M nhãn $Y = [y_1, y_2, \ldots, y_M]$.

Khung CORECT xử lý hội thoại bằng cách trước tiên trích xuất **đặc trưng đơn phương thức ở mức câu nói**. Với mỗi u_i, ta thu được các embedding đặc trưng theo từng phương thức: $x_i^a,\ x_i^v,\ x_i^l$ cho âm thanh, hình ảnh và ngôn ngữ (sử dụng các tầng FC cho a, v và bộ mã hóa Transformer cho l). Để kết hợp thông tin người nói, mỗi embedding được tăng cường bằng véc-tơ người nói s_i, tạo thành: $X_i^\tau = x_i^\tau + \eta s_i, \quad \tau \in \{a, v, l\}$.

Để nắm bắt cấu trúc hội thoại cục bộ, CORECT xây dựng một đồ thị đa phương thức $\mathcal{G}(\mathcal{V}, \mathcal{R}, \mathcal{E})$, trong đó mỗi lượt lời tạo ra ba nút $u_i^a,\ u_i^v,\ u_i^l$, và các cạnh mã hóa (1) *quan hệ đa phương thức* $\mathcal{R}_{multi}$ liên kết các phương thức khác nhau của cùng một lượt lời, và (2) *quan hệ thời gian* $\mathcal{R}_{temp}$ liên kết các lượt lời trước–sau trong một cửa sổ thời gian. Mạng Tích chập Đồ thị Quan hệ–Thời gian (RT-GCN) tạo ra các biểu diễn có ngữ

Bảng 2.1: Kết quả trên IEMOCAP (6-way).

Methods	IEMOCAP (6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)
bc-LSTM	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10
CMN	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13
ICON	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN	47.10	**80.88**	58.71	66.08	70.97	61.21	65.54	65.04
MMGCN	45.45	77.53	61.99	66.70	72.04	64.12	65.56	65.71
DialogueCRN	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
COGMEN	55.76	80.17	63.21	61.69	**74.91**	63.90	67.04	67.27
CORECT (Ours)	**59.30**	80.53	**66.94**	**69.59**	72.69	**68.50**	**69.93** ($\uparrow$ 2.89)	**70.02** ($\uparrow$ 2.75)

cảnh: $g_i^\tau, \quad \tau \in \{a, v, l\}$. Một tầng Graph Transformer tiếp tục tinh chỉnh các biểu diễn này, thu được: o_i^τ.

Để nắm bắt các phụ thuộc liên phương thức ở cấp độ toàn cục, CORECT áp dụng một bộ biến đổi liên phương thức theo cặp (pairwise cross-modal transformer). Với mỗi cặp phương thức α và β, mô-đun này tính toán các tương tác dựa trên cơ chế chú ý: $\text{CM}^{\beta \to \alpha}$, và xếp chồng D tầng để tạo ra đặc trưng liên phương thức cuối cùng: $Z_{\alpha \rightleftarrows \beta}^{[D]}$. Quy trình này được lặp lại cho tất cả các cặp phương thức (a, v, l). Biểu diễn cuối cùng của mỗi lượt lời nói được tạo bằng cách kết hợp: $H_i = \text{Fusion}(o_i^\tau, \ Z_{\alpha \rightleftarrows \beta}^{[D]})$. Một bộ phân loại dự đoán nhãn cảm xúc: $\hat{y}_i = \arg\max(\text{softmax}(W H_i + b))$.

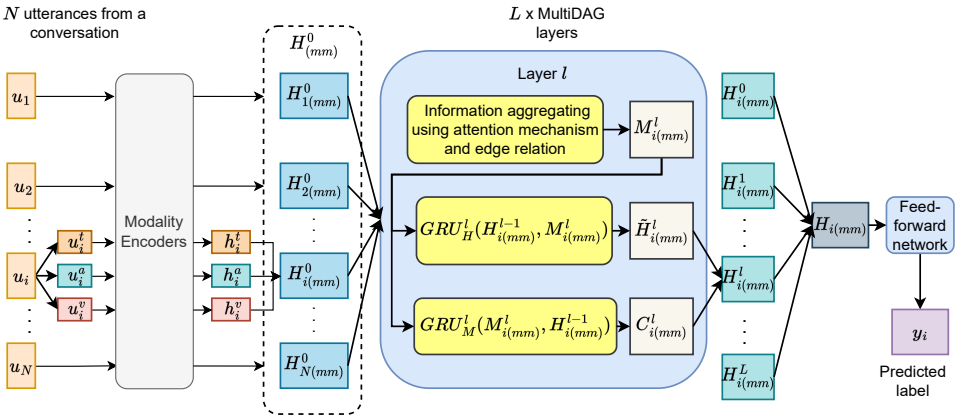
Kết quả. Trên bộ dữ liệu IEMOCAP, CORECT liên tục vượt trội hơn các mô hình trước đây, đạt mức cải thiện đáng kể ở cả thiết lập 6 lớp và 4 lớp. Trên CMU-MOSEI, CORECT cũng thể hiện hiệu năng mạnh mẽ trong các nhiệm vụ phân loại cảm xúc (2 lớp và 7 lớp) cũng như phân loại cảm xúc tinh tế, đạt hoặc vượt mức của các mô hình hiện đại nhất. Các kết quả này khẳng định tính hiệu quả và độ bền vững của CORECT trong bài toán hiểu hội thoại đa phương thức.

Bảng 2.3: Kết quả trên bộ dữ liệu CMU-MOSEI so sánh với các nghiên cứu trước.

Methods	Sentiment Classification Accuracy (%)		Emotion Classification (Binary, 1 vs. all) weighted F1-score (%)					
	2 Class	7 Class	Happiness	Sadness	Angry	Fear	Disgust	Surprise
Multilouge-Net	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE	82.40	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN	82.95	45.22	70.88	70.91	74.20	*87.79*	81.83	*86.05*
CORECT (Ours)	**83.66**	**46.31**	**71.35**	**72.86**	**76.77**	**87.90**	**84.26**	**86.48**

2.2 Kết hợp đa phương thức với mô hình hoá đồ thị có hướng không chu trình và học theo giáo trình

Với một đoạn hội thoại $C = \{u_1, \ldots, u_N\}$, trong đó mỗi lượt lời u_i được biểu diễn bằng các đặc trưng đa phương thức $u_i = \{u_i^a, u_i^v, u_i^l\}$, mục tiêu của MultiDAG+CL là thực hiện nhận diện cảm xúc trong hội thoại đa phương thức (multimodal ERC) thông qua việc kết hợp **(1) MultiDAG**, một mô hình đồ thị có hướng không chu trình (directed acyclic graph) dùng để suy luận theo ngữ cảnh, và **(2) Curriculum Learning (CL)**, một chiến lược huấn luyện dựa trên độ khó của mẫu.



Hình 2.7: Tổng quan kiến trúc của MultiDAG.

Mỗi modality được mã hoá bằng các bộ mã hoá chuyên biệt: $h_i^a = Enc_A(u_i^a)$, $h_i^v = Enc_V(u_i^v)$, $h_i^l = Enc_L(u_i^l)$, và sau đó được nối lại để tạo thành biểu diễn đa phương thức của lượt lời $H_{i(mm)}^0 = h_i^a \oplus h_i^v \oplus h_i^l$.

Đoạn hội thoại được mô hình hoá dưới dạng một DAG, trong đó mỗi lượt lời chỉ nhận thông tin từ các lượt xuất hiện trước nó. Một DAG-GNN cập nhật trạng thái ẩn qua từng tầng bằng cách sử dụng attention lên các nút phía trước và tích hợp dựa trên GRU, tạo ra các biểu diễn có ngữ cảnh $H_{i(mm)} = \|_{l=0}^{L}(\tilde{H}_{i(mm)}^l + C_{i(mm)}^l)$. Cách xử lý này cho phép mô hình nắm bắt được cả phụ thuộc cục bộ lẫn dài hạn theo dòng chảy có hướng của hội thoại. Biểu diễn cuối cùng $H_{i(mm)}$ được đưa vào một mạng Feed-Forward để dự đoán cảm xúc.

Thước đo độ khó (Difficulty Measure). Hàm đo độ khó (DMF) ước lượng độ khó của mỗi đoạn hội thoại dựa trên tần suất *chuyển đổi cảm xúc*: $\mathcal{DIF}(c_i) = \frac{N_{\text{shift}}(c_i)+N_{\text{sp}}(c_i)}{N_u(c_i)+N_{\text{sp}}(c_i)}$. Mức độ biến thiên cảm xúc càng lớn thì hội thoại càng khó.

Bộ lập lịch huấn luyện (Training Scheduler). Các hội thoại được sắp xếp theo

độ khó và chia thành các nhóm tăng dần $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$. Quá trình huấn luyện bắt đầu từ nhóm dễ nhất và dần bổ sung các nhóm khó hơn, tạo thành một lộ trình học có thứ tự: $\mathcal{D}^{train} \leftarrow \mathcal{D}_1 \rightarrow \mathcal{D}_1 \cup \mathcal{D}_2 \rightarrow \cdots \rightarrow \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_k$. Sau khi tất cả các nhóm được đưa vào, giai đoạn huấn luyện chuẩn được tiếp tục.

Kết quả. Bảng 2.8 trình bày so sánh toàn diện với các mô hình ERC đa phương thức hiện đại. Phương pháp *MultiDAG+CL* đạt hiệu suất tốt nhất trên cả IEMOCAP và MELD, vượt mức SOTA trước đó lần lượt 1.05% và 0.34%. Mô hình cũng mang lại cải thiện ổn định trên hầu hết các lớp cảm xúc, đặc biệt ở các lớp *Sad*, *Neutral*, và *Angry*.

Bảng 2.8: Kết quả trên IEMOCAP and MELD.

Mô hình	**IEMOCAP**								**MELD**	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)	Acc. (%)	w-F1 (%)
bc-LSTM	33.82	78.76	56.75	64.35	60.25	60.75	60.51	60.42	59.62	57.29
MFN	48.19	73.41	56.28	63.04	64.11	61.82	61.24	61.60	60.80	57.80
ICON	32.80	74.40	60.60	68.20	68.40	66.20	64.00	63.50	58.20	56.30
DialogueRNN	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	60.31	57.66
DialogueGCN	**51.57**	80.48	57.69	53.95	72.81	57.33	63.22	62.89	58.62	56.36
DAG-ERC	47.59	79.83	69.36	66.67	66.79	**68.66**	67.53	68.03	61.04	63.66
MMGCN	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	60.42	58.31
CTNet	51.3	79.9	65.8	67.2	**78.7**	58.8	68.0	67.5	62.0	60.5
DAG-ERC+HCL	-	-	-	-	-	-	-	68.73	-	63.89
COGMEN	-	-	-	-	-	-	68.2	67.6	-	-
MultiDAG (Ours)	49.65	79.83	66.40	67.59	71.78	67.90	68.30	68.45	64.29	63.87
MultiDAG+HCL (Ours)	45.26	**81.40**	**69.53**	**70.33**	71.61	66.94	**69.11**	**69.08**	**64.41**	**64.00**

2.3 Tổng kết chương

Trong chương này, chúng tôi đã giải quyết trọng tâm nghiên cứu thứ nhất của luận án. Mục 2.1 giới thiệu **CORECT**, mô hình hoá cấu trúc hội thoại theo quan hệ và thời gian thông qua RT-GCN và mô-đun P-CM, từ đó tăng cường kết hợp đa phương thức trong điều kiện dữ liệu đầy đủ phương thức. Mục 2.2 trình bày **MultiDAG+CL**, kết hợp mô hình hoá ngữ cảnh dựa trên đồ thị có hướng không chu trình (DAG) với chiến lược học theo giáo trình (curriculum learning) để xử lý tốt hơn các chuyển dịch cảm xúc và mức độ khó khác nhau của hội thoại.

Các thực nghiệm trên IEMOCAP và CMU-MOSEI (đối với CORECT), cũng như trên IEMOCAP và MELD (đối với MultiDAG+CL), cho thấy mô hình đều đạt được mức cải thiện ổn định so với các đường cơ sở mạnh, qua đó trực tiếp hỗ trợ **Mục tiêu O1** bằng cách cung cấp các cơ chế kết hợp hội thoại đa phương thức có cấu trúc cho MERC. Nhìn chung, các phương pháp này hình thành nền tảng về kết hợp và mô hình hoá ngữ cảnh của luận án, đồng thời mở đường cho chương tiếp theo, nơi chúng tôi nghiên cứu các cơ chế kết hợp đa phương thức vững chắc trong điều kiện dữ liệu thực tế chất lượng thấp.

Chương 3

Nhận diện cảm xúc đa phương thức trong hội thoại với mô thức không đầy đủ

Giới thiệu. Giao tiếp của con người vốn mang tính đa phương thức, kết hợp ngôn ngữ, giọng nói, biểu cảm khuôn mặt và nhiều tín hiệu hành vi khác, khiến MERC và MSA trở thành những nhiệm vụ cốt lõi trong affective computing. Tuy nhiên, trong thực tế, các hệ thống hiếm khi tiếp cận được đầy đủ tất cả các phương thức. Do đó, dữ liệu đa phương thức không đầy đủ là một thách thức phổ biến, nhưng nhiều mô hình hiện tại vẫn giả định rằng mọi phương thức luôn có sẵn (đầy đủ) hoặc chỉ xử lý sự thiếu hụt một cách hời hợt.

Các giải pháp hiện tại thường rơi vào ba nhóm chính nhưng vẫn chưa đủ cho bài toán nhận diện cảm xúc trong hội thoại. Nhiều hạn chế quan trọng vẫn tồn tại: phần lớn các phương pháp giả định chỉ một phương thức bị thiếu tại một thời điểm; nhiều mô hình loại bỏ hoàn toàn mẫu dữ liệu không đầy đủ hoặc chỉ áp dụng các heuristic đơn giản; và rất ít phương pháp đồng thời khai thác tái tạo đặc trưng, suy luận liên phương thức và mô hình hóa ngữ cảnh dựa trên đồ thị. Đáng chú ý, hội thoại đa phương thức chứa các quan hệ phụ thuộc phong phú mà các kiến trúc hiện tại vẫn chưa khai thác đầy đủ, đặc biệt trong bối cảnh thiếu phương thức.

Để giải quyết những thách thức này, chúng tôi giới thiệu **Mi-CGA** (Hình 3.2), một khung thống nhất dựa trên đồ thị dành cho nhận diện cảm xúc đa phương thức trong hội thoại dưới điều kiện thiếu phương thức. Mi-CGA hoạt động theo hai giai đoạn. Giai đoạn Incomplete Multimodal Representation (IMR) mô phỏng điều kiện thiếu dữ liệu và học các biểu diễn đa phương thức bền vững có thể hoạt động với bất kỳ tập con phương thức nào. Giai đoạn thứ hai, Cross-modal Graph Attention Network (CGA-Net), xây dựng một kiến trúc suy luận đa phương thức nhạy cảm với ngữ cảnh.

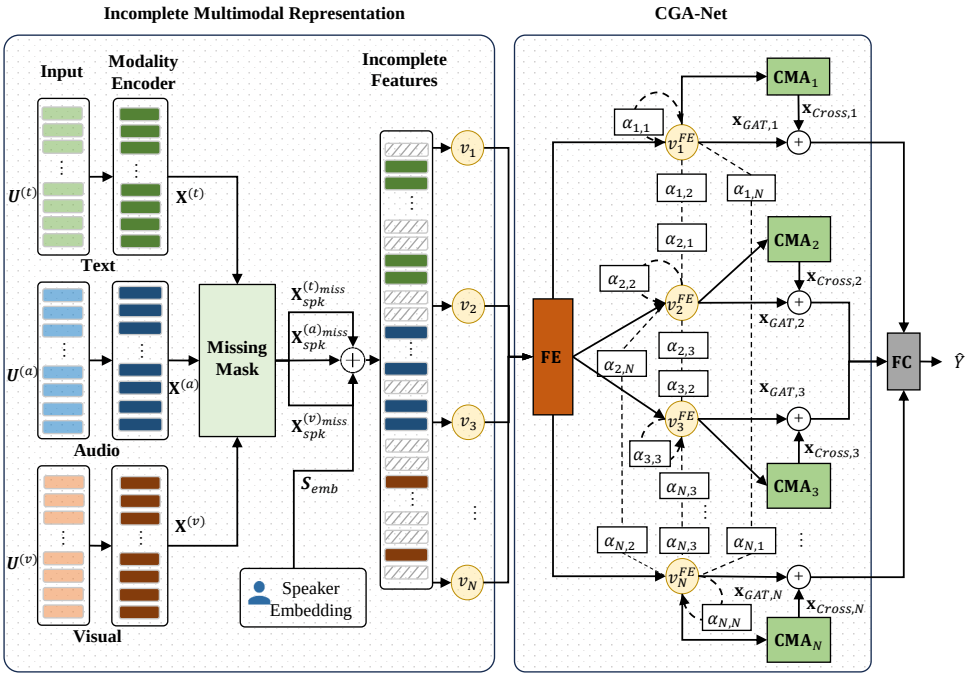
Nghiên cứu này thúc đẩy **Objective O2** của luận án bằng cách giải quyết tính bền

vững trước dữ liệu đa phương thức chất lượng thấp, đặc biệt là trường hợp thiếu phương thức. Hơn nữa, Mi-CGA đóng góp trực tiếp cho **RQ3** thông qua việc tái tạo đặc trưng bị thiếu bằng mô hình hóa dựa trên đồ thị và attention; đồng thời hỗ trợ **RQ4** bằng cách cho phép học đa phương thức ổn định và chính xác trong các kịch bản thiếu phương thức hoặc thay đổi số lượng phương thức. Công trình này đã được công bố tại Neurocomputing (SCIE Q1, IF 6.5) năm 2025 [**VanNTC 3**].

Mi-CGA: Mạng Tập trung Đồ thị Liên phương thức cho Nhận diện Cảm xúc Bền vững dưới Điều kiện Thiếu Phương thức. Xét một hội thoại C gồm chuỗi N lượt phát ngôn $\{U_1, U_2, \ldots, U_N\}$. Mỗi lượt U_i trong C gắn với ba phương thức dữ liệu gồm âm thanh (a), hình ảnh (v) và văn bản (t), tức là $U_i = \{u_i^{(a)}, u_i^{(v)}, u_i^{(t)}\}$.

Giả sử tồn tại một tỉ lệ thiếu phương thức tổng thể $\rho \in [0, 1]$ áp dụng lên dữ liệu đa phương thức. Khi đó, ta ký hiệu $x_i^{(a)miss} \in \mathbb{R}^{d_a}$, $x_i^{(v)miss} \in \mathbb{R}^{d_v}$, và $x_i^{(t)miss} \in \mathbb{R}^{d_t}$ là biểu diễn không đầy đủ của $u_i^{(a)}$, $u_i^{(v)}$ và $u_i^{(t)}$ với các kích thước tiềm ẩn tương ứng d_a, d_t, d_v.

Trong bài toán nhận diện cảm xúc đa phương thức dưới điều kiện thiếu phương thức, Mi-CGA nhận vào bộ ba $(x_i^{(a)miss}, x_i^{(v)miss}, x_i^{(t)miss})$ và dự đoán nhãn cảm xúc tương ứng $\hat{y}_i$ của U_i từ tập nhãn định nghĩa trước $E = \{y_1, y_2, \ldots, y_{|E|}\}$.



Hình 3.2: Kiến trúc tổng thể của mô hình Mi-CGA.

Mi-CGA được thiết kế để giải quyết bài toán nhận diện cảm xúc đa phương thức trong hội thoại dưới các điều kiện thiếu phương thức. Với hội thoại gồm N lượt phát ngôn $\{U_1, \ldots, U_N\}$, mỗi U_i chứa đặc trưng âm thanh, hình ảnh và văn bản $U_i = \{u_i^{(a)}, u_i^{(v)}, u_i^{(t)}\}$,

một tỉ lệ thiếu phương thức toàn cục ρ được áp dụng nhằm mô phỏng tình huống thiếu dữ liệu ở cả giai đoạn huấn luyện và suy luận. Bộ đặc trưng bị che $(x_i^{(a)miss}, x_i^{(v)miss}, x_i^{(t)miss})$ sau đó được sử dụng để dự đoán nhãn cảm xúc $\hat{y}_i$.

Mô hình gồm hai giai đoạn chính: (1) Giai đoạn I - Incomplete Multimodal Representation (IMR): Giai đoạn này học các đặc trưng biểu diễn bền vững ở mức lượt phát ngôn, có thể hoạt động với bất kỳ dạng thiếu phương thức nào. Các bộ mã hoá theo phương thức sinh ra embedding ngữ cảnh: $x_i^{(a)}$, $x_i^{(v)}$, và $x_i^{(t)}$ (BiLSTM cho văn bản và mạng fully-connected cho âm thanh và hình ảnh). Mặt nạ thiếu $\mathcal{M}$ được áp dụng theo tỉ lệ thiếu ρ, tạo ra đặc trưng bị che $\mathbf{X}^{miss} = \mathbf{X} \odot \mathcal{M}$. Embedding người nói S_{emb} được ghép vào để tăng giàu thông tin: $\mathbf{X}_{spk}^{miss} = \eta S_{emb} \oplus \mathbf{X}^{miss}$. (2) Giai đoạn II - Cross-modal Graph Attention Network (CGA-Net): Giai đoạn này kết hợp đặc trưng được tái tạo, cấu trúc hội thoại và tương tác liên phương thức. Feature Estimation (FE) tái tạo phương thức bị thiếu bằng bộ mã hoá GCN f_ϕ và bộ giải mã MLP g_θ để thu được: $\mathbf{X}^{coarse} = g_\theta(f_\phi(\mathbf{X}_{spk}^{miss}, \tilde{\mathbf{A}}))$.

Bước làm mượt kết hợp giá trị tái tạo và quan sát: $\mathbf{X}^{FE} = (1-\lambda)\mathbf{X}^{coarse} + \lambda\mathbf{X}_{spk}^{miss}$. MulGAT mô hình hóa phụ thuộc trong hội thoại. Hệ số chú ý α_{ij} được tính trên các lượt lân cận và tổng hợp bằng đa đầu: $\mathbf{X}_{GAT} = \text{MulGAT}(\mathbf{X}^{FE})$. CMA nắm bắt tương tác liên phương thức chi tiết bằng attention hai chiều cho mọi cặp phương thức, sinh ra: $\mathbf{X}^{a\rightleftarrows t}$, $\mathbf{X}^{t\rightleftarrows v}$, $\mathbf{X}^{v\rightleftarrows a}$.

Tất cả được ghép thành đặc trưng liên phương thức toàn cục: $\mathbf{X}_{Cross}$. Biểu diễn cuối cùng của mỗi lượt phát ngôn là: $\mathbf{X}_{Final} = [\mathbf{X}_{GAT}, \mathbf{X}_{Cross}]$. Một bộ phân loại feed-forward + softmax dự đoán: $\hat{y}_i$. Mô hình được huấn luyện với hàm mất mát kép: $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rct}$, trong đó $\mathcal{L}_{cls}$ là cross-entropy, còn $\mathcal{L}_{rct}$ là mất mát tái tạo (MSE hoặc KL) tại các vị trí đã che.

Kết quả. Bảng 3.2 cho thấy Mi-CGA vượt trội hơn tất cả các mô hình SOTA trên nhiều bộ dữ liệu trong điều kiện thiếu phương thức. Trên IEMOCAP (4-class), Mi-CGA đạt mức cải thiện +6.30% w-F1 trung bình so với GCNet; và trên IEMOCAP (6-class), mô hình đạt SOTA mới với độ chính xác 62.43% (+6.25%). Các mức cải thiện tương tự trên CMU-MOSI và CMU-MOSEI tiếp tục khẳng định tính vững chắc của Mi-CGA trong bối cảnh ERC với dữ liệu đa phương thức không đầy đủ.

Ở mọi mức missing rate, từ dữ liệu đầy đủ đến thiếu hụt nghiêm trọng, Mi-CGA đều cho kết quả vượt trội so với các phương pháp trước. Đáng chú ý, Mi-CGA suy giảm hiệu năng rất ít khi ρ tăng. Xu hướng này xuất hiện nhất quán trên tất cả các bộ dữ liệu, chứng minh khả năng thích ứng và độ bền vững của Mi-CGA trong cả điều kiện đầy đủ lẫn thiếu hụt phương thức.

Bảng 3.2: So sánh với các phương pháp hiện tại ở nhiều mức độ thiếu phương thức khác nhau.

Dataset	Models	Missing Rates								
		0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**Average**
IEMOCAP (4-way)	CPM-Net	58.00	55.29	53.65	52.52	51.01	49.09	47.38	44.76	51.46
	AE	74.82	71.36	67.40	62.02	57.24	50.56	43.04	39.86	58.29
	CRA	76.26	71.28	67.34	62.24	57.04	49.86	43.22	38.56	58.23
	MMIN	74.94	71.84	69.36	66.34	63.30	60.54	57.52	55.44	64.91
	GCNet	78.36	77.48	77.34	76.22	75.14	73.80	71.88	71.38	75.20
	Mi-CGA	**83.42**	**82.83**	**82.27**	**81.50**	**83.17**	**80.08**	**79.96**	**79.35**	**81.50**
	Δ	5.06	5.35	4.93	5.28	8.03	6.28	8.08	7.97	6.30
IEMOCAP (6-way)	CPM-Net	41.05	37.33	36.22	35.73	35.11	33.64	32.26	31.25	35.32
	AE	56.76	52.82	48.66	42.26	35.18	29.12	25.08	23.18	39.13
	CRA	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN	56.96	53.94	51.46	48.42	45.60	42.82	40.18	37.84	47.15
	GCNet	58.64	58.50	57.64	57.08	56.12	54.40	53.60	53.46	56.18
	Mi-CGA	**66.04**	**65.83**	**64.07**	**63.08**	**61.72**	**59.96**	**59.52**	**59.18**	**62.65**
	Δ	7.36	7.33	6.43	6.00	5.60	5.56	5.92	5.72	6.47
CMU-MOSI	CPM-Net	71.90	68.91	71.12	70.59	64.95	65.88	64.02	61.79	67.77
	AE	56.76	52.82	48.66	42.26	35.18	29.12	25.08	23.18	39.13
	CRA	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN	85.20	81.91	78.22	74.60	70.14	67.72	64.04	61.53	72.92
	GCNet	85.01	82.54	80.17	78.54	76.48	73.45	69.46	68.35	76.75
	DiCMoR	85.60	83.90	82.00	80.20	77.70	76.40	73.00	70.08	78.70
	IMDer	85.60	84.80	**83.40**	81.00	78.50	75.90	74.00	71.20	79.30
	Mi-CGA	**87.21**	**85.02**	83.28	**81.83**	**79.56**	**78.62**	**75.63**	**73.05**	**80.05**
	Δ	1.61	0.22	-0.12	0.83	1.06	2.22	1.63	1.85	0.75
CMU-MOSEI	CPM-Net	78.47	74.79	74.48	73.81	72.39	70.43	68.73	67.07	72.52
	AE	86.66	84.37	82.58	80.57	78.80	76.43	74.26	72.81	79.56
	CRA	86.48	84.19	82.25	80.12	78.55	75.85	74.07	72.46	79.25
	MMIN	85.78	83.77	81.85	79.77	77.63	75.36	72.95	71.18	78.54
	GCNet	87.12	**86.50**	85.50	84.53	83.55	82.44	80.27	80.20	83.76
	DiCMoR	85.10	83.50	81.50	79.30	77.40	75.80	73.70	72.20	78.60
	IMDer	85.10	84.60	82.40	80.70	78.10	77.40	75.50	74.60	79.80
	Mi-CGA	**87.61**	86.21	**85.80**	**84.81**	**84.26**	**84.82**	**82.85**	**81.56**	**83.92**
	Δ	0.49	-0.29	0.30	0.28	0.71	2.38	2.58	1.36	0.16

3.1 Chapter summary

Trong chương này, chúng tôi đã giải quyết thách thức học từ dữ liệu đa phương thức không đầy đủ trong bối cảnh dữ liệu chất lượng thấp bằng cách giới thiệu Mi-CGA, một framework chuyên biệt cho bài toán Nhận diện Cảm xúc trong Hội thoại (MER). Bằng cách kết hợp hiệu quả cấu trúc nội phương thức và liên phương thức thông qua mô hình hóa dựa trên đồ thị, Mi-CGA trực tiếp đóng góp vào **Objective O2** và giải quyết **RQ3** thông qua việc cung cấp một giải pháp có nguyên lý cho tái tạo đặc trưng phương thức bị thiếu, đồng thời đảm bảo khả năng hiểu cảm xúc ổn định ngay cả khi dữ liệu đa phương thức không đầy đủ.

Chương 4

Nhận diện cảm xúc đa phương thức trong hội thoại với phương thức mất cân bằng

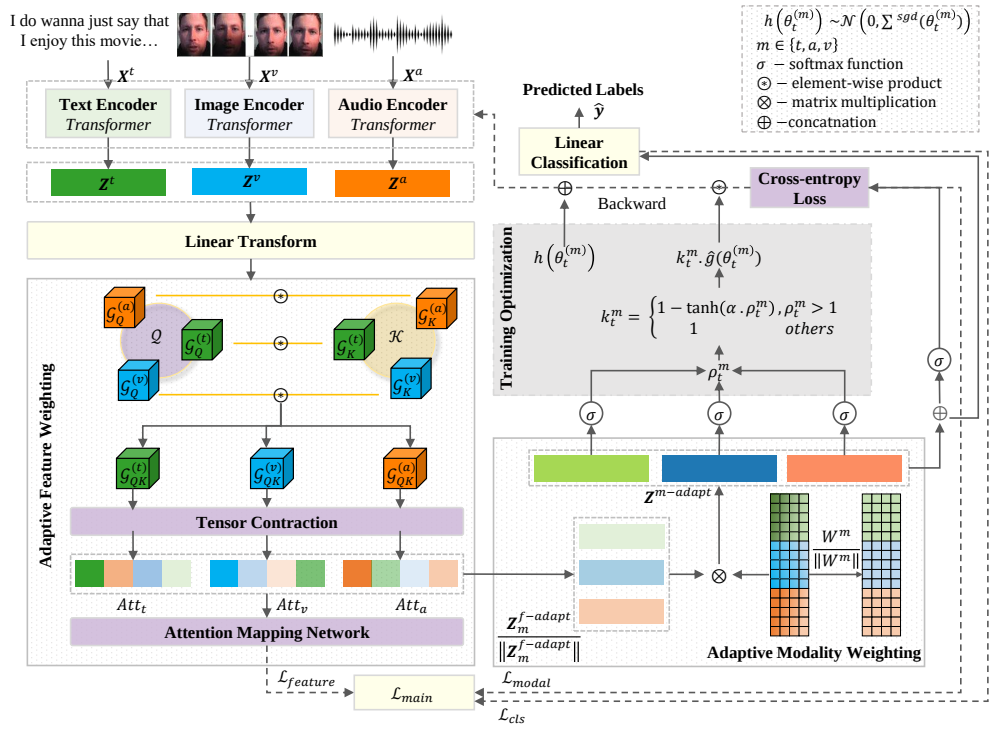
Trọng tâm thứ hai của luận án trong hướng nghiên cứu về dữ liệu đa phương thức chất lượng thấp là **học cân bằng đa phương thức**, với mục tiêu đảm bảo rằng mọi phương thức đều đóng góp một cách có ý nghĩa trong quá trình huấn luyện thay vì để một phương thức mạnh lấn át các phương thức còn lại. Thách thức này—thường được gọi là *mất cân bằng phương thức*—xuất hiện phổ biến trong bài toán ERC đa phương thức và các nhiệm vụ liên quan, khi tín hiệu văn bản thường chi phối quá trình tối ưu hoá, trong khi các phương thức âm thanh và hình ảnh bị học kém hơn.

Để giải quyết vấn đề này, luận án đề xuất hai hướng tiếp cận bổ trợ: (1) **Ada2I**, một mô hình end-to-end tái cân bằng việc học đa phương thức ở cả cấp độ đặc trưng và cấp độ phương thức; và (2) **SPCL**, một mô-đun nhẹ dạng plug-and-play giúp giảm dần sự mất cân bằng phương thức và có thể tích hợp linh hoạt vào nhiều kiến trúc đa phương thức khác nhau. Hai đóng góp này trực tiếp hỗ trợ **Objective O2**, với mục tiêu xây dựng các chiến lược học đa phương thức vững chắc trong điều kiện dữ liệu không hoàn hảo hoặc chất lượng thấp. Cụ thể: (i) bằng cách tăng cường hoặc khôi phục các đặc trưng của phương thức bị học kém, chúng giải quyết **RQ3** về cải thiện biểu diễn đa phương thức khi tín hiệu suy giảm hoặc không đầy đủ; và (ii) bằng cách đưa ra các cơ chế cân bằng thích ứng dựa trên curriculum ở các mức độ đặc trưng, phương thức, cảm xúc và hội thoại, chúng cải thiện tính ổn định và khả năng khái quát hoá của mô hình, đáp ứng yêu cầu trong **RQ4**.

Các kết quả này đã được kiểm chứng thông qua công bố tại *ACM Multimedia 2024 (CORE A*)* đối với Ada2I [**VanNTC 4**], và một nghiên cứu mở rộng về cân bằng phương thức dựa trên Self-Paced Curriculum Learning (SPCL), hiện đang trong giai đoạn major revision tại *Neural Computing and Applications (Scopus Q1)* [**VanNTC 5**].

4.1 Ada2I: Tăng cường cân bằng phương thức cho nhận diện cảm xúc hội thoại đa phương thức

Ada2I là một framework end-to-end được thiết kế nhằm giảm thiểu hiện tượng *mất cân bằng phương thức* trong bài toán ERC đa phương thức bằng cách đảm bảo rằng mọi phương thức đều đóng góp một cách tương xứng trong quá trình học biểu diễn. Mô hình bao gồm ba giai đoạn chính: (1) *Mã hoá phương thức (Modality Encoding)*, (2) *Cân bằng đặc trưng thích ứng (Adaptive Feature Weighting – AFW)*, và (3) *Cân bằng phương thức thích ứng (Adaptive Modality Weighting – AMW)*, cùng với một chiến lược huấn luyện giúp cân bằng động các gradient giữa các phương thức.



Hình 4.2: Kiến trúc tổng quan của Ada2I

Với một hội thoại gồm các chuỗi đơn phương thức $\mathbf{X}^m$ cho $m \in \{t, a, v\}$, bộ mã hoá Transformer $\phi(\theta^{(m)})$ tạo ra các biểu diễn đã được ngữ cảnh hoá theo từng phương thức dưới dạng $\mathbf{Z}^m = \phi(\theta^{(m)}, \mathbf{X}^m)$. Đây là các đặc trưng đơn phương thức ban đầu phục vụ cho các bước cân bằng tiếp theo.

Để hiệu chỉnh sự mất cân bằng tại *mức đặc trưng*, Ada2I mô hình hoá tương tác đa chiều giữa các phương thức thông qua phân rã vòng tensor (tensor-ring decomposition). Mỗi phương thức sinh ra các lõi truy vấn và khoá dạng vòng tensor $\mathcal{G}_Q^m$ và $\mathcal{G}_K^m$, sau đó được hợp nhất thành các tensor bậc cao hơn $\mathcal{Q}$ và $\mathcal{K}$. Các hệ số chú ý Θ^m được tính toán

thông qua tương tác theo từng phần tử có chuẩn hoá tỉ lệ, và được rút gọn thành các ma trận pooling $\mathbf{A}^{(m)}$. Từ đó thu được trọng số chú ý theo đặc trưng Att_m, và biểu diễn đặc trưng đã được cân bằng được tính như sau: $\mathbf{Z}_m^{f\text{-}adapt} = Att_m \mathbf{Z}^m + \beta \mathbf{Z}^m$, trong đó β điều chỉnh mức đóng góp còn lại từ đặc trưng gốc.

Để ngăn các phương thức mạnh (ví dụ: văn bản) lấn át các phương thức yếu hơn, Ada2I áp dụng chuẩn hoá L2 theo phương thức và tái trọng số dựa trên cosine. Với mỗi phương thức m, các logits đã chuẩn hoá được tính bằng $\mathbf{Z}_m^{f\text{-}adapt}$ và trọng số đầu ra W^m, và biểu diễn đa phương thức đã được cân bằng được tổng hợp như sau: $\mathbf{Z}^{m\text{-}adapt} = \sum_m \frac{W^m \mathbf{Z}_m^{f\text{-}adapt}}{\|W^m\| \, \|\mathbf{Z}_m^{f\text{-}adapt}\|} + b$. Vector thu được được đưa qua một MLP để dự đoán nhãn cảm xúc $\hat{y}$.

Ada2I sử dụng hàm mất mát kết hợp, bao gồm mất mát phân loại $\mathcal{L}_{cls}$, mất mát căn chỉnh mức đặc trưng $\mathcal{L}_{feature}$, và mất mát cân bằng phương thức $\mathcal{L}_{modal}$. Để xử lý mất cân bằng tối ưu hoá, Ada2I mở rộng tỉ lệ chênh lệch lên ba phương thức bằng cách tính s_t^m (từ độ tương đồng cosine giữa logits và đặc trưng) và ước lượng mức độ chi phối tương đối: $\rho_t^m = \frac{s_t^m}{\min_j s_t^j}$. Hệ số điều biến $k_t^m = 1 - \tanh(\alpha \rho_t^m)$ được dùng để giảm gradient của các phương thức chiếm ưu thế trong quá trình cập nhật bộ mã hoá, đảm bảo rằng các phương thức yếu hơn nhận được mức tối ưu hoá đầy đủ xuyên suốt quá trình huấn luyện.

Biểu diễn đa phương thức đã được cân bằng $\mathbf{Z}^{m\text{-}adapt}$ sau đó được đưa vào một MLP để sinh nhãn cảm xúc dự đoán $\hat{y}$. Tổng thể, AFW, AMW, và cơ chế điều biến gradient cho phép Ada2I duy trì quá trình học đa phương thức ổn định, cân bằng và cải thiện hiệu quả nhận diện cảm xúc trên tất cả các phương thức.

Kết quả. Như thể hiện trong Bảng 4.2, Ada2I liên tục vượt trội hơn so với các mô hình SOTA trước đây trên tất cả các tổ hợp phương thức ở cả hai bộ dữ liệu. Đáng chú ý, với cặp phương thức yếu nhất (A+V) trên MELD, Ada2I đạt mức cải thiện đáng kể: tăng 10.77% về w-F1 và 6.98% về Accuracy, từ đó thu hẹp đáng kể khoảng cách so với các thiết lập vốn có ưu thế của phương thức văn bản.

Bảng 4.2: So sánh kết quả trong thiết lập đa phương thức của Ada2I với mô hình cơ sở đã được cân bằng phương thức thông qua FAGM (ký hiệu †).

IEMOCAP								
Methods	T+A+V		T+A		T+V		A+V	
	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc
DialogueRNN†	61.31	61.61	61.90	61.98	60.19	59.95	48.31	50.71
DialogueGCN†	62.76	63.22	64.36	64.39	61.25	62.23	49.20	49.85
BiDDIN†	58.81	58.84	58.88	58.16	59.04	58.96	46.36	46.77
MM-DFN†	64.92	64.57	63.91	64.20	61.02	60.60	54.48	55.03
MMGCN†	64.53	64.51	63.25	63.40	61.02	61.06	54.14	54.90
Ada2I (Ours)	**68.97**	**68.76**	**66.91**	**67.28**	**65.48**	**65.43**	**55.16**	**55.64**
$\Delta(\%)$	*↑4.05*	*↑4.19*	*↑2.55*	*↑2.89*	*↑4.23*	*↑3.20*	*↑0.68*	*↑0.61*

MELD								
Methods	T+A+V		T+A		T+V		A+V	
	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc
DialogueRNN†	56.42	58.05	56.46	58.01	55.67	57.39	40.46	45.39
DialogueGCN†	54.61	58.96	54.80	57.28	55.26	57.10	10.02	44.44
BiDDIN†	57.47	59.18	56.56	58.05	56.93	58.10	44.39	48.62
MM-DFN†	55.75	60.80	57.10	60.00	57.73	60.65	42.05	48.66
MMGCN†	58.48	61.15	57.59	60.69	57.14	59.46	43.49	48.43
Ada2I (Ours)	**60.38**	**63.03**	**60.08**	**62.64**	**58.62**	**61.95**	**55.16**	**55.64**
$\Delta(\%)$	*↑1.90*	*↑1.88*	*↑2.49*	*↑1.95*	*↑0.89*	*↑1.30*	*↑10.77*	*↑6.98*

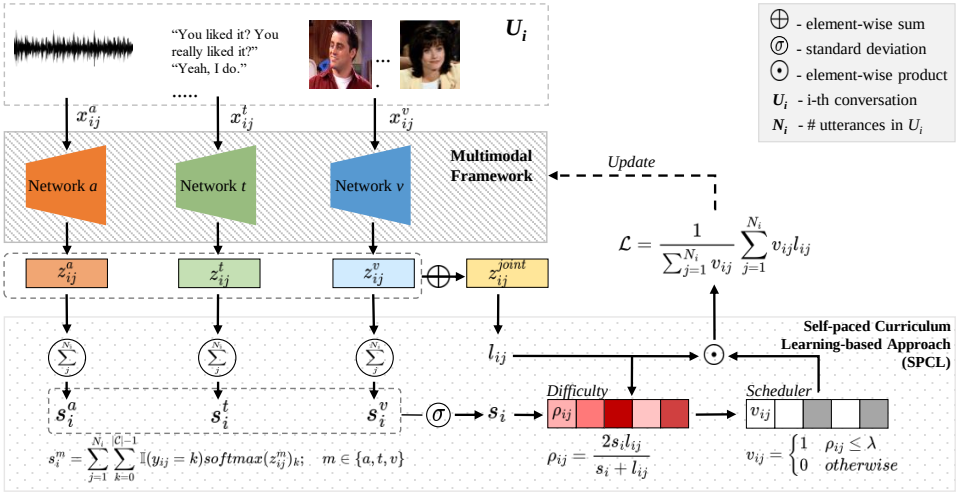
4.2 SPCL: Ứng dụng Self-Paced Curriculum Learning để tăng cường cân bằng phương thức trong nhận diện cảm xúc hội thoại đa phương thức

Nhiều hướng tiếp cận đã được đề xuất nhằm giảm thiểu tình trạng mất cân bằng giữa các phương thức, từ việc cải thiện chất lượng đặc trưng đơn phương thức, bổ sung các mục tiêu học phụ trợ, cho đến cân bằng gradient giữa các phương thức. Tuy nhiên, các phương pháp này thường đòi hỏi tiền huấn luyện quy mô lớn, các hàm mất mát đặc thù theo tác vụ, hoặc thao tác gradient phức tạp, khiến khả năng tổng quát hóa và mở rộng bị hạn chế.

Vượt qua những giới hạn đó, chúng tôi giới thiệu **Self-Paced Curriculum Learning (SPCL)**, một chiến lược huấn luyện đơn giản nhưng thích ứng cao, giúp cân bằng đóng góp giữa các phương thức bằng cách hướng dẫn mô hình học từ mẫu dễ đến mẫu khó theo cách tiến triển. SPCL bao gồm: (1) một **Bộ đo độ khó (Difficulty Measurer)** đánh giá độ phức tạp của mẫu dựa trên hiệu năng nhận dạng và mức độ sai lệch giữa các phương thức, và (2) một **Bộ lập lịch học (Learning Scheduler)** lựa chọn mẫu theo cách thích ứng để đảm bảo các phương thức yếu hơn nhận được tối ưu hóa đầy đủ.

Bằng cách kết hợp hai thành phần này, SPCL mang đến một giải pháp thống nhất,

không phụ thuộc kiến trúc, để xử lý mất cân bằng phương thức; đồng thời cải thiện tính ổn định trong huấn luyện và tăng cường đóng góp từ các phương thức vốn bị tối ưu hóa kém. Hình 4.7 minh họa quy trình tổng thể của SPCL và cách tích hợp liền mạch của nó vào các mô hình MER hiện có.



Hình 4.7: Khung tích hợp SPCL

Với một tập dữ liệu hội thoại đa phương thức $\mathcal{D}$ và tập nhãn cảm xúc $\mathcal{C}$, mỗi phát ngôn x_{ij} được biểu diễn thông qua ba phương thức (âm thanh, văn bản và hình ảnh). Phương pháp của chúng tôi trước hết sinh ra các logit đơn phương thức thông qua các bộ dự đoán chuyên biệt cho từng phương thức, sau đó hợp nhất chúng thành một logit đa phương thức phục vụ nhận dạng cảm xúc. Để tiếp tục xử lý hiện tượng mất cân bằng phương thức trong quá trình huấn luyện, chúng tôi tích hợp mô-đun Self-Paced Curriculum Learning (SPCL), mô-đun này sẽ chọn mẫu huấn luyện một cách động dựa trên độ khó của chúng, qua đó khuyến khích quá trình tối ưu diễn ra cân bằng hơn giữa các phương thức. Đối với mỗi phát ngôn x_{ij}, các bộ dự đoán đơn phương thức sinh ra logit theo: $z_{ij}^m = \phi_m(x_{ij}^m; \theta^m)$ với $m \in \{a, t, v\}$. Logit đa phương thức cuối cùng được tính bằng tổng các logit đơn phương thức: $z_{ij}^{joint} = \sum_m z_{ij}^m$, và hàm mất mát ở mức phát ngôn được xác định bởi: $l_{ij} = -\log(\text{softmax}(z_{ij}^{joint})_{y_{ij}})$.

SPCL giới thiệu cơ chế chọn mẫu thích ứng thông qua hai thành phần:

(1) *Difficulty Measurer*: Mỗi phát ngôn được gán một điểm độ khó ρ_{ij}, bao gồm: một thành phần ở mức phát ngôn l_{ij} phản ánh độ khó phân loại, và một thành phần ở mức hội thoại s_i được tính từ độ lệch chuẩn của các điểm đơn phương thức, dùng để ước lượng mức độ sai lệch giữa các phương thức. Độ khó cuối cùng được tính bằng trung bình điều hòa: $\rho_{ij} = \frac{2 s_i l_{ij}}{s_i + l_{ij}}$, bảo đảm rằng không thành phần nào chiếm ưu thế tuyệt đối.

Bảng 4.5: So sánh hiệu suất của các mô hình cơ sở với mô-đun SPCL của chúng tôi và các phương pháp bổ trợ khác trên IEMOCAP.

Mô hình	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [**?**]								
Baseline	60.43	60.54	61.61	61.72	59.19	59.48	47.89	48.49
+ RNA loss	58.43	58.47	57.42	57.73	56.23	56.62	47.40	49.29
+ OGM-GE	57.16	57.24	59.30	59.52	55.88	56.13	43.71	44.98
+ OPM	58.89	59.72	57.02	57.55	60.48	60.54	49.80	51.76
+ FAGM	62.76	63.22	64.36	64.39	61.25	62.23	49.20	49.85
+ **SPCL**	**66.99**† $\pm$1.03	**67.03**† $\pm$0.95	**65.32**† $\pm$0.99	**65.46**† $\pm$1.15	**64.47**† $\pm$0.21	**64.46**† $\pm$0.20	**57.89**† $\pm$1.00	**58.59**† $\pm$0.49
Δ	*4.23*	*3.81*	*0.96*	*1.07*	*3.22*	*2.23*	*8.09*	*6.83*
Δ_{Base}	*6.56*	*6.49*	*3.71*	*3.74*	*5.28*	*4.98*	*10.00*	*10.10*
BiDDIN [**?**]								
Baseline	58.29	58.20	58.73	58.67	58.57	57.93	45.35	46.03
+ RNA loss	58.63	58.55	58.02	57.92	57.29	57.24	42.54	44.82
+ OGM-GE	58.06	57.98	57.71	57.73	57.58	57.55	39.84	40.42
+ OPM	56.27	56.62	57.82	57.60	52.59	52.60	37.72	40.48
+ FAGM	58.81	58.84	58.88	58.16	59.04	58.96	**46.36**	46.77
+ **SPCL**	**59.90**† $\pm$0.13	**60.73**† $\pm$0.56	**60.24**† $\pm$1.11	**60.43**† $\pm$0.99	**61.10**† $\pm$0.82	**61.91**† $\pm$0.58	46.34 $\pm$0.43	**49.11**† $\pm$0.71
Δ	*1.09*	*1.89*	*1.36*	*1.76*	*2.06*	*2.95*	*-0.02*	*2.34*
Δ_{Base}	*1.61*	*2.53*	*1.51*	*1.76*	*2.53*	*3.98*	*0.99*	*3.08*
MMGCN								
Baseline	62.67	62.67	62.66	62.72	58.99	59.14	47.22	49.23
+ RNA loss	63.13	63.28	59.25	59.27	56.30	56.50	50.35	51.20
+ OGM-GE	62.42	62.69	62.33	62.42	58.83	59.03	51.90	53.54
+ OPM	64.60	64.10	62.30	62.70	59.70	59.60	50.60	52.00
+ FAGM	64.53	64.51	63.25	63.40	61.02	61.06	**54.14**	**54.90**
+ **SPCL**	**67.66**† $\pm$0.57	**67.71**† $\pm$0.64	**66.75**† $\pm$0.42	**66.51**† $\pm$0.42	**65.00**† $\pm$1.05	**65.09**† $\pm$1.11	53.70 $\pm$0.71	54.04 $\pm$0.94
Δ	*3.06*	*3.20*	*3.50*	*3.11*	*3.98*	*4.03*	*-0.44*	*-0.86*
Δ_{Base}	*5.00*	*5.04*	*4.09*	*3.79*	*6.01*	*5.95*	*6.48*	*4.81*
MM-DFN								
Baseline	61.54	61.72	61.98	62.12	59.78	59.93	48.42	49.11
+ RNA loss	60.23	60.49	60.18	60.41	57.74	57.92	45.63	46.32
+ OGM-GE	59.92	60.13	60.57	60.69	58.33	58.49	44.98	45.51
+ OPM	63.30	62.91	64.43	64.45	64.06	63.89	53.55	53.79
+ FAGM	63.45	63.72	63.83	63.94	61.58	61.72	50.35	51.02
+ **SPCL**	**67.16**† $\pm$0.67	**67.08**† $\pm$0.54	**66.03**† $\pm$0.86	**66.09**† $\pm$0.65	**64.31**† $\pm$0.66	**64.70**† $\pm$0.63	**53.38**† $\pm$0.67	**53.47**† $\pm$0.93
Δ	*3.71*	*3.36*	*1.60*	*1.64*	*0.25*	*0.81*	*-0.17*	*-0.32*
Δ_{Base}	*5.62*	*5.36*	*4.05*	*3.97*	*4.53*	*4.77*	*4.96*	*4.36*

(2) *Learning Scheduler*: Một bộ điều phối học với *hard regularizer* sẽ chọn các mẫu thỏa mãn $\rho_{ij} \leq \lambda$ thông qua mặt nạ nhị phân $v_{ij} \in \{0, 1\}$, giúp mô hình tập trung vào các mẫu dễ và cân bằng hơn trong các epoch đầu. Ngưỡng λ được tăng dần theo quy luật $\lambda^{(t)} = \alpha\lambda^{(t-1)}$, từ đó từng bước đưa các mẫu khó hơn vào quá trình học khi mô hình đã ổn định. SPCL điều chỉnh hàm loss MERC chuẩn bằng cách nhân trọng số v_{ij} vào các giá trị mất mát ở mức phát ngôn: $\mathcal{L}_{SPCL} = \frac{1}{\sum v_{ij}} \sum v_{ij}\, l_{ij}$. Các tham số của mô hình được cập nhật dựa trên gradient của $\mathcal{L}_{SPCL}$, giúp tối ưu hóa diễn ra theo hướng tiếp nhận dần các mẫu khó hơn nhưng vẫn đảm bảo cân bằng giữa các phương thức.

Kết quả. Trên cả hai bộ dữ liệu IEMOCAP và MELD, SPCL liên tục cải thiện hiệu suất so với các mô hình backbone và các kỹ thuật cân bằng hiện có, đặc biệt trong thiết lập đủ ba phương thức TAV. Trên IEMOCAP, SPCL mang lại mức cải thiện rõ rệt cho tất cả các mô hình, đạt mức tăng trung bình +0.85% so với phương pháp tốt thứ hai và +2.25% so với backbone. Trên MELD, SPCL cũng cho thấy mức cải thiện đáng kể, đặc biệt trong các kiến trúc transformer như MM-DFN, nơi SPCL vượt OPM +0.42%. Tuy nhiên, với DialogueGCN, mức cải thiện khiêm tốn hơn (ví dụ +0.14% so với OGM-GE), có khả năng xuất phát từ cấu trúc hội thoại ngắn và phân mảnh của MELD, vốn không hoàn toàn phù hợp với lịch trình học tiến dần của SPCL. Dù vậy, SPCL vẫn chứng tỏ

Bảng 4.6: So sánh hiệu suất của các mô hình cơ sở với mô-đun SPCL của chúng tôi và các phương pháp bổ trợ khác trên MELD.

Model	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN								
Baseline	53.11	55.08	51.99	54.22	54.22	56.07	43.54	44.54
+ RNA loss	56.65	58.47	54.21	58.35	53.78	58.12	**43.64**	**47.32**
+ OGM-GE	57.73	57.36	**56.38**	**58.81**	56.15	57.78	42.05	46.51
+ OPM	54.47	57.12	53.26	56.17	53.21	57.66	40.52	43.64
+ FAGM	54.61	58.96	54.80	57.28	55.26	57.10	40.02	44.44
+ **SPCL**	**57.87**$\pm$1.49	**60.77**$^{\dagger}\pm$1.21	58.04$\pm$0.56	60.84$\pm$0.69	**56.18**$\pm$1.38	**58.61**$^{\dagger}\pm$1.43	42.28$\pm$0.79	46.64$\pm$1.38
Δ	*0.14*	*1.81*	*1.66*	*2.03*	*0.03*	*0.49*	*-1.36*	*-0.68*
Δ_{Base}	*4.76*	*5.69*	*6.05*	*6.62*	*1.96*	*2.54*	*-1.26*	*2.10*
BiDDIN								
Baseline	56.41	58.54	56.23	57.85	56.46	58.06	43.07	47.35
+ RNA loss	52.18	49.16	53.21	50.31	52.59	49.43	41.05	44.60
+ OGM-GE	55.27	53.41	51.96	47.74	52.18	48.58	43.03	46.97
+ OPM	53.87	57.62	54.73	58.58	56.25	59.77	40.69	47.39
+ FAGM	57.47	59.18	56.56	58.05	56.93	58.10	**44.39**	**48.62**
+ **SPCL**	**57.60**$^{\dagger}\pm$0.25	**60.86**$^{\dagger}\pm$0.30	**58.08**$^{\dagger}\pm$0.30	**61.22**$^{\dagger}\pm$0.26	**58.10**$^{\dagger}\pm$0.43	**61.00**$^{\dagger}\pm$0.74	42.30$\pm$0.23	48.15$\pm$0.43
Δ	*0.13*	*1.68*	*1.52*	*2.64*	*1.17*	*1.23*	*-2.09*	*-0.47*
Δ_{Base}	*1.19*	*2.32*	*1.85*	*3.37*	*1.64*	*2.94*	*-0.77*	*1.12*
MMGCN								
Baseline	57.71	59.95	57.29	59.79	56.73	59.31	42.38	**49.12**
+ RNA loss	56.94	58.62	56.00	57.59	55.48	57.70	41.84	46.91
+ OGM-GE	57.59	59.92	56.80	59.77	56.20	59.08	42.20	48.81
+ OPM	55.78	57.24	56.27	59.77	55.29	59.23	42.72	47.20
+ FAGM	58.48	61.15	57.59	60.69	57.14	59.46	43.49	48.43
+ **SPCL**	**59.11**$^{\dagger}\pm$0.48	**61.32**$^{\dagger}\pm$0.48	**58.93**$^{\dagger}\pm$0.29	**61.65**$^{\dagger}\pm$0.39	**58.14**$^{\dagger}\pm$1.17	**60.64**$^{\dagger}\pm$1.81	**43.79**$^{\dagger}\pm$0.31	49.10$\pm$0.28
Δ	*0.63*	*0.17*	*1.34*	*0.96*	*1.00*	*1.18*	*0.30*	*-0.02*
Δ_{Base}	*1.40*	*1.37*	*1.64*	*1.86*	*1.41*	*1.33*	*1.41*	*-0.02*
MM-DFN								
Baseline	57.52	59.90	57.11	59.47	57.46	59.68	40.04	43.91
+ RNA loss	56.02	58.20	54.13	55.59	54.13	55.59	36.39	47.54
+ OGM-GE	56.53	58.39	55.86	59.08	56.25	58.24	40.60	48.43
+ OPM	58.75	61.42	57.67	61.38	58.28	61.49	**42.51**	47.16
+ FAGM	57.55	60.80	57.10	60.00	57.73	60.65	42.05	**48.66**
+ **SPCL**	**59.17**$^{\dagger}\pm$0.30	**61.91**$^{\dagger}\pm$0.90	**59.11**$^{\dagger}\pm$0.32	**62.31**$^{\dagger}\pm$0.32	**58.91**$^{\dagger}\pm$0.17	**61.94**$^{\dagger}\pm$0.34	43.32$\pm$0.57	48.59$^{\dagger}\pm$0.55
Δ	*0.42*	*0.49*	*1.44*	*0.93*	*0.63*	*0.45*	*0.81*	*-0.07*
Δ_{Base}	*1.65*	*2.01*	*2.00*	*2.84*	*1.45*	*2.26*	*3.28*	*4.68*

tính ổn định, mang lại các cải thiện nhất quán và có ý nghĩa trên cả hai benchmark.

4.3 Tổng kết chương

Trong chương này, chúng tôi đã giải quyết trọng tâm thứ hai của luận án: **học đa phương thức cân bằng cho dữ liệu đa phương thức chất lượng thấp**. Trong Mục 4.1, chúng tôi giới thiệu **Ada2I**, một framework end-to-end được thiết kế để tái cân bằng quá trình học ở cả hai mức độ—*đặc trưng* và *phương thức*—nhằm giảm thiểu hiện tượng phương thức chiếm ưu thế. Trong Mục 4.2, chúng tôi trình bày **SPCL**, một module nhẹ dạng plug-and-play giúp hướng dẫn mô hình học từ các mẫu dễ đến khó, đồng thời giảm dần sự mất cân bằng giữa các phương thức theo cách thích ứng, có thể tích hợp linh hoạt vào nhiều kiến trúc mô hình khác nhau. Chúng tôi cũng tiến hành các thí nghiệm mở rộng trên các bộ dữ liệu chuẩn như IEMOCAP, MELD và CMU-MOSEI, cho thấy cả Ada2I và SPCL đều cải thiện hiệu quả mô hình trong các điều kiện học mất cân bằng. Tổng thể, các đóng góp này thúc đẩy việc hiện thực hóa **Objective O2** bằng cách giải quyết hai khía cạnh cốt lõi của học đa phương thức cân bằng.

Kết luận

Luận án này trình bày các công trình nghiên cứu về MERC, đồng thời xử lý cả bài toán mô hình hoá hiểu biết hội thoại đa phương thức và độ vững của quá trình học trong các điều kiện dữ liệu thực tế chất lượng thấp. Thay vì xem xét từng thách thức một cách tách biệt, luận án xây dựng một khung nghiên cứu tiến triển, kết nối kết hợp đa phương thức, mô hình hoá ngữ cảnh hội thoại và động lực học học máy vững chắc trong một phạm vi phương pháp luận thống nhất.

Tóm tắt các đóng góp. Để đạt được **Mục tiêu O1**, luận án nghiên cứu cách cần mô hình hoá biểu diễn đa phương thức và ngữ cảnh hội thoại để nhận diện cảm xúc hiệu quả trong đối thoại. Cụ thể, CORECT giải quyết RQ1 bằng cách đưa ra một khung đồ thị quan hệ–thời gian cho phép kết hợp đa phương thức có nhận thức ngữ cảnh trong khi vẫn giữ được các biểu diễn đặc trưng riêng của từng phương thức. Trên nền tảng đó, MultiDAG+CL giải quyết RQ2 bằng cách đi sâu vào cách học các phụ thuộc hội thoại theo thời gian và phụ thuộc người nói dưới các mức độ phức tạp hội thoại khác nhau, thông qua mô hình đồ thị có hướng không chu trình và tối ưu hoá dựa trên học theo giáo trình. Kết hợp lại, các phương pháp này thiết lập một cách tiếp cận có cơ sở cho kết hợp đa phương thức và suy luận ngữ cảnh trong nhận diện cảm xúc hội thoại.

Để hoàn thành **Mục tiêu O2**, luận án tiếp tục mở rộng nghiên cứu sang MERC trong các điều kiện dữ liệu thực tế chất lượng thấp. Mi-CGA giải quyết RQ3 bằng cách cho phép kết hợp đa phương thức vững hơn khi một hoặc nhiều phương thức bị thiếu, thông qua truyền thông tin dựa trên đồ thị và suy luận chéo giữa các phương thức. Bên cạnh đó, Ada2I và SPCL giải quyết RQ4 bằng cách điều chỉnh mất cân bằng phương thức từ các góc nhìn bổ trợ: Ada2I tập trung tái cân bằng đóng góp của các phương thức ở mức biểu diễn, trong khi SPCL ổn định động lực học học thông qua tối ưu hoá dựa trên giáo trình. Nhìn chung, các phương pháp này tạo nên một khung tiếp cận nhất quán để cải thiện độ vững và độ ổn định của mô hình MERC trong các bối cảnh hội thoại thực tế.

Hạn chế. Mặc dù các phương pháp đề xuất cho thấy hiệu quả rõ rệt, vẫn còn một số hạn chế. Thứ nhất, nhiều phương pháp dựa trên các giả định sẵn có về cấu trúc hội thoại hoặc tính sẵn sàng của phương thức, điều này có thể hạn chế tính linh hoạt trong các kịch bản hội thoại tự phát, nhiễu hoặc miền mở. Thứ hai, các chiến lược xây dựng đồ thị và thiết kế giáo trình còn phụ thuộc vào các tiêu chí mang tính kinh nghiệm hoặc đặc thù bộ dữ liệu, và hiệu quả của chúng có thể thay đổi giữa các miền ứng dụng. Thứ ba, các đánh giá thực nghiệm chủ yếu được thực hiện trên các bộ dữ liệu chuẩn, do đó cần thêm các kiểm chứng trong những hệ thống hội thoại thực tế quy mô lớn để đánh giá khả năng mở rộng và tính khả thi khi triển khai.

Hướng nghiên cứu tương lai. Các hạn chế trên gợi mở một số hướng nghiên cứu đầy tiềm năng trong tương lai. Những mở rộng khả dĩ bao gồm các chiến lược xây dựng đồ thị và giáo trình thích ứng, có khả năng phản ứng động trước độ phức tạp của hội thoại; sự tích hợp chặt chẽ hơn giữa các cơ chế kết hợp và các cơ chế tăng cường độ vững trong các khung học thống nhất; và việc mở rộng MERC sang các bài toán trí tuệ hội thoại rộng hơn như hệ thống đối thoại giàu tính đồng cảm hoặc tương tác người–AI. Nghiên cứu các mô hình MERC có khả năng mở rộng và đáng tin cậy cho triển khai thực tế cũng sẽ là một hướng quan trọng trong tương lai.

Tổng thể, luận án góp phần nâng cao hiểu biết về MERC bằng cách đưa ra một khung nghiên cứu thống nhất và tiến triển, kết nối kết hợp đa phương thức, mô hình hoá hội thoại và độ vững dưới các điều kiện dữ liệu chất lượng thấp. Các phương pháp được đề xuất hướng tới việc phát triển các hệ thống nhận diện cảm xúc đa phương thức tin cậy, nhận thức ngữ cảnh và có khả năng ứng dụng thực tiễn cao.

Danh sách công bố khoa học

- [**VanNTC 1**] "Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 15154–15167, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.937 – **CORE Rank A* Conference**

- [**VanNTC 2**] "Curriculum Learning Meets Directed Acyclic Graph for Multimodal Emotion Recognition." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4259–4265, Torino, Italy. ELRA and ICCL. – **CORE Rank B Conference**

- [**VanNTC 3**] "Mi-CGA: Cross-modal Graph Attention Network for Robust Emotion Recognition in the Presence of Incomplete Modalities". *Neurocomputing*, 623: 129342. https://doi.org/10.1016/j.neucom.2025.129342 – **SCIE Q1 Journal, Impact Factor: 6.5**

- [**VanNTC 4**] "Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition". In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM 2024)*, pages 9330–9339. https://doi.org/10.1145/3664647.3681648 – **CORE Rank A* Conference**

- [**VanNTC 5**] "Leveraging Self-Paced Curriculum Learning for Enhanced Modality Balance in Multimodal Conversational Emotion Recognition". *Neural Computing and Applications*. https://doi.org/10.1007/s00521-026-12160-6 – **Scopus Q1 Journal**