KẾT HỢP BIỂU DIỄN THỊ GIÁC ĐA BACKBONE VÀ NGỮ NGHĨA MỞ RỘNG TỪ MÔ HÌNH NGÔN NGỮ LỚN TRONG CHÚ THÍCH ẢNH

Nguyễn Trung Tính, Nguyễn Văn Thịnh*

Khoa Công nghệ thông tin, Trường Đại học Sư phạm TP. Hồ Chí Minh

4801104136@student.hcmue.edu.vn, thinhnv@hcmue.edu.vn

TÓM TẮT— Chú thích ảnh tự động vẫn là một bài toán khó do mô hình phải khai thác đồng thời thông tin thị giác và ngữ nghĩa để sinh ra các câu mô tả chính xác, tự nhiên. Bài báo này đề xuất một phương pháp chú thích ảnh kết hợp biểu diễn thị giác đa backbone với ngữ nghĩa mở rộng từ mô hình ngôn ngữ lớn (LLM). Ở nhánh thị giác, ảnh đầu vào được mã hóa bởi nhiều backbone nhằm tận dụng tính bổ sung giữa đặc trưng cục bộ và bối cảnh toàn cục. Ở nhánh ngữ nghĩa, ngữ cảnh thị giác cấu trúc gồm đối tượng, tương tác người–vật và quan hệ giữa các đối tượng được sử dụng để dẫn hướng LLM sinh ra các mô tả ngữ nghĩa mức câu. Các mô tả này được đánh giá bằng BLIP Image–Text Matching, sau đó được mã hóa thành biểu diễn liên tục và hợp nhất với nhánh thị giác thông qua cross-attention trước khi đưa vào bộ giải mã Transformer để sinh chú thích cuối cùng. Thực nghiệm trên tập MS COCO đạt 74.9 BLEU-1, 31.8 BLEU-4, 25.8 METEOR, 56.8 ROUGE-L và 101.9 CIDEr. Kết quả cho thấy sự kết hợp giữa đặc trưng thị giác đa backbone và thông tin ngữ nghĩa từ LLM giúp mô hình sinh ra các chú thích chính xác hơn, giàu ngữ cảnh hơn và phù hợp hơn với nội dung ảnh.

Từ khóa— Chú thích ảnh, biểu diễn thị giác đa backbone, mở rộng ngữ nghĩa ảnh bằng LLM, chú ý chéo, Transformer.

I. GIỚI THIỆU

Chú thích ảnh là bài toán kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên, mục tiêu là sinh ra các câu mô tả bằng ngôn ngữ tự nhiên phản ánh đúng nội dung của ảnh đầu vào [1]. Bài toán này có ý nghĩa trong nhiều ứng dụng thực tiễn như hỗ trợ người khiếm thị tiếp cận và hiểu nội dung hình ảnh [2, 3], chẩn đoán bệnh trên cây trồng [4], kiểm tra an toàn công trình xây dựng [5] và hỗ trợ kiểm duyệt nội dung [6]. Trong các ứng dụng này, mô hình chú thích ảnh không chỉ cần nhận diện đúng đối tượng mà còn phải mô tả được bối cảnh, hành động và quan hệ giữa các đối tượng trong ảnh.

Trong những năm gần đây, các mô hình chú thích ảnh dựa trên kiến trúc encoder–decoder đã đạt được nhiều tiến bộ đáng kể [7, 8]. Ở các phương pháp thế hệ đầu, ảnh thường được mã hóa bằng mạng nơ-ron tích chập (CNN), còn phần ngôn ngữ được sinh bởi RNN hoặc LSTM [9, 10]. Tuy nhiên, các mô hình này vẫn bộc lộ những hạn chế khi xử lý các cảnh phức tạp hoặc ngữ cảnh đa dạng, do phụ thuộc mạnh vào đặc trưng cục bộ từ CNN và khả năng mô hình hóa phụ thuộc dài hạn còn hạn chế của RNN/LSTM [11]. Sự ra đời của Transformer và các biến thể thị giác như Vision Transformer (ViT) [12] hay Swin Transformer [13] đã cải thiện đáng kể khả năng mô hình hóa phụ thuộc dài hạn và quan hệ toàn cục giữa các vùng ảnh. Đồng thời, Transformer decoder cho phép khai thác hiệu quả cơ chế self-attention và cross-attention để kết hợp ngữ cảnh ngôn ngữ với thông tin thị giác trong quá trình sinh chú thích, qua đó mở ra hướng phát triển mới cho các hệ thống chú thích ảnh hiện đại [3, 14].

Mặc dù vậy, phần lớn các mô hình hiện nay vẫn chủ yếu khai thác đặc trưng thị giác từ một backbone đơn lẻ hoặc tập trung vào đặc trưng vùng ảnh mà chưa tận dụng đầy đủ các nguồn ngữ nghĩa cấp cao [15-17]. Một số nghiên cứu đã tích hợp đồ thị cảnh, đồ thị quan hệ hoặc tri thức ngoài như ConceptNet nhằm bổ sung ngữ cảnh cho quá trình sinh chú thích [18-20]. Tuy nhiên, các biểu diễn đó thường hoặc thiên về quan hệ hình học và thuộc tính bề mặt, hoặc đòi hỏi quy trình tiền xử lý phức tạp. Bên cạnh đó, một số công trình gần đây bắt đầu đưa mô hình ngôn ngữ lớn vào bài toán chú thích ảnh, nhưng nếu để mô hình ngôn ngữ lớn tham gia trực tiếp trong quá trình suy luận chính thì chi phí tính toán lớn và việc tích hợp sâu vào kiến trúc mã hóa–giải mã chuyên biệt cho bài toán chú thích ảnh vẫn còn nhiều khó khăn [21, 22].

Xuất phát từ những khoảng trống trên, bài báo này đề xuất một phương pháp chú thích ảnh theo hướng kết hợp biểu diễn đặc trưng thị giác đa backbone với biểu diễn ngữ nghĩa mở rộng từ mô hình ngôn ngữ lớn. Cụ thể, ở nhánh thị giác, nhiều backbone được khai thác đồng thời để tạo ra biểu diễn có tính bổ sung cao hơn so với cách tiếp cận đơn backbone. Ở nhánh ngữ nghĩa, ngữ cảnh ảnh được tổng hợp từ ba nguồn tri thức thị giác gồm đối tượng, tương tác người–vật và quan hệ giữa các đối tượng, sau đó được đưa vào LLM (Qwen3-4B [23]) để sinh ra các mô tả ngữ nghĩa mức câu. Các mô tả này tiếp tục được BLIP ITM chấm điểm mức độ phù hợp ảnh–văn bản trước khi được mã hóa thành biểu diễn ngữ nghĩa. Hai nguồn biểu diễn sau đó được tích hợp thông qua khối hợp nhất thị giác–ngữ nghĩa dựa trên cross-attention, rồi được sử dụng bởi bộ giải mã Transformer để sinh chú thích cuối cùng.

Đóng góp chính của bài báo có thể được tóm tắt như sau:

* Corresponding Author

- Đưa ra một khung mã hóa thị giác đa backbone cho chú thích ảnh, cho phép khai thác tính bổ sung giữa các bộ trích xuất đặc trưng khác nhau nhằm tạo ra biểu diễn thị giác phong phú và toàn diện hơn.
- Đề xuất một chiến lược mở rộng ngữ nghĩa ảnh dựa trên LLM, trong đó các mô tả ngữ nghĩa mức câu được sinh từ ngữ cảnh ảnh và được sàng lọc bằng BLIP-ITM trước khi tích hợp vào mô hình, nhằm bảo đảm độ phù hợp ngữ nghĩa giữa ảnh và thông tin bổ sung.
- Thiết kế một cơ chế hợp nhất thị giác–ngữ nghĩa có chọn lọc dựa trên cross-attention nhằm tích hợp hiệu quả đặc trưng thị giác và biểu diễn ngữ nghĩa sinh bởi LLM, qua đó nâng cao chất lượng biểu diễn đầu vào cho bộ giải mã Transformer, từ đó cải thiện khả năng sinh chú thích chính xác và giàu ngữ nghĩa hơn.
- Kết quả thực nghiệm trên bộ dữ liệu MS COCO cho thấy phương pháp đề xuất đạt hiệu quả tốt, qua đó khẳng định tính khả thi và hiệu quả của việc kết hợp đặc trưng thị giác đa backbone, ngữ nghĩa mở rộng từ LLM và cơ chế hợp nhất chéo đa phương thức trong bài toán chú thích ảnh.

Phần còn lại của bài báo được tổ chức như sau. Mục II trình bày các công trình liên quan. Mục III mô tả chi tiết mô hình đề xuất. Mục IV trình bày thiết lập thực nghiệm, kết quả và bàn luận. Cuối cùng, Mục V đưa ra kết luận và định hướng nghiên cứu tiếp theo.

II. CÔNG TRÌNH LIÊN QUAN

Phần lớn các nghiên cứu hiện nay về chú thích ảnh vẫn được xây dựng trên khung encoder–decoder, trong đó ảnh được mã hóa thành biểu diễn trung gian và sau đó được giải mã thành chuỗi từ [11]. Sự khác biệt giữa các hướng tiếp cận chủ yếu nằm ở thành phần được tăng cường trong khung này. Nhìn chung, các công trình gần đây có thể được phân thành ba hướng chính: (i) nâng cao chất lượng biểu diễn thị giác ở bộ mã hóa, (ii) bổ sung ngữ nghĩa trung gian thông qua quan hệ giữa các đối tượng hoặc tri thức ngoài, và (iii) khai thác mô hình ngôn ngữ lớn hoặc mô hình ngôn ngữ–thị giác để tăng cường khả năng diễn đạt ngữ nghĩa trong quá trình sinh mô tả [7].

Một hướng nghiên cứu quan trọng tập trung vào việc nâng cao chất lượng biểu diễn thị giác trong khung encoder–decoder. Các công trình như Show and Tell [9] và Show, Attend and Tell [10] đặt nền móng cho hướng tiếp cận mã hóa ảnh rồi giải mã thành câu mô tả. Các biến thể gần đây hơn như Bi-LS-AttM [15], Topic FRCNN-LSTM [17] và Xception-YOLO Attention [16] tiếp tục cải thiện chất lượng mô tả thông qua attention, đặc trưng vùng ảnh hoặc thông tin đối tượng. Khi Transformer được sử dụng trong chú thích ảnh, các backbone như ViT [12], Swin Transformer [13] và các mô hình như Swin-Trans-Cap [3] cho thấy khả năng mô hình hóa ngữ cảnh toàn cục tốt hơn so với CNN truyền thống. Bên cạnh đó, các mô hình như Meshed-Memory Transformer [24] và Context-Aware Transformer [25] cho thấy Transformer decoder có thể tích hợp hiệu quả ngữ cảnh thị giác và ngữ cảnh ngôn ngữ trong quá trình sinh chú thích. Tuy nhiên, phần lớn các phương pháp trong nhóm này vẫn chủ yếu dựa trên một backbone chính hoặc một nguồn đặc trưng thị giác chi phối, nên chưa khai thác đầy đủ tính bổ sung giữa đặc trưng cục bộ và bối cảnh toàn cục.

Song song với hướng trên, nhiều nghiên cứu đã tìm cách bổ sung ngữ nghĩa trung gian vào khung encoder–decoder thông qua quan hệ giữa các đối tượng hoặc tri thức bên ngoài tập dữ liệu. Các hướng tiếp cận dựa trên đồ thị cảnh, đồ thị quan hệ hoặc tri thức thường thức cho thấy việc mô hình hóa cấu trúc cảnh có thể cải thiện đáng kể khả năng mô tả. Trong nhóm này, OD-VR-Cap [18], GIC-SSF [26] khai thác quan hệ giữa các đối tượng để xây dựng đồ thị phục vụ sinh chú thích, trong khi RGTranCNet [19] và SemanticGuideAtt [27] tích hợp thêm tri thức ngữ nghĩa từ ConceptNet vào mô hình Transformer nhằm cải thiện khả năng diễn đạt. Một số nghiên cứu khác cũng cho thấy việc đưa tri thức ngoài tập dữ liệu vào mô hình có thể nâng cao chất lượng và tính đa dạng của câu mô tả [21, 28]. Tuy nhiên, các biểu diễn ngữ nghĩa trong các công trình này vẫn chủ yếu ở dạng nhãn, thuộc tính, bộ ba quan hệ hoặc đồ thị, tức là còn mang tính rời rạc và chưa thực sự gần với không gian ngôn ngữ tự nhiên mà bộ giải mã cần thao tác trực tiếp.

Gần đây, sự phát triển của mô hình ngôn ngữ lớn và mô hình ngôn ngữ–thị giác đã mở ra khả năng tăng cường ngữ nghĩa ở mức cao hơn trong khung encoder–decoder. Các mô hình như BLIP, BLIP-2 và các biến thể kết hợp CLIP–GPT [22] cho thấy LLM có thể đóng vai trò quan trọng trong việc nâng cao khả năng diễn đạt ngữ nghĩa của mô hình. Tuy nhiên, nhiều phương pháp trong nhóm này hoặc sử dụng LLM như một bộ sinh mô tả trực tiếp, hoặc dựa trên các quy trình tiền huấn luyện quy mô lớn với chi phí tính toán cao. Trong khi đó, hướng sử dụng LLM như một thành phần suy diễn ngữ nghĩa trung gian để sinh ra nguồn thông tin bổ trợ cho mô hình chú thích ảnh, thay vì thay thế trực tiếp bộ giải mã, vẫn chưa được khai thác một cách tường minh và hiệu quả.

Từ các thách thức trên có thể thấy rằng, mặc dù khung encoder–decoder trong chú thích ảnh đã được cải thiện đáng kể ở cả nhánh mã hóa thị giác lẫn tăng cường ngữ nghĩa, vẫn còn ba hạn chế đáng chú ý. Thứ nhất, việc khai thác đồng thời nhiều backbone để tạo ra biểu diễn thị giác có tính bổ sung cao vẫn còn hạn chế. Thứ hai, các phương pháp tăng cường ngữ nghĩa trước đây chủ yếu dựa trên nhãn, thuộc tính, đồ thị hoặc tri thức tĩnh, trong

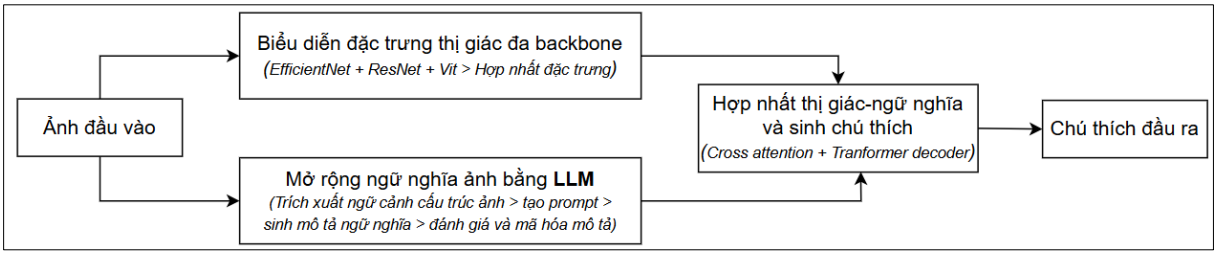
khi việc sử dụng mô tả ngữ nghĩa mức câu gần với không gian ngôn ngữ tự nhiên mà bộ giải mã cần thao tác vẫn chưa được khai thác đầy đủ. Thứ ba, dù mô hình ngôn ngữ lớn đã cho thấy tiềm năng mạnh mẽ, hướng sử dụng LLM như một thành phần trung gian sinh thông tin ngữ nghĩa bổ trợ cho khung encoder–decoder, thay vì thay thế trực tiếp mô hình chú thích ảnh vẫn còn tương đối mở. Những khoảng trống này tạo cơ sở cho việc phát triển các phương pháp kết hợp biểu diễn thị giác đa backbone với thông tin ngữ nghĩa mức cao trong chú thích ảnh.

III. PHƯƠNG PHÁP CHÚ THÍCH ẢNH ĐỀ XUẤT

Mục này trình bày chi tiết phương pháp chú thích ảnh đề xuất, từ kiến trúc tổng quan đến các thành phần chính trong quá trình mã hóa thị giác, mở rộng ngữ nghĩa bằng mô hình ngôn ngữ lớn, và hợp nhất đa nguồn thông tin để sinh câu chú thích cuối cùng.

A. KIẾN TRÚC TỔNG QUAN

Hình 1 mô tả kiến trúc tổng quan của phương pháp chú thích ảnh đề xuất. Mô hình được tổ chức thành ba phần chính. Phần thứ nhất là biểu diễn đặc trưng thị giác đa backbone, trong đó ảnh đầu vào được đưa qua nhiều backbone khác nhau nhằm khai thác đồng thời đặc trưng cục bộ và ngữ cảnh toàn cục. Phần thứ hai là mở rộng ngữ nghĩa ảnh bằng LLM, trong đó ngữ cảnh ảnh được chuyển thành các mô tả ngữ nghĩa mức câu để bổ sung thông tin diễn giải ở mức cao hơn cho mô hình. Phần thứ ba là hợp nhất thị giác–ngữ nghĩa và sinh chú thích, trong đó hai nguồn biểu diễn được tích hợp thông qua cơ chế cross-attention trước khi được sử dụng bởi bộ giải mã Transformer để sinh câu mô tả cuối cùng.



Hình 1. *Kiến trúc tổng quan của phương pháp đề xuất gồm ba phần chính: biểu diễn đặc trưng thị giác đa backbone, biểu diễn ngữ nghĩa mở rộng từ mô hình ngôn ngữ lớn, và hợp nhất thị giác–ngữ nghĩa để sinh chú thích.*

Cho ảnh đầu vào I mô hình sinh ra câu chú thích $Y = (y_1, y_2, \dots, y_{N_T})$, trong đó N_T là độ dài của chuỗi đầu ra. Gọi V_I là biểu diễn đặc trưng thị giác đa backbone và T_I là biểu diễn ngữ nghĩa mở rộng từ LLM của ảnh I. Khi đó, xác suất sinh toàn bộ câu chú thích được mô hình hóa như sau:

$$P(Y \mid I) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, V_I, T_I),$$

trong đó $y_{<t}$ là chuỗi các từ đã sinh trước thời điểm t. Biểu thức trên cho thấy quá trình sinh từ tại mỗi bước phụ thuộc đồng thời vào ngữ cảnh ngôn ngữ, thông tin thị giác của ảnh và biểu diễn ngữ nghĩa mở rộng.

Khác với các cách tiếp cận chỉ khai thác một nguồn đặc trưng thị giác hoặc chỉ dựa trên thông tin trực quan thuần túy, phương pháp đề xuất tận dụng đồng thời hai nguồn thông tin có tính bổ sung. Nếu phần biểu diễn thị giác cung cấp nền tảng trực quan của ảnh, thì phần biểu diễn ngữ nghĩa mở rộng từ LLM bổ sung thông tin diễn giải ở mức cao hơn. Trên cơ sở đó, phần hợp nhất thị giác–ngữ nghĩa đóng vai trò liên kết hai không gian biểu diễn trước khi chuyển chúng thành câu mô tả bằng bộ giải mã Transformer.

B. BIỂU DIỄN ĐẶC TRƯNG THỊ GIÁC ĐA BACKBONE

Trong bài toán chú thích ảnh, chất lượng của biểu diễn thị giác ảnh hưởng trực tiếp đến khả năng sinh mô tả của mô hình. Nếu đặc trưng ảnh không phản ánh đầy đủ nội dung trực quan, bộ giải mã khó có thể tạo ra các câu chú thích chính xác và tự nhiên. Tuy nhiên, mỗi backbone thường chỉ mạnh ở một số khía cạnh nhất định: các kiến trúc dựa trên mạng tích chập có ưu thế trong việc khai thác đặc trưng cục bộ và cấu trúc không gian ổn định, trong khi các mô hình thị giác dựa trên Transformer hiệu quả hơn trong việc nắm bắt bối cảnh toàn cục và các phụ thuộc xa. Vì vậy, việc chỉ sử dụng một backbone đơn lẻ có thể làm hạn chế khả năng biểu diễn đầy đủ nội dung ảnh, đặc biệt trong các cảnh phức tạp.

Để tận dụng tính bổ sung giữa đặc trưng cục bộ và ngữ cảnh toàn cục, phương pháp đề xuất xây dựng biểu diễn thị giác đa backbone thông qua ba nhánh trích xuất đặc trưng song song. Với ảnh đầu vào $I \in \mathbb{R}^{3\times256\times256}$, **EfficientNet-B0** tạo đặc trưng $F^{(eff)} \in \mathbb{R}^{64\times1280}$, **ResNet-101** tạo đặc trưng $F^{(res)} \in \mathbb{R}^{64\times2048}$, trong khi **ViT-Base** nhận ảnh đã được nội suy về kích thước 224×224 và chuẩn hóa để sinh đặc trưng $F^{(vit)} \in \mathbb{R}^{64\times768}$ từ 64 patch token đầu tiên của $last_hidden_state$ sau khi loại bỏ token $[CLS]$.

Ba chuỗi đặc trưng này được nối theo chiều cuối để tạo thành biểu diễn thị giác đa nguồn:

$$F_I = \left[F^{(eff)}; F^{(res)}; F^{(vit)}\right] \in \mathbb{R}^{64\times 4096}$$

Cách kết hợp này cho phép mô hình đồng thời khai thác các đặc trưng cục bộ chi tiết từ hai backbone CNN và đặc trưng ngữ cảnh toàn cục từ backbone Transformer. Tiếp theo, F_I được chiếu vào không gian nhúng chung có chiều $d_{model} = 768$ thông qua một mạng chiếu tuyến tính:

$$F_I = \text{LayerNorm}\left(\text{Linear}\left(\text{ReLU}(\text{Linear}(\text{LayerNorm}(F_I)))\right)\right) \in \mathbb{R}^{64\times d_{model}}$$

Mạng này gồm hai tầng tuyến tính theo ánh xạ $4096 \to 2048 \to d_{model}$, xen kẽ hàm kích hoạt $ReLU$, $dropout$ với xác suất 0.1 và chuẩn hóa $LayerNorm$. Bước chiếu này giúp đưa ba nguồn đặc trưng thị giác về cùng một không gian biểu diễn thống nhất trước khi tích hợp vào mô hình chú thích ảnh.

Sau bước chiếu, ba cơ chế mã hóa vị trí hai chiều được áp dụng đồng thời trên các phần chiều khác nhau của F_I: RoPE 2D trên 256 chiều đầu, Learned 2D Positional Encoding trên 256 chiều tiếp theo, và Sinusoidal 2D Positional Encoding trên $d_{model} - 512$ còn lại. Các thành phần sau mã hóa vị trí được nối lại để tạo thành biểu diễn thị giác cuối cùng:

$$V_I \in \mathbb{R}^{64\times d_{model}}$$

Về bản chất, V_I không chỉ là kết quả của phép ghép cơ học nhiều đặc trưng khác nhau, mà là đầu ra của một quá trình chuẩn hóa và hợp nhất nhằm đưa các đặc trưng trực quan về một không gian biểu diễn chung có thể khai thác hiệu quả trong các bước sau. Nhờ đó, mô hình có thể đồng thời tận dụng các đặc trưng cục bộ như biên, kết cấu, hình dạng và các đặc trưng toàn cục như bố cục cảnh, bối cảnh và quan hệ không gian giữa các đối tượng. Đây là một lợi thế quan trọng đối với bài toán chú thích ảnh, bởi một câu mô tả tốt không chỉ cần đúng về mặt nhận dạng đối tượng mà còn cần phản ánh được bối cảnh và cấu trúc của cảnh.

C. MỞ RỘNG NGỮ NGHĨA ẢNH BẰNG LLM

Phần này trình bày quy trình mở rộng ngữ nghĩa ảnh bằng LLM, bao gồm xây dựng ngữ cảnh đầu vào, sinh mô tả ngữ nghĩa, đánh giá chất lượng mô tả và mã hóa chúng thành biểu diễn liên tục.

1. XÂY DỰNG NGỮ CẢNH ĐẦU VÀO CHO LLM

Để định hướng quá trình sinh ngữ nghĩa, trước hết mô hình xây dựng một ngữ cảnh thị giác cấu trúc từ ba nguồn tri thức bổ sung lẫn nhau, gồm đối tượng xuất hiện trong ảnh, tương tác người–vật, và quan hệ giữa các đối tượng. Với ảnh đầu vào I, ngữ cảnh này được ký hiệu là:

$$\mathcal{C}(I) = \{\mathcal{O}(I), \mathcal{H}(I), \mathcal{R}(I)\},$$

trong đó $\mathcal{O}(I)$, $\mathcal{H}(I)$ và $\mathcal{R}(I)$ lần lượt là tập đối tượng, tương tác người–vật và quan hệ giữa các đối tượng. Cụ thể, $\mathcal{O}(I)$ được trích xuất bằng YOLOv12 [29], một bộ phát hiện đối tượng thời gian thực theo hướng attention-centric; $\mathcal{H}(I)$ được thu nhận bằng HOICLIP [30] dưới dạng các bộ ba $< human, action, object >$ cho phép mô tả các tương tác giữa người và vật; còn $\mathcal{R}(I)$ được suy ra bằng RelTR [31] dưới dạng các bộ ba $< subject, predicate, object >$, nhằm biểu diễn quan hệ ngữ nghĩa và không gian giữa các đối tượng trong ảnh.

Từ ngữ cảnh $\mathcal{C}(I)$, một prompt văn bản có cấu trúc được xây dựng để đưa vào LLM, trong đó ba thành phần $Objects$ $(\mathcal{O})$, $Actions(\mathcal{H})$ và $Relations$ $(\mathcal{R})$ được liệt kê tường minh nhằm dẫn hướng mô hình sinh ra các mô tả ngữ nghĩa mức câu, kèm chỉ thị yêu cầu mô hình sinh đúng K mô tả ngữ nghĩa, mỗi mô tả dài tối thiểu N từ, bảo đảm tính đa dạng và bám sát nội dung ảnh. Cách tổ chức này giúp LLM không hoạt động như một bộ sinh chú thích độc lập, mà như một thành phần suy diễn ngữ nghĩa trung gian được ràng buộc bởi bằng chứng thị giác đã trích xuất trước đó. Nhờ vậy, các mô tả sinh ra có xu hướng bám sát hơn với nội dung ảnh, đồng thời phản ánh tốt hơn các yếu tố bối cảnh, hành động và quan hệ giữa các đối tượng.

Để minh họa cách xây dựng ngữ cảnh đầu vào cho LLM, ví dụ dưới đây trình bày một prompt được tạo từ ngữ cảnh thị giác của một ảnh cụ thể trong tập dữ liệu huấn luyện.

```
For image below, generate exactly 5 semantic descriptions.
Each description must contain more than 5 words, be specific, diverse,
and faithful to the visual context.
Objects (YOLO): person 3, umbrella 2, car 1
Actions (HOI): human - hold - umbrella; human - stand under - umbrella
Relations (RelTR): woman - holding - umbrella; man - holding - umbrella;
man - wearing - shirt; person - holding - umbrella; woman - wearing - shirt
OUTPUT FORMAT:
description 1 | description 2 | description 3 | description 4 | description
```

Prompt trong ví dụ trên cho phép LLM suy diễn các mô tả ở mức ngữ nghĩa cao hơn từ ngữ cảnh thị giác đã cấu trúc hóa, chẳng hạn bối cảnh thời tiết, không gian đường phố hoặc bố cục cảnh, trong khi vẫn duy trì sự ràng buộc

với nội dung trực quan của ảnh. Nhờ đó, các mô tả sinh ra vừa giàu thông tin hơn so với các nhãn đầu vào, vừa phù hợp hơn với vai trò nguồn thông tin ngữ nghĩa bổ trợ cho mô hình chú thích ảnh.

2. SINH MÔ TẢ NGỮ NGHĨA

Từ prompt được xây dựng từ ngữ cảnh $\mathcal{C}(I)$, mô hình ngôn ngữ lớn sinh ra một tập mô tả ngữ nghĩa:

$$\mathcal{S}(I) = \{s_1, s_2, \dots, s_K\},$$

trong đó K là số lượng mô tả và mỗi mô tả s_i có ít nhất N từ. Điều kiện độ dài tối thiểu này nhằm bảo đảm mô tả đủ giàu thông tin, tránh các phát biểu quá ngắn hoặc quá rời rạc. Các mô tả này không được dùng như chú thích cuối cùng mà đóng vai trò là tri thức ngữ nghĩa mở rộng cho ảnh đầu vào.

Mỗi mô tả phản ánh một khía cạnh quan trọng của ảnh, chẳng hạn đối tượng trung tâm, hành động đang diễn ra, bối cảnh chung hoặc mối liên hệ nổi bật giữa các đối tượng. So với nhãn hoặc thuộc tính rời rạc, mô tả ở mức câu có ưu thế ở chỗ bảo toàn được cấu trúc ngôn ngữ tự nhiên và biểu đạt tốt hơn các yếu tố ngữ cảnh tiềm ẩn. Vì vậy, việc sinh các mô tả ngữ nghĩa từ mô hình ngôn ngữ lớn không chỉ nhằm bổ sung thêm thông tin, mà còn nhằm làm giàu không gian biểu diễn theo hướng phù hợp hơn với quá trình sinh ngôn ngữ.

Đối với prompt minh họa ở Mục C.1, từ cùng một ngữ cảnh thị giác gồm các đối tượng "*person*", "*umbrella*", "*car*", các tương tác "*hold umbrella*", "*stand under umbrella*" và các quan hệ như "*woman holding umbrella*" hay "*man holding umbrella*", LLM có thể sinh ra các mô tả như "*a rainy day scene with two people holding umbrellas and a car in the background*" hoặc "*a street scene featuring two individuals holding umbrellas and a car parked nearby*". Các mô tả này cho thấy LLM không chỉ lặp lại thông tin đầu vào, mà còn tổng hợp chúng để suy ra thông tin ngữ nghĩa bậc cao hơn, như bối cảnh thời tiết, không gian đường phố và bố cục cảnh.

Vì vậy, bước sinh mô tả ngữ nghĩa không chỉ nhằm bổ sung thêm thông tin cho ảnh, mà còn làm giàu biểu diễn trung gian theo hướng phù hợp hơn với quá trình sinh ngôn ngữ. Tập mô tả $\mathcal{S}(I)$ sau đó được đưa vào bước đánh giá chất lượng bằng BLIP ITM trước khi được mã hóa thành biểu diễn ngữ nghĩa liên tục để hợp nhất với nhánh thị giác.

3. ĐÁNH GIÁ CHẤT LƯỢNG MÔ TẢ BẰNG BLIP ITM

Các mô tả ngữ nghĩa do LLM sinh ra có thể không đồng đều về chất lượng. Nguyên nhân là LLM không trực tiếp quan sát ảnh, mà chỉ suy diễn từ ngữ cảnh văn bản được xây dựng từ các thông tin thị giác đầu vào. Vì vậy, trước khi đưa các mô tả này vào bước mã hóa ngữ nghĩa, cần có một cơ chế đánh giá khách quan để đo mức độ tương thích giữa từng mô tả và ảnh gốc. Trong nghiên cứu này, nhiệm vụ đó được thực hiện bằng BLIP với nhánh Image–Text Matching (ITM). Cụ thể, phiên bản **BLIP-ITM-Base-COCO**, đã được huấn luyện trước trên tập dữ liệu MS COCO, nhận đầu vào là một cặp gồm ảnh I và câu mô tả s_i, sinh ra điểm số tương thích hình ảnh-văn bản:

$$\alpha_i = P(match \mid I, s_i),$$

trong đó $\alpha_i \in [0,1]$ phản ánh mức độ phù hợp giữa ảnh I và mô tả s_i. Điểm số này được sử dụng như một trọng số ngữ nghĩa ở bước tiếp theo: các mô tả có α_i cao sẽ đóng góp nhiều hơn vào biểu diễn ngữ nghĩa, trong khi ảnh hưởng của các mô tả kém phù hợp được giảm một cách tự nhiên. Nhờ đó, BLIP ITM giúp giảm nhiễu ngữ nghĩa và tăng độ tin cậy của nguồn thông tin bổ trợ do mô hình ngôn ngữ lớn cung cấp.

Đối với ví dụ minh họa ở Mục C.1–C.2, các mô tả sinh ra từ cùng một ngữ cảnh thị giác nhận các điểm ITM khác nhau. Chẳng hạn, mô tả *a rainy day scene with two people holding umbrellas and a car in the background* đạt điểm cao hơn các mô tả còn lại, cho thấy nội dung của câu này phù hợp hơn với ảnh gốc. Kết quả đó cho thấy BLIP ITM không chỉ đóng vai trò sàng lọc, mà còn cung cấp một cơ chế định lượng để ưu tiên những mô tả vừa giàu ngữ nghĩa vừa bám sát bằng chứng thị giác.

Vì vậy, bước đánh giá bằng BLIP ITM giữ vai trò trung gian quan trọng giữa quá trình sinh mô tả của LLM và quá trình mã hóa ngữ nghĩa sau đó. Thay vì sử dụng trực tiếp toàn bộ các mô tả do LLM sinh ra với cùng mức đóng góp, phương pháp đề xuất gán trọng số cho từng mô tả theo độ phù hợp thực tế với ảnh, qua đó làm cho biểu diễn ngữ nghĩa mở rộng trở nên tin cậy và ổn định hơn trước khi hợp nhất với nhánh thị giác.

4. MÃ HÓA MÔ TẢ NGỮ NGHĨA THÀNH BIỂU DIỄN NGỮ NGHĨA

Sau khi các câu mô tả ngữ nghĩa được sinh bởi LLM và được đánh giá chất lượng bằng BLIP ITM, bước tiếp theo là chuyển chúng sang dạng biểu diễn liên tục để có thể tích hợp với nhánh thị giác trong mô hình chú thích ảnh. Mục tiêu của bước này là biến tập các câu mô tả rời rạc thành một biểu diễn ngữ nghĩa mở rộng có cấu trúc, đồng thời bảo toàn được cả nội dung ngôn ngữ của từng câu lẫn mức độ phù hợp của câu đó với ảnh đầu vào. Cụ thể, mỗi câu mô tả s_i trước hết được đưa qua lớp nhúng dùng chung với bộ giải mã Transformer để ánh xạ các token ngôn ngữ sang không gian nhúng có chiều d_{model}:

$$E_i = Embed(s_i) \in \mathbb{R}^{L \times d_{model}},$$

trong đó L là số lượng token của câu mô tả và d_{model} là số chiều không gian nhúng. Để thu được vector đại diện cho toàn bộ câu, phương pháp áp dụng mean-pooling có mặt nạ trên chiều token:

$$\bar{e}_i = \frac{\sum_{l=1}^{L} \mathbb{1}[w_l \neq 0] \cdot E_k[l]}{\sum_{l=1}^{L} \mathbb{1}[w_l \neq 0]}$$

Việc sử dụng lớp nhúng chung với bộ giải mã giúp biểu diễn ngữ nghĩa và chuỗi chú thích nằm trong cùng một không gian biểu diễn, qua đó tạo điều kiện thuận lợi cho bước hợp nhất về sau.

Tiếp theo, véc-tơ được gán trọng số theo điểm tương thích ảnh-văn bản do BLIP ITM cung cấp:

$$\hat{e}_i = \alpha_i \bar{e}_i,$$

trong đó, α_i là điểm số đánh giá mức độ phù hợp giữa ảnh I và mô tả s_i. Cơ chế này cho phép các mô tả có độ tin cậy cao đóng góp nhiều hơn vào biểu diễn ngữ nghĩa, trong khi ảnh hưởng của các mô tả kém phù hợp được giảm một cách tự nhiên. Sau đó, véc-tơ có trọng số được chiếu sang không gian ngữ nghĩa cuối cùng thông qua một lớp tuyến tính kèm chuẩn hóa:

$$T_i = \text{LayerNorm}(\text{Linear}(\hat{e}_i)) \in \mathbb{R}^{d_{model}}$$

Cuối cùng, các vector ngữ nghĩa của toàn bộ tập mô tả được ghép lại để tạo thành biểu diễn ngữ nghĩa mở rộng:

$$T = [T_1, T_2, \dots, T_K] \in \mathbb{R}^{K \times d_{model}}$$

Biểu diễn T là đầu ra của nhánh ngữ nghĩa và sẽ được sử dụng trong bước hợp nhất với biểu diễn thị giác thông qua cơ chế cross-attention. Theo cách này, nhánh ngữ nghĩa không chỉ bảo toàn nội dung ngôn ngữ của các mô tả do LLM sinh ra, mà còn điều chỉnh mức đóng góp của từng mô tả theo độ phù hợp thực tế với ảnh đầu vào, từ đó tạo ra nguồn thông tin bổ trợ ổn định và giàu ngữ nghĩa hơn cho mô hình chú thích ảnh.

D. HỢP NHẤT THỊ GIÁC-NGỮ NGHĨA VÀ SINH CHÚ THÍCH

Sau khi thu được hai nguồn biểu diễn từ nhánh thị giác và nhánh ngữ nghĩa, phần này trình bày cơ chế hợp nhất chúng trong một không gian chung, đồng thời mô tả cách bộ giải mã Transformer khai thác biểu diễn hợp nhất đó để sinh chú thích ảnh.

1. HỢP NHẤT BIỂU DIỄN BẰNG CROSS-ATTENTION

Phương pháp đề xuất sử dụng cross-attention để cho phép biểu diễn ngữ nghĩa truy xuất có chọn lọc thông tin từ biểu diễn thị giác. Cụ thể, gọi $V \in \mathbb{R}^{N_M \times d_{model}}$ là biểu diễn thị giác và $T \in \mathbb{R}^{N_K \times d_{model}}$ là biểu diễn ngữ nghĩa thu được từ mục C.1 và C.4. Khi đó, với T đóng vai trò $Query$ và V đóng vai trò $Key/Value$, đầu ra của cơ chế hợp nhất được xác định như sau:

$$A = MultiHeadCrossAttention(Q = T, K = V, V = V),$$

trong đó $A \in \mathbb{R}^{N_K \times d_{model}}$ biểu thị phần thông tin thị giác mà từng mô tả ngữ nghĩa truy xuất được từ ảnh. Thiết kế này cho phép mỗi mô tả mức câu được "neo" vào các vùng ảnh liên quan thay vì chỉ ghép cơ học hai nguồn đặc trưng. Nói cách khác, mô hình không giả định mọi mô tả ngữ nghĩa đều liên quan như nhau đến toàn bộ ảnh, mà học cách xác định những vùng thị giác phù hợp nhất với từng biểu diễn ngữ nghĩa.

Sau bước cross-attention, biểu diễn ngữ nghĩa đã được tăng cường bởi thông tin thị giác được kết hợp với biểu diễn thị giác ban đầu để tạo thành chuỗi ngữ cảnh hợp nhất:

$$Z = [V; A] \in \mathbb{R}^{(N_M + N_K) \times d_{model}}$$

Chuỗi Z sau đó được đưa qua bộ mã hóa Transformer để thu được đầu ra ngữ cảnh:

$$H = Encoder(Z),$$

với $H \in \mathbb{R}^{(N_M + N_K) \times d_{model}}$ đóng vai trò là bộ nhớ ngữ cảnh cho bộ giải mã ở bước tiếp theo. Theo cách này, bộ nhớ đầu vào của decoder không chỉ chứa thông tin trực quan của ảnh mà còn bao gồm các biểu diễn ngữ nghĩa đã được căn chỉnh với ảnh thông qua cơ chế attention. Điều này giúp tạo ra một biểu diễn hợp nhất vừa giàu ngữ cảnh, vừa có tính chọn lọc cao, phù hợp với mục tiêu sinh chú thích chính xác và mạch lạc hơn.

2. SINH CHÚ THÍCH BẰNG BỘ GIẢI MÃ TRANSFORMER

Trên cơ sở bộ nhớ ngữ cảnh H, mô hình sinh câu chú thích bằng bộ giải mã Transformer theo cơ chế tự hồi quy. Tại mỗi thời điểm t, chuỗi các token đã sinh trước đó $y_{<t} = (y_1, y_2, \dots, y_{t-1})$ được đưa qua lớp nhúng, sau đó bộ giải mã ước lượng xác suất của từ tiếp theo:

$$P(y_t \mid y_{<t}, I) = P(y_t \mid y_{<t}, H).$$

Quá trình này được lặp lại cho đến khi mô hình sinh ra token kết thúc chuỗi, tạo thành câu chú thích hoàn chỉnh:

$$Y = \{y_1, y_2, \dots, y_{N_T}\}.$$

Xét trên toàn bộ chuỗi, xác suất sinh chú thích được mô hình hóa như sau:

$$P(Y \mid I) = \prod_{t=1}^{N_T} P(y_t \mid y_{<t}, H).$$

Biểu thức trên cho thấy việc sinh từng từ trong câu chú thích phụ thuộc đồng thời vào ngữ cảnh ngôn ngữ trước đó và bộ nhớ hợp nhất giữa thông tin thị giác với ngữ nghĩa mở rộng. Nhờ vậy, mô hình có thể khai thác đồng thời bằng chứng trực quan của ảnh và các thông tin diễn giải do LLM cung cấp trong suốt quá trình sinh câu.

Để huấn luyện mô hình, hàm mất mát sinh chú thích được xác định theo cross-entropy trên toàn bộ chuỗi:

$$\mathcal{L} = -\sum_{t=1}^{N_T} \log P\,(y_t^* \mid y_{<t}^*, H),$$

trong đó y_t^* là từ mục tiêu tại thời điểm t. Hàm mất mát này trực tiếp tối ưu khả năng sinh chú thích của mô hình trên cơ sở bộ nhớ ngữ cảnh đã tích hợp đồng thời thông tin thị giác và ngữ nghĩa. Trong giai đoạn suy luận, câu chú thích cuối cùng được sinh bằng beam search với độ rộng được thiết lập, nhằm lựa chọn chuỗi đầu ra có xác suất tổng thể cao nhất.

Tóm lại, phần hợp nhất thị giác–ngữ nghĩa và sinh chú thích giữ vai trò trung tâm trong phương pháp đề xuất. Thông qua việc sử dụng cross-attention để liên kết các mô tả ngữ nghĩa với thông tin thị giác liên quan, sau đó đưa biểu diễn hợp nhất vào bộ mã hóa và bộ giải mã Transformer, mô hình có khả năng sinh ra các câu chú thích vừa bám sát nội dung trực quan, vừa giàu ý nghĩa hơn về mặt diễn đạt.

IV. THỰC NGHIỆM

Mục này trình bày thiết lập thực nghiệm, các độ đo đánh giá, phân tích chi phí tính toán và kết quả thu được nhằm kiểm chứng hiệu quả của phương pháp đề xuất trên bộ dữ liệu chuẩn cho bài toán chú thích ảnh.

A. DỮ LIỆU VÀ THIẾT LẬP THỰC NGHIỆM

Để bảo đảm việc đánh giá mô hình được thực hiện rõ ràng và có khả năng đối sánh với các nghiên cứu trước, phần này trình bày dữ liệu sử dụng cùng các thiết lập liên quan đến mô hình, quá trình huấn luyện và suy luận.

1. DỮ LIỆU THỰC NGHIỆM

Mô hình được huấn luyện và đánh giá trên tập dữ liệu chuẩn MS COCO [32], một trong những bộ dữ liệu benchmark được sử dụng rộng rãi nhất trong bài toán chú thích ảnh. Tập dữ liệu này gồm khoảng 123.000 ảnh, mỗi ảnh đi kèm 5 câu chú thích tham chiếu do con người tạo ra. Để bảo đảm tính công bằng khi so sánh với các công trình trước, nghiên cứu sử dụng phân chia Karpathy [33], gồm 82.783 ảnh huấn luyện, 5.000 ảnh kiểm định và 5.000 ảnh kiểm tra. Đối với dữ liệu văn bản, các câu chú thích được tiền xử lý bằng tokenizer Moses; những từ xuất hiện dưới 5 lần bị loại bỏ nhằm giảm nhiễu từ vựng. Chiều dài tối đa của mỗi câu chú thích được đặt là 35 token.

2. THIẾT LẬP MÔ HÌNH VÀ HUẤN LUYỆN

Trong cấu hình thực nghiệm, khối mã hóa ảnh tạo ra biểu diễn đặc trưng có chiều 4096, sau đó được chiếu về không gian nhúng chung với kích thước $d_{model} = 768$. Bộ giải mã Transformer sử dụng 12 attention heads, chiều của tầng feed-forward là 4096, dropout = 0.1, và hàm kích hoạt GeLU. Đối với nhánh ngữ nghĩa, mỗi ảnh được gán 5 câu mô tả, với chiều dài tối thiểu của mỗi câu là 5 token. Trong nghiên cứu này, Qwen3-4B [23] được sử dụng để sinh các mô tả ngữ nghĩa mức câu từ ngữ cảnh thị giác cấu trúc. Mô hình này được lựa chọn vì là LLM mở trọng số hoàn toàn, cho phép triển khai cục bộ mà không phụ thuộc vào API thương mại, đồng thời duy trì được sự cân bằng phù hợp giữa chất lượng sinh ngôn ngữ và chi phí tính toán trong điều kiện phần cứng thực nghiệm sử dụng GPU NVIDIA Tesla P100 16 GB. Mô hình được tối ưu bằng thuật toán Adam với $learning\ rate = 1 \times 10^{-4}$. Hàm mất mát chính là CrossEntropyLoss với label smoothing = 0.1. Kích thước batch được đặt là 32, trong khi giai đoạn suy luận sử dụng beam search với beam width = 5. Toàn bộ quá trình thực nghiệm được triển khai trên nền tảng Kaggle với GPU NVIDIA Tesla P100 16 GB, sử dụng PyTorch, thư viện Transformers của Hugging Face cho các mô hình tiền huấn luyện, và pycocoevalcap để tính các độ đo đánh giá.

B. ĐỘ ĐO ĐÁNH GIÁ

Hiệu quả của mô hình được đánh giá bằng năm độ đo chuẩn trong bài toán chú thích ảnh, gồm BLEU-1 (B@1), BLEU-4 (B@4), METEOR (M), ROUGE-L (R) và CIDEr (C). Trong đó, B@1 và B@4 phản ánh mức độ trùng khớp n-gram giữa câu sinh ra và câu tham chiếu; M đánh giá mức độ linh hoạt ngôn ngữ thông qua đối sánh theo gốc từ

và từ đồng nghĩa; R đo độ tương đồng theo chuỗi con chung dài nhất; còn C là độ đo quan trọng nhất vì phản ánh mức độ đồng thuận ngữ nghĩa giữa câu sinh ra và nhiều câu tham chiếu, đồng thời giảm ảnh hưởng của các biểu đạt quá phổ biến nhờ cơ chế trọng số TF-IDF.

C. PHÂN TÍCH CHI PHÍ TÍNH TOÁN

Bên cạnh các độ đo đánh giá chất lượng chú thích, việc phân tích chi phí tính toán của mô hình cũng cần thiết nhằm làm rõ hơn sự đánh đổi giữa hiệu năng và khả năng triển khai thực tế. Do phương pháp đề xuất kết hợp nhiều backbone thị giác, các mô-đun trích xuất thông tin ngữ nghĩa từ ảnh và mô hình ngôn ngữ lớn, bài báo bổ sung thêm đánh giá về số lượng tham số, FLOPs và thời gian suy luận của các thành phần chính trong pipeline. Kết quả được trình bày trong Bảng 1.

Bảng 1. *Số lượng tham số, FLOPs và thời gian suy luận của các thành phần trong mô hình đề xuất*

Nhóm thành phần	Thành phần	Parameters (M)	FLOPs (G)	Thời gian suy luận
Nhánh thị giác	EfficientNet-B0	4.01	1.08	7.64 ms/img
	ResNet-101	42.50	20.54	10.47 ms/img
	ViT-Base/16	86.39	17.60	13.82 ms/img
	Tổng	132.90	39.22	31.93 ms/img
Mô hình chú thích ảnh	CaptionModel (Encoder+Decoder)	66.83	3.71	4.97 ms/img
Nhánh ngữ nghĩa	YOLOv12-X	59.22	200.3	75.70 ms/img
	RelTR	41.57	117.7	45.26 ms/img
	HOICLIP	162.79	197.2	85.79 ms/img
	BLIP-ITM-base-coco	223.74	-	53.67 ms/sentence
	Qwen3-4B	4022	-	926.7 ms/sentence

Ghi chú FLOPs của BLIP-ITM-base-coco và Qwen3-4B không được báo cáo do độ dài chuỗi vào/ra thay đổi, khiến việc ước lượng FLOPs bằng các công cụ hiện có kém tin cậy. Thời gian suy luận được tính theo **ms/ảnh** đối với các thành phần xử lý ảnh và **ms/câu** đối với các thành phần xử lý mô tả mức câu.

Bảng 1 cho thấy phần gia tăng về độ phức tạp chủ yếu tập trung ở nhánh bổ sung ngữ cảnh, đặc biệt tại các thành phần dùng để phát hiện đối tượng, dự đoán mối quan hệ, nhận diện tương tác và xử lý biểu diễn ngôn ngữ mức cao. Trong khi đó, nhánh thị giác và bộ sinh chú thích có mức tiêu tốn tài nguyên thấp hơn đáng kể. Kết quả này phản ánh đúng định hướng của mô hình, trong đó đặc trưng ảnh được tăng cường thêm bằng thông tin bối cảnh nhằm cải thiện chất lượng chú thích đầu ra.

Cần nhấn mạnh rằng các thành phần thuộc nhánh thị giác và nhánh bổ sung ngữ cảnh đều là các mạng đã huấn luyện sẵn và được giữ cố định trong nghiên cứu này; quá trình huấn luyện chủ yếu áp dụng cho bộ sinh chú thích. Mặc dù cách thiết kế này làm tăng độ phức tạp tính toán, kết quả CIDEr đạt 101.9 cho thấy lượng tài nguyên bổ sung đã mang lại cải thiện đáng kể về chất lượng mô tả. Do phần lớn các công trình hiện nay chưa công bố đầy đủ số lượng tham số, FLOPs và thời gian suy luận, vì vậy bài báo tập trung báo cáo minh bạch các chỉ số của phương pháp đề xuất và sử dụng chúng để làm rõ hơn sự đánh đổi giữa tài nguyên tính toán và hiệu quả chú thích.

D. KẾT QUẢ VÀ BÀN LUẬN

Phần này trình bày các kết quả thực nghiệm theo cả góc độ định lượng và định tính, qua đó phân tích hiệu quả của phương pháp đề xuất khi so sánh với các công trình gần đây cũng như khi đánh giá ảnh hưởng của từng thành phần và các tham số trong mô hình.

1. SO SÁNH VỚI CÁC CÔNG BỐ GẦN ĐÂY

Bảng 2 trình bày kết quả so sánh giữa mô hình đề xuất (MB-LLM) và một số phương pháp chú thích ảnh tiêu biểu trên tập MS COCO theo phân chia Karpathy. Kết quả cho thấy MB-LLM đạt 74.9 B@1, 31.8 B@4, 25.8 M, 56.8 R và 101.9 C, là cấu hình có kết quả tốt nhất trong số các phương pháp được đối sánh trên các độ đo được báo cáo trong bảng. Đặc biệt, việc đạt 101.9 C cho thấy mô hình không chỉ sinh câu đúng về mặt từ vựng mà còn có mức đồng thuận ngữ nghĩa cao với các chú thích tham chiếu.

So với các mô hình kinh điển dựa trên khung CNN–LSTM như Show, Attend and Tell, phương pháp đề xuất cho thấy mức cải thiện rõ rệt. Cụ thể, so với biến thể Hard-ATT (2015), mô hình tăng 3.1 điểm B@1 và 6.8 điểm B@4;

so với Soft-ATT (2015), mức tăng tương ứng là 4.2 điểm B@1 và 7.5 điểm B@4. Kết quả này cho thấy việc kết hợp biểu diễn thị giác đa backbone với thông tin ngữ nghĩa từ LLM mang lại ưu thế rõ rệt so với các kiến trúc chú thích ảnh truyền thống chỉ dựa trên một nguồn đặc trưng thị giác.

Bảng 2. *So sánh hiệu suất của các phương pháp chú thích ảnh trên tập MS COCO (Karpathy test)*

Phương pháp	B@1	B@4	M	R	C
Show, attend and tell (Hard-ATT 2015) [10]	71.8	25.0	23.0	-	-
Show, attend and tell (Soft-ATT 2015) [10]	70.7	24.3	23.9	-	-
Xception-YOLO Attention (2022) [16]	49.2	10.1	16.3	35.8	39.0
Bi-LS-AttM (2023) [15]	68.8	25.2	21.5	-	41.2
OD-VR-Cap (2024) [18]	72.6	28.3	24.8	53.4	85.1
Topic FRCNN-LSTM (2025) [17]	63.8	30.4	22.3	48.4	82.7
Swin-Trans-Cap (2025) [3]	67.3	27.9	25.2	-	70.3
Đề xuất (**MB-LLM**)	**74.9**	**31.8**	**25.8**	**56.8**	**101.9**

Khi so sánh với các phương pháp gần đây hơn, ưu thế của MB-LLM vẫn được duy trì ổn định. So với Bi-LS-AttM (2023), mô hình tăng 6.1 điểm B@1, 6.6 điểm B@4, 4.3 điểm M và 60.7 điểm C. So với OD-VR-Cap (2024), mức cải thiện đạt 2.3 điểm B@1, 3.5 điểm B@4, 1.0 điểm M, 3.4 điểm R và đặc biệt 16.8 điểm C. So với Topic FRCNN-LSTM (2025), mô hình đề xuất tiếp tục vượt trội trên toàn bộ các chỉ số, với chênh lệch lớn nhất nằm ở C (19.2 điểm). Đáng chú ý, khi đối sánh với Swin-Trans-Cap (2025), một phương pháp cũng dựa trên kiến trúc Transformer, MB-LLM vẫn đạt kết quả cao hơn rõ rệt, với mức tăng 7.6 điểm B@1, 3.9 điểm B@4, 0.6 điểm M và 31.6 điểm C. Kết quả này cho thấy lợi ích của việc kết hợp đặc trưng thị giác đa backbone với thông tin ngữ nghĩa mức câu từ LLM, thay vì chỉ dựa vào một kiến trúc Transformer thị giác đơn lẻ.

Nhìn chung, kết quả ở Bảng 2 khẳng định rằng MB-LLM đạt hiệu quả cạnh tranh mạnh trên tập MS COCO. Mức cải thiện nổi bật ở B@4 và C cho thấy phương pháp không chỉ nâng cao độ chính xác ở cấp cụm từ dài mà còn tăng mức độ phù hợp ngữ nghĩa giữa câu sinh ra và tập chú thích tham chiếu. Đây là một chỉ dấu quan trọng trong bài toán chú thích ảnh, nơi chất lượng đầu ra không chỉ phụ thuộc vào độ đúng về mặt từ vựng mà còn ở khả năng nắm bắt nội dung đặc trưng và bối cảnh của ảnh.

2. ĐÁNH GIÁ ẢNH HƯỞNG CỦA CÁC THÀNH PHẦN VÀ THAM SỐ TRONG MÔ HÌNH

Mục này trình bày các thực nghiệm phân tích thành phần và đánh giá độ nhạy của một số tham số chính, nhằm làm rõ đóng góp của từng thành phần trong kiến trúc đề xuất và mức độ ảnh hưởng của chúng đến chất lượng chú thích đầu ra.

a) Ảnh hưởng của đặc trưng đa backbone

Bảng 3 trình bày kết quả đánh giá ảnh hưởng của các cấu hình đặc trưng thị giác khác nhau. Trong nhóm backbone đơn, ResNet-101 cho kết quả tốt nhất về B@1 (73.7), B@4 (29.7) và R (55.5), trong khi ViT đạt giá trị cao nhất về M (24.6) và C (91.3). Sự khác biệt này phản ánh đặc tính biểu diễn khác nhau của hai kiến trúc. Với cơ chế tích chập sâu, ResNet-101 có ưu thế trong việc khai thác các đặc trưng cục bộ như biên, kết cấu, hình dạng và cấu trúc không gian ổn định của các vùng ảnh. Nhờ đó, mô hình dễ nhận diện chính xác các đối tượng và sinh ra các từ hoặc cụm từ có mức trùng khớp cao với câu tham chiếu, nên đạt kết quả tốt hơn trên BLEU-1, BLEU-4 và ROUGE-L. Trong khi đó, ViT sử dụng self-attention toàn cục, cho phép mô hình hóa tốt hơn ngữ cảnh tổng thể, các phụ thuộc xa và quan hệ giữa các vùng ảnh không liền kề, nên chú thích sinh ra phù hợp hơn ở mức ngữ nghĩa; điều này được phản ánh rõ hơn qua METEOR và CIDEr.

Bảng 3. *Đánh giá ảnh hưởng của đặc trưng đa backbone*

Đặc trưng	B@1	B@4	M	R	C
CNN (EfficientNet-B0)	71.4	26.9	22.5	53.4	79.8
CNN (ResNet-101)	73.7	29.7	24.4	55.5	90.5
ViT	72.7	29.0	24.6	50.0	91.3
CNN + ViT (Đề xuất)	**74.5**	**30.3**	**25.3**	**56.3**	**95.0**

Khi kết hợp đặc trưng theo cấu hình CNN + ViT, mô hình đạt 74.5 B@1, 30.3 B@4, 25.3 M, 56.3 R và 95.0 C, vượt tất cả các cấu hình backbone đơn. So với EfficientNet-B0, mức tăng tương ứng là 3.1 điểm B@1, 3.4 điểm B@4, 2.8 điểm M, 2.9 điểm R và 15.2 điểm C. So với ResNet-101, mức cải thiện đạt 0.8 điểm B@1, 0.6 điểm B@4, 0.9 điểm M, 0.8 điểm R và 4.5 điểm C. So với ViT, mô hình tiếp tục tăng 1.8 điểm B@1, 1.3 điểm B@4, 0.7 điểm M, 6.3 điểm R và 3.7 điểm C. Những kết quả này cho thấy CNN và ViT mang tính bổ sung lẫn nhau, và việc kết hợp hai loại backbone giúp mô hình đồng thời khai thác hiệu quả đặc trưng cục bộ và bối cảnh toàn cục, từ đó tạo ra biểu diễn thị giác giàu thông tin hơn cho bài toán chú thích ảnh.

b) Ảnh hưởng của nhánh ngữ nghĩa từ LLM

Bảng 4. *Đánh giá ảnh hưởng của biểu diễn ngữ nghĩa mở rộng từ LLM*

Phương pháp	B@1	B@4	M	R	C
Đề xuất (**Đa backbone**)	74.5	30.3	25.3	56.3	95.0
Đề xuất (**Đa backbone + LLM**)	**74.9**	**31.8**	**25.8**	**56.8**	**101.9**

Bảng 4 đánh giá tác động của việc bổ sung nhánh ngữ nghĩa từ LLM trên nền cấu hình đa backbone. Khi chỉ sử dụng nhánh thị giác đa backbone, mô hình đạt 74.5 B@1, 30.3 B@4, 25.3 M, 56.3 R và 95.0 C. Sau khi bổ sung nhánh ngữ nghĩa mức câu từ LLM, mô hình MB-LLM đạt 74.9 B@1, 31.8 B@4, 25.8 M, 56.8 R và 101.9 C. Như vậy, cấu hình có LLM cải thiện đồng thời trên toàn bộ các độ đo, trong đó mức tăng nổi bật nhất là 1.5 điểm B@4 và 6.9 điểm C. Kết quả này cho thấy đóng góp chính của nhánh LLM không nằm ở việc làm tăng mức độ trùng khớp từ vựng ở bề mặt, mà ở khả năng bổ sung thông tin ngữ cảnh và quan hệ ngữ nghĩa giúp câu chú thích trở nên cụ thể hơn và phù hợp hơn với nội dung ảnh. Điều này là hợp lý vì BLEU và ROUGE-L chủ yếu phản ánh mức độ tương đồng hình thức giữa câu sinh và câu tham chiếu, trong khi CIDEr là một trong những độ đo có tính đại diện cao trong bài toán chú thích ảnh, do nhấn mạnh mức độ đồng thuận ngữ nghĩa với nhiều câu tham chiếu. Vì vậy, mức tăng 6.9 điểm CIDEr cho thấy nhánh LLM đã mang lại cải thiện thực chất về chất lượng đầu ra, đặc biệt ở phương diện mô tả đúng nội dung đặc trưng của ảnh.

c) Đánh giá độ nhạy của tham số K và N trong sinh mô tả ngữ nghĩa

Để đánh giá ảnh hưởng của cấu hình sinh mô tả ngữ nghĩa từ LLM, bài báo khảo sát độ nhạy của hai tham số chính là số lượng mô tả ngữ nghĩa cho mỗi ảnh (K) và độ dài tối thiểu của mỗi mô tả (N). Thí nghiệm được thực hiện bằng cách thay đổi từng tham số trong khi cố định tham số còn lại, nhằm làm rõ mức độ tác động của chúng đến chất lượng chú thích đầu ra.

Bảng 5. *Ảnh hưởng của độ dài tối thiểu (N) của mô tả ngữ nghĩa khi cố định $K = 5$*

Độ dài tối thiểu N token	B@1	B@4	M	R	C
$N = 3$	74.3	31.2	25.4	56.2	99.6
$N = 5$	74.9	**31.8**	**25.8**	**56.8**	**101.9**
$N = 7$	**75.1**	**31.8**	25.6	56.4	101.5

Kết quả trong Bảng 5 cho thấy hiệu năng của hệ thống thay đổi ở mức vừa phải theo tham số N. Khi $N = 3$, các mô tả ngữ nghĩa có xu hướng ngắn, nên thông tin bổ sung chủ yếu dừng ở mức liệt kê đối tượng hoặc hành động đơn lẻ, chưa phản ánh đầy đủ bối cảnh và quan hệ giữa các đối tượng trong ảnh. Vì vậy, các độ đo đều thấp hơn so với cấu hình $N = 5$, trong đó CIDEr giảm 2.3 điểm. Khi tăng lên $N = 5$, mỗi mô tả đạt độ dài đủ để hình thành một câu tự nhiên, qua đó biểu diễn tốt hơn nội dung và ngữ cảnh của ảnh; cấu hình này đạt kết quả tốt nhất trên hầu hết các độ đo. Trong khi đó, khi $N = 7$, B@1 tăng nhẹ lên 75.1 nhưng METEOR, ROUGE-L và CIDEr đều giảm. Điều này cho thấy việc ép mô tả dài hơn không nhất thiết làm giàu thêm thông tin hữu ích, mà có thể làm tăng các thành phần diễn đạt dư thừa, khiến biểu diễn ngữ nghĩa trở nên kém tập trung hơn.

Bảng 6. *Ảnh hưởng của số lượng mô tả ngữ nghĩa (K) khi cố định ($N = 5$)*

Số câu mô tả K	B@1	B@4	M	R	C
$K = 3$	73.8	30.9	25.6	56.0	99.2
$K = 5$	**74.9**	31.8	**25.8**	**56.8**	**101.9**
$K = 7$	**74.9**	**31.9**	25.5	56.4	100.7

Bảng 6 cho thấy số lượng mô tả ngữ nghĩa cũng ảnh hưởng đáng kể đến hiệu năng của hệ thống. Với $K = 3$, số lượng mô tả còn hạn chế nên chưa bao quát đầy đủ các khía cạnh khác nhau của ảnh, chẳng hạn như đối tượng

trung tâm, hành động, quan hệ và bối cảnh không gian. Do đó, cấu hình này cho kết quả thấp hơn rõ rệt so với $K = 5$, đặc biệt CIDEr giảm 2.7 điểm. Khi tăng lên $K = 5$, mô hình có đủ số lượng mô tả để phản ánh đa dạng hơn nội dung ảnh, đồng thời cơ chế trọng số BLIP-ITM giúp ưu tiên các mô tả phù hợp hơn với ảnh đầu vào. Đây là cấu hình đạt kết quả tốt nhất hoặc gần tốt nhất trên toàn bộ các độ đo. Ngược lại, khi tăng lên $K = 7$, B@4 chỉ tăng rất nhẹ 0.1 điểm trong khi METEOR, ROUGE-L và CIDEr đều giảm. Kết quả này cho thấy việc bổ sung thêm mô tả sau một ngưỡng nhất định không còn mang lại lợi ích tương xứng, do thông tin mới tăng thêm ít nhưng nguy cơ lặp lại hoặc đưa vào chi tiết kém liên quan lại cao hơn.

Nhìn chung, kết quả ở Bảng 5 và Bảng 6 cho thấy mô hình có độ nhạy vừa phải với hai tham số K và N, nhưng không biến động lớn trong phạm vi khảo sát. Điều này phản ánh tính ổn định tương đối của kiến trúc, nhờ cơ chế trọng số BLIP-ITM và quá trình cross-attention có chọn lọc giữa nhánh thị giác và nhánh ngữ nghĩa. Trong các cấu hình đã khảo sát, $K = 5$ và $N = 5$ cho sự cân bằng tốt nhất giữa lượng thông tin bổ sung và mức độ nhiễu, do đó được chọn làm cấu hình chính cho các thực nghiệm còn lại.

d) Đánh giá độ nhạy của beam width trong Beam Search

Để đánh giá ảnh hưởng của chiến lược giải mã Beam Search trong giai đoạn suy luận, chúng tôi khảo sát độ nhạy của beam width với các giá trị 3, 5, 7 và 10. Kết quả được trình bày trong Bảng 7.

Bảng 7. *Ảnh hưởng của beam width đến chất lượng chú thích sinh ra*

Beam width	B@1	B@4	M	R	C
3	**74.9**	31.5	25.6	56.7	101.2
5	**74.9**	**31.8**	**25.8**	**56.8**	**101.9**
7	74.7	31.7	**25.8**	56.7	101.7
10	74.7	31.6	**25.8**	56.7	101.6

Kết quả trong Bảng 7 cho thấy beam width = 5 đạt hiệu năng tốt nhất hoặc gần tốt nhất trên toàn bộ các độ đo. Khi beam width = 3, không gian tìm kiếm còn hẹp nên mô hình chưa khai thác hết các giả thuyết chuỗi có chất lượng cao, dẫn đến B@4 và CIDEr thấp hơn so với cấu hình chuẩn. Trong khi đó, khi tăng beam width lên 7 và 10, các độ đo không tiếp tục cải thiện mà giảm nhẹ, cho thấy việc mở rộng thêm không gian tìm kiếm không đem lại lợi ích tương ứng về chất lượng chú thích. Nhìn chung, hiệu năng của hệ thống tương đối ổn định trong khoảng khảo sát, và beam width = 5 là cấu hình cân bằng tốt nhất giữa khả năng tìm kiếm và chất lượng đầu ra.

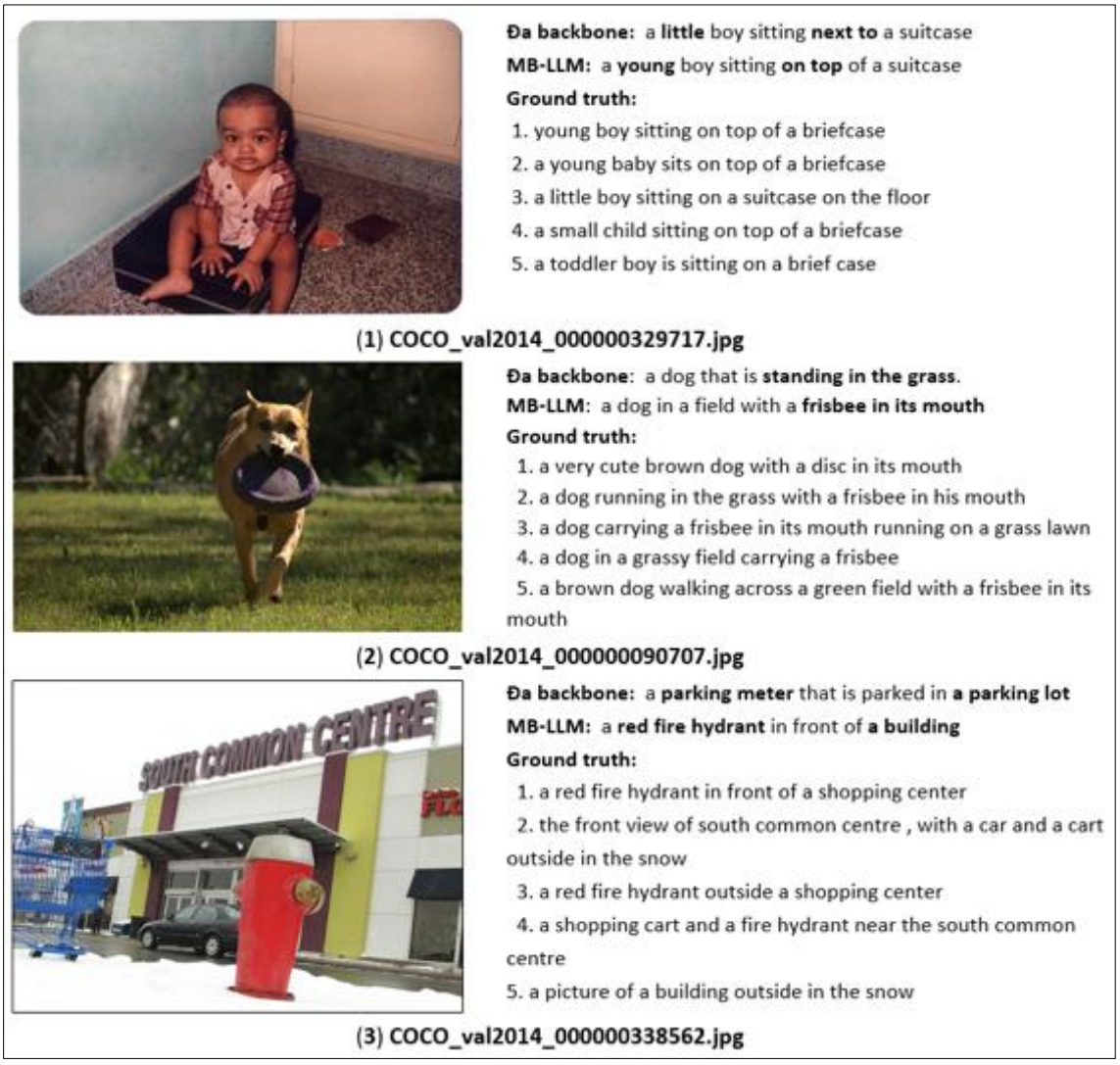
3. ĐÁNH GIÁ ĐỊNH TÍNH

Hình 2 trình bày kết quả đánh giá định tính trên ba ảnh thuộc tập kiểm tra của MS COCO, nhằm so sánh chú thích sinh bởi mô hình chỉ sử dụng nhánh thị giác đa backbone với mô hình kết hợp đa backbone và LLM (MB-LLM). Kết quả cho thấy việc bổ sung nhánh ngữ nghĩa từ LLM giúp câu chú thích đầu ra bám sát hơn với nội dung chính của ảnh và gần hơn với các câu chú thích tham chiếu. Ở ảnh thứ nhất, mô hình đa backbone sinh câu "*a little boy sitting next to a suitcase*", trong khi MB-LLM sinh câu "*a young boy sitting on top of a suitcase*". So với các câu tham chiếu, mô tả thứ hai phù hợp hơn vì phản ánh đúng quan hệ không gian giữa đứa trẻ và chiếc va-li, cụ thể là hành động ngồi trên thay vì chỉ ngồi bên cạnh. Kết quả này cho thấy nhánh ngữ nghĩa từ LLM đã hỗ trợ mô hình nhận diện chính xác hơn tương tác giữa đối tượng và vật thể trong ảnh.

Ở ảnh thứ hai, mô hình đa backbone tạo chú thích "*a dog that is standing in the grass*", trong khi MB-LLM sinh câu "*a dog in a field with a frisbee in its mouth*". Có thể thấy câu thứ hai giàu thông tin hơn rõ rệt, bởi ngoài việc nhận diện đúng đối tượng trung tâm là con chó, mô hình còn bổ sung chính xác chi tiết quan trọng là chiếc "*frisbee*" trong miệng. So với các câu ground truth, đây cũng là mô tả gần hơn về mặt nội dung ngữ nghĩa, cho thấy nhánh LLM giúp mô hình tập trung tốt hơn vào các thực thể nổi bật và đặc trưng của cảnh.

Ở ảnh thứ ba, mô hình đa backbone sinh câu "*a parking meter that is parked in a parking lot*", trong khi MB-LLM sinh câu "*a red fire hydrant in front of a building*". Trong trường hợp này, mô hình không sử dụng LLM mắc lỗi nhận dạng đối tượng khi nhầm "*fire hydrant*" thành "*parking meter*", đồng thời tạo ra một diễn đạt thiếu tự nhiên. Ngược lại, MB-LLM không chỉ nhận diện đúng đối tượng trung tâm mà còn mô tả đúng quan hệ không gian "*in front of a building*", phù hợp hơn với các câu tham chiếu. Ví dụ này cho thấy thông tin ngữ nghĩa từ LLM có thể góp phần giảm sai lệch ngữ nghĩa và cải thiện độ chính xác của chú thích trong các cảnh có bố cục phức tạp.

Tổng hợp ba ví dụ minh họa cho thấy MB-LLM có xu hướng sinh ra các chú thích cụ thể hơn, chính xác hơn về quan hệ giữa các đối tượng, và gần hơn với nội dung tham chiếu so với mô hình chỉ sử dụng nhánh thị giác. Kết quả này nhất quán với phần đánh giá định lượng, qua đó củng cố nhận định rằng việc bổ sung nhánh ngữ nghĩa mức câu từ LLM là hữu ích đối với bài toán chú thích ảnh.



Hình 2. *So sánh định tính kết quả chú thích ảnh trên ba ảnh thuộc tập kiểm tra của MS COCO giữa mô hình đa backbone và MB-LLM. Mô hình tích hợp LLM (MB-LLM) cho thấy khả năng sinh chú thích gần hơn với ground truth về đối tượng, quan hệ và ngữ cảnh.*

V. KẾT LUẬN

Bài báo đã đề xuất một phương pháp chú thích ảnh theo hướng kết hợp biểu diễn thị giác đa backbone với mở rộng ngữ nghĩa ảnh bằng mô hình ngôn ngữ lớn. Ở nhánh thị giác, ảnh đầu vào được mã hóa đồng thời bởi nhiều backbone nhằm tận dụng tính bổ sung giữa đặc trưng cục bộ và bối cảnh toàn cục. Ở nhánh ngữ nghĩa, ngữ cảnh thị giác cấu trúc gồm đối tượng, tương tác người–vật và quan hệ giữa các đối tượng được sử dụng để dẫn hướng LLM sinh ra các mô tả ngữ nghĩa mức câu. Các mô tả này tiếp tục được đánh giá bằng BLIP Image–Text Matching, sau đó được mã hóa thành biểu diễn liên tục và hợp nhất với nhánh thị giác thông qua cơ chế cross-attention trước khi đưa vào bộ giải mã Transformer để sinh chú thích cuối cùng. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt hiệu quả tốt và ổn định; đáng chú ý, mô hình MB-LLM đạt 101.9 C trên tập MS COCO, cho thấy hiệu quả rõ rệt của việc kết hợp đặc trưng thị giác đa backbone với thông tin ngữ nghĩa từ LLM. Bên cạnh đó, các phân tích thành phần và đánh giá định tính cũng xác nhận rằng mỗi thành phần trong kiến trúc đều đóng góp tích cực vào hiệu quả chung của hệ thống. Nhìn chung, nghiên cứu khẳng định rằng hướng kết hợp đa backbone thị giác với ngữ nghĩa mở rộng từ LLM là một giải pháp khả thi và hiệu quả cho bài toán chú thích ảnh.

Một hạn chế của nghiên cứu là chưa khảo sát độ nhạy của hệ thống đối với các họ LLM khác nhau. Trong các nghiên cứu tiếp theo, cần thực hiện so sánh có hệ thống giữa các LLM khác nhau về quy mô và kiến trúc để đánh giá đầy đủ hơn ảnh hưởng của mô hình ngôn ngữ đến chất lượng biểu diễn ngữ nghĩa và hiệu năng chú thích ảnh. Bên cạnh đó, phương pháp có thể được mở rộng theo hướng đánh giá trên nhiều bộ dữ liệu hơn và sử dụng các mô hình ngôn ngữ lớn hoặc mô hình thị giác–ngôn ngữ mạnh hơn để nâng cao chất lượng mô tả ngữ nghĩa. Về khả năng mở rộng, các nguồn tri thức có cấu trúc như AMR hoặc ConceptNet có thể được tích hợp ngay tại bước xây

dựng ngữ cảnh đầu vào cho LLM. Cụ thể, AMR có thể hỗ trợ tổ chức lại thông tin về đối tượng, tương tác và quan hệ dưới dạng cấu trúc ngữ nghĩa có hướng và vai nghĩa rõ ràng hơn, trong khi ConceptNet có thể bổ sung thêm tri thức thường thức liên quan đến các thực thể đã phát hiện. Nhờ đó, ngữ cảnh đầu vào cho LLM có thể được làm giàu mà không cần thay đổi kiến trúc tổng thể của pipeline hiện tại.

VI. TÀI LIỆU THAM KHẢO

[1] Ghandi, T., H. Pourreza, and H. Mahyar (2023), *Deep learning approaches on image captioning: A review.* ACM Computing Surveys. **56**(3): p. 1-39.

[2] Kavitha, R., et al. (2023). *Deep learning-based image captioning for visually impaired people*. in *E3S Web of Conferences*. EDP Sciences.

[3] Văn, T.N., et al. (2025), *Chú thích ảnh cho người khiếm thị sử dụng Transformer.* Tạp chí Khoa học HUFLIT. **9**(3): p. 14-14.

[4] Lee, D.I., et al. (2023), *Crop disease diagnosis with deep learning-based image captioning and object detection.* Applied Sciences. **13**(5): p. 3148.

[5] Tsai, W.-L., et al. (2025), *Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention.* Automation in Construction. **169**: p. 105863.

[6] Kearns, L. (2025), *Content moderation assistance through image caption generation.* Intelligent Systems with Applications. **25**: p. 200489.

[7] Sharma, A., H. Singh, and M. Pant (2025), *Pixels to Prose: A Comprehensive Survey of Image Captioning Techniques with Deep Learning and Generative Artificial Intelligence.* Neurocomputing, 132385.

[8] Van Thinh, N. and T. Van Lang (2025), *Integrating abstract meaning representation to enhance transformer-based image captioning.* IEEE Access.

[9] Vinyals, O., et al. (2015), *Show and tell: A neural image caption generator*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[10] Xu, K., et al. (2015), *Show, attend and tell: Neural image caption generation with visual attention*. in *International conference on machine learning*. PMLR.

[11] Agarwal, L. and B. Verma (2024), *From methods to datasets: A survey on Image-Caption Generators.* Multimedia Tools and Applications. **83**(9): p. 28077-28123.

[12] Dosovitskiy, A., et al. (2020), *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.

[13] Liu, Z., et al. (2021), *Swin transformer: Hierarchical vision transformer using shifted windows*. in *Proceedings of the IEEE/CVF international conference on computer vision*.

[14] Al-Shamayleh, A.S., et al. (2024), *A comprehensive literature review on image captioning methods and metrics based on deep learning technique.* Multimedia Tools and Applications. **83**(12): p. 34219-34268.

[15] Xie, T., et al. (2023), *Bi-LS-AttM: A bidirectional LSTM and attention mechanism model for improving image captioning.* Applied Sciences. **13**(13): p. 7916.

[16] Al-Malla, M.A., A. Jafar, and N. Ghneim (2022), *Image captioning model using attention and object features to mimic human image understanding.* Journal of Big Data. **9**(1): p. 20.

[17] Yeh, J.-F., K.-M. Lin, and C.-C. Chen (2025), *Image Captioning Using Topic Faster R-CNN-LSTM Networks.* Information. **16**(9): p. 726.

[18] Van Thinh, N. and T. Van Lang (2024), *OD-VR-CAP: Image captioning based on detecting and predicting relationships between objects.* Journal of Computer Science and Cybernetics. **40**(4): p. 327-346.

[19] Van Thinh, N. and T. Van Lang (2025), *RGTranCNet: Effective image captioning model using cross-attention and semantic knowledge.* Vietnam Journal of Science and Technology.

[20] Văn, T.N., et al. (2025), *Chú thích ảnh dựa vào việc khai thác mối quan hệ giữa các đối tượng trong ảnh.* Tạp chí Khoa học HUFLIT. **9**(3).

[21] Thinh, N.V., T.V. Lang, and N.M. Hai (2025). *CLIP-AMR-GPT: Enhancing Image Captioning via Cross-Modal Semantics Fusion and GPT-Based Re-ranking*. in *International Conference on Multi-disciplinary Trends in Artificial Intelligence*. Springer.

[22] Mishra, S., et al. (2024). *Image caption generation using vision transformer and gpt architecture*. in *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE.

[23] Yang, A., et al. (2025), *Qwen3 technical report.* arXiv preprint arXiv:2505.09388.

[24] Cornia, M., et al. (2020). *Meshed-memory transformer for image captioning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[25] Yang, X., et al. (2023), *Context-aware transformer for image captioning.* Neurocomputing. **549**: p. 126440.

[26] Parseh, M.J. and S. Ghadiri (2025), *Graph-based image captioning with semantic and spatial features.* Signal Processing: Image Communication. **133**: p. 117273.

[27] Hafeth, D.A., S. Kollias, and M. Ghafoor (2023), *Semantic representations with attention networks for boosting image captioning.* IEEE Access. **11**: p. 40230-40239.

[28] Li, Z., Q. Su, and T. Chen (2023), *External knowledge-assisted transformer for image captioning.* Image and Vision Computing. **140**: p. 104864.
[29] Tian, Y., Q. Ye, and D. Doermann (2025), *Yolov12: Attention-centric real-time object detectors.* arXiv preprint arXiv:2502.12524.
[30] Ning, S., et al. (2023). *Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
[31] Cong, Y., M.Y. Yang, and B. Rosenhahn (2023), *Reltr: Relation transformer for scene graph generation.* IEEE Transactions on Pattern Analysis and Machine Intelligence. **45**(9): p. 11169-11183.
[32] Lin, T.-Y., et al. (2014). *Microsoft coco: Common objects in context*. in *European conference on computer vision*. Springer.
[33] Karpathy, A. and L. Fei-Fei (2015). *Deep visual-semantic alignments for generating image descriptions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

INTEGRATING MULTI-BACKBONE VISUAL REPRESENTATIONS AND LLM-ENHANCED SEMANTICS FOR IMAGE CAPTIONING

Nguyen Trung Tinh, Nguyen Van Thinh*

ABSTRACT—Automatic image captioning remains a challenging task because a model must jointly exploit visual and semantic information to generate accurate and natural descriptions. This paper proposes an image captioning approach that combines multi-backbone visual representations with LLM-enhanced semantics. In the visual branch, the input image is encoded by multiple backbones to leverage the complementarity between local features and global context. In the semantic branch, a structured visual context, including objects, human–object interactions, and inter-object relations, is used to guide a large language model (LLM) in generating sentence-level semantic descriptions. These descriptions are evaluated using BLIP Image–Text Matching model, subsequenly encoded into continuous representations and fused with the visual branch through cross-attention before being fed into a Transformer decoder to produce the final caption. Experiments on the MS COCO dataset achieve 74.9 BLEU-1, 31.8 BLEU-4, 25.8 METEOR, 56.8 ROUGE-L, and 101.9 CIDEr. The results demonstrate that combining multi-backbone visual features with semantic information derived from an LLM enables the model to generate captions that are more accurate, more context-aware, and better aligned with image content.

Keywords— Image captioning, multi-backbone visual representation, LLM-based image semantic expansion, cross-attention, Transformer.



Nguyễn Trung Tính là sinh viên ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh; lĩnh vực quan tâm: học máy và ứng dụng.



Nguyễn Văn Thịnh nhận bằng thạc sĩ ngành Hệ thống thông tin tại Trường Đại học Khoa học tự nhiên, Đại học Quốc gia TP. Hồ Chí Minh năm 2012. Hiện nay, ông là nghiên cứu sinh ngành Khoa học máy tính tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, đồng thời là giảng viên tại Khoa Công nghệ thông tin, Trường Đại học Sư phạm TP. HCM. Nghiên cứu của ông tập trung vào trí tuệ nhân tạo và ứng dụng, đặc biệt là học sâu đa phương thức và các mô hình ngôn ngữ đa phương thức, hướng đến các bài toán truy vấn thông tin, chú thích ảnh tự động và hỏi-đáp dựa trên hình ảnh.