NGHIÊN CỨU XÂY DỰNG KHO DỮ LIỆU PHỤC VỤ HỎI ĐÁP HIỆN THÂN ĐA PHƯƠNG THỨC

RESEARCH ON BUILDING A DATASETS TO SERVE MULTIMODEL EMBODIED QUESTION ANSWERING

Võ Trung Hùng*

Trường Đại học Sư phạm Kỹ thuật - Đại học Đà Nẵng, Việt Nam[1]

*Tác giả liên hệ / Corresponding author: vthung@ute.udn.vn

(Nhận bài / Received: 28/7/2025; Sửa bài / Revised: 11/8/2025; Chấp nhận đăng / Accepted: 05/9/2025)

DOI: 10.31130/ud-jst.2025.23(9A).384

Tóm tắt - Hỏi đáp hiện thân (Embodied Question Answering - EQA) là một nhiệm vụ AI mới. Trong đó, một tác nhân được sinh ra tại một vị trí ngẫu nhiên trong môi trường 3D và được đưa ra các câu hỏi tuỳ theo tình huống. Để trả lời câu hỏi, trước tiên, tác nhân phải điều hướng thông minh để khám phá môi trường, thu thập thông tin qua các góc nhìn khác nhau. Vì vậy, EQA đòi hỏi một loạt các kỹ năng AI như hiểu ngôn ngữ, nhận dạng trực quan, nhận thức chủ động, điều hướng theo mục tiêu, suy luận… Bài báo này trình bày kết quả nghiên cứu về EQA, đặc biệt tập trung vào việc xây dựng một tập dữ liệu đủ lớn để phục vụ các nghiên cứu về EQA. Tác giả đã thử nghiệm bộ dữ liệu trên một số mô hình và kết quả cho thấy bộ dữ liệu này cho kết quả tốt trên tất cả các mô hình thử nghiệm. Kết quả này là tiền đề quan trọng để tiếp tục triển khai các nghiên cứu liên quan đến EQA và EAI.

Từ khóa - Hỏi đáp hiện thân; kho dữ liệu; đa phương thức; mô hình; AI hiện thân

Abstract - Embodied Question Answering (EQA) is a new AI task in which an agent is spawned at a random location in a 3D environment and asked questions depending on the situation. To answer the questions, the agent must first navigate intelligently to explore the environment, collect information from different perspectives. Therefore, EQA requires a series of AI skills such as language understanding, visual recognition, active perception, goal-directed navigation, inference, etc. This paper presents the research results on EQA, focusing in particular on building a large enough dataset to serve EQA research. We tested the dataset on several models and the results showed that this dataset gave good results on all the tested models. This result is an important premise for further research related to EQA and EAI.

Key words - Embodied question answering; dataset; multimodality; model; Embodied AI

1. Đặt vấn đề

Trong lĩnh vực trí tuệ nhân tạo, "Embodied AI" (AI hiện thân) là một lĩnh vực nghiên cứu tập trung vào việc phát triển các hệ thống AI có khả năng tương tác với thế giới thực một cách trực tiếp, thay vì chỉ hoạt động trong môi trường ảo. AI hiện thân đề cập đến các hệ thống trí tuệ nhân tạo có thể tương tác và học hỏi từ môi trường của chúng bằng cách sử dụng một bộ công nghệ bao gồm cảm biến, động cơ, máy học và xử lý ngôn ngữ tự nhiên. Một số ví dụ nổi bật về trí tuệ nhân tạo hiện thân là xe tự hành, rô bốt hình người và máy bay không người lái. Thuật ngữ "embodied" là một từ mang nghĩa rộng, chỉ việc hiện thực hóa, thể hiện một ý tưởng, phẩm chất hay khái niệm trừu tượng nào đó thành hình thức cụ thể và dễ hiểu hơn [1].

Hỏi đáp hiện thân (Embodied Question Answering - EQA) là một nhiệm vụ AI mới, trong đó một tác nhân được sinh ra tại một vị trí ngẫu nhiên trong môi trường 3D và được hỏi một câu hỏi (ví dụ "Chiếc xe kia màu gì?"). Để trả lời, trước tiên, tác nhân phải điều hướng thông minh để khám phá môi trường, thu thập thông tin thông qua góc nhìn ngôi thứ nhất (egocentric), sau đó trả lời câu hỏi (ví dụ "màu cam"). Embodied QA đòi hỏi một loạt các kỹ năng AI như hiểu ngôn ngữ, nhận dạng trực quan, nhận thức chủ động, điều hướng theo mục tiêu, suy luận, bộ nhớ dài hạn và đưa ngôn ngữ vào hành động [2].

Liên quan đến lĩnh vực hỏi đáp hiện thân, đã có nhiều nghiên cứu cả về lý thuyết và thực nghiệm. Khi xây dựng hệ thống EmbodiedQA, các tác giả đã xây dựng một tập dữ liệu quy mô lớn và một khuôn khổ kết hợp học tăng cường (Reinforcement Learning - RL), nhận thức thị giác và hiểu ngôn ngữ. Các tác nhân được đào tạo trong môi trường AI2-THOR và House3D [2]. Một hướng nghiên cứu khác là sử dụng các mô hình ngôn ngữ lớn (Large Language Models - LLM), nó cho phép kết hợp nhiều mô-đun lại với nhau để tạo ra các chương trình có thể thực hiện các tác vụ suy luận phức tạp trên hình ảnh. Tiêu biểu cho hướng nghiên cứu này là hệ thống TANGO (Training-free Embodied AI Agents for Open-world Tasks), một phương pháp mở rộng việc biên soạn chương trình thông qua các LLM đã được quan sát cho hình ảnh, nhằm mục đích tích hợp các khả năng đó vào các tác nhân hiện thân có khả năng quan sát và hoạt động trong thế giới thực [3]. Nhằm tăng cường hiệu quả, một mô hình mới đang được tập trung nghiên cứu là sử dụng bản đồ Nơ-ron và mạng lưới bộ nhớ theo giai đoạn (Neural Maps and Episodic Memory Networks). Về bản chất, đây là các mô hình tính toán tận dụng bộ nhớ không gian và ngữ nghĩa để nâng cao khả năng định hướng và lập luận trong các hệ thống trí tuệ nhân tạo. Bản đồ Nơ-ron tạo ra một biểu diễn tôpô của môi trường, cho phép các tác nhân định hướng và lập luận về các mối quan hệ không gian. Mặt khác, Mạng lưới bộ nhớ theo giai

[1] The University of Danang - University of Technology and Education, Vietnam (Vo Trung Hung)

đoạn lưu trữ và truy xuất cho phép sử dụng các trải nghiệm trong quá khứ để cung cấp thông tin cho các hành động và quyết định trong tương lai, dựa trên cả thông tin không gian và ngữ nghĩa [4], [5]. ViLBERT, VLN-BERT và các mô hình ngôn ngữ thị giác khác đã tạo ra các bộ chuyển đổi được huấn luyện trước và đã được điều chỉnh cho các thiết lập được thể hiện để kết hợp nhận thức và tín hiệu ngôn ngữ [6], [7]. Một số nghiên cứu khác tập trung vào phát triển các mô hình điều hướng đa tác nhân gồm: các chính sách điều hướng hiệu quả đã được phát triển, cải thiện khả năng của các tác nhân trong việc tiếp cận các đối tượng mục tiêu được mô tả bằng ngôn ngữ… [8].

EQA có thể được chia thành hai nhánh nghiên cứu dựa trên các tương tác cụ thể: nhánh đầu tiên tập trung vào một tác nhân, giống như một rô-bốt ảo, điều hướng để trả lời các câu hỏi bằng lời, chỉ kết hợp các truy vấn bằng lời [2]. Nhánh thứ hai bao gồm các biểu thức đa phương thức, trong đó con người tương tác với môi trường bằng cách sử dụng cả lời nói và cử chỉ bằng lời [9]. Vấn đề đặt ra là làm thế nào để thiết kế các nhiệm vụ EQA để hiểu các câu hỏi sử dụng các biểu thức đa phương thức (lời nói và cử chỉ phi ngôn ngữ) trong các bối cảnh cụ thể. Ví dụ, một nhiệm vụ EQA có thể liên quan đến việc chỉ vào một vật thể và hỏi "vật thể đó là gì?", đòi hỏi phải lý luận về các biểu thức đa phương thức để trả lời câu hỏi. Ngoài ra, một hạn chế đáng chú ý trong nhiều tập dữ liệu EQA hiện có là góc nhìn đơn lẻ (của người nói hoặc người quan sát) về các phát ngôn bằng lời, không giống như các tương tác trong thế giới thực, nơi mọi người sử dụng cả hai góc nhìn một cách hoán đổi cho nhau. Ví dụ, câu hỏi của người nói, *"Vật thể bên phải chiếc cốc là gì?"* có thể được hiểu là bên trái chiếc cốc theo góc nhìn của người quan sát. Việc thiếu nhiều góc nhìn này trong các tập dữ liệu hiện có sẽ cản trở sự phát triển của các mô hình QA.

Để giải quyết những thiếu sót của các tập dữ liệu EQA hiện có, tác giả đã mở rộng một trình mô phỏng hiện thân để phát triển một tập dữ liệu mới quy mô lớn. Tác giả đã giải quyết những hạn chế của các phương pháp hợp nhất đa phương thức hiện có và phát triển một mô hình huấn luyện đa phương thức cho các tác vụ EQA.

2. Các nghiên cứu liên quan

2.1. Hỏi đáp trực quan

Hỏi đáp trực quan (Visual Question Answering - VQA) là một nhiệm vụ đa ngành trong trí tuệ nhân tạo, kết hợp thị giác máy tính và xử lý ngôn ngữ tự nhiên. Trong VQA, hệ thống được cung cấp một hình ảnh và một câu hỏi ngôn ngữ tự nhiên về hình ảnh đó. Mục tiêu của hệ thống là tạo ra một câu trả lời đúng dựa trên nội dung trực quan của hình ảnh và ngữ nghĩa của câu hỏi. Ví dụ, khi chúng ta có một hình ảnh là một cậu bé đang cầm một quả bóng bay màu đỏ, đặt ra cho hệ thống một câu hỏi *"Quả bóng bay màu gì?"* và sau đó hệ thống sẽ trả lời *"Đỏ"*.

Kiến trúc tổng quát của hệ thống hỏi đáp trực quan như Hình 1.

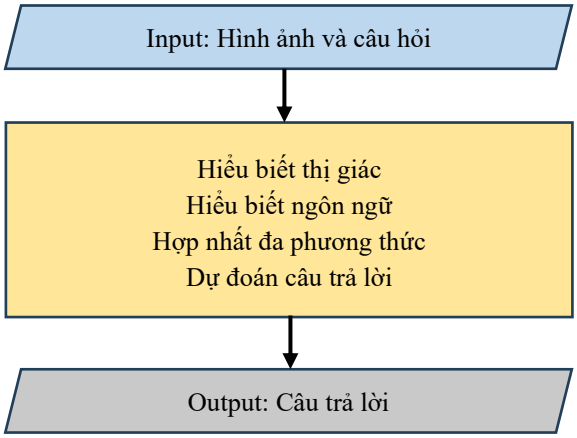
Các thành phần cốt lõi của một hệ thống hỏi đáp trực quan sẽ bao gồm:

- Hiểu biết thị giác (Visual Understanding): trích xuất các đặc điểm từ hình ảnh bằng mạng nơ-ron tích chập (CNN), bộ phát hiện đối tượng (như Faster R-CNN), hoặc bộ chuyển đổi thị giác (ViT).

- Hiểu biết ngôn ngữ (Language Understanding): phân tích cú pháp và nhúng câu hỏi bằng RNN, GRU, LSTM hoặc bộ chuyển đổi (ví dụ: BERT).

- Hợp nhất đa phương thức (Multimodal Fusion): kết hợp các đặc điểm thị giác và văn bản thông qua các kỹ thuật như cơ chế chú ý, gộp song tuyến tính hoặc bộ chuyển đổi đa phương thức.

- Dự đoán câu trả lời (Answer Prediction): dự đoán câu trả lời như một nhiệm vụ phân loại (từ một tập hợp cố định các câu trả lời chung) hoặc một nhiệm vụ tạo sinh (cho các câu trả lời mở).



Hình 1. *Kiến trúc hệ thống hỏi đáp trực quan*

Để phục vụ cho hệ thống, nhiều tập dữ liệu đã được phát triển để nghiên cứu các nhiệm vụ trả lời câu hỏi trực quan. Các tập dữ liệu này chủ yếu liên quan đến việc trả lời các câu hỏi bằng lời nói bằng cách sử dụng cảnh trực quan làm bối cảnh. Một số mô hình ngôn ngữ trực quan (VL) đã được phát triển cho các nhiệm vụ VQA và được đánh giá theo kết quả trên các tập dữ liệu này [10].

2.2. Hỏi đáp hiện thân

EQA là một nhiệm vụ AI nâng cao, trong đó tác nhân phải chủ động khám phá môi trường 3D để trả lời câu hỏi thay vì dựa vào hình ảnh tĩnh như trong VQA truyền thống. Tác nhân được "trực quan hóa", nghĩa là nó có sự hiện diện mô phỏng hoặc vật lý trong một không gian và phải sử dụng khả năng định hướng, thị giác và lý luận để tìm ra câu trả lời.

- Hiểu ngôn ngữ (Language Understanding): phân tích câu hỏi và hiểu đối tượng/vị trí mục tiêu.

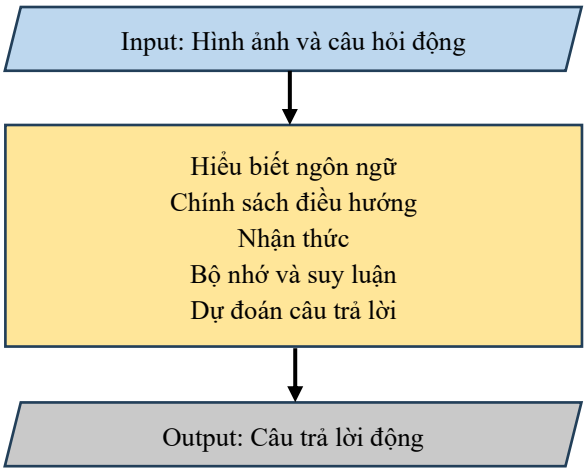
- Chính sách điều hướng (Navigation Policy): quyết định hướng đi dựa trên các quan sát hiện tại và câu hỏi.

- Nhận thức (Perception): sử dụng đầu vào thị giác (các cảm biến giống camera) để phát hiện và nhận dạng vật thể, kết cấu, màu sắc, v.v.

- Bộ nhớ và suy luận (Memory and reasoning): xây dựng nhận thức không gian và ghi nhớ những địa điểm đã đến hoặc vật thể đã nhìn thấy trước đó.

- Tạo câu trả lời (Answer generation): tổng hợp các phát hiện thành câu trả lời bằng ngôn ngữ tự nhiên.

Một số mô hình đã được phát triển cho các nhiệm vụ EQA hiện có. Ví dụ, Das và cộng sự [2] đã giới thiệu một mô hình mô-đun để học chính sách nhằm điều hướng và trả lời câu hỏi bằng lời, trong khi Gao và cộng sự [11] đã sử dụng một mô hình dựa trên bộ biến đổi để tạo ra các mã thông báo bộ nhớ làm manh mối khám phá. Các mô hình này nhằm mục đích phát triển một chính sách điều hướng để trả lời các câu hỏi bằng lời.



Hình 2. *Kiến trúc hệ thống hỏi đáp hiện thân*

Hầu hết các nghiên cứu VQA và EQA hiện tại tập trung vào việc hiểu các câu hỏi bằng lời, trái ngược với mục tiêu của tác giả là hiểu biết một cách toàn diện các biểu thức đa phương thức (lời nói và cử chỉ) trong các bối cảnh cụ thể. Hơn nữa, các mô hình hiện tại hợp nhất các cấu trúc nhúng khác nhau (biểu diễn bằng lời liên tục và rời rạc), có khả năng dẫn đến các biểu diễn ngôn ngữ trực quan VL (Visual-Language) không tối ưu.

3. Giải pháp đề xuất

3.1. Các nhiệm vụ liên quan đến EQA

Trên cơ sở nghiên cứu các kết quả đã có, tác giả đề xuất 8 nhiệm vụ EQA mới bao gồm: dự đoán sự tồn tại của đối tượng, xác định đối tượng, xác định góc nhìn để nhận biết đối tượng, xác định số lượng đối tượng, truy vấn thuộc tính đối tượng, so sánh thuộc tính đối tượng, xác định nền tảng quan điểm và nền tảng quan hệ. Các nhiệm vụ ở đây cũng tương tự như các nhiệm vụ đã được các tác giả/nhóm nghiên cứu khác phát triển trong các công trình trước đây [10]. Tuy nhiên, những nhiệm vụ đó chỉ liên quan đến các câu hỏi bằng lời. Trong nghiên cứu này tác giả đề xuất việc thiết kế các nhiệm vụ QA trong các bối cảnh có thực, trong đó đối tượng đại diện của con người được phép đặt câu hỏi bằng cách sử dụng các câu nói bằng lời và cử chỉ không bằng lời trong môi trường ảo.

Các nhiệm vụ đề xuất cụ thể như sau:

- Dự đoán sự tồn tại (Existence Prediction - EP): nhiệm vụ này liên quan đến việc xác định xem trong khung hình có chứa một đối tượng cụ thể với các thuộc tính cụ thể nào đó (ví dụ: màu sắc, hình dáng, vị trí) hay không. Để thực hiện nhiệm vụ này, đòi hỏi kiến thức về hình dạng của đối tượng cũng như hiểu biết toàn diện về khung hình.

- Nền tảng đối tượng (Object Grounding - OG): trong nhiệm vụ này, danh mục đối tượng được xác định dựa trên câu hỏi bằng cách sử dụng các biểu thức đa phương thức. Nhiệm vụ này cũng liên quan đến việc hiểu câu hỏi được hỏi từ góc nhìn nào (tức là người nói, người quan sát, người trung lập).

- Xác định góc nhìn đối lượng (Perspective-Aware Object Grounding - POG): tương tự như nhiệm vụ nền tảng đối tượng, nhiệm vụ này liên quan đến việc xác định đối tượng nào đang được nhắc đến. Tuy nhiên, nhiệm vụ này cũng bao gồm góc nhìn bằng lời trong câu hỏi (người nói, người quan sát, người trung lập). Điều này được thực hiện có chủ đích để xác định mức độ tác động của góc nhìn bằng lời đến hiệu suất xác định đúng đối tượng.

- Đếm đối tượng (Object Counting - OC): trong nhiệm vụ này, số lượng đối tượng trong một khung hình được hỏi dựa trên các mối quan hệ không gian khác nhau. Để hiểu điều này, phải chú ý đến các đối tượng khác nhau trong khung hình trực quan và phải sử dụng các mối quan hệ không gian được đưa ra trong câu hỏi bằng lời để xác định xem một số đối tượng nhất định có thuộc tính đó hay không trước khi đếm.

- Truy vấn thuộc tính đối tượng (Object Attribute Query - OAQ): nhiệm vụ này liên quan đến việc xác định các thuộc tính (hình dáng, màu sắc, vị trí…) của một đối tượng nhất định được truy vấn. Điều này có thể hữu ích trong các tình huống mà con người muốn tìm hiểu các đặc điểm cụ thể của một đối tượng. Vị trí không gian và màu sắc của đối tượng phải được xác định bằng cách sử dụng các biểu thức bằng lời và không bằng lời đã cho trước.

- So sánh thuộc tính đối tượng (Object Attribute Compare - OAC): nhiệm vụ này bao gồm việc so sánh các thuộc tính của hai đối tượng, kể cả việc chỉ vào một đối tượng và truy vấn tính tương đồng của chúng trong các thuộc tính cho trước.

- Nền tảng quan điểm (Perspective Grounding - PG): hiểu được quan điểm bằng lời của con người là rất quan trọng đối với giao tiếp hiệu quả giữa con người và hệ thống AI, vì con người mô tả các đối tượng từ các góc nhìn khác nhau. Chúng ta có thể mô phỏng điều này bằng cách sử dụng ba góc nhìn (trung lập, người nói và người quan sát), giao cho mô hình nhiệm vụ xác định góc nhìn của một câu hỏi nhất định.

- Nền tảng quan hệ (Relation Grounding - RG): nhiệm vụ này bao gồm việc xác định xem một phát ngôn bằng lời và cử chỉ không bằng lời có đề cập đến cùng một đối tượng hay không. Vì một biểu thức tham chiếu có thể được diễn giải khác nhau từ các góc nhìn trực quan và bằng lời khác nhau, nên việc hiểu nhiệm vụ này đòi hỏi phải có suy luận phức tạp về góc nhìn và mối quan hệ không gian.

3.2. Xây dựng tập dữ liệu EQA

Để thực hiện được các nhiệm vụ trên, thì dữ liệu đóng vai trò cực kỳ quan trọng. Dữ liệu đóng vai trò trung tâm trong hệ thống EQA theo ba cách chính:

- Huấn luyện tác nhân: các tác nhân cần dữ liệu phong phú, được chú thích để có thể: hiểu được ngôn ngữ (hiểu các câu hỏi và sinh câu trả lời); giúp nhận thức môi trường (hình ảnh, khung cảnh 3D); hỗ trợ thực hiện hành động (điều hướng/di chuyển); kết hợp nhận thức và suy luận để đưa ra câu trả lời phù hợp.

- Giúp nhận biết nền tảng ngôn ngữ trong môi trường: dữ liệu cho phép nền tảng ngôn ngữ có thể kết nối các từ và cụm từ (như "nhà bếp" hoặc "phía sau ghế") với các đối tượng trực quan, không gian và địa điểm.

- Đánh giá các tác nhân: bộ dữ liệu EQA cũng được sử dụng để đánh giá mức độ hiệu quả của tác nhân trong việc thực hiện các tác vụ đầu cuối như điều hướng và trả lời chính xác các câu hỏi.

Sau khi bộ dữ liệu được tạo, nó được sử dụng trong các giai đoạn huấn luyện, xác thực và kiểm tra:

- Huấn luyện: sử dụng các phương pháp học có giám sát (Supervised Learning) hoặc học tăng cường (Reinforcement Learning) để giúp các tác nhân học cách liên kết các câu hỏi và đầu vào trực quan với các hành động điều hướng và đưa ra trả lời. Bộ dữ liệu giúp liên kết ngôn ngữ với các tín hiệu thị giác và chuyển động.

- Mô phỏng trong quá trình huấn luyện: mô phỏng tương tác của tác nhân trong môi trường bằng cách sử dụng bộ dữ liệu và đưa ra sự công nhận/phần thưởng (rewards) khi thiết bị đến đúng vị trí và đưa ra câu trả lời đúng.

- Đánh giá: kiểm tra tác nhân với các câu hỏi/môi trường chưa có trong bộ dữ liệu.

- Đo lường: dùng để đo lường sự thành công điều hướng (có đến đúng vị trí hay không), độ chính xác của câu trả lời và hiệu quả (các bước đã thực hiện, thời gian).

- Tinh chỉnh và so sánh: so sánh các mô hình sử dụng bộ dữ liệu EQA tiêu chuẩn như: EQA v1 (AI Facebook), EmbodiedQA (AI2), CLEVR-Nav, THOR-QA, Room-to-Room (R2R) …

Có nhiều cách khác nhau để xây dựng bộ dữ liệu EQA và nó là một quá trình phức tạp và đa phương thức. Quá trình này bao gồm:

- Môi trường mô phỏng hoặc thực tế: sử dụng các trình mô phỏng 3D (ví dụ: AI2-THOR, Habitat, Matterport3D) để tạo nhà, văn phòng ảo và các môi trường khác.

- Chú thích cảnh: gắn nhãn cho các đối tượng, vị trí và thuộc tính của chúng (ví dụ: "cốc đỏ trên bàn", vị trí của đối tượng).

- Thiết kế câu hỏi: các câu hỏi phải đòi hỏi sự khám phá (không chỉ quan sát từ góc nhìn tĩnh). Ví dụ: "Gối trên ghế sofa màu gì?", "Phòng ngủ có TV không?",…

- Chú thích đường dẫn và hành động: chú thích các đường dẫn điều hướng tối ưu, vị trí đối tượng và góc nhìn cần thiết để trả lời câu hỏi.

- Chú thích câu trả lời: đưa ra câu trả lời đúng cho mỗi câu hỏi dựa trên bằng chứng trực quan thực tế.

- Tính biến thiên và cân bằng: bao gồm nhiều môi trường, loại câu hỏi và loại câu trả lời khác nhau để tránh sai lệch dữ liệu.

Để xây dựng tập dữ liệu EQA, tác giả đã tiến hành tổng hợp từ một số nguồn có sẵn như Bảng 1.

Trong nghiên cứu này, tác giả đã kết hợp các bộ dữ liệu này, chọn lọc và gán nhãn lại để tạo dữ liệu cho các tác vụ EQA khác nhau. Để tăng khả năng khái quát hóa của tập dữ liệu, tác giả đã sử dụng nhiều môi trường khác nhau về góc nhìn camera, vị trí đối tượng và biểu cảm phi ngôn ngữ/bằng lời. Trong mỗi cảnh trực quan, tác giả tạo ra bốn tình huống khác nhau:

- Tình huống không có con người và do đó không có biểu cảm phi ngôn ngữ;

- Tình huống có ánh mắt nhìn của con người;

- Tình huống có cử chỉ chỉ của con người;

- Tình huống liên quan đến con người, ở đó sử dụng ánh mắt nhìn của con người và cử chỉ chỉ tay để phục vụ điều hướng.

Các biểu cảm phi ngôn ngữ được tạo ra bao gồm cử chỉ chỉ tay và ánh mắt. Cử chỉ chỉ tay được tạo theo thủ tục bằng cách sử dụng động học nghịch đảo thông qua công cụ Unity.

Sau khi tiến hành xây dựng bộ dữ liệu, tác giả đã phân chia các dữ liệu thành 03 tập dùng để huấn luyện, kiểm chứng và đánh giá với kết quả cụ thể như Bảng 2.

Bảng 1. *Các nguồn dữ liệu và mô tả*

STT	Tên dữ liệu và mô tả
1	**EmbodiedQA (Facebook AI Research)** - Môi trường: AI2-THOR / House3D; - Nhiệm vụ: Điều hướng và trả lời câu hỏi trực quan; - Câu hỏi: Dựa trên mẫu; - Kích thước: ~5.000 câu hỏi trên ~750 cảnh; - Người chơi: Bắt đầu ở một vị trí ngẫu nhiên; phải điều hướng để trả lời; - Định dạng: Bao gồm chế độ xem RGB, không gian hành động, câu hỏi và câu trả lời.
2	**Matterport3D (R2R Stanford)** - Môi trường: Không gian trong nhà được quét thực tế (Matterport3D) - Nhiệm vụ: Điều hướng bằng thị giác và ngôn ngữ (VLN) - Câu hỏi: Hướng dẫn điều hướng - Kích thước: ~21.000 quỹ đạo điều hướng - Sử dụng trong EQA: Có thể mở rộng để điều hướng dựa trên câu hỏi
3	**ALFRED – Viện Allen về AI** - Môi trường: AI2-THOR (môi trường tương tác) - Nhiệm vụ: Thực hiện theo hướng dẫn với thao tác đối tượng - Thành phần QA: Không phải QA truyền thống, nhưng bao gồm hiểu ngôn ngữ và tư duy trực quan - Sử dụng trong EQA: Phù hợp cho việc lập kế hoạch và tư duy nhiệm vụ cụ thể

Bảng 2. *Dữ liệu theo các nhiệm vụ EQA*

Loại dữ liệu	Dành cho nhiệm vụ EQA							
	EP	OG	POG	OC	OAQ	OAC	PG	RG
Huấn luyện	106	106	106	106	106	28	78	34
Kiểm chứng	12	12	12	12	12	18	35	15
Đánh giá	12	12	12	12	12	18	35	15

Trong bộ dữ liệu này, bao gồm các khung hình, thông tin chú thích về các góc nhìn và các câu hỏi cho từng nhiệm vụ.

Ví dụ, các khung hình với các góc nhìn khác nhau như Hình 3.

Hình 3. *Ví dụ các khung hình theo các góc nhìn khác nhau*

Tương ứng với các khung hình, sẽ có các câu hỏi theo từng nhiệm vụ được thiết kế. Ví dụ, dưới đây là một số câu hỏi tương ứng với các khung hình trên, Bảng 3.

Bảng 3. *Ví dụ các câu hỏi kèm theo khung hình*

Nhiệm vụ	Ví vụ về câu hỏi
EP	Có quả dưa chuột nào trong cảnh này không? Có quả dưa chuột xanh nào trong cảnh này không?
OG	Tên của vật đó là gì? Vật màu vàng kia tên là gì? Vật màu vàng bên phải tên là gì? Vật màu vàng bên phải bên phải miếng pho mát màu vàng tên là gì?
POG	Xét theo góc nhìn của người quan sát, tên của vật thể đó là gì? Xét theo góc nhìn của người quan sát, tên của vật màu vàng đó là gì? Xét theo góc nhìn của người nói, tên của vật màu vàng bên phải đó là gì?
OC	Có bao nhiêu vật thể ở phía trên vật thể đó? Còn lại bao nhiêu vật thể của vật màu vàng?
OAQ	Màu sắc của vật/đồ vật đó là gì? Màu sắc của hộp đựng xà phòng rửa tay là gì?
OAC	Màu của thứ đó có giống màu của miếng pho mát không? Màu của hộp đựng xà phòng rửa tay đó có giống màu của chai nước ngọt không?
PG	Bình đựng xà phòng rửa tay phía trên chai nước ngọt. Vật thể này được mô tả theo góc nhìn nào?
RG	Bình đựng xà phòng rửa tay phía trên quả dưa chuột, liệu vật thể này có được nhắc đến một cách chính xác không? Xét theo góc nhìn của người quan sát, bình đựng xà phòng rửa tay phía dưới quả dưa chuột, liệu vật thể này có được nhắc đến một cách chính xác không? Xét theo góc nhìn của người quan sát, bình đựng xà phòng rửa tay bên cạnh máy pha cà phê, liệu vật thể này có được nhắc đến một cách chính xác không?

4. Thử nghiệm và đánh giá

Để triển khai thử nghiệm, tác giả đã chuẩn bị dữ liệu và các môi trường, chương trình tương ứng được lưu trữ ở các thư mục sau:

- Dữ liệu: bao gồm các khung nhìn theo các góc nhìn khác nhau, mô tả thông tin đi kèm theo các khung hình và dữ liệu bao gồm các câu hỏi và phương án trả lời.

- Huấn luyện: Chúng ta cần phải tái tạo các thiết lập thử nghiệm và môi trường huấn luyện trên bộ dữ liệu được xây dựng. Ở địa chỉ dưới đây, tác giả chia sẻ nội dung Singularity Container được xây dựng từ cùng một Docker để sử dụng cho thử nghiệm. Singularity Container ở đây là một công nghệ ảo hóa nhẹ (lightweight container) được thiết kế chuyên biệt cho mục đích nghiên cứu và thử nghiệm.

- Kiểm tra: thử nghiệm và đánh giá mô hình đã được huấn luyện cho các nhiệm vụ dự đoán cho trước với một số ngữ cảnh cụ thể. Việc này khi tiến hành thực nghiệm tương đối khó vì phải cài đặt môi trường và chương trình thử nghiệm tương ứng.

Tác giả đã thử nghiệm bộ dữ liệu trên nhiều mô hình khác nhau để so sánh hiệu suất. Kết quả là khác nhau khi thử nghiệm trên nhiều mô hình khác nhau như Dual Encoder [12], CLIP [13] và ViLT [14]. Các mô hình ngôn ngữ trực quan (Visual Language) hiện có cho các tác vụ QA được thiết kế để trả lời một câu hỏi bằng một ngữ cảnh trực quan duy nhất.

Đối với mô hình Dual-Encoder (ViT+BERT), tác giả trích xuất độc lập các biểu diễn trực quan cho từng chế độ xem bằng cách sử dụng một mô hình ViT được chia sẻ và các biểu diễn bằng lời nói bằng cách sử dụng một mô hình BERT. Tác giả hợp nhất các biểu diễn bằng lời nói và trực quan này để tạo ra các biểu diễn nhiệm vụ.

Đối với các mô hình CLIP, tác giả ghép từng chế độ xem trực quan với một câu hỏi bằng lời và chuyển câu hỏi này qua mô hình để trích xuất nhiều biểu diễn bằng lời nói và trực quan rồi hợp nhất chúng để tạo ra các biểu diễn nhiệm vụ.

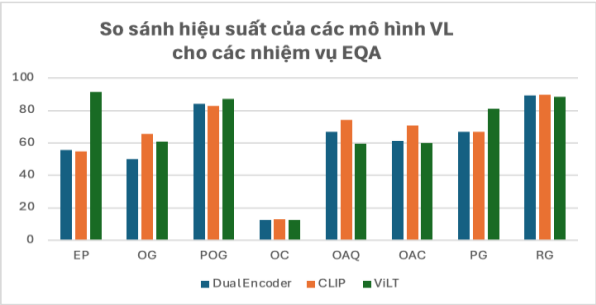
Đối với ViLT, tác giả sử dụng ResNet-101 để trích xuất các biểu diễn trực quan được truyền qua mô hình với các nhúng bằng lời để tạo ra các biểu diễn nhiệm vụ. Kết quả thu được biểu diễn qua Bảng 4.

Bảng 4. *Kết quả trả lời đúng khi thử nghiệm trên các mô hình*

	EP	OG	POG	OC	OAQ	OAC	PG	RG
Dual Encoder	55,7	49,9	84,2	12,3	66,9	61,4	66,7	89,3
CLIP	54,7	65,4	82,7	13,1	74,3	70,5	66,9	89,9
ViLT	91,5	61,0	87,4	12,5	59,4	60,1	81,2	88,6

Hình 4 so sánh hiệu suất khi dùng bộ dữ liệu trên các mô hình khác nhau.

Kết quả này cho thấy, hiệu suất là khá tương đồng trên các mô hình khác nhau. Tuy nhiên, việc đếm các đối tượng cho kết quả chính xác khá thấp.



Hình 4. *Đồ thị so sánh độ chính xác theo mô hình*

5. Kết luận

Bài báo này trình bày những nghiên cứu quan trọng trong lĩnh vực EQA, một nhiệm vụ AI mới, trong đó một tác nhân phải điều hướng và khám phá môi trường 3D một cách thông minh để thu thập thông tin và trả lời các câu hỏi phụ thuộc vào tình huống. EQA đòi hỏi một tập hợp phức tạp các kỹ năng AI, bao gồm hiểu ngôn ngữ, nhận dạng hình ảnh, nhận thức chủ động, điều hướng theo mục tiêu, suy luận, trí nhớ dài hạn và ngôn ngữ nền tảng trong hành động.

Đóng góp cốt lõi của nghiên cứu này là việc phát triển một tập dữ liệu quy mô lớn được thiết kế đặc biệt để giải quyết những thiếu sót của các tập dữ liệu EQA hiện có. EQA và hỏi đáp trực quan (VQA) truyền thống thường tập trung vào các câu hỏi bằng lời nói và quan điểm đơn lẻ. Tập dữ liệu của tác giả mở rộng hơn nữa bằng cách kết hợp các biểu thức đa phương thức (lời nói và cử chỉ phi ngôn ngữ) và xem xét rõ ràng nhiều góc nhìn (người nói, người quan sát, người trung lập, quan điểm của tác nhân, quan điểm bên ngoài, quan điểm từ trên xuống) để hiểu toàn diện hơn về các tương tác trong thế giới thực.

Để tạo điều kiện thuận lợi cho việc này, tác giả đã đề xuất tám nhiệm vụ EQA mới đòi hỏi khả năng suy luận phức tạp về các yếu tố đầu vào đa phương thức và các mối quan hệ không gian gồm dự đoán sự tồn tại (EP), xác định đối tượng (OG), xác định đối tượng nhận thức góc nhìn (POG), đếm Đối tượng (OC, truy vấn thuộc tính đối tượng (OAQ, so sánh thuộc tính đối tượng (OAC, nền tảng quan điểm (PG) và nền tảng quan hệ (RG.

Việc xây dựng tập dữ liệu này bao gồm tổng hợp và dán nhãn lại dữ liệu từ các nguồn hiện có như EmbodiedQA (Nghiên cứu AI của Facebook), Matterport3D (R2R Stanford) và ALFRED (Viện Allen về AI). Tác giả tạo ra bốn tình huống riêng biệt cho mỗi cảnh thị giác, tích hợp các biểu hiện phi ngôn ngữ như ánh mắt của con người và các cử chỉ trỏ được tạo theo thủ tục bằng cách sử dụng động học nghịch đảo. Tập dữ liệu bao gồm các khung hình được chú thích từ nhiều góc nhìn, câu hỏi và câu trả lời tương ứng, cung cấp dữ liệu phong phú cho việc đào tạo và đánh giá tác nhân.

Đánh giá thử nghiệm của tác giả đã chứng minh tính hữu ích của tập dữ liệu này trên nhiều mô hình ngôn ngữ thị giác (VL) khác nhau, bao gồm bộ mã hóa kép, CLIP và ViLT. Mặc dù kết quả phần lớn tương đương nhau giữa các mô hình này, cho thấy khả năng ứng dụng rộng rãi của tập dữ liệu, nhưng kết quả quan sát cho thấy các tác vụ đếm đối tượng mang lại độ chính xác thấp đáng kể. Điều này cho thấy một thách thức cụ thể cần được nghiên cứu sâu hơn và tinh chỉnh mô hình.

Nhìn chung, việc phát triển tập dữ liệu EQA quy mô lớn, đa phương thức và đa góc nhìn này là tiền đề quan trọng cho việc tiếp tục nghiên cứu liên quan đến EQA và trí tuệ nhân tạo hiện thân. Nó cung cấp một nền tảng vững chắc để phát triển và đánh giá các tác nhân AI có khả năng tương tác tinh vi hơn và giống con người hơn trong môi trường 3D phức tạp.

TÀI LIỆU THAM KHẢO

[1] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A Survey of Embodied AI: From Simulators to Research Tasks", *in IEEE Transactions on Emerging Topics in Computational Intelligence,* Vol. 6, No. 2, pp. 230-244, 2022, https://doi.org/10.48550/arXiv.2103.04918

[2] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh and D. Batra, "Embodied Question Answering", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA*, 2018, pp. 2135-2159, https://doi.org/10.48550/arXiv.1711.11543

[3] F. Ziliotto, T. Campari, L. Serafini, L. Ballan, "TANGO: Training-free Embodied AI Agents for Open-world Tasks", *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR),* 2025. pp. 24603-24613, https://doi.org/10.48550/arXiv.2412.10402

[4] C. L. Fan, H. M. Sokolowski, R. S. Rosenbaum, B. Levine, "What about "space" is important for episodic memory?", *WIREs Cogn Sci.,* Vol. 14, Issue3, Special Issue: Autobiographical Memory, pp. 1-21, 2023, https://doi.org/10.1002/wcs.1645

[5] J. Whittington, W. Dorrell, T. Behrens, S. Ganguli, and M. El-Gaby, "A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps", *Neuron*, Vol. 113, no. 2, pp. 321-333, 2025, https://doi.org/10.1016/j.neuron.2024.10.017

[6] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks", *Proceedings of the 33rd International Conference on Neural Information Processing Systems,* 2019, pp. 13–23, https://dl.acm.org/doi/10.5555/3454287.3454289

[7] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo and S. Gould, "VLN-BERT: A Recurrent Vision-and-Language BERT for Navigation", *Proceedings 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2021, pp. 1643-1653, https://10.1109/CVPR46437.2021.00169

[8] Z. Wang, M. Wu, Y. Cao, Y. Ma, M. Chen, and T. Tuytelaars, "*Navigating the Nuances: A Fine-grained Evaluation of Vision-Language Navigation",* Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, pp. 4681–4704, 2024, https://doi.org/10.48550/arXiv.2409.17313

[9] Y. Chen et al., "Embodied reference understanding with language and gesture", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1385–1395, https://yixchen.github.io/YouRefIt

[10] B. Kim, J. Kim, D. Lee, and B. Jang, "Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges", *ACM Computing Surveys,* Vol. 57, No. 10, pp. 1-35, 2025, https://doi.org/10.1145/3728635

[11] C. Gao, J. Chen, S. Liu, L. Wang, Q. Zhang, and Q. Wu, "Room-and-object aware knowledge reasoning for remote embodied referring expression", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3064–3073, https://openaccess.thecvf.com/content/CVPR2021

[12] A. Dosovitskiy *et al.,* "An image is worth 16x16 words: Transformers for image recognition at scale", *Proceedings of the International Conference on Learning Representations (ICLR),* 2021, https://arxiv.org/pdf/2010.11929/1000

[13] A. Radford *et al.,* "Learning Transferable Visual Models From Natural Language Supervision", *Proceedings of the 38th International Conference on Machine Learning, PMLR*, Vol. 139, pp. 8748-8763, 2021, https://proceedings.mlr.press/v139/radford21a

[14] S. Kim, J. Nam, and B. Ko, "Vit-net: Interpretable vision transformers with neural tree decoder", *International conference on machine learning, PMLR,* Vol. 162, pp. 11162-11172, 2022, https://proceedings.mlr.press/v162/kim22g.html