EVCAP: Retrieval-Augmented Image Captioning with External Visual–Name Memory for Open-World Comprehension

Jiaxuan Li[1*], Duc Minh Vo[1*], Akihiro Sugimoto[2], Hideki Nakayama[1]
[1]The University of Tokyo, Japan [2]National Institute of Informatics, Japan
{li,vmduc}@nlab.ci.i.u-tokyo.ac.jp sugimoto@nii.ac.jp nakayama@ci.i.u-tokyo.ac.jp

Abstract

*Large language models (LLMs)-based image captioning has the capability of describing objects not explicitly observed in training data; yet novel objects occur frequently, necessitating the requirement of sustaining up-to-date object knowledge for open-world comprehension. Instead of relying on large amounts of data and/or scaling up network parameters, we introduce a highly effective retrieval-augmented image captioning method that prompts LLMs with object names retrieved from External Visual–name memory (*EVCAP*). We build ever-changing object knowledge memory using objects' visuals and names, enabling us to (i) update the memory at a minimal cost and (ii) effortlessly augment LLMs with retrieved object names by utilizing a lightweight and fast-to-train model. Our model, which was trained only on the COCO dataset, can adapt to out-of-domain without requiring additional fine-tuning or re-training. Our experiments conducted on benchmarks and synthetic commonsense-violating data show that* EVCAP*, with only 3.97M trainable parameters, exhibits superior performance compared to other methods based on frozen pre-trained LLMs. Its performance is also competitive to specialist SOTAs that require extensive training.*

1. Introduction

Advanced image captioning based on large language models (LLMs) [3, 8, 9, 25] has focused on the approach using big-scale models trained on ever-increasingly large-scale datasets, which is no longer viable. This is because the computational cost to train the models increases exponentially and, more importantly, updating training data is almost impossible to keep pace with the growth of novel objects in our daily lives. Sustaining ever-changing object knowledge with a reasonable cost is a pressing concern in LLMs-based models to truly unlock open-world comprehension.

Retrieval-augmented image captioning [20, 35] is emerging as an alternative since it considerably reduces training costs in both time and data while producing encouraging results. Nonetheless, with their huge datastore, it is obvious that LLMs would imitate the given texts, limiting their ability to describe open-world objects properly. For instance, SmallCap [35] considers the words *"skateboard"* and *"wooden floor"* to be a pair regardless of visual appearances containing a commonsense-violating pair

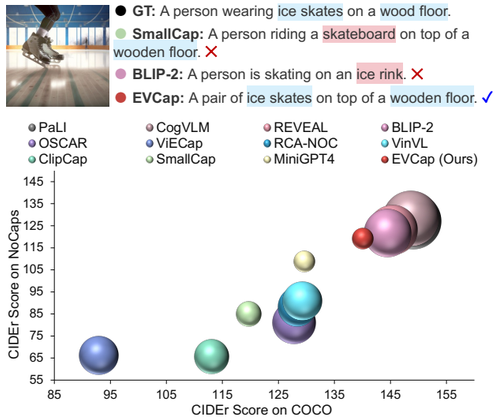


Figure 1. Overall comparison of our EVCAP and SOTAs. (Upper) Generated captions by SmallCap, BLIP-2, and our EVCAP for a commonsense-violating image from the WHOOPS dataset. × and ✓ indicate incorrect and correct predictions, respectively. Incorrect objects in captions are highlighted in red, while correct ones are in blue. SmallCap and BLIP-2 give incorrect predictions for *"ice skates"* and *"wood floor"*, respectively, while our EVCAP utilizes an external visual–name memory to enhance attention to objects within the image, leading to superior performance for image captioning. (Lower) Comparison of the number of trainable parameters, CIDEr score on COCO and NoCaps datasets. The size of each circle reflects the log number of trainable parameters. EVCAP (3.97M) has less trainable parameters than others while achieving comparable results with SOTAs at scale.

*Equal contributions. Code is available at https://jiaxuan-li.github.io/EVCap.

EVCAP: Chú thích hình ảnh tăng cường truy xuất với Bộ nhớ Tên-Hình ảnh Bên ngoài để Hiểu Thế giới Mở

Gia Huyền Lý1 , Đức Minh Võ1 , Akihiro Sugimoto2 , Hideki Nakayama1
1Đại học Tokyo, Nhật Bản 2Viện Tin học Quốc gia, Nhật Bản
{li,vmduc}@nlab.ci.iu-tokyo.ac.jp sugimoto@nii.ac.jp nakayama@ci.iu-tokyo.ac.jp

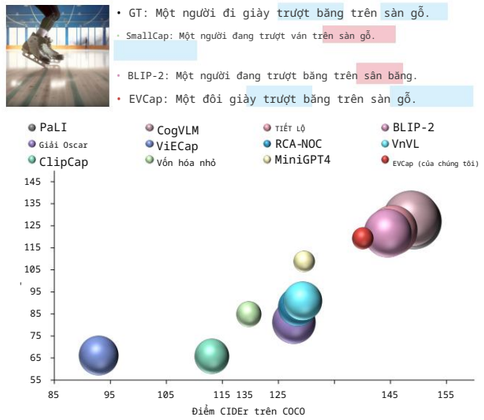
Tóm tắt

Chú thích hình ảnh dựa trên mô hình ngôn ngữ lớn (LLM)
có khả năng mô tả các đối tượng không được quan sát rõ ràng trong dữ
liệu đào tạo; tuy nhiên các đối tượng mới thường xuyên xuất hiện,
đòi hỏi phải duy trì kiến thức về đối tượng được cập nhật để hiểu thế
giới mở. Thay vì
dựa vào lượng lớn dữ liệu và/hoặc mở rộng các tham số mạng, chúng tôi
giới thiệu một phương pháp chú thích hình ảnh tăng cường truy xuất có
hiệu quả cao giúp thúc đẩy LLM
với tên đối tượng được lấy từ External Visual-name
bộ nhớ (EVCAP). Chúng tôi xây dựng bộ nhớ kiến thức đối tượng luôn thay
đổi bằng cách sử dụng hình ảnh và tên của đối tượng, cho phép chúng tôi
(i) cập nhật bộ nhớ với chi phí tối thiểu và (ii) tăng cường LLM một
cách dễ dàng với các tên đối tượng được truy xuất bằng cách sử dụng mô
hình nhẹ và đào tạo nhanh. Mô hình của chúng tôi,
được đào tạo chỉ trên tập dữ liệu COCO, có thể thích ứng
ra khỏi miền mà không cần phải tinh chỉnh thêm
hoặc đào tạo lại. Các thí nghiệm của chúng tôi được tiến hành trên các chuẩn mực
và dữ liệu vi phạm lẽ thường tổng hợp cho thấy EV- CAP, với chỉ 3,97
triệu tham số có thể đào tạo, thể hiện hiệu suất vượt trội so với các
phương pháp khác dựa trên
LLM được đào tạo trước đông lạnh. Hiệu suất của nó cũng cạnh tranh với
các SOTA chuyên ngành đòi hỏi đào tạo chuyên sâu.

1. Giới thiệu

Việc chú thích hình ảnh nâng cao dựa trên các mô hình ngôn ngữ lớn
(LLM) [3, 8, 9, 25] đã tập trung vào phương pháp sử dụng
các mô hình quy mô lớn được đào tạo trên quy mô ngày càng lớn
bộ dữ liệu, không còn khả thi nữa. Điều này là do chi phí tính toán để
đào tạo các mô hình tăng theo cấp số nhân
và quan trọng hơn, việc cập nhật dữ liệu đào tạo gần như không thể
theo kịp tốc độ phát triển của các đối tượng mới trong
cuộc sống hàng ngày của chúng ta. Duy trì kiến thức về đối tượng luôn thay đổi
với chi phí hợp lý là mối quan tâm cấp bách trong LLM dựa trên
mô hình để thực sự mở khóa sự hiểu biết về thế giới mở.

Chú thích hình ảnh được tăng cường truy xuất [20, 35] là nổi lên như một giải pháp thay thế vì nó làm giảm đáng kể
chi phí đào tạo về cả thời gian và dữ liệu trong khi vẫn tạo ra kết
quả đáng khích lệ. Tuy nhiên, với kho dữ liệu khổng lồ của họ,
rõ ràng là các LLM sẽ bắt chước các văn bản đã cho, hạn chế khả năng
mô tả chính xác các đối tượng trong thế giới mở.
Ví dụ, SmallCap [35] coi các từ "ván trượt" và "sàn gỗ" là một cặp bất
kể hình thức trực quan có chứa một cặp vi phạm lẽ thường



Hình 1. So sánh tổng thể EVCAP và SOTA của chúng tôi. (Phía trên)
Tạo chú thích bằng SmallCap, BLIP-2 và EVCAP của chúng tôi cho
hình ảnh vi phạm lẽ thường từ tập dữ liệu WHOOPS. \times và
\checkmark chỉ ra các dự đoán không chính xác và đúng, tương ứng. Các đối
tượng không chính xác trong chú thích được tô sáng màu đỏ, trong khi những cái đúng
có màu xanh lam. SmallCap và BLIP-2 đưa ra dự đoán không chính xác cho
"giày trượt băng" và "sàn gỗ", tương ứng, trong khi EVCAP của chúng tôi
sử dụng trí nhớ hình ảnh bên ngoài-tên để tăng cường sự chú ý đến
các đối tượng trong hình ảnh, dẫn đến hiệu suất vượt trội cho chú thích hình
ảnh. (Thấp hơn) So sánh số lượng có thể đào tạo
tham số, điểm CIDEr trên các tập dữ liệu COCO và NoCaps.
Kích thước của mỗi vòng tròn phản ánh số logarit của các tham số có thể đào tạo được.
EVCAP (3,97M) có ít thông số có thể đào tạo hơn những thông số khác trong khi
đạt được kết quả tương đương với SOTA ở quy mô lớn.

*Đóng góp ngang nhau. Mã có sẵn tại https://jiaxuan-li.github.io/EVCap.

of *"ice skates"* and *"wood floor"* (Fig. 1, upper). Additionally, prompting the LLMs given a lot of retrieved texts becomes cumbersome, requiring more trainable parameters. Fig. 1 (lower) shows that the CIDEr scores obtained by a lightweight SmallCap [35] with 43M trainable parameters are far away from those obtained by a heavy REVEAL [20] with 2.1B trainable parameters. Beyond that, due to the frequent occurrence of new objects, access to their sample texts is not always feasible, making the memory utilized in [20, 35] difficult to grow. We thus aim to streamline the external memory used in previous work [20, 35] by storing a sufficiently small amount of object information. And, of course, not only does the model not stereotype the example sentences, but the number of trainable parameters would be reduced drastically as a result of the causation (Fig. 1).

We follow [13, 40] to construct a key-value memory where the key is represented by object's features, and the value corresponds to object's name. Unlike [13, 40], which rely on object definition as the key, our method leverages the visual appearance of the object as the key because of the abundance of object images readily available on the internet. We propose an external visual–name memory tailored for ease of expansion and cost-effectiveness in upholding up-to-date object information. We present a highly effective retrieval-augmented LLMs-based image captioning method, called EVCAP, that prompts frozen LLMs with object names retrieved from our proposed memory for open-world comprehension. EVCAP contains a frozen image encoder ViT [14] and Q-Former [25] with *trainable* image query tokens for object retrieval, an attentive fusion module, a *trainable* linear layer for mapping between vision and language latent spaces, and a frozen LLM decoder [10] for generating captions. Specifically, the attentive fusion module feeds retrieved object names and visual features into a customized frozen Q-Former using *trainable* object name query tokens to implicitly reduce the presence of superfluous object names. As a result, EVCAP amounts to only 3.97M trainable parameters. Once trained, the model can be adapted to new domains and large-scale data without further fine-tuning or re-training. Our contributions are as follows:

- We provide an extensible external visual–name memory with minimal but useful object information, which enables LLMs-based models to comprehend the open world.
- We present a highly efficacious retrieval-augmented image captioning EVCAP with 3.97M trainable parameters.

On in-/out-domain benchmarks and synthetic commonsense-violating dataset, EVCAP trained solely on COCO dataset competes with other lightweight methods by a margin while being on par with other specialist SOTAs.

2. Related Work

Image captioning aims to describe the contents of a given image. It can be roughly divided into two approaches: non-LLMs-based methods and LLMs-based ones. The former approaches [4, 22, 42] typically employ a visual encoder and a language decoder in an end-to-end fashion to generate captions. However, they are incapable of describing open-world objects. The latter one leverages pre-trained large-scale vision models (CLIP [32], ViT [12]) and LLMs (GPTs [7, 31], T5 [33], LLaMA [37]) by bridging the gap between two modalities using either pre-training with large-scale data or the learned mapper or prompt techniques. LLMs-based models [8, 9, 25, 29] demonstrate advancements in image captioning challenges, allowing the capacity to describe anything as long as pre-trained vision models can recognize it. Our method belongs to the LLMs-based approaches, but instead of relying fully on the pre-trained vision model, we use object names retrieved from the external memory to augment LLMs-based image captioning.

Novel object captioning is a branch of image captioning that describes images containing objects that were not seen during training. Non-LLMs-based methods explore more objects by learning from unpaired image-sentence sources (DCC [19], NOC [39]) or rely on novel object detectors to recognize novel concepts (NBT [28], OSCAR [26] and VinVL [45]). LLMs-based methods such as ViECap [15] leverage the pre-trained CLIP [32] to obtain object entities Nevertheless, the cut-off in training time of the pre-trained object detector or CLIP prevents it from detecting novel objects that arise quickly in reality. Unlike earlier work, we can readily update our recognition of novel concepts by adding them to external memory, ensuring that we keep any new objects from the past and even the future.

Retrieval-augmented image captioning is a recently popular approach that augments the captioning model with retrieved information for better open-world understanding. AoANet [16] uses a memory bank of image-sentence pairs and target words. SmallCap [35] employs image-to-text retrieval to obtain sampled captions from a captions datastore. RA-CM3 [44] retrieves documents from an external memory of a mixture of text and image via a dense multimodal retriever. EXTRA [34] and Re-ViLM [43] exploit the similarity of the input image and vision candidates to retrieve captions. Different from the previous methods, our external memory contains visual–name pairs to avoid redundant information in the external captions/documents. In addition, we use an attentive fusion module to mitigate the effects of irrelevant retrieved object names on caption generation.

3. Proposed EVCAP

3.1. Idea of EVCAP

We aim to build a retrieval-augmented LLMs-based image captioning model with a sufficiently small yet informative external memory. It involves two challenges: (1) constructing an expandable external memory and (2) building an effective LLMs-based model using retrieved object names.

của "giày trượt băng" và "sàn gỗ" (Hình 1, phía trên). Ngoài ra, việc nhắc
nhở LLM khi có nhiều văn bản được lấy ra trở nên cồng kềnh, đòi hỏi nhiều
thông số để đào tạo hơn.
Hình 1 (dưới) cho thấy điểm CIDEr thu được bằng
SmallCap nhẹ [35] với 43M tham số có thể đào tạo
rất xa so với những thứ thu được bằng một REVEAL nặng nề [20]
với 2,1B tham số có thể đào tạo được. Ngoài ra, do
sự xuất hiện thường xuyên của các đối tượng mới, truy cập vào mẫu của chúng
văn bản không phải lúc nào cũng khả thi, làm cho bộ nhớ được sử dụng
trong [20, 35] khó phát triển. Do đó, chúng tôi hướng tới mục tiêu hợp lý hóa
bộ nhớ ngoài được sử dụng trong công trình trước đây [20, 35] bằng cách lưu trữ
một lượng thông tin đối tượng đủ nhỏ. Và, của
Tất nhiên, mô hình không chỉ không rập khuôn ví dụ
câu, nhưng số lượng các tham số có thể đào tạo sẽ là
giảm mạnh do nguyên nhân gây ra (Hình 1).

Chúng tôi theo [13, 40] để xây dựng bộ nhớ khóa-giá trị
trong đó khóa được biểu diễn bằng các đặc điểm của đối tượng và
giá trị tương ứng với tên của đối tượng. Không giống như [13, 40],
dựa vào định nghĩ a đối tượng như là chìa khóa, phương pháp của chúng tôi tận dụng
sự xuất hiện trực quan của đối tượng như là chìa khóa vì
sự phong phú của hình ảnh đối tượng có sẵn trên internet. Chúng tôi đề xuất
một bộ nhớ hình ảnh bên ngoài được thiết kế riêng
để dễ dàng mở rộng và tiết kiệm chi phí trong việc duy trì
thông tin đối tượng cập nhật. Chúng tôi trình bày một chú thích hình ảnh
dựa trên LLMs được tăng cường truy xuất có hiệu quả cao
phương pháp, được gọi là EVCAP, nhắc nhở các LLM đông lạnh với tên đối
tượng được lấy từ bộ nhớ đề xuất của chúng tôi để hiểu thế giới mở. EVCAP
chứa một bộ mã hóa hình ảnh đông lạnh ViT [14] và Q-Former [25] với hình
ảnh có thể đào tạo
truy vấn mã thông báo để truy xuất đối tượng, một mô-đun hợp nhất chú ý,
một lớp tuyến tính có thể đào tạo để lập bản đồ giữa tầm nhìn và
không gian tiềm ẩn ngôn ngữ và bộ giải mã LLM đông lạnh [10] cho
tạo chú thích. Cụ thể, mô-đun hợp nhất chú ý đưa tên đối tượng đã truy xuất
và các đặc điểm trực quan vào một
Q-Former đông lạnh tùy chỉnh bằng cách sử dụng tên đối tượng có thể đào tạo
truy vấn mã thông báo để ngầm giảm sự hiện diện của các tên đối tượng thừa.
Kết quả là, EVCAP chỉ có giá trị
3,97M tham số có thể đào tạo. Sau khi được đào tạo, mô hình có thể được
được điều chỉnh cho phù hợp với các miền mới và dữ liệu quy mô lớn mà không cần thêm
tinh chỉnh hoặc đào tạo lại. Đóng góp của chúng tôi như sau:

- Chúng tôi cung cấp bộ nhớ tên-hình ảnh bên ngoài có thể mở rộng
  với thông tin đối tượng tối thiểu nhưng hữu ích, cho phép các mô hình
  dựa trên LLM hiểu được thế giới mở.
- Chúng tôi trình bày một EVCAP chú thích hình ảnh được tăng cường khả năng
  truy xuất có hiệu quả cao với 3,97 triệu tham số có thể đào tạo được.

Trên các chuẩn mực trong/ngoài miền và tổng hợp
tập dữ liệu vi phạm lẽ thường, EVCAP được đào tạo chỉ trên
Bộ dữ liệu COCO cạnh tranh với các phương pháp nhẹ khác bằng cách
một biên độ nhất định trong khi vẫn ngang bằng với các SOTA chuyên ngành khác.

2. Công trình liên quan

Chú thích hình ảnh nhằm mục đích mô tả nội dung của một
hình ảnh. Nó có thể được chia thành hai cách tiếp cận: không

Các phương pháp dựa trên LLM và các phương pháp dựa trên LLM. Cái trước
các cách tiếp cận [4, 22, 42] thường sử dụng bộ mã hóa trực quan
và một bộ giải mã ngôn ngữ theo kiểu đầu cuối để tạo ra phụ đề. Tuy nhiên,
chúng không có khả năng mô tả
các đối tượng thế giới mở. Cái sau tận dụng các đối tượng được đào tạo trước
mô hình thị giác quy mô lớn (CLIP [32], ViT [12]) và LLM
(GPT [7, 31], T5 [33], LLaMA [37]) bằng cách thu hẹp khoảng cách
giữa hai phương thức sử dụng phương pháp đào tạo trước với dữ liệu quy mô
lớn hoặc phương pháp lập bản đồ hoặc kỹ thuật nhắc nhở đã học.
Các mô hình dựa trên LLM [8, 9, 25, 29] chứng minh những tiến bộ trong các
thách thức về chú thích hình ảnh, cho phép khả năng
để mô tả bất cứ điều gì miễn là các mô hình thị giác được đào tạo trước
có thể nhận ra nó. Phương pháp của chúng tôi thuộc về LLMs-based
cách tiếp cận, nhưng thay vì hoàn toàn dựa vào các phương pháp được đào tạo trước
mô hình thị giác, chúng tôi sử dụng tên đối tượng lấy từ bộ nhớ ngoài để
tăng cường chú thích hình ảnh dựa trên LLM.
Chú thích đối tượng mới là một nhánh của chú thích hình ảnh
mô tả hình ảnh chứa các vật thể không được nhìn thấy
trong quá trình đào tạo. Các phương pháp không dựa trên LLM khám phá thêm
các đối tượng bằng cách học từ các nguồn câu hình ảnh không ghép đôi
(DCC [19], NOC [39]) hoặc dựa vào các máy dò đối tượng mới
để nhận ra các khái niệm mới lạ (NBT [28], OSCAR [26] và
VinVL [45]). Các phương pháp dựa trên LLM như ViECap [15]
tận dụng CLIP được đào tạo trước [32] để có được các thực thể đối tượng
Tuy nhiên, thời gian đào tạo bị cắt giảm của những người được đào tạo trước
máy dò đối tượng hoặc CLIP ngăn không cho nó phát hiện ra vật thể mới lạ
các đối tượng phát sinh nhanh chóng trong thực tế. Không giống như công việc trước đó,
chúng ta có thể dễ dàng cập nhật sự công nhận của chúng ta về các khái niệm mới bằng cách
thêm chúng vào bộ nhớ ngoài, đảm bảo rằng chúng ta giữ lại bất kỳ
những đồ vật mới từ quá khứ và thậm chí cả tương lai.
Chú thích hình ảnh tăng cường truy xuất là một phương pháp phổ biến gần đây
giúp tăng cường mô hình chú thích bằng thông tin được truy xuất lại để hiểu
rõ hơn về thế giới mở.
AoANet [16] sử dụng một ngân hàng bộ nhớ của các cặp hình ảnh-câu
và các từ mục tiêu. SmallCap [35] sử dụng phương pháp truy xuất hình ảnh
thành văn bản để lấy các chú thích mẫu từ kho dữ liệu chú thích.
RA-CM3 [44] lấy các tài liệu từ bộ nhớ ngoài gồm hỗn hợp văn bản và hình
ảnh thông qua một đa phương thức dày đặc
người lấy lại. EXTRA [34] và Re-ViLM [43] khai thác tính tương đồng của hình
ảnh đầu vào và ứng viên thị giác để lấy lại
chú thích. Khác với các phương pháp trước đây, bên ngoài của chúng tôi
bộ nhớ chứa các cặp tên-hình ảnh để tránh thông tin trùng lặp trong các chú
thích/tài liệu bên ngoài. Ngoài ra,
chúng tôi sử dụng một mô-đun hợp nhất chú ý để giảm thiểu tác động của
tên đối tượng không liên quan được lấy ra khi tạo chú thích.

3. Đề xuất EVCAP

3.1. Ý tưởng của EVCAP

Chúng tôi hướng đến mục tiêu xây dựng một hình ảnh dựa trên LLM được tăng cường khả năng truy xuất
mô hình chú thích với kích thước đủ nhỏ nhưng vẫn cung cấp nhiều thông tin
bộ nhớ ngoài. Nó bao gồm hai thách thức: (1) xây dựng bộ nhớ ngoài có thể
mở rộng và (2) xây dựng mô hình hiệu quả dựa trên LLM bằng cách sử dụng tên
đối tượng đã truy xuất.

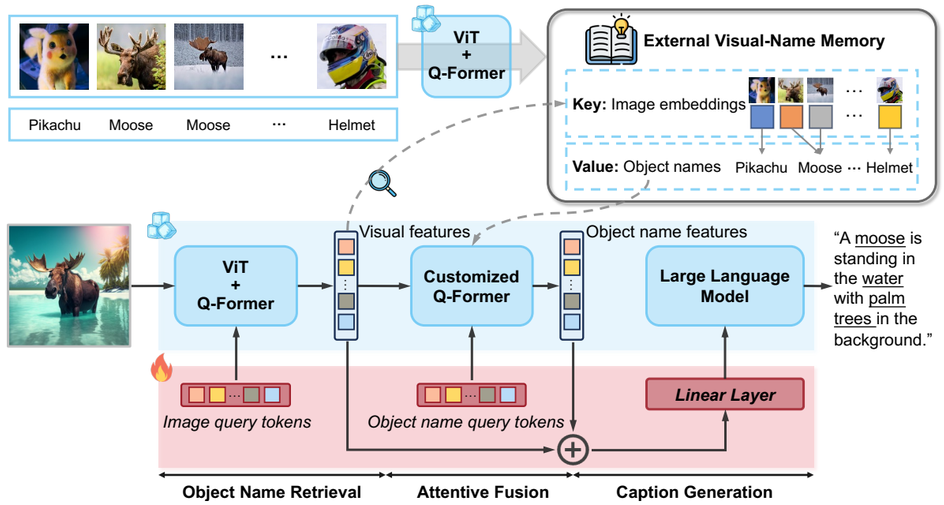


Figure 2. Schematic of our proposed EVCAP. It consists of an external visual–name memory with image embeddings and object names (upper), a frozen ViT and Q-Former equipped with *trainable* image query tokens, an attentive fusion module developed by a customized frozen Q-Former and *trainable* object name query tokens, and a frozen LLM with a *trainable* linear layer (lower). The ViT and Q-Former extract learned visual features from the input image, which are then used to retrieve object names from the external memory. These retrieved object names and learned visual features undergo cross-attention in the customized Q-Former, creating refined object name features. Finally, the object name features combined with visual features are fed into the LLM post a linear layer for generating captions.

As discussed above, challenge (1) can be resolved by utilizing the visual appearance of objects. However, if we restrict our memory to only a visual–name pair for each object, our memory will be lacking in diversity. Therefore, we gather several images for each target object. Additionally, we keep the synthetic images in our memory to avoid the harm that synthetic images might cause to our method, as pointed out in [18]. With the capability to collect images from the internet, EVCAP can be easily expanded to include novel objects from the real world effortlessly.

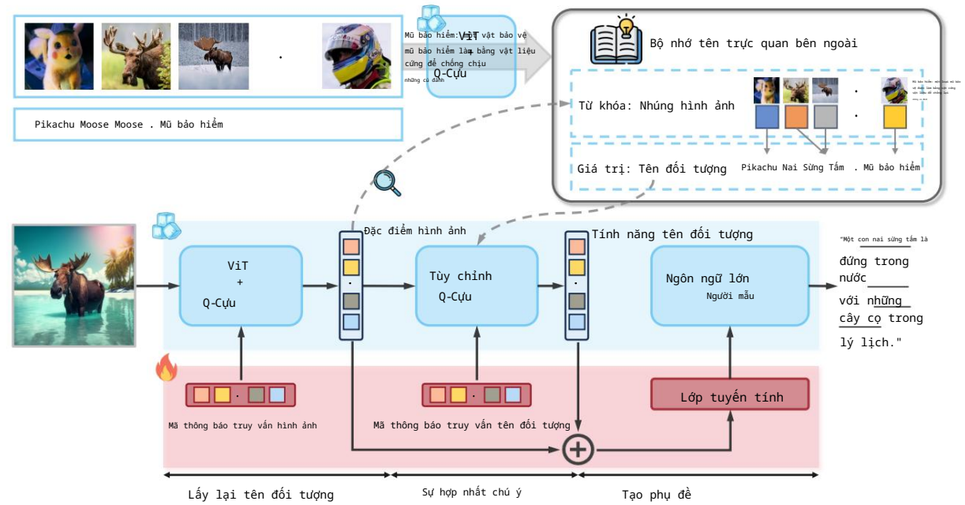
We base our method on the frozen pre-trained vision model and LLM with several trainable layers (Fig. 2), giving in a model that is cheap to train. To guide the LLM, we adopt a recently popular approach called prompting as in [11, 25, 29, 35, 46]. We begin by matching the learned visual features from the input image with image embeddings stored in memory, retrieving object names. We also introduce an attentive fusion module designed to implicitly remove irrelevant retrieved names. Finally, following the attentive fusion, we combine the learned visual features and object name features to form a prompt for the LLM to generate a caption, thus addressing challenge (2).

3.2. External visual–name memory

To build the external visual–name memory, we first collect image–name pairs from the external data source. After that, we encode these images into image embeddings, which serve as keys in memory, and use their names as values.

External data source. We utilize object images from LVIS dataset [17] to construct our external visual–name memory $\mathcal{M}$. Specifically, we use 1203 objects in LVIS, where we randomly select from one to ten images for each object, amounting to 8581 object images. Furthermore, as mentioned in Sec. 3.1, we also incorporate synthetic images in our memory construction. Using stable diffusion [36], we generate five additional images for each object, with a prompt of "*a photo of* {*object name*}", resulting in a total of $M = 14596$ $(8581 + 5 \times 1203)$ images. Each object image X^i is associated with an object name v^i. Note that many object images may share the same object name. For the sake of simplicity, we may regard each image as corresponding to a single name. In summary, we have M image–name pairs $\{(X^i, v^i)\}_{i=1}^{M}$ for external memory construction.

External memory construction. For each image X^i, we use a frozen vision encoder $\mathcal{E}(\cdot)$ (see Sec. 3.3 for detail) to project it into 32 embeddings with the size of 1×768 each: $\{\mathbf{k}_1^i, \mathbf{k}_2^i, \cdots, \mathbf{k}_{32}^i\} = \mathcal{E}(X^i)$. We then average 32 embeddings to produce a single embedding $\mathbf{k}^i$ (1×768) that serves as the key (visual) in $\mathcal{M}$. The paired object name v^i acts as its value (name). Consequently, we have the visual–name memory $\mathcal{M} = \{(\mathbf{k}^i, v^i)\}_{i=1}^{M}$ which is indexed using FAISS [21], facilitating rapid searches based on sim-



Hình 2. Sơ đồ EVCAP được đề xuất của chúng tôi. Nó bao gồm bộ nhớ tên trực quan bên ngoài với nhúng hình ảnh và tên đối tượng (phía trên), một ViT đông lạnh và Q-Former được trang bị mã thông báo truy vấn hình ảnh có thể đào tạo, một mô-đun hợp nhất được phát triển bởi một mã thông báo truy vấn tên đối tượng có thể đào tạo và Q-Former đông lạnh, và LLM đông lạnh với lớp tuyến tính có thể đào tạo (phía dưới). ViT và Q-Former trích xuất các đặc điểm trực quan đã học được từ hình ảnh đầu vào, sau đó được sử dụng để truy xuất tên đối tượng từ bộ nhớ ngoài. Những tên đối tượng được lấy lại và các đặc điểm trực quan đã học được trải qua sự chú ý chéo trong Q-Former tùy chỉnh, tạo ra tên đối tượng được tinh chỉnh tính năng. Cuối cùng, các tính năng tên đối tượng kết hợp với các tính năng trực quan được đưa vào LLM sau một lớp tuyến tính để tạo chú thích.

Như đã thảo luận ở trên, thách thức (1) có thể được giải quyết bằng cách sử dụng hình ảnh trực quan của các đối tượng. Tuy nhiên, nếu chúng ta giới hạn lại trí nhớ của mình chỉ với một cặp hình ảnh-tên cho mỗi đối tượng, trí nhớ của chúng ta sẽ thiếu tính đa dạng. Do đó, chúng tôi thu thập một số hình ảnh cho mỗi đối tượng mục tiêu. Ngoài ra, chúng tôi giữ các hình ảnh tổng hợp trong bộ nhớ của chúng tôi để tránh tác hại mà hình ảnh tổng hợp có thể gây ra cho phương pháp của chúng tôi, như đã chỉ ra trong [18]. Với khả năng thu thập hình ảnh từ internet, EVCAP có thể dễ dàng được mở rộng để dễ dàng đưa vào các đối tượng mới lạ từ thế giới thực.

Chúng tôi dựa phương pháp của mình vào tầm nhìn được đào tạo trước đông lạnh mô hình và LLM với một số lớp có thể đào tạo (Hình 2), cung cấp một mô hình có chi phí đào tạo thấp. Để hướng dẫn LLM, chúng tôi áp dụng một cách tiếp cận phổ biến gần đây được gọi là nhắc nhở như trong [11, 25, 29, 35, 46]. Chúng tôi bắt đầu bằng cách khớp các đặc điểm trực quan đã học được từ hình ảnh đầu vào với nhúng hình ảnh được lưu trữ trong bộ nhớ, truy xuất tên đối tượng. Chúng tôi cũng giới thiệu một mô-đun hợp nhất chú ý được thiết kế để xóa ngầm các tên đã truy xuất không liên quan. Cuối cùng, sau khi hợp nhất chú ý, chúng tôi kết hợp các tính năng trực quan đã học và các tính năng tên đối tượng để tạo lời nhắc cho LLM tạo chú thích, do đó giải quyết được thách thức (2).

3.2. Bộ nhớ hình ảnh bên ngoài-tên

Để xây dựng bộ nhớ hình ảnh-tên bên ngoài, trước tiên chúng ta thu thập các cặp hình ảnh-tên từ nguồn dữ liệu bên ngoài. Sau đó rằng chúng tôi mã hóa những hình ảnh này thành các hình ảnh nhúng, đóng vai trò là khóa trong bộ nhớ và sử dụng tên của chúng làm giá trị. Nguồn dữ liệu bên ngoài. Chúng tôi sử dụng hình ảnh đối tượng từ LVIS tập dữ liệu [17] để xây dựng bộ nhớ tên-hình ảnh bên ngoài của chúng tôi M. Cụ thể, chúng tôi sử dụng 1203 đối tượng trong LVIS, nơi chúng tôi chọn ngẫu nhiên từ một đến mười hình ảnh cho mỗi đối tượng, lên tới 8581 hình ảnh đối tượng. Hơn nữa, như đã đề cập trong Mục 3.1, chúng tôi cũng kết hợp các hình ảnh tổng hợp trong cấu trúc bộ nhớ của chúng ta. Sử dụng sự khuếch tán ổn định [36], chúng tôi tạo ra năm hình ảnh bổ sung cho mỗi đối tượng, với nhắc nhở "một bức ảnh của {tên đối tượng}", dẫn đến tổng số của M = 14596 (8581 + 5 × 1203) hình ảnh. Mỗi đối tượng im-age Xi được liên kết với một tên đối tượng v . Lưu ý rằng nhiều hình ảnh đối tượng có thể chia sẻ cùng một tên đối tượng. Vì lợi ích của sự đơn giản, chúng ta có thể coi mỗi hình ảnh là tương ứng đến một tên duy nhất. Tóm lại, chúng ta có M image-name cặp {(Xi , vi)}M để xây dựng bộ nhớ ngoài.

Xây dựng bộ nhớ ngoài. Đối với mỗi hình ảnh Xi , chúng tôi sử dụng bộ mã hóa thị giác đông lạnh E(·) (xem mục 3.3 để biết chi tiết) để chiếu nó vào 32 nhúng có kích thước 1 × 768 mỗi: {k , , ..., 32} = E(Xi). Sau đó chúng ta tính trung bình 32 nhúng để tạo ra một nhúng k duy nhất (1 × 768) đóng vai trò là chìa khóa (hình ảnh) trong M. Đối tượng ghép nối tên v hoạt động như giá trị của nó (tên). Do đó, chúng ta có bộ nhớ hình ảnh-tên M = {(k ,vi)}M được lập chỉ mục sử dụng FAISS [21], tạo điều kiện tìm kiếm nhanh dựa trên mô phỏng

ilarity measures. Our memory can be expanded effortlessly by gathering additional visual–name pairs (see Sec. 5.3).

3.3. Object names retrieval

Image encoding. We feed a frozen vision encoder $\mathcal{E}$ image X and image query tokens $\mathbf{T}_{\text{img}}$ to produce visual features $\mathcal{Q}$. To enable the retrieval process controllable, we make image query tokens to be trainable. Thus, the image encoding process can be summarized as $\mathcal{Q} = \mathcal{E}(X, \mathbf{T}_{\text{img}})$. We use the BLIP-2 pre-trained vision encoder [25], which consists of a pre-trained vision transformer ViT-g [14] outputting image features (257×1408), and a Q-Former receiving image features producing $|\mathcal{Q}| = 32$ learned visual features (1×768 each). We denote $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{32}\}$.
Retrieval. Having obtained $\mathcal{Q}$, we calculate the cosine similarity between the query $\mathbf{q}_j \in \mathcal{Q}$ and the key $\mathbf{k}^i \in \mathcal{M}$. The similarity calculation is given by $\text{SIM}(\mathbf{q}_j, \mathbf{k}^i) = \frac{\mathbf{q}_j^\top \mathbf{k}^i}{\|\mathbf{q}_j\| \|\mathbf{k}^i\|}$, where $i \in [1, M]$, $j \in [1, 32]$. Given each $\mathbf{q}_j$, we select one key with the highest similarity score, resulting in 32 key–value candidates $\{\mathbf{k}_j^{\text{best}}, v_j^{\text{best}}\}_{j=1}^{32}$.

After that, we filter out candidates with repeated object names (values), and then select the top-K values. In particular, we determine the index j from the key that has the highest SIM score. These selected values v_j^{best} are redefined as the new notation v_l in the retrieved top-K object names for the input image, which can be summarized as follows:

$$\{\mathbf{k}_j^{\text{best}}, v_j^{\text{best}}\} = \arg\max_{\mathbf{k}^i} \text{SIM}\left(\mathbf{q}_j, \mathbf{k}^i\right),$$

$$j = \arg\max_{j} \text{SIM}(\mathbf{q}_j, \mathbf{k}_j^{\text{best}}), v_l \leftarrow v_j^{\text{best}},$$

where $l \in [1, \text{K}]$. As a result, the retrieved top-K object names are $\{v_l\}_{l=1}^{\text{K}}$.

3.4. Attentive fusion

Since the object names obtained from the retrieval process may be redundant, we develop an attentive fusion module to selectively distill object name features.

The retrieved object names $\{v_l\}_{l=1}^{\text{K}}$ are concatenated together into a sequence $\mathcal{S}$, each separated by a delimiter: $\mathcal{S} = \{v_1, \texttt{[SEP]}, v_2, \texttt{[SEP]}, \cdots, \texttt{[SEP]}, v_{\text{K}}\}$. The sequence $\mathcal{S}$ and visual features $\mathcal{Q}$ are fed into a customized Q-Former $\mathcal{F}(\cdot)$, which is constructed from the frozen pre-trained Q-Former as we used in vision encoder $\mathcal{E}$. Nonetheless, in order to enable object names to get attention from visual features, we switch the image embedding port and the text instruction port (see the supplement for architecture detail). Like in the image encoding process in Sec. 3.3, we make the object name query tokens $\mathbf{T}_{\text{obj}}$ learnable during training to assist in learning object name features related to the caption. The size of $\mathbf{T}_{\text{obj}}$ is $P \times 768$, where P indicates the number of object name query tokens. We get the object name features $\mathcal{V} = \mathcal{F}(\mathcal{S}, \mathcal{Q}, \mathbf{T}_{\text{obj}})$.

3.5. Caption generation

Before inputting the visual features $\mathcal{Q}$ and object name features $\mathcal{V}$ into the LLM decoder, we concatenate ($\oplus$) them and use a linear layer $\phi(\cdot)$ to project them into the input latent space of the LLM as $\phi(\mathcal{Q} \oplus \mathcal{V})$. The LLM used for caption generation in this work is the pre-trained Vicuna-13B [10], an open-source chatbot constructed from LLaMA [37]. During training and evaluation, we design a prompt in a conversational format, that is similar to [46]:

```
###Human: <Img><ProjFeature></Img>
Describe this image in detail.
###Assistant:
```

in which, `ProjFeature` denotes the projected feature $\phi(\mathcal{Q} \oplus \mathcal{V})$ after the linear layer. In training phase, given input caption tokens $\{c_i\}_{i=1}^{L}$, the LLM decoder concatenates the embedded prompt $\{\mathbf{w}_i\}_{i=1}^{N}$ and the embedded caption tokens $\{\mathbf{c}_i\}_{i=1}^{L}$ as input, and predicts the caption tokens in an autoregressive fashion, while in the evaluation phase, we only need to input the embedded prompt. We train EVCAP by minimizing the cross-entropy loss in an end-to-end way: $\mathcal{L}_\theta = -\sum_{i=1}^{L} \log p_\theta\left(c_i \mid \mathbf{w}_1, ...\mathbf{w}_N, \mathbf{c}_1, ..., \mathbf{c}_{i-1}\right)$, in which θ indicates the trainable parameters.

4. Experimental Settings

4.1. Training setup

Implementation. EVCAP uses the same image encoder as in BLIP-2 [25], consisting of a ViT-g [14] and their pre-trained Q-Former. Since we intend to obtain object name features through cross-attention between retrieved object names and visual features, we develop a customized Q-Former, which consists of BERT [23] with cross-attention layers inserted at every other transformer block. We use a frozen Vicuna-13B [10] as the caption generator.
Training dataset. For all experiments, we exclusively train EVCAP using the training set of **COCO** dataset [27], consisting of 82k images and 5 captions per images. The entire training process takes about 3 hours on 4 A6000 GPUs, using mixed precisions (more details in the supplementary).

4.2. Evaluation setup

Evaluation dataset. We evaluate EVCAP, trained using the COCO training set, across four datasets: its test set, two challenging benchmarks – NoCaps validation set and Flickr30k test set, and a synthetic commonsense-violating dataset – WHOOPS. We adhere follow prior work [15, 41] to use the same images of Karpathy split [22] on **COCO** test set, **NoCaps** [2] validation set, and Karpathy split on **Flickr30k** [30] test set. In addition, **WHOOPS** [6] is a synthetic image captioning dataset comprising 500 synthetic commonsense-violating images and 2500 paired captions.

biện pháp ilarity. Bộ nhớ của chúng ta có thể được mở rộng một cách dễ dàng
bằng cách thu thập thêm các cặp hình ảnh-tên (xem Mục 5.3).

3.3. Lấy lại tên đối tượng

Mã hóa hình ảnh. Chúng tôi cung cấp một bộ mã hóa tầm nhìn đông lạnh E hình ảnh
Mã thông báo truy vấn hình ảnh và X Timg để tạo ra các tính năng trực quan
Q. Để cho phép quá trình truy xuất có thể kiểm soát được, chúng tôi thực hiện
mã thông báo truy vấn hình ảnh có thể được đào tạo. Do đó, quá trình mã
hóa hình ảnh có thể được tóm tắt như sau: Q = E(X, Timg).
Chúng tôi sử dụng bộ mã hóa thị giác được đào tạo trước BLIP-2 [25],
bao gồm một bộ biến đổi thị giác được đào tạo trước ViT-g [14] đưa ra
các đặc điểm hình ảnh (257 × 1408) và một Q-Former nhận các đặc điểm
hình ảnh tạo ra |Q| = 32 thị giác đã học
tính năng (mỗi tính năng có 1 × 768). Chúng tôi biểu thị Q = {q1, q2, ..., q32}.
Truy xuất. Sau khi có được Q, chúng tôi tính toán độ tương tự cosin giữa
truy vấn qj Q và khóa k ~ M. Các
tính toán độ tương đồng được đưa ra bởi SIM(qj , k ~) = $\frac{q_j \cdot k_i}{\|q_j\| \|k_i\|}$,
trong đó i [1, M], j [1, 32]. Với mỗi qj , chúng ta chọn
một khóa có điểm tương đồng cao nhất, dẫn đến 32
ứng viên khóa-giá trị {k $^{\text{tốt nhất}}_{j}$, v$^{\text{tốt nhất}}_{j}$ }$^{32}_{j=1}$.
Sau đó, chúng tôi lọc ra các ứng viên có đối tượng lặp lại
tên (giá trị), sau đó chọn các giá trị top-K. Cụ thể, chúng tôi xác định chỉ số j từ khóa có giá trị
cao nhất
điểm SIM. Các giá trị được chọn này v $_{j}$ được định nghĩa lại là
ký hiệu mới vl trong các tên đối tượng top-K được lấy lại cho
hình ảnh đầu vào có thể được tóm tắt như sau:

trong đó l [1, K]. Do đó, đối tượng top-K được lấy lại
tên là {vl} $^{K}_{l=1}$.

3.4. Sự hợp nhất chú ý

Vì tên đối tượng thu được từ quá trình truy xuất
có thể là thừa, chúng tôi phát triển một mô-đun hợp nhất chu đáo
để chắt lọc có chọn lọc các đặc điểm tên đối tượng.
Tên đối tượng được lấy lại {vl} $^{K}_{\text{chuỗi}}$ được nối với nhau thành một
S, mỗi chuỗi được phân tách bằng một dấu phân cách:
S = {v1, [SEP], v2, [SEP], · · · , [SEP], vK}. Chuỗi S và các tính năng
trực quan Q được đưa vào một
Q-Former F(·), được xây dựng từ Q-Former được đào tạo trước đông lạnh
như chúng tôi đã sử dụng trong bộ mã hóa thị giác E. Tuy nhiên, để cho
phép tên đối tượng thu hút sự chú ý từ
các tính năng trực quan, chúng tôi chuyển đổi cổng nhúng hình ảnh và
cổng hướng dẫn văn bản (xem phần bổ sung về kiến trúc
chi tiết). Giống như trong quá trình mã hóa hình ảnh ở Mục 3.3, chúng tôi
làm cho các mã thông báo truy vấn tên đối tượng Tobj có thể học được trong
đào tạo để hỗ trợ việc học các đặc điểm tên đối tượng liên quan đến
chú thích. Kích thước của Tobj là P ×768, trong đó P biểu thị
số lượng mã thông báo truy vấn tên đối tượng. Chúng tôi nhận được đối tượng
tên các tính năng V = F(S, Q, Tobj).

3.5. Tạo chú thích

Trước khi nhập các đặc điểm trực quan Q và các đặc điểm tên đối tượng V
vào bộ giải mã LLM, chúng tôi nối ( ) chúng
và sử dụng một lớp tuyến tính φ(·) để chiếu chúng vào không gian tiềm ẩn
đầu vào của LLM dưới dạng φ(Q V). LLM
được sử dụng để tạo chú thích trong tác phẩm này là được đào tạo trước
Vicuna-13B [10], một chatbot nguồn mở được xây dựng từ
LLaMA [37]. Trong quá trình đào tạo và đánh giá, chúng tôi thiết kế một
nhắc nhở theo định dạng đàm thoại, tương tự như [46]:

```
###Con người: <Img><ProjFeature></Img>
Hãy mô tả chi tiết hình ảnh này.
###Trợ lý:
```

trong đó, ProjFeature biểu thị tính năng được chiếu
φ(Q V) sau lớp tuyến tính. Trong giai đoạn đào tạo, đưa ra các mã
thông báo chú thích đầu vào {ci} $^{L}_{\text{tôi}=1}$, bộ giải mã LLM nối lại
lời nhắc nhúng {wi} $^{N}_{\text{tôi}=1}$ và chú thích nhúng
token {ci} $^{L}_{\text{tôi}=1}$ như đầu vào và dự đoán các mã thông báo chú thích trong
một cách tự hồi quy, trong khi ở giai đoạn đánh giá, chúng tôi
chỉ cần nhập lời nhắc nhúng. Chúng tôi đào tạo EV- CAP bằng cách giảm
thiểu tổn thất entropy chéo trong một end-to-end
cách: Lθ = $^{L}_{\text{tôi}=1}$ log pθ (ci | wi, ...wN , c1, ..., ci 1), trong
trong đó θ biểu thị các tham số có thể đào tạo được.

4. Cài đặt thử nghiệm

4.1. Thiết lập đào tạo

Triển khai. EVCAP sử dụng cùng một bộ mã hóa hình ảnh như
trong BLIP-2 [25], bao gồm ViT-g [14] và Q-Former được đào tạo trước của
chúng. Vì chúng tôi dự định lấy tên đối tượng
các tính năng thông qua sự chú ý chéo giữa các đối tượng được lấy lại
tên và các tính năng trực quan, chúng tôi phát triển một Q-Former tùy
chỉnh, bao gồm BERT [23] với sự chú ý chéo
các lớp được chèn vào mỗi khối biến áp khác. Chúng tôi sử dụng một
Vicuna-13B đông lạnh [10] làm trình tạo chú thích.
Tập dữ liệu đào tạo. Đối với tất cả các thí nghiệm, chúng tôi chỉ đào tạo
EVCAP sử dụng bộ dữ liệu đào tạo COCO [27], bao gồm 82k hình ảnh và 5
chú thích cho mỗi hình ảnh. Toàn bộ
quá trình đào tạo mất khoảng 3 giờ trên 4 GPU A6000, sử dụng độ chính
xác hỗn hợp (thông tin chi tiết hơn trong phần bổ sung).

4.2. Thiết lập đánh giá

Bộ dữ liệu đánh giá. Chúng tôi đánh giá EVCAP, được đào tạo bằng
bộ đào tạo COCO, trên bốn tập dữ liệu: bộ thử nghiệm của nó,
hai chuẩn mực đầy thách thức - bộ xác thực NoCaps và
Bộ kiểm tra Flickr30k và một bộ vi phạm lẽ thường tổng hợp
tập dữ liệu - WHOOPS. Chúng tôi tuân thủ theo công trình trước đây [15, 41]
sử dụng cùng một hình ảnh của Karpathy chia tách [22] trên COCO
bộ kiểm tra, bộ xác thực NoCaps [2] và phân tách Karpathy trên
Bộ kiểm tra Flickr30k [30] . Ngoài ra, WHOOPS [6] là một tập dữ liệu chú
thích hình ảnh tổng hợp bao gồm 500
hình ảnh vi phạm lẽ thường và 2500 chú thích ghép nối.

Table 1. Quantitative comparison against SOTA methods on three common image captioning benchmarks. * denotes using a **memory bank**. We report the size of training data and parameters; BLEU@4 (B@4), METEOR (M), CIDEr (C), and SPICE (S) scores on COCO test set; C and S scores on in-domain, near-domain, out-domain and overall data of NoCaps validation set; C and S scores on Flickr30k test set. Higher score is better. **Bold** indicates the best results among compared methods, normal indicates the second best results.

Method	Training Data	Training Para.	COCO Test B@4	COCO Test M	COCO Test C	COCO Test S	NoCaps val In-domain C	NoCaps val In-domain S	NoCaps val Near-domain C	NoCaps val Near-domain S	NoCaps val Out-domain C	NoCaps val Out-domain S	NoCaps val Overall C	NoCaps val Overall S	Flickr30k Test C	Flickr30k Test S
Heavyweight-training models																
VinVL [45]	8.9M	110M	38.2	30.3	129.3	23.6	96.8	13.5	90.7	13.1	87.4	11.6	90.9	12.8	–	–
AoANet+MA* [16]	COCO	–	38.0	28.7	121.0	21.8	–	–	–	–	–	–	–	–	–	–
NOC-REK* [40]	COCO	110M	–	–	–	–	104.7	14.8	100.2	14.1	100.7	13.0	100.9	14.0	–	–
RCA-NOC* [13]	COCO	110M	37.4	29.6	128.4	23.1	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	–	–
ViECap GPT2 [15]	COCO	124M	27.2	24.8	92.9	18.2	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5	47.9	13.6
InstructBLIP Vicuna-13B [11]	129M	188M	–	–	–	–	–	–	–	–	–	–	121.9	–	82.8	–
OSCAR [26]	4.1M	338M	37.4	30.7	127.8	23.5	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7	–	–
BLIP [24]	129M	446M	40.4	–	136.7	–	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	–	–
BLIP-2 FlanT5-XL [25]	129M	1.2B	**42.4**	–	144.5	–	**123.7**	**16.3**	**120.2**	**15.9**	**124.8**	**15.1**	121.6	**15.8**	–	–
REVEAL* T5 [20]	1.3B	2.1B	–	–	**145.4**	–	–	–	–	–	–	–	**123.0**	–	–	–
Lightweight-training models																
MiniGPT4 Vicuna-13B [46]	5M	**3.94M**	38.0	29.6	129.6	23.4	99.0	14.8	106.9	15.3	110.8	14.9	108.8	15.1	78.4	16.9
SmallCap* GPT2 [35]	COCO	7M	37.0	27.9	119.7	21.3	–	–	–	–	–	–	–	–	60.6	–
ClipCap GPT2 [29]	COCO	43M	33.5	27.5	113.1	21.1	84.9	12.1	66.8	10.9	49.1	9.6	65.8	10.9	–	–
EVCAP* Vicuna-13B	COCO	3.97M	41.5	**31.2**	140.1	**24.7**	111.7	15.3	119.5	15.6	116.5	14.7	119.3	15.3	**84.4**	**18.0**
Specialist SOTAs																
Qwen-VL Qwen-7B [5]	1.4B	9.6B	–	–	–	–	–	–	–	–	–	–	121.4	–	85.8	–
CogVLM Vicuna-7B [41]	1.5B	6.5B	–	–	148.7	–	–	–	–	–	132.6	–	128.3	–	94.9	–
PaLI mT5-XXL [9]	1.6B	17B	–	–	149.1	–	–	–	–	–	–	–	127.0	–	–	–
PaLI-X UL2-32B [8]	2.2B	55B	–	–	149.2	–	–	–	–	–	–	–	126.3	–	–	–

Compared methods. We compare EVCAP with several SOTAs. According to the trainable parameters size, they can be divided into 1) Heavyweight-training (between 100M to 5B): VinVL [45], AoANet [16], NOC-REK [40], RCA-NOC [13], ViECap [15], InstructBLIP [11], OSCAR [26], BLIP [24], BLIP-2 [25], REVEAL [20]; 2) Lightweight-training (less than 100M): MiniGPT4 [46], SmallCap [35], ClipCap [29]; and also 3) Specialist SOTAs with huge trainable parameters (larger than 5B): Qwen-VL [5], CogVLM [41], PaLI [9], PaLI-X [8]. Among these methods, AoANet, NOC-REK, RCA-NOC, REVEAL, and SmallCap are retrieval-augmented captioning methods.

5. Experimental Results

5.1. Results on in-/out-domain benchmarks

We assess EVCAP against SOTAs on both in-domain and out-domain benchmarks. The COCO test set can be considered as in-domain data as we only train our model on the COCO training set. Out-domain benchmarks are the NoCaps validation set and the Flickr30k test set.

Quantitative results. Tab. 1 details our EVCAP's performance in comparison with SOTA methods. We first evaluate training costs in terms of training data sizes and parameters. Similar to various heavyweight-training models that exclude LLMs and the majority of lightweight-training models, EVCAP is trained solely on the COCO training set. It utilizes only 3.97M trainable parameters, positioning it as the second smallest, slightly larger than MiniGPT4 with 3.94M. Among lightweight-training models, our approach outperforms others, achieving the highest scores on all benchmarks. Despite using less training data and nearly identical trainable parameters as MiniGPT4, EVCAP significantly surpasses it, with a marked improvement of 10.5, 10.5, and 6.0 in CIDEr scores for each benchmark. When further compared with heavyweight-training models, the performance of EVCAP stands out among million-level models, nearly matching InstructBLIP, except in NoCaps. Note that since BLIP-2 does not include Vicuna checkpoints, InstructBLIP performs pre-training with Vicuna using the same procedure as BLIP-2, whereas EVCAP does not involve pre-training. Against REVEAL, which also uses external memory, our EVCAP utilizes about 1/3000 training data and 1/500 training parameters yet yields comparable results. Moreover, EVCAP's performance is on par with BLIP-2, the top-performing model with 1.2B trainable parameters. This highlights EVCAP's efficiency and effectiveness despite its significantly smaller training cost, thanks to our external visual–name memory. Regarding specialist SOTAs, they use billion-level training data and over 5B trainable parameters, so it is acceptable that they can achieve exceptionally strong performance, surpassing EVCAP by nearly 10 on all benchmarks in CIDEr scores.

Qualitative results. Fig. 3 presents a comparison of captions generated by our EVCAP and three SOTA models across three benchmarks. The captions of SmallCap are generated by its publicly accessible demo [1]. We generate captions of MiniGPT4 and BLIP-2 using their respec-

Bảng 1. So sánh định lượng với các phương pháp SOTA trên ba tiêu chuẩn chú thích hình ảnh phổ biến. * biểu thị việc sử dụng bộ nhớ ngân hàng. Chúng tôi báo cáo kích thước của dữ liệu đào tạo và các tham số; BLEU@4 (B@4), METEOR (M), CIDEr (C) và SPICE (S) điểm trên COCO bộ kiểm tra; Điểm C và S trên dữ liệu trong miền, gần miền, ngoài miền và dữ liệu tổng thể của bộ xác thực NoCaps; Điểm C và S trên bài kiểm tra Flickr30k thiết lập. Điểm cao hơn là tốt hơn. In đậm chỉ ra kết quả tốt nhất trong số các phương pháp được so sánh, bình thường chỉ ra kết quả tốt thứ hai.

Phương pháp	Dữ liệu đào tạo	Para.	COCO Test B@4	COCO Test M	COCO Test C	COCO Test S	NoCaps giá trị Trong miền C	Trong miền S	Gần miền C	Gần miền S	Ngoài miền C	Ngoài miền S	Tổng thể C	Tổng thể S	Flickr30k C	Flickr30k S
Người mẫu tập tạ nặng																
VinVL [45]	8,9M	110M	38,2	30,3	129,3	23,6	96,8	13,5	90,7	13,1	87,4	11,6	90,9	12,8	-	-
AoANet+MA* [16]	COCO	–	38.0	28.7	121.0	21.8	–	–	–	–	–	–	–	–	–	–
NOC-REK* [40]	COCO	110M	–	–	–	–	104,7	14,8	100,2	14.1	100,7	13.0	100,9	14.0	-	-
RCA-NOC* [13]	COCO	110M	37.4	29.6	128.4	23.1	92,2	12,9	87,8	12.6	87,5	11.5	88,3	12.4	-	-
ViECap GPT2 [15]	COCO	124M	27.2	24.8	92.9	18.2	61,1	10,4	64,3	9.9	65,0	8.6	66,2	9.5	47,9	13.6
InstructBLIP Vicuna-13B [11]	129M	188M	–	–	–	–	–	–	–	–	–	–	121,9	–	82,8	–
OSCAR [26]	4,1M	338M	37,4	30,7	127,8	23,5	83,4	12,0	81,6	12.0	77,6	10,6	81,1	11,7	-	-
BLIP [24]	129M	446M	40.4	BLIP-2 FlanT5-	136,7	-	114,9	15,2	112,1	14.9	115,3	14,4	113,2	14,8	-	-
BLIP-2 FlanT5-XL [25]	129M	1,2B	42,4	-	144,5	-	123,7	16,3	120,2	15.9	124,8	15,1	121,6	15,8	-	-
REVEAL* T5 [20]	1,3 tỷ	2.1B	-	-	145,4	-	-	-	-	-	-	-	123.0	-	-	-
Các mô hình tập luyện nhẹ																
MiniGPT4 Vicuna-13B [46]	5M	3,94M	38,0	29,6	COCO 129,6	23,4	99.0	14.8	106,9	15.3	110,8	14,9	108,8	15.1	78,4	16,9
GPT2 nhỏ* [35]	COCO	43M	33,5	27,5	COCO 119,7	21,3	-	-	-	-	-	-	-	-	60,6	-
ClipCap GPT2 [29]					113,1	21,1	84,9	12.1	66,8	10,9	49,1	9,6	65,8	10,9	-	-
EVCAP* Vicuna-13B					140,1	24,7	111,7	15.3	119,5	15,6	116,5	14,7	119,3	15,3	84,4	18.0
Chuyên gia SOTA																
Qwen-VL Qwen-7B [5]	1,4B	9,6B	-	-	-	-	-	-	-	-	-	-	121,4	-	85,8	-
CogVLM Vicuna-7B [41]	1,5B	6,5B	-	-	148,7	-	-	-	-	-	132,6	-	128,3	-	94,9	-
PaLI mT5-XXL [9]	1,6B	17B	-	-	149,1	-	-	-	-	-	-	-	127,0	-	-	-
PaLI-X UL2-32B [8]	2,2B	55B	-	-	149,2	-	-	-	-	-	-	-	126,3	-	-	-

So sánh các phương pháp. Chúng tôi so sánh EVCAP với một số SOTA. Theo kích thước tham số có thể đào tạo, chúng có thể được chia thành 1) Tập tạ nặng (giữa 100M đến 5B): VinVL [45], AoANet [16], NOC-REK [40], RCA-NOC [13], ViECap [15], InstructBLIP [11], OS-CAR [26], BLIP [24], BLIP-2 [25], REVEAL [20]; 2) Huấn luyện nhẹ (dưới 100M): MiniGPT4 [46], SmallCap [35], ClipCap [29]; và 3) SO-TA chuyên biệt với các tham số có thể đào tạo lớn (lớn hơn 5B): Qwen-VL [5], CogVLM [41], PaLI [9], PaLI-X [8]. Trong số này các phương pháp, AoANet, NOC-REK, RCA-NOC, REVEAL và SmallCap là phương pháp chú thích tăng cường khả năng truy xuất.

5. Kết quả thực nghiệm

5.1. Kết quả trên các chuẩn mực trong/ngoài miền

Chúng tôi đánh giá EVCAP so với SOTA trên cả miền trong và chuẩn mực ngoài miền. Bộ kiểm tra COCO có thể được coi là dữ liệu trong miền vì chúng tôi chỉ đào tạo mô hình của mình trên Bộ đào tạo COCO. Các chuẩn mực ngoài miền là bộ xác thực No-Caps và bộ thử nghiệm Flickr30k.

Kết quả định lượng. Bảng 1 nêu chi tiết hiệu suất EVCAP của chúng tôi khi so sánh với các phương pháp SOTA. Đầu tiên, chúng tôi đánh giá chi phí đào tạo theo kích thước dữ liệu đào tạo và các tham số. Tương tự như nhiều mô hình đào tạo hạng nặng khác loại trừ LLM và phần lớn các chương trình đào tạo nhẹ mô hình, EVCAP được đào tạo hoàn toàn trên chương trình đào tạo COCO thiết lập. Nó chỉ sử dụng 3,97M tham số có thể đào tạo, định vị nó là nhỏ thứ hai, lớn hơn một chút so với MiniGPT4 với 3,94M. Trong số các mô hình đào tạo nhẹ, phương pháp của chúng tôi vượt trội hơn những phương pháp khác, đạt được điểm số cao nhất trên tất cả các điểm chuẩn. Mặc dù sử dụng ít dữ liệu đào tạo hơn và các thông số có thể đào tạo gần như giống hệt như MiniGPT4, EVCAP vượt trội đáng kể, với sự cải thiện rõ rệt Điểm CIDEr là 10,5, 10,5 và 6,0 cho mỗi tiêu chuẩn.

Khi so sánh thêm với các mô hình tập luyện tạ nặng, hiệu suất của EVCAP nổi bật giữa hàng triệu cấp các mô hình, gần giống với InstructBLIP, ngoại trừ NoCaps. Lưu ý rằng vì BLIP-2 không bao gồm các điểm kiểm tra Vicuna, InstructBLIP thực hiện đào tạo trước với Vicuna bằng cùng một quy trình như BLIP-2, trong khi EVCAP thì không . không liên quan đến việc đào tạo trước. Chống lại REVEAL, cũng sử dụng bộ nhớ ngoài, EVCAP của chúng tôi sử dụng khoảng 1/3000 dữ liệu đào tạo và 1/500 tham số đào tạo nhưng vẫn mang lại kết quả tương đương. Hơn nữa, hiệu suất của EVCAP ngang bằng với BLIP-2, mô hình có hiệu suất cao nhất với 1,2B tham số có thể đào tạo được. Điều này làm nổi bật hiệu quả của EVCAP và hiệu quả mặc dù chi phí đào tạo nhỏ hơn đáng kể, nhờ vào trí nhớ hình ảnh bên ngoài của chúng ta-tên. Về các SOTA chuyên gia, họ sử dụng dữ liệu đào tạo cấp độ hàng tỷ và hơn 5B thông số có thể đào tạo được, vì vậy có thể chấp nhận được rằng chúng có thể đạt được hiệu suất cực kỳ mạnh mẽ, vượt trội EVCAP tăng gần 10 điểm trên tất cả các tiêu chuẩn về điểm số CIDEr.

Kết quả định tính. Hình 3 trình bày sự so sánh các chú thích được tạo ra bởi EVCAP của chúng tôi và ba mô hình SOTA trên ba chuẩn mực. Các chú thích của SmallCap là được tạo ra bởi bản demo có thể truy cập công khai của nó [1]. Chúng tôi tạo ra các chú thích của MiniGPT4 và BLIP-2 bằng cách sử dụng

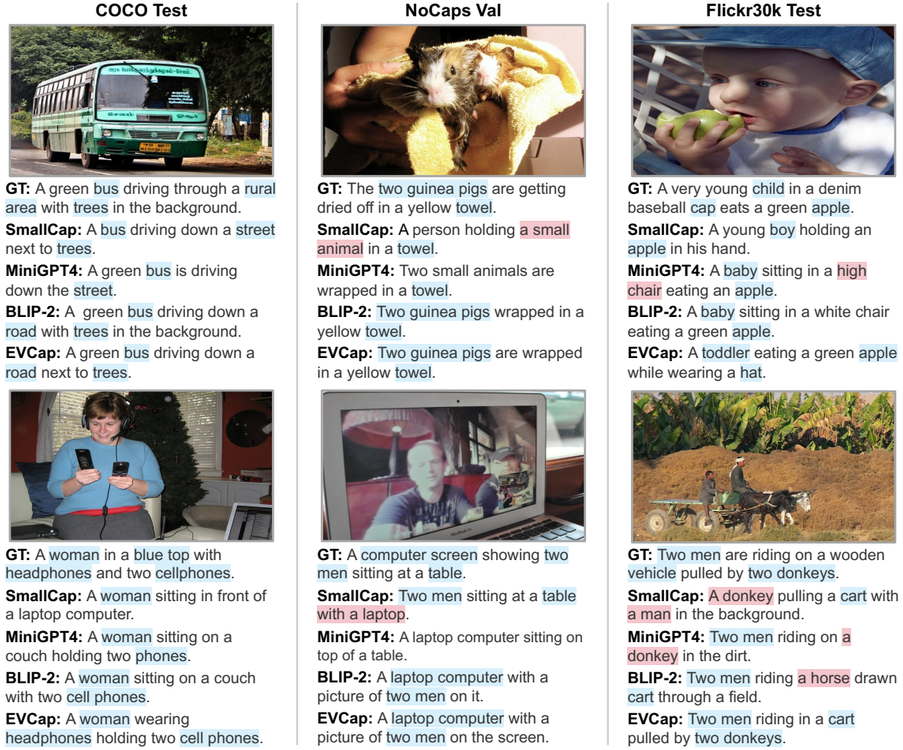


Figure 3. Examples of captions generated by our EVCAP and three SOTA methods on COCO test set, NoCaps validation set, and Flickr30k test set. GT refers to the Ground Truth captions. Incorrect objects in captions are highlighted in red, while correct ones are in blue. Our EVCAP correctly generates captions across different datasets, showing performance comparable to BLIP-2.

tive pre-trained models. As a lightweight and retrieval-augmented captioning method, SmallCap struggles to produce accurate captions for given images, primarily because it relies on retrieved captions laden with extraneous information. MiniGPT4, though aligned with the primary content of images, sometimes misses certain objects like *"trees"* and *"headphones"*. This oversight stems from its focus on the main objects in images, without integrating additional cues for other objects provided by the retrieved object names. In contrast, the captions generated by our EVCAP are comparable to those of BLIP-2.

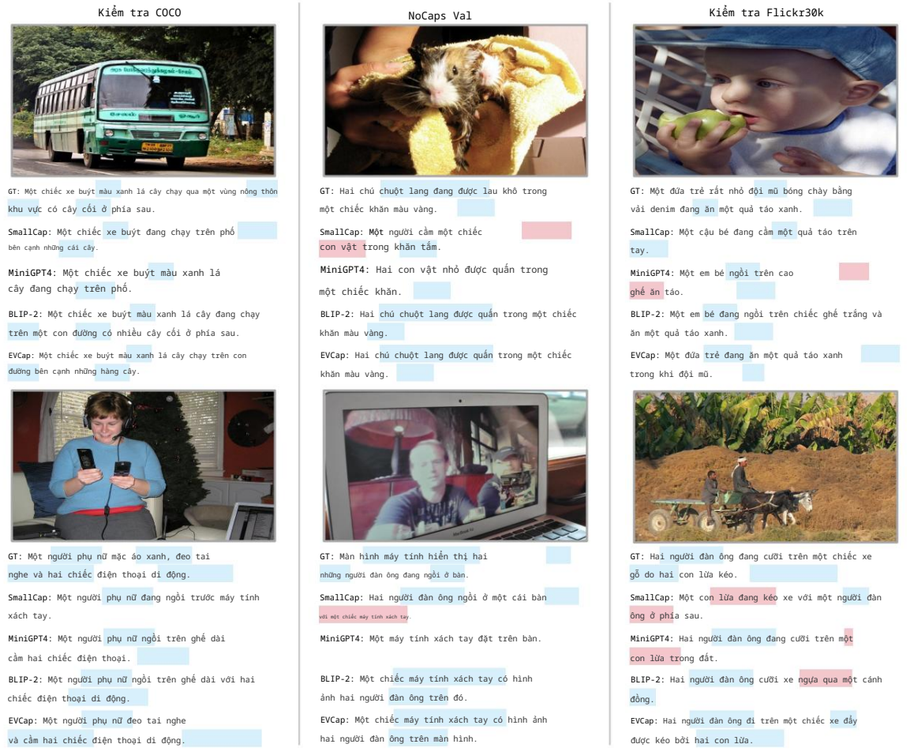
5.2. Results on commonsense-violating data

To explore our EVCAP's capability in describing contents in open-word settings, we further evaluate it on WHOOPS dataset, which contains commonsense-violating images.

Quantitative results. In Tab. 2, we compare the performance of EVCAP, MiniGPT4, BLIP, and BLIP-2 on

Figure 4. Examples of captions generated by our EVCAP, EVCAP (w/ WHOOPS), and three SOTAs on WHOOPS dataset. Incorrect objects are highlighted in red, while correct ones are in blue.

WHOOPS dataset. This dataset is particularly challenging due to its inclusion of unusual objects [6]. Initially, as an end-to-end trained model, our EVCAP exhibits performance similar to MiniGPT4. However, there is a noticeable improvement in the CIDEr score, after the external memory is enriched with 2396 new objects from the WHOOPS dataset, each represented by 5 synthesized images generated using stable diffusion [36]. It highlights the effectiveness of



Hình 3. Ví dụ về chú thích được tạo bởi EVCAP và ba phương pháp SOTA của chúng tôi trên bộ kiểm tra COCO, bộ xác thực NoCaps và Flickr30k bộ kiểm tra. GT đề cập đến chú thích Ground Truth. Các đối tượng không chính xác trong chú thích được đánh dấu màu, trong khi những cái đúng có màu xanh lam. đỏ EVCAP tạo chú thích chính xác trên các tập dữ liệu khác nhau, cho thấy hiệu suất tương đương với BLIP-2.

các mô hình được đào tạo trước. Là một phương pháp chú thích nhẹ và được tăng cường khả năng truy xuất, SmallCap gặp khó khăn trong việc tạo ra các chú thích chính xác cho các hình ảnh nhất định, chủ yếu là vì nó dựa vào các chú thích đã lấy được có chứa thông tin không liên quan. MiniGPT4, mặc dù được liên kết với chính nội dung của hình ảnh, đôi khi thiếu một số đối tượng như "cây cối" và "tai nghe". Sự giám sát này bắt nguồn từ tập trung vào các đối tượng chính trong hình ảnh, mà không tích hợp các tín hiệu bổ sung cho các đối tượng khác được cung cấp bởi các đối tượng được lấy lại tên đối tượng. Ngược lại, các chú thích được tạo ra bởi EVCAP có thể so sánh với BLIP-2.

5.2. Kết quả về dữ liệu vi phạm lẽ thường

Để khám phá khả năng mô tả nội dung của EVCAP trong các thiết lập từ ngữ mở, chúng tôi tiếp tục đánh giá nó trên WHOOPS tập dữ liệu có chứa những hình ảnh vi phạm lẽ thường. Kết quả định lượng. Trong Tab. 2, chúng tôi so sánh hiệu suất của EVCAP, MiniGPT4, BLIP và BLIP-2 trên

Hình 4. Ví dụ về chú thích được tạo bởi EVCAP của chúng tôi, EVCAP (w/ WHOOPS) và ba SOTA trên tập dữ liệu WHOOPS. Không đúng các đối tượng được tô sáng màu đỏ, trong khi những cái đúng thì có màu xanh.

Bộ dữ liệu WHOOPS. Bộ dữ liệu này đặc biệt khó khăn do bao gồm các đối tượng bất thường [6]. Ban đầu, như một mô hình được đào tạo từ đầu đến cuối, EVCAP của chúng tôi thể hiện hiệu suất tương tự như MiniGPT4. Tuy nhiên, có một điều đáng chú ý cải thiện điểm số CIDEr sau khi bộ nhớ ngoài được làm giàu với 2396 đối tượng mới từ WHOOPS tập dữ liệu, mỗi tập được biểu diễn bằng 5 hình ảnh tổng hợp được tạo ra sử dụng sự khuếch tán ổn định [36]. Nó làm nổi bật hiệu quả của

Table 2. Quantitative results on commonsense-violating data – WHOOPS dataset. EVCAP (w/ WHOOPS) denotes EVCAP using the memory expanded by WHOOPS objects. The results reveal the open-world comprehension ability and expandability of EVCAP.

Method	B@4	M	C	S
Only pre-trained models				
BLIP [24] (from [6])	13	–	65	–
BLIP-2 FlanT5-XXL [25] (from [6])	31	–	120	–
BLIP-2 FlanT5-XXL [25] (reproduced)	28	26.7	93.1	17.9
Finetuned models on COCO				
MiniGPT4 [46]	24.2	26.7	84.8	18.2
BLIP [24]	22.9	25.0	79.3	17.1
BLIP-2 FlanT5-XL [25]	25.8	27.0	89.1	18.3
End-to-end trained models on COCO				
EVCAP	24.1	26.1	85.3	17.7
EVCAP (w/ WHOOPS)	24.4	26.1	86.3	17.8

Table 3. Ablation study on components prior to the LLM decoder in EVCAP. The result of "+ Attentive fusion" demonstrates the substantial impact of the external visual–name memory.

Method	COCO test		NoCaps val		Flickr30k test	
	C	S	C	S	C	S
ViT + Q-Former (Baseline)	134.4	23.9	108.8	14.2	76.8	17.3
+ Image query tokens (Baseline+)	134.1	23.8	109.0	14.3	77.3	17.2
+ Attentive fusion (EVCAP)	140.1	24.7	119.3	15.3	84.4	18.0

GT: A hamster on a blanket with a hand behind it.
Baseline: A hamster sitting on a couch with a person's hand near it.
Baseline+: A close up of a small hamster on a bed.
EVCap: A small hamster sitting on a blanket next to a hand.
(pillow, hamster, grizzly, gameboard, kitten, spectacles, cat, blanket, giraffe, pug-dog)

Figure 5. Visualization of the captions generated from ablation study on the NoCaps validation set. We also show the retrieved object names by EVCAP, presented in gray. Incorrect objects in captions are highlighted in red, while correct ones are in blue.

our idea of incorporating an expandable external memory into the captioning model for open-world comprehension.

Qualitative results. Fig. 4 illustrates the captions generated by EVCAP, EVCAP (w/WHOOPS), and three SOTAs for one image from the WHOOPS dataset. Similar to other methods except for BLIP-2, EVCAP can not recognize *"blue cartoon character"* as *"Pikachu"*, while EVCAP (w/WHOOPS) successfully predicts it because of the updated memory. SmallCap and MiniGPT4 tend to generate captions with hallucinatory objects, a result of commonsense-violating contents present in the images.

5.3. Detailed analysis

Ablation study. We assess the contribution of each component prior to the LLM decoder in EVCAP by incrementally integrating the image query tokens and the attentive fusion module into our baseline model. The baseline model comprises a ViT+Q-Former, a linear layer, and a LLM decoder.

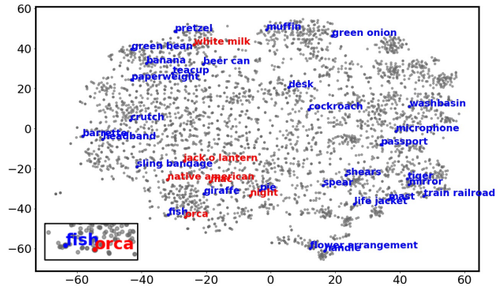


Figure 6. Visualization of the visual features in external memory using t-SNE. For visual features in LVIS dataset's objects (**blue**), the related objects fall in the same cluster. After adding more visual features of synthesized images from WHOOPS' objects, new objects (**red**) are located at appropriate clusters (zoom-in view).

The quantitative results are shown in Tab. 3. When employing only the baseline model (Baseline), CIDEr scores drop notably by 5.7, 10.5, and 7.6 on COCO, NoCaps, and Flickr30k, respectively. The inclusion of trainable image query tokens (Baseline+) brings a marginal improvement on NoCaps and Flickr30k. However, the performance is significantly enhanced with the addition of attentive fusion (along with the introduction of external memory), indicating the pivotal role of the external visual–name memory in the overall effectiveness of EVCAP. This is further corroborated by the qualitative results in Fig. 5, where captions from Baseline and Baseline+ inaccurately include objects like *"couch"* and *"bed"*, and Baseline+ overlooks *"hand"*.

Exploration for external memory expandability. To demonstrate the scalability of the external memory in EVCAP, we visualize the visual features stored in LVIS external memory, and newly synthesized data from objects appearing in the WHOOPS dataset. We employ t-SNE [38] to plot visual features after reducing their dimensions to 2-D (Fig. 6). For clear visualization, we only randomly display 3649 visual features in LVIS memory, and add 479 visual features from WHOOPS objects. Among them, 35 samples are randomly labeled. The result shows a clear clustering of LVIS objects (blue) in the external memory, as well as the successful integration and appropriate localization of new objects from WHOOPS (red) into these clusters. This pattern not only confirms the distinctiveness of visual features already present in the memory but also demonstrates the potential to accurately incorporate and differentiate new objects introduced from updated data. These findings highlight our external memory's ability to expand and maintain its effectiveness even as new data is incorporated.

Impact of external memory size. We examine the impact of external memory size in Tab. 4. On the one hand, we randomly remove 30%, 60%, and 90% data in the external

Bảng 2. Kết quả định lượng về dữ liệu vi phạm lẽ thường – Bộ dữ liệu WHOOPS. EVCAP (w/ WHOOPS) biểu thị EVCAP bằng cách sử dụng bộ nhớ được mở rộng bởi các đối tượng WHOOPS. Kết quả cho thấy khả năng hiểu thế giới mở và khả năng mở rộng của EVCAP.

Phương pháp	B@4	MCS		
Chỉ có các mô hình được đào tạo trước				
BLIP [24] (từ [6])	13	-		-
BLIP-2 FlanT5-XXL [25] (từ [6])	31	-	65	-
BLIP-2 FlanT5-XXL [25] (đã tái tạo)	28	120	26,7	93,1 17,9
Các mô hình tinh chỉnh trên COCO				
MiniGPT4 [46]	24,2	26,7	84,8	18,2
BLIP [24]	22,9	25,0	79,3	17,1
BLIP-2 FlanT5-XL [25]	25,8	27,0	89,1	18,3
Các mô hình được đào tạo đầu cuối trên COCO				
EVCAP 24,1 26,1 85,3 17,7				
EVCAP (có WHOOPS) 24,4 26,1 86,3 17,8				

Bảng 3. Nghiên cứu cắt bỏ các thành phần trước bộ giải mã LLM trong EVCAP. Kết quả của "+ Sự hợp nhất chú ý" chứng minh tác động đáng kể của trí nhớ hình ảnh-tên bên ngoài.

Phương pháp	Kiểm tra COCO		NoCaps val		Kiểm tra Flickr30k
	C	S	C	SC	S
ViT + Q-Former (Cơ bản)	134,4	23,9	108,8	14,2 76,8	17.3
+ Mã thông báo truy vấn hình ảnh (Baseline+)	134,1	23,8	109,0	14,3 77,3	17.2
+ Hợp nhất chú ý (EVCAP)	140,1	24,7	119,3	15,3 84,4	18.0

GT: Một con chuột lang trên chân với một bàn tay ở phía sau.
Đường cơ sở: Một chú chuột lang đang ngồi trên ghế dài với bàn tay của một người ở gần nó.
Baseline+: Cận cảnh một chú chuột hamster nhỏ trên giường.
EVCap: Một chú chuột hamster nhỏ ngồi trên chăn bên cạnh một bàn tay.
(gối, chuột đồng, gấu xám, bàn cờ, mèo con, kính, mèo, chăn, hươu cao cổ, chó pug)

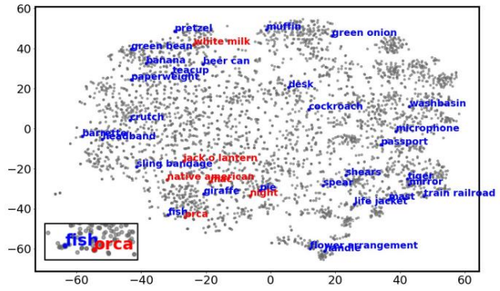
Hình 5. Hình ảnh hóa các chú thích được tạo ra từ quá trình cắt bỏ nghiên cứu về bộ xác thực NoCaps. Chúng tôi cũng hiển thị các dữ liệu đã thu thập được tên đối tượng theo EVCAP, được trình bày bằng màu xám. Các đối tượng không chính xác trong chú thích được tô sáng màu đỏ , trong khi những cái đúng thì có màu xanh .

ý tưởng của chúng tôi về việc kết hợp bộ nhớ ngoài có thể mở rộng vào mô hình chú thích để hiểu thế giới mở.

Kết quả định tính. Hình 4 minh họa các chú thích được tạo ra bởi EVCAP, EVCAP (có WHOOPS) và ba SO-TA cho một hình ảnh từ tập dữ liệu WHOOPS. Tương tự như các phương pháp khác ngoại trừ BLIP-2, EVCAP không thể nhận ra "nhân vật hoạt hình màu xanh" là "Pikachu", trong khi EVCAP (w/WHOOPS) dự đoán thành công vì bộ nhớ được cập nhật. SmallCap và MiniGPT4 có xu hướng tạo chú thích với các đối tượng ảo giác, kết quả của nội dung vi phạm lẽ thường có trong hình ảnh.

5.3. Phân tích chi tiết

Nghiên cứu cắt bỏ. Chúng tôi đánh giá sự đóng góp của từng thành phần trước bộ giải mã LLM trong EVCAP bằng cách gia tăng tích hợp các mã thông báo truy vấn hình ảnh và sự hợp nhất chú ý mô-đun vào mô hình cơ sở của chúng tôi. Mô hình cơ sở bao gồm ViT+Q-Former, một lớp tuyến tính và bộ giải mã LLM.



Hình 6. Hình ảnh trực quan của các đặc điểm hình ảnh trong bộ nhớ ngoài sử dụng t-SNE. Đối với các tính năng trực quan trong các đối tượng của tập dữ liệu LVIS (màu xanh lam), các đối tượng liên quan nằm trong cùng một nhóm. Sau khi thêm nhiều tính năng trực quan hơn của hình ảnh tổng hợp từ các đối tượng của WHOOPS, các đối tượng (màu đỏ) được đặt ở các cụm thích hợp (chế độ xem phóng to).

Kết quả định lượng được thể hiện trong Tab. 3. Khi chỉ sử dụng mô hình cơ sở (Baseline), CIDEr đạt điểm giảm đáng kể 5,7, 10,5 và 7,6 trên COCO, NoCaps và Flickr30k, tương ứng. Việc đưa vào hình ảnh có thể đào tạo mã thông báo truy vấn (Baseline+) mang lại sự cải thiện nhỏ trên NoCaps và Flickr30k. Tuy nhiên, hiệu suất là được tăng cường đáng kể với việc bổ sung sự hợp nhất chu đáo (cùng với sự ra đời của bộ nhớ ngoài), chỉ ra vai trò quan trọng của bộ nhớ hình ảnh-tên bên ngoài trong hiệu quả tổng thể của EVCAP. Điều này được xác nhận thêm được diễn đạt bằng các kết quả định tính trong Hình 5, trong đó có chú thích từ Baseline và Baseline+ bao gồm các đối tượng không chính xác giống như "ghế dài" và "giường", và Baseline+ bỏ qua "bàn tay".

Khám phá khả năng mở rộng bộ nhớ ngoài. Để chứng minh khả năng mở rộng của bộ nhớ ngoài trong EV- CAP, chúng tôi hình dung các tính năng trực quan được lưu trữ trong bộ nhớ ngoài LVIS và dữ liệu mới tổng hợp từ các đối tượng xuất hiện trong tập dữ liệu WHOOPS. Chúng tôi sử dụng t-SNE [38] để vẽ các đặc điểm trực quan sau khi giảm kích thước của chúng xuống 2 chiều (Hình 6). Để hình dung rõ ràng, chúng tôi chỉ hiển thị ngẫu nhiên 3649 tính năng trực quan trong bộ nhớ LVIS và thêm 479 tính năng trực quan các tính năng từ các đối tượng WHOOPS. Trong số đó, 35 mẫu được dán nhãn ngẫu nhiên. Kết quả cho thấy một cụm rõ ràng các đối tượng LVIS (màu xanh lam) trong bộ nhớ ngoài, cũng như như sự tích hợp thành công và bản địa hóa phù hợp của các đối tượng mới từ WHOOPS (màu đỏ) vào các cụm này. Điều này mẫu không chỉ xác nhận tính đặc biệt của các đặc điểm thị giác đã có trong bộ nhớ mà còn chứng minh khả năng kết hợp và phân biệt chính xác các sản phẩm mới các đối tượng được đưa vào từ dữ liệu được cập nhật. Những phát hiện này làm nổi bật khả năng mở rộng và duy trì bộ nhớ ngoài của chúng ta hiệu quả của nó ngay cả khi dữ liệu mới được đưa vào.

Tác động của kích thước bộ nhớ ngoài. Chúng tôi kiểm tra tác động của kích thước bộ nhớ ngoài trong Tab. 4. Một mặt, chúng tôi xóa ngẫu nhiên 30%, 60% và 90% dữ liệu trong bộ nhớ ngoài

Table 4. Impact of the external memory size on the performance of EVCAP by evaluation under CIDEr scores. Changes in the size of external memory result in changes in performance.

Method	NoCaps val				Flickr30k
	In	Near	Out	Overall	Test
LVIS objects (EVCAP)	111.7	119.5	116.5	119.3	84.4
– 30% LVIS	112.0	119.2	115.3	118.8	85.0
– 60% LVIS	111.4	119.1	116.2	119.0	85.1
– 90% LVIS	110.6	118.2	115.8	118.3	83.6
+ WHOOPS	110.7	118.9	116.7	119.0	84.9

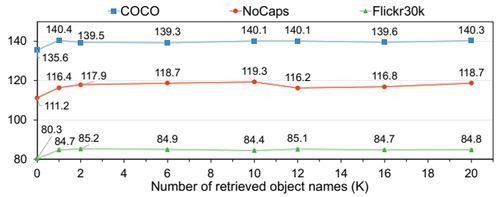


Figure 7. CIDEr scores after training EVCAP with the number of retrieved object names K from 0 to 20. The results indicate that the performance is relatively optimal when K is set to be 10.

memory constructed from LVIS objects. The results show the performance gradually degrades on NoCaps as reducing 30% and 90% LVIS. Despite some unexpected increases in certain results on NoCaps (5th row) and Flickr30k (4th - 5th rows), they do not alter the overall downward trend. Similar phenomena are also noted in SmallCap [35], we speculate it is due to data distribution. On the other hand, as we infuse WHOOPS knowledge into LVIS memory, there is a slight improvement on NoCaps (out) and Flickr30k. These observations validate the model's capability to effectively retrieve object names from an updated memory, enhancing its performance in generating captions.

Impact of the number of retrieved object names. We investigate how the number of retrieved object names K (Sec. 3.3) affect EVCAP in Fig. 7. We train the model with K from 0 to 20 and evaluate the performance under CIDEr on all three benchmarks. From the results, we can find that the model works worst on the out-domain dataset (NoCaps) with zero object names. It confirms that the performance boost from Baseline+ to EVCAP (Tab. 3) is primarily attributed to the retrieval-augmented mechanism, but not the customized Q-Former itself. With more object names, performance fluctuates but improves. Furthermore, we observe that setting K to 10 yields relatively optimal overall performance, validating the choice of $K = 10$ in EVCAP.

Analysis with different decoders. To explore the influence of different LLMs decoders on our EVCAP, we experiment by substituting Vicuna-13B with GPT2 and Vicuna-7B, as detailed in Tab. 5. With GPT2 as the decoder, EVCAP still markedly surpasses other GPT2-based models, achieving

Table 5. Analysis with different LLM decoders including GPT2, Vicuna-7B, and Vicuna-13B. The results reveal EVCAP is effective when applying it in different LLM decoders.

Method	LLM	COCO test		NoCaps val		Flickr30k test	
		C	S	C	S	C	S
SmallCap [35]	GPT2	119.7	21.3	–	–	60.6	–
ViECap [15]	GPT2	92.9	18.2	66.2	9.5	47.9	13.6
EVCAP	GPT2	131.0	23.2	97.6	13.3	70.6	16.1
MiniGPT4 [46]	Vicuna-7B	119.4	23.5	108.7	15.7	73.9	17.2
InstructBLIP [11]	Vicuna-7B	–	–	123.1	–	82.4	–
EVCAP	Vicuna-7B	139.0	24.7	116.8	15.3	82.7	18.0
MiniGPT4 [46]	Vicuna-13B	129.6	23.4	108.8	15.1	78.4	16.9
InstructBLIP [11]	Vicuna-13B	–	–	121.9	–	82.8	–
EVCAP	Vicuna-13B	140.1	24.7	119.3	15.3	84.4	18.0

impressive gains of 11.3 and 10.0 under CIDEr on COCO and Flickr30k, compared to SmallCap. When employing Vicuna-7B, the comparison of performance trends mirrors those observed with Vicuna-13B, further attesting to the robustness and adaptability of EVCAP across different LLM decoders. Notably, both SmallCap, which retrieves captions, and our GPT2-based EVCAP, which retrieves object names, use the same GPT2 decoder. Therefore, their comparison also underscores the effectiveness of our method's object name retrieval and attentive fusion strategy.

Limitations. First, EVCAP cannot retrieve all objects that appear in the given image due to the memory coverage limits, leading to incomplete image descriptions (Fig. 4). We will investigate integrating object detection with image captioning to enhance completeness. Second, our focus on object representation restricts consideration of other crucial captioning elements, affecting overall performance. Similar to all models trained with COCO dataset, EVCAP has limitations in generating varied styles, which is reflected in our relatively modest performance improvements in Tab. 2, compared to MiniGPT4. We will overcome it by exploring methodologies that encourage style diversity in the future.

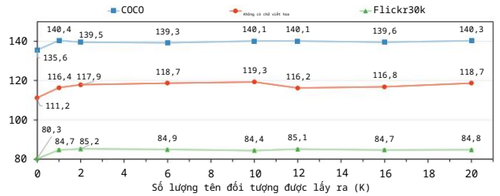
6. Conclusion

We further advance image captioning in real-world scenarios by introducing EVCAP, a novel image captioning model with object names retrieved from an external visual–name memory. The external memory is easily expandable, allowing for effortless updates with new object visuals and names. We extensively compare EVCAP with SOTAs on various benchmarks and commonsense-violating data, demonstrating its significant superiority in performance.

Acknowledgements. This work was supported by JSPS/MEXT KAKENHI Grant Numbers JP22H05015, JP23H03449, JP23KJ0404, and JP22K17947, and the commissioned research (No. 225) by the National Institute of Information and Communications Technology (NICT), Japan, and the Institute of AI and Beyond of the University of Tokyo, and ROIS NII Open Collaborative Research 2023-23FC01, 2024-24S1201.

Bảng 4. Tác động của kích thước bộ nhớ ngoài đến hiệu suất của EVCAP bằng cách đánh giá theo điểm số CIDEr. Những thay đổi về kích thước của bộ nhớ ngoài dẫn đến những thay đổi về hiệu suất.

Phương pháp	NoCaps giá trị		Flickr30k
	Gần Ra Tổng thể		
Đối tượng LVIS (EVCAP) 111,7 119,5 116,5 - 30% LVIS 112,0	119,3		84,4
119,2 115,3 - 60% LVIS 111,4 119,1 116,2 - 90% LVIS	118,8		85.0
110,6 118,2 115,8 + WHOOPS 110,7 118,9 116,7	119,0		85,1
	118,3		83,6
	119,0		84,9



Hình 7. Điểm số của CIDEr sau khi đào tạo EVCAP với số lượng đã lấy tên đối tượng K từ 0 đến 20. Kết quả cho thấy rằng hiệu suất tương đối tối ưu khi K được đặt bằng 10.

bộ nhớ được xây dựng từ các đối tượng LVIS. Kết quả cho thấy hiệu suất giảm dần trên NoCaps khi giảm 30% và 90% LVIS. Mặc dù có một số sự gia tăng bất ngờ trong một số kết quả trên NoCaps (hàng thứ 5) và Flickr30k (hàng thứ 4 - thứ 5) hàng), chúng không làm thay đổi xu hướng giảm chung. Tương tự hiện tượng cũng được ghi nhận trong SmallCap [35], chúng tôi suy đoán nó là do sự phân phối dữ liệu. Mặt khác, khi chúng ta đưa kiến thức WHOOPS vào bộ nhớ LVIS, có một cải thiện nhẹ trên NoCaps (ra) và Flickr30k. Những các quan sát xác nhận khả năng của mô hình để có hiệu quả lấy tên đối tượng từ bộ nhớ được cập nhật, nâng cao hiệu suất của nó trong việc tạo phụ đề.

Tác động của số lượng tên đối tượng được lấy lại. Chúng tôi điều tra cách số lượng tên đối tượng được lấy ra K (Phần 3.3) ảnh hưởng đến EVCAP trong Hình 7. Chúng tôi đào tạo mô hình với K từ 0 đến 20 và đánh giá hiệu suất theo CIDEr trên cả ba chuẩn mực. Từ kết quả, chúng ta có thể thấy rằng mô hình hoạt động kém nhất trên tập dữ liệu miền ngoài (NoCaps) với số không tên đối tượng. Nó xác nhận rằng hiệu suất sự tăng cường từ Baseline+ lên EVCAP (Tab. 3) chủ yếu là do cơ chế tăng cường truy xuất, nhưng không phải Q-Former tùy chỉnh. Với nhiều tên đối tượng hơn, hiệu suất dao động nhưng được cải thiện. Hơn nữa, chúng tôi quan sát việc thiết lập K thành 10 mang lại hiệu suất tổng thể tương đối tối ưu, xác nhận sự lựa chọn K = 10 trong EVCAP.

Phân tích với các bộ giải mã khác nhau. Để khám phá ảnh hưởng của các bộ giải mã LLM khác nhau trên EVCAP của chúng tôi, chúng tôi thử nghiệm bằng cách thay thế Vicuna-13B bằng GPT2 và Vicuna-7B, như được trình bày chi tiết trong Tab. 5. Với GPT2 là bộ giải mã, EVCAP vẫn vượt trội đáng kể so với các mô hình dựa trên GPT2 khác, đạt được

Bảng 5. Phân tích với các bộ giải mã LLM khác nhau bao gồm GPT2, Vicuna-7B và Vicuna-13B. Kết quả cho thấy EVCAP có hiệu quả khi áp dụng trong các bộ giải mã LLM khác nhau.

Phương pháp	Thạc sĩ Luật	Kiểm tra COCO		NoCaps val Kiểm tra Flickr30k			
		C	S	C	S	C	S
Vốn hóa nhỏ [35]	GPT2	119,7 21,3 - 60,6 92,9 18,2					-
ViECap [15]	GPT2	66,2 9,5 47,9 131,0 23,2 97,6					13.6
EVCAP	GPT2	13,3 70,6					16.1
MiniGPT4 [46]	Vicuna-7B 119,4 23,5 108,7 15,7 73,9 - 123,1						17.2
Hướng dẫnBLIP [11] Vicuna-7B		-		82,4 Vicuna-7B-139,0			-
EVCAP	24,7 116,8 15,3 82,7						18.0
MiniGPT4 [46]	Vicuna-13B 129,6 23,4 108,8 15,1 78,4 Hướng dẫnBLIP [11]						16,9
Vicuna-13B - 121,9 82,8 EVCAP Vicuna-13B 140,Ĩ 24,7 119,3 15,3 84,4				-			-
							18.0

mức tăng ấn tượng 11,3 và 10,0 theo CIDEr trên COCO và Flickr30k, so với SmallCap. Khi sử dụng Vicuna-7B, sự so sánh các xu hướng hiệu suất phản ánh những quan sát được với Vicuna-13B, tiếp tục chứng minh tính mạnh mẽ và khả năng thích ứng của EVCAP trên các LLM khác nhau bộ giải mã. Đáng chú ý là cả SmallCap, lấy phụ đề và EVCAP dựa trên GPT2 của chúng tôi, lấy đối tượng tên, sử dụng cùng một bộ giải mã GPT2. Do đó, việc so sánh của họ cũng nhấn mạnh tính hiệu quả của phương pháp của chúng tôi tìm kiếm tên đối tượng và chiến lược hợp nhất có chủ đích.

Hạn chế. Đầu tiên, EVCAP không thể truy xuất tất cả các đối tượng xuất hiện trong hình ảnh đã cho do giới hạn phạm vi bộ nhớ, dẫn đến mô tả hình ảnh không đầy đủ (Hình 4). Chúng tôi sẽ điều tra việc tích hợp phát hiện đối tượng với chú thích hình ảnh để tăng cường tính hoàn chỉnh. Thứ hai, sự tập trung của chúng tôi vào biểu diễn đối tượng hạn chế việc xem xét các yếu tố quan trọng khác các yếu tố chú thích, ảnh hưởng đến hiệu suất tổng thể. Tương tự như tất cả các mô hình được đào tạo với tập dữ liệu COCO, EVCAP có những hạn chế trong việc tạo ra các phong cách đa dạng, điều này được phản ánh trong cải thiện hiệu suất tương đối khiêm tốn của chúng tôi trong Tab. 2, so với MiniGPT4. Chúng ta sẽ khắc phục nó bằng cách khám phá phương pháp khuyến khích sự đa dạng về phong cách trong tương lai.

6. Kết luận

Chúng tôi tiếp tục cải tiến chú thích hình ảnh trong các tình huống thực tế bằng cách giới thiệu EVCAP, một chú thích hình ảnh mới mô hình với tên đối tượng được lấy từ hình ảnh bên ngoài- bộ nhớ tên. Bộ nhớ ngoài có thể dễ dàng mở rộng, cho phép cập nhật dễ dàng với hình ảnh đối tượng mới và tên. Chúng tôi so sánh rộng rãi EVCAP với SOTA trên nhiều chuẩn mực khác nhau và dữ liệu vi phạm lẽ thường, chứng minh tính ưu việt đáng kể của nó về hiệu suất.

Lời cảm ơn. Công trình này được hỗ trợ bởi JSPS/MEXT Số tài trợ của KAKENHI JP22H05015, JP23H03449, JP23KJ0404, và JP22K17947, và nghiên cứu được ủy quyền (Số 225) của Viện Thông tin và Công nghệ truyền thông (NICT), Nhật Bản và Viện AI và hơn thế nữa của Đại học Tokyo và ROIS NII Open Nghiên cứu hợp tác 2023-23FC01, 2024-24S1201.

References

[1] https://huggingface.co/spaces/RitaParadaRamos/SmallCapDemo. 5

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 5

[6] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 4, 6, 7

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[8] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. PaLI-X: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 1, 2, 5

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 5

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 2, 4

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3, 5, 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[13] Jiashuo Fan, Yaoyuan Liang, Leyao Liu, Shaolun Huang, and Lei Zhang. Rca-noc: Relative contrastive alignment for novel object captioning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 5

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4

[15] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 4, 5, 8

[16] Zhengcong Fei. Memory-augmented image captioning. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2, 5

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[18] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[19] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Saenko Kate, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[20] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 5

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 3

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 4

[23] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of*

Tài liệu tham khảo

[1] https://hugingface.co/không gian/RitaParadaRamos/SmallCapDemo. 5

[2] Agrawal khắc nghiệt, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Ste-fan Lee và Peter Anderson. Nocaps: Chú thích đối tượng mới lạ ở quy mô lớn. Trong Proc. Hội nghị quốc tế IEEE về Tầm nhìn máy tính (ICCV), 2019. 4

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A mô hình ngôn ngữ trực quan cho việc học ít cảnh quay. Trong Tiến bộ trong Hệ thống xử lý thông tin thần kinh (NeurIPS), 2022. 1

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. Sự chú ý từ dưới lên và từ trên xuống cho chú thích hình ảnh và trả lời câu hỏi trực quan. Trong Proc. IEEE Conference on Tầm nhìn máy tính và nhận dạng mẫu (CVPR), 2018. 2

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Chu và Jingren Zhou. Qwen-vl: Một mô hình ngôn ngữ tầm nhìn lớn biên giới với khả năng đa dạng. bản in trước arXiv arXiv:2308.12966, 2023. 5

[6] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Lud-wig Schmidt, Yuval Elovici, Gabriel Stanovsky và Roy Schwartz. Phá vỡ lẽ thường: Ồ! Một chuẩn mực về tầm nhìn và ngôn ngữ của hình ảnh tổng hợp và sáng tác. Trong Proc. Hội nghị quốc tế IEEE về tầm nhìn máy tính (ICCV), 2023. 4, 6, 7

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Sub-biah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakan-tan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Các mô hình ngôn ngữ là những người học ít lần. Trong Những tiến bộ trong thần kinh Hệ thống xử lý thông tin (NeurIPS), 2020. 2

[8] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Se-bastian Goodman, Xiao Wang, Yi Tay, et al. PaLI-X: Bật mở rộng mô hình ngôn ngữ và tầm nhìn đa ngôn ngữ. arXiv bản in trước arXiv:2305.18565, 2023. 1, 2, 5

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weichen Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby và Radu Soricut. PaLI: Mô hình ngôn ngữ-hình ảnh đa ngôn ngữ được chia tỷ lệ chung. Trong Hội nghị quốc tế về Biểu diễn học tập (ICLR), 2023. 1, 2, 5

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez và cộng sự. Vicuna: Một nguồn mở chatbot gây ấn tượng với gpt-4 với chất lượng chatgpt 90%*. Xem https://vicuna.lmsys.org (truy cập ngày 14 tháng 4 năm 2023), 2023. 2, 4

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung và Steven Hoi. Instructblip: Hướng tới các mô hình ngôn ngữ thị giác mục đích chung với điều chỉnh hướng dẫn. Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh (NeurIPS), 2023. 3, 5, 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, Jakob Uszkoreit và Neil Houlsby. Một hình ảnh là giá trị 16x16 từ: Máy biến áp để nhận dạng hình ảnh tại quy mô. Trong Hội nghị quốc tế về Biểu diễn học tập (ICLR), 2021. 2

[13] Jiashuo Fan, Yaoyuan Liang, Leyao Liu, Shaolun Huang, và Lei Zhang. Rca-noc: Căn chỉnh tương phản tương đối cho chú thích đối tượng mới lạ. Trong Proc. Hội nghị quốc tế IEEE về tầm nhìn máy tính (ICCV), 2023. 2, 5

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Vương Hành Cương, Hoàng Thiết Quân, Vương Xinlong và Nhạc Cao. Eva: Khám phá giới hạn của việc học biểu diễn hình ảnh có mặt nạ ở quy mô lớn. Trong Proc. IEEE Conference on Comp-computer Vision and Pattern Recognition (CVPR), 2023. 2, 4

[15] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang và Feng Zheng. Giải mã có thể chuyển giao bằng hình ảnh các thực thể cho chú thích hình ảnh zero-shot. Trong Proc. IEEE Hội nghị quốc tế về tầm nhìn máy tính (ICCV), 2023. 2, 4, 5, 8

[16] Trịnh Công Phi. Chú thích hình ảnh tăng cường trí nhớ. Trong Proc. Hội nghị AAAI về Trí tuệ nhân tạo (AAAI), 2021. 2, 5

[17] Agrim Gupta, Piotr Dollar và Ross Girshick. LVIS: A bộ dữ liệu để phân đoạn trường hợp từ vựng lớn. Trong Proc. Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2019. 3

[18] Ryuichiro Hataya, Han Bao và Hiromi Arai. Các mô hình tạo sinh quy mô lớn có làm hỏng các tập dữ liệu trong tương lai không? Trong Proc. Hội nghị quốc tế IEEE về tầm nhìn máy tính (ICCV), 2023. 3

[19] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Saenko Kate và Trevor Dar-rell. Chú thích thành phần sâu: Mô tả các danh mục đối tượng mới mà không có dữ liệu đào tạo được ghép nối. Trong Proc. IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2016. 2

[20] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross và Alireza Fathi. Tiết lộ: Ngôn ngữ hình ảnh tăng cường khả năng truy xuất đào tạo trước với bộ nhớ kiến thức đa phương thức đa nguồn. Trong Proc. Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu (CVPR), 2023. 1, 2, 5

[21] Jeff Johnson, Matthijs Douze và Herve J egou. Tìm kiếm sự tương đồng ở quy mô tỷ với GPU. Giao dịch IEEE về Big Dữ liệu, 7(3):535–547, 2019. 3

[22] Andrej Karpathy và Li Fei-Fei. Sự sắp xếp ngữ nghĩa thị giác sâu sắc để tạo ra các mô tả hình ảnh. Trong Proc. IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2015. 2, 4

[23] Jacob Devlin Ming-Wei Chang Kenton và Lee Kristina Toutanova. Bert: Đào tạo trước các bộ chuyển đổi song hướng sâu để hiểu ngôn ngữ. Trong Biên bản báo cáo

North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019. 4

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. International conference on machine learning (ICML)*, 2022. 5, 7

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. International conference on machine learning (ICML)*, 2023. 1, 2, 3, 4, 5, 7

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 4

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[29] Ron Mokady, Amir Hertz, and Amit H Bermano. Clip-cap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 3, 5

[30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 4

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International conference on machine learning (ICML)*, 2021. 2

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(1):5485–5551, 2020. 2

[34] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. In *Proc. Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023. 2

[35] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 5, 8

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 4

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research (JMLR)*, 9(11), 2008. 7

[39] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[40] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[41] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 4, 5

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. International conference on machine learning (ICML)*. 2

[43] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-ViLM: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023. 2

[44] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[45] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5

[46] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 4, 5, 7, 8

Chi nhánh Bắc Mỹ của Hiệp hội Ngôn ngữ học tính toán: Công nghệ ngôn ngữ của con người (NAACL-HLT), 2019. 4

[24] Junnan Li, Dongxu Li, Caiming Xiong, và Steven Hoi. Blip: Khởi động quá trình đào tạo trước hình ảnh ngôn ngữ cho hợp nhất hiểu và tạo ra thị giác-ngôn ngữ. Trong Proc. Hội nghị quốc tế về máy học (ICML), 2022. 5, 7

[25] Junnan Li, Dongxu Li, Silvio Savarese, và Steven Hoi. Blip-2: Khởi động quá trình đào tạo trước hình ảnh ngôn ngữ với bộ mã hóa hình ảnh đóng băng và mô hình ngôn ngữ lớn. Trong Proc. Hội nghị quốc tế về máy học (ICML), 2023. 1, 2, 3, 4, 5, 7

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Đối tượng-ngữ nghĩa được căn chỉnh trước khi đào tạo cho nhiệm vụ ngôn ngữ thị giác. Trong Proc. Hội nghị Châu Âu về Tầm nhìn máy tính (ECCV), 2020. 2, 5

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar và C Lawrence Zitnick. Microsoft coco: Các đối tượng phổ biến trong ngữ cảnh. Trong Proc. Hội nghị Châu Âu về Tầm nhìn Máy tính (ECCV), 2014. 4

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, và Devi Parikh. Trò chuyện của trẻ sơ sinh thần kinh. Trong Proc. Hội nghị IEEE về máy tính Nhận dạng thị giác và mẫu (CVPR), 2018. 2

[29] Ron Mokady, Amir Hertz và Amit H Bermano. Clip-cap: Tiền tố clip cho chú thích hình ảnh. bản in trước arXiv arXiv:2111.09734, 2021. 2, 3, 5

[30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier và Svetlana Lazeb-nik. Các thực thể Flickr30k: Thu thập các tương ứng giữa vùng và cụm từ để tạo ra các mô hình hình ảnh và câu phong phú hơn. Trong Proc. Hội nghị quốc tế IEEE về tầm nhìn máy tính (ICCV), 2015. 4

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Các mô hình ngôn ngữ là những người học đa nhiệm không được giám sát. Blog OpenAI, 1(8):9, 2019. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Học các mô hình trực quan có thể chuyển giao từ tầm nhìn siêu ngôn ngữ tự nhiên. Trong Proc. Hội nghị quốc tế về học máy (ICML), 2021. 2

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Chu, Wei Li, và Peter J Liu. Khám phá giới hạn của việc học chuyển giao với một bộ chuyển đổi văn bản sang văn bản thống nhất. Tạp chí nghiên cứu học máy (JMLR), 21(1):5485–5551, 2020. 2

[34] Rita Ramos, Desmond Elliott và Bruno Martins. Chú thích hình ảnh được tăng cường khả năng truy xuất. Trong Proc. Hội nghị của Chi hội Châu Âu của Hiệp hội Ngôn ngữ học tính toán (EACL), 2023. 2

[35] Rita Ramos, Bruno Martins, Desmond Elliott và Yova Kementchedjhieva. Smallcap: Chú thích hình ảnh nhẹ được nhắc đến với sự tăng cường truy xuất. Trong Proc. IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2023. 1, 2, 3, 5, 8

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser và Bjorn Ommer. Tổng hợp hình ảnh độ phân giải cao với các mô hình khuếch tán tiềm ẩn. Trong Proc. IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2022. 3, 6

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, và những người khác. Llama: Mô hình ngôn ngữ nền tảng mở và hiệu quả. bản in trước arXiv arXiv:2302.13971, 2023. 2, 4

[38] Laurens Van der Maaten và Geoffrey Hinton. Hình dung dữ liệu bằng cách sử dụng t-sne. Tạp chí nghiên cứu học máy (JMLR), 9(11), 2008. 7

[39] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell và Kate Saenko. Chú thích hình ảnh với nhiều đối tượng khác nhau. Trong Proc. Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu (CVPR), 2017. 2

[40] Đức Minh Võ, Hong Chen, Akihiro Sugimoto, và Hideki Nakayama. Noc-rek: Chú thích đối tượng mới lạ với các mục đã lấy từ vựng từ kiến thức bên ngoài. Trong Proc. IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2022. 2, 5

[41] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding và Jie Tang. Cogvlm: Chuyên gia trực quan cho đào tạo trước mô hình ngôn ngữ. bản in trước arXiv arXiv:2311.03079, 2023. 4, 5

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Cho Kyunghyun, Aaron Courville, Ruslan Salakhudinov, Rich Zemel và Yoshua Bengio. Hiển thị, tham dự và kể: Tạo chú thích hình ảnh thần kinh với sự chú ý trực quan. Trong Proc. Hội nghị quốc tế về máy học (ICML). 2

[43] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-ViLM: Mô hình ngôn ngữ trực quan tăng cường truy xuất cho chú thích hình ảnh không có và ít ảnh. Trong Phát hiện của Hiệp hội Ngôn ngữ học tính toán: EMNLP, 2023. 2

[44] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer và Wen-tau Yih. Tăng cường khả năng truy xuất mô hình ngôn ngữ đa phương thức. Trong Hội nghị quốc tế về Biểu diễn Học tập (ICLR), 2023. 2

[45] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lôi Chương, Lijuan Wang, Yejin Choi và Jianfeng Gao. Vinvl: Xem lại các biểu diễn trực quan trong ngôn ngữ thị giác mô hình. Trong Proc. Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu (CVPR), 2021. 2, 5

[46] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, và Mo-hamed Elhoseiny. Minigpt-4: Nâng cao ngôn ngữ thị giác hiểu biết với các mô hình ngôn ngữ lớn tiên tiến. arXiv bản in trước arXiv:2304.10592, 2023. 3, 4, 5, 7, 8