PathRecall: Memory-Augmented Visual Question Answering for Assistive Indoor Navigation

Pei Du
pdu@pdx.edu
Portland State University
Portland, Oregon, USA

Nirupama Bulusu
nbulusu@pdx.edu
Portland State University
Portland, USA

Abstract

Accessibility is crucial to the human experience in the built environment. This paper presents PathRecall, a memory-augmented visual question answering (VQA) system designed to support visually impaired (VI) users in navigating indoor environments. PathRecall integrates three key components: (1) a captioning module that extracts question-relevant keywords and generates semantic descriptions of past visual scenes; (2) a question-aware reranking model, MemoryNet, which jointly embeds the user's question and image features to identify relevant frames. The module enables users to ask free-form questions and returns the most relevant visual frame from recent video history to support situational awareness; and (3) a modified BLIP-2-based VQA model which uses the generated captions as prompts to enhance its ability to understand and interpret the retrieved frame, allowing it to generate more accurate and relevant answers to the questions posed. Unlike traditional retrieval systems relying solely on visual similarity, MemoryNet is trained to find the most relevant image based on the specific question, using labeled examples as ground truth. We evaluate PathRecall on a custom dataset of egocentric indoor videos with associated user questions and annotated relevant frames. Results demonstrate that MemoryNet significantly improves top-1 recall over baseline visual-only retrieval methods, showcasing the importance of question-aware memory in assistive navigation. The integration of captioning and BLIP-2 further boosts VQA accuracy by enhancing scene understanding. Our findings highlight the potential of lightweight, modular AI systems for situational awareness and visual memory augmentation in smart indoor spaces, offering practical support for independent navigation by VI users. PathRecall illustrates how foundation models can be adapted for accessibility, human-centered analytics, and embodied intelligence in the built environment. While it does not rely on traditional sensor signals, it complements smart building infrastructure by enabling semantic querying of past observations, bridging the gap between raw sensor data and memory support.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Indoor Navigation, Edge Intelligence, Visual Question Answering, Assistive Systems.

ACM Reference Format:
Pei Du and Nirupama Bulusu. 2018. PathRecall: Memory-Augmented Visual Question Answering for Assistive Indoor Navigation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Over four decades of scientific research have documented the impact of buildings and urban design on human health and performance. One important factor in improving the human experience is designing buildings with accessibility as a first order design principle, and in particular providing accessibility, not only in accessible outdoor paths, but also in providing safe and effective indoor navigation for VI individuals. Simultaneously, rapid advances in artificial intelligence, provide new opportunities to tackle challenging problems in assistive technologies. Despite advancements in physical infrastructure, many public indoor spaces—such as transportation hubs, office buildings, and shopping centers continue to present significant accessibility challenges including inconsistent signage, poor spatial cues, and a lack of intuitive layout design [1–3]. These barriers often hinder situational awareness and safe, autonomous mobility.

Unlike outdoor settings, indoor environments frequently lack standardized spatial cues such as tactile paving, audible signals, or consistent wayfinding signage. Architectural elements like irregular layouts, unmarked doors, transparent walls, or unexpected obstacles further complicate spatial orientation. These environments are typically designed with sighted users in mind, leading to information asymmetry and increased cognitive load for VI users. Consequently, even with mobility aids like white canes or guide dogs, navigating unfamiliar indoor spaces remains a persistent and unsolved challenge - particularly in the absence of GPS and other real-time, context-aware assistive systems.

Visual Question Answering (VQA) has emerged as a promising solution, allowing VI individuals to ask free-form questions about their surroundings and receive meaningful responses based on image understanding. However, existing VQA systems [4, 5] face

several limitations, including inaccurate image understanding, poor context comprehension, and difficulty providing detailed and useful responses to VI users. Existing VQA models are typically designed to operate on a single image input, limiting their ability to reason over temporally extended visual context. This poses challenges in indoor navigation scenarios, where users may refer to recently encountered landmarks or past visual observations. To address these limitations, our objective is to develop a memory-augmented VQA system that enables VI users to ask free-form questions such as "Did I pass the exit?" and retrieve the most relevant frame from the recent egocentric video history. The retrieved frame, paired with question-aware caption, is used to generate accurate answers grounded in visual content and question context.

To overcome these challenges, a more effective VQA system should not only interpret the current image but also access the user's visual history, understand the semantic intent behind the question, and provide contextually grounded answers. This calls for a shift from static image-based models to memory-augmented frameworks that retrieve the most relevant past observations and reason over them in light of the query. Inspired by how humans recall recent experiences to answer situational questions, we propose a system that integrates memory retrieval, semantic understanding, and question-aware reasoning.

Building on this motivation, we present PathRecall which is designed as a modular system that aligns visual memory with natural language queries. It incorporates three key components:

MemoryNet, a lightweight multimodal reranker, identifies semantically relevant past frames by jointly encoding both the user's question and visual features. This enables retrieval to focus on the user's intent.

A Caption Generator enhances the interpretability of the retrieved frame by extracting salient keywords from the question and generating targeted semantic captions, bridging the gap between vision and language representations.

Finally, a modified BLIP-2 model integrates the generated caption as a contextual prompt and uses bi-directional cross attention mechanism, allowing for more accurate and question-aware answer generation.

This pipeline moves beyond conventional VQA by grounding answers in question-aware memory retrieval, enabling more personalized and contextually aligned responses in complex indoor spaces.

To our knowledge, this is the first work to combine memory-based frame retrieval with question-aware VQA for assistive indoor navigation. The key contributions of our research are:

(1) A novel AI-driven VQA system that supports real-time, question-aware memory retrieval in indoor environments.

(2) A question-aware captioning mechanism built upon the BLIP-2 architecture, which enhances image comprehension by extracting key elements from user questions and generating context-relevant prompts. This improves visual grounding and guides the VQA model toward more accurate and context-aware responses.

(3) A MemoryNet, a lightweight visual-language reranker trained to identify question-relevant frames from video memory.

We evaluated a custom image dataset extracted from egocentric videos, demonstrating improved recall accuracy over FAISS-like baselines and we also evaluated our system on: (1) quantitative accuracy on the A-OKVQA [6] benchmark to assess general VQA performance, and (2) qualitative analysis on a custom egocentric indoor dataset to demonstrate the system's capability in realistic navigation scenarios.

The rest of this paper is organized as follows: Section 2 reviews related work in the field of AI-driven assistive technologies for indoor navigation and VQA models. Section 3 details the methodology behind our proposed system. Section 4 presents experimental results and performance evaluations. Finally, Section 5 discusses the implications of our findings and concludes the paper with future research directions.

2 Related Work

2.1 Visual Question Answering (VQA)

Visual Question Answering (VQA) is an interdisciplinary field combining computer vision and natural language processing to enable AI models to answer questions based on visual inputs. Innovations in deep learning have significantly improved VQA performance, with models leveraging convolutional neural networks (CNNs) for image feature extraction and transformers for contextual understanding [7, 8]. Notable VQA models [9], [10] have demonstrated significant accuracy improvement in answering general questions about images. However, existing VQA models primarily focus on generic datasets rather than addressing specific needs of visually impaired (VI) users in real-world indoor environments[11, 12]. Additionally, they often struggle with contextual understanding, leading to inaccurate or ambiguous responses[13, 14]. These challenges highlight the need for specialized VQA approaches tailored to assist VI users in daily navigation and understand their surroundings.

2.2 Assistive Navigation in Built Environments

AI-driven assistive technologies have gained attention for improving accessibility for VI individuals. Traditional assistive tools, such as screen readers and braille displays, provide limited interaction with the surrounding environment [15, 16]. Recent advancements include AI-based navigation applications, object recognition systems, and wearable smart devices [17, 18]. For example, Seeing AI [19] by Microsoft and Be My Eyes [20] use AI to describe the environment to users. However, these tools primarily rely on static image descriptions rather than dynamic, question-based interactions, limiting their effectiveness in complex indoor settings where users require more specific and interactive assistance [9, 11, 12].

2.3 Image Captioning and VLM Models

Image captioning plays a crucial role in improving image comprehension, particularly for assistive applications. Traditional approaches, such as Show and Tell [21] and Show, Attend, and Tell [22], leverage deep learning to generate textual descriptions from images. More recent models, such as BLIP-2, incorporate multimodal transformers to improve image-text alignment. The integration of captioning with VQA models has shown potential in enhancing answer accuracy by providing additional context. However, few studies have explored the direct application of these models to aid VI individuals in understanding their surroundings [23].

2.4 Memory in VQA and Visual Dialog

Several recent works have investigated memory-enhanced or retrieval-augmented architectures to support temporal reasoning in vision-language tasks. MemVid [24]uses a four-stage cognitive-inspired pipeline—memorization, reasoning, retrieval, and answer generation—to better handle long videos, achieving strong performance on long-video understanding benchmarks. Similarly, ReWind [25]adopts a learnable memory module with read-perceive-write cycles, coupled with adaptive frame selection for downstream VQA, showing significant gains on MovieChat-1K and Charades-STA datasets. Another recent method, ReRe [26], introduces retrieval-augmented natural language reasoning for VQA with explanations, leveraging external retrieved context to improve both answer accuracy and interpretability. In video question answering (video QA), temporal reasoning is addressed by processing video sequences directly. For example, models designed for datasets like EgoVQA or Ego4D [27] use transformers or hierarchical encoders to process long egocentric videos. While effective in encoding temporally continuous frames, such methods are typically trained end-to-end, are computationally expensive, and require dense temporal supervision (e.g., ground-truth frame spans or clips).

Despite significant advancements in VQA, assistive technologies, and memory-augmented architectures, a gap endures in integrating these technologies to support VI users [28] navigate in the indoor environment effectively. Current VQA systems are not optimized for accessibility needs [29], existing assistive tools lack interactive question-answering capabilities, and memory-augmented architecture do not retrieve user-encountered image frames in response to a specific, indoor navigation question. Our work addresses this gap by integrating a question-aware captioning mechanism and a modified BLIP-2 backbone with a lightweight, question-guided retrieval module, enabling context-aware, accessible VQA grounded in egocentric visual memory.

3 PathRecall System

3.1 Overview

The proposed system is a memory-augmented visual question answering (VQA) system designed to assist visually impaired (VI) users in understanding and navigating indoor environments. Unlike traditional VQA systems, which operate on a single image and often lack contextual grounding, PathRecall enables users to ask questions about recent path and retrieves the most relevant frame from recent video history for interpretation.

This multi-stage AI pipeline enhances accuracy and contextual understanding, allowing VI individuals to receive relevant and precise responses to queries about recently encountered scenes. By integrating memory-based retrieval and multimodal reasoning, the system generates detailed and accurate responses tailored to the user's needs, offering situational awareness over traditional image-only VQA systems.

3.2 System Architecture

Our AI-driven assistive system comprises three components (Fig.1).

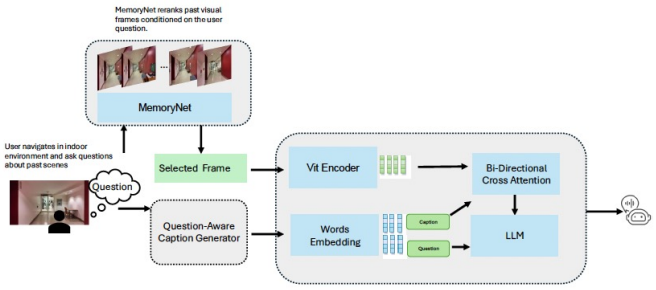


Figure 1: System Architecture: ChatGPT-based Question Generator, Caption Generator, and BLIP-2-based VQA Module.

3.2.1 MemoryNet: Question-Aware Frame Reranker. To retrieve the most relevant visual frame from recent egocentric videos, we introduce MemoryNet, a lightweight multimodal reranker that scores past frames based on their semantic alignment with the user's question. MemoryNet is designed to overcome the limitations of retrieval based on visual similarities by incorporating user's questions.

Given a user question and five candidate frames sampled from egocentric video, MemoryNet jointly embeds both modalities. The question is encoded via a MiniLM sentence transformer, while image features are extracted from BLIP-2's visual encoder. These embeddings are projected into a shared latent space, where a tanh-based fusion layer combines the two signals. A feedforward scoring module then computes a scalar relevance score for each image.

We construct a training dataset in which each set of five candidate frames is labeled with binary indicators (1 = relevant, 0 = irrelevant) based on manual annotations. Each question is paired with a query image and five previous candidate images, and annotators select which frame(s) contain information to answer the question. MemoryNet is trained using binary cross-entropy loss to maximize scores for relevant images and suppress scores of irrelevant ones.

By learning to align visual content with natural language intent, MemoryNet significantly improves retrieval accuracy compared to cosine similarity or FAISS-like baseline. It outputs a ranked list of candidate frames, from which the top-1 is selected and passed to the downstream VQA module.

3.2.2 Question-aware caption generator. Given generated questions, the Caption Generator processes them by extracting relevant keywords [30] from the question and generating textual descriptions of the image. This step is crucial for improving contextual understanding by identifying keywords that can be useful details to VI users. For instance, the model extracts wooden, window, trees from the question "Did I pass by a wooden table near the window with a view of trees?" and generates a caption of the image related to the question with essential details. The caption could be like "A wooden table is positioned near large windows with a clear view of green trees outside. The table is located in a quiet corner past the hallway." Without this step, the caption could be "A hallway with a wall-mounted poster and an emergency phone. The hallway opens up to a room with large windows, a few wooden chairs, and tables." without taking VI users' needs into consideration.

3.2.3 BLIP-2-based VQA Module. BLIP-2 serves as the core of the VQA system, integrating multimodal information from the image and generated captions to answer user queries with enhanced accuracy. The model operates as follows. The generated caption will be passed to the word embedding module and then will be passed to the Bi-Directional Cross Attention Module [31] in Fig.2 along with the raw image features. The module consists of two primary attention pathways:

1. Text-to-Image Attention: This pathway enables textual embeddings (e.g., caption tokens) to attend to visual features extracted from the image encoder (e.g., ViT [32]). $T \in \mathbb{R}^{L_T \times d_v}$ where L_T is the sequence length and d_v is the hidden dimension. The visual embedding can be represented as $V \in \mathbb{R}^{L_V \times d_v}$ where L_V is the number of image tokens and K_V, V_V represent the key and the value separately. Multi-head attention computes a weighted aggregation of visual features based on the relevance to the text as in Eq.1:

$$Q_f = Attention(Q_T, K_V, V_V) = softmax(\frac{Q_T K_V^T}{\sqrt{d_v}})V_V \quad (1)$$

The attention mechanism learns a weighted alignment between textual queries and relevant visual regions, increasing the model's ability to focus on contextually significant parts of the image.

2. Image-to-Text Attention: In this pathway, visual embeddings attend to textual representations, allowing the model to refine image features based on semantic cues from the text. This improves scene understanding by integrating linguistic priors with visual content, ensuring better alignment between the two modalities. Similarly, we can represent this step in a mathematical expression as in Eq.2 where image embeddings serve as queries Q_V and textual embeddings T act as key (K_T) and value (V_T).

$$I_r = Attention(Q_V, K_T, V_T) = softmax(\frac{Q_V K_T^T}{\sqrt{d_t}})V_T \quad (2)$$

Next we obtain a global reverse representation by averaging over the image tokens:

$$R = \frac{1}{L_V}\sum_{i=1}^{L_V} I_r(i) \quad (3)$$

And now R is expanded along the query token dimension to match the shape of Q_f. Then we concatenate the forward output and the expanded reverse output R_v and pass it to a 2-layer MLP to compute a gate $G \in \mathbb{R}^{(L_T,d)}$. The output is fused by an element wise weighted sum and the output will be passed to a FFN:

$$Output = G \odot I_r + (1 - G) \odot R_v \quad (4)$$

4 Experiment

4.1 Dataset

Since the system contains three modules, and it's not end-to-end training, we take two steps to prepare dataset for each MemoryNet and Blip2-based VQA module.

For VQA module, We evaluated our model using the A-OKVQA dataset, a benchmark designed for open-ended visual question answering (VQA) tasks that require diverse knowledge. This dataset is particularly valuable as it extends beyond simple object recognition, incorporating common sense reasoning, external knowledge,

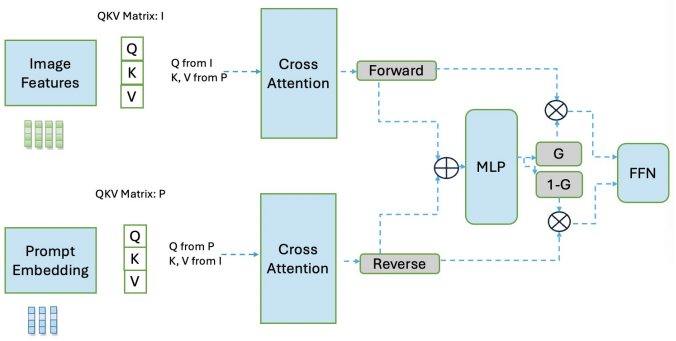


Figure 2: Bi-Directional Cross-Attention Module.

and contextual understanding, making it well-suited for testing the generalization capabilities of VQA models. The dataset consists of images sourced from the COCO dataset, multiple-choice questions with four possible answer options, and ten ground-truth answers per question to allow for a more flexible evaluation of answers.

To adapt the dataset for visually impaired (VI) users, we also construct a custom dataset comprising indoor environment images that are captured from real-world scenarios, such as library, airport and public buildings. The questions are carefully designed to align with real-life challenges faced by VI individuals, such as identifying object locations, understanding room layouts, or determining navigation obstacles that they recently passed. For example, instead of generic VQA questions like "What is the person holding?", our dataset includes more functionally relevant queries such as "Did I see a workspace with white tables and screens on them recently?" or "Was I recently in a room with several round tables and desktop computers?".

For training the MemoryNet, we constructed a custom dataset around 1500 pictures by video collection, frame extraction and candidate frame selection and labeling. We captured egocentric videos using a handheld DJI Osmo Pocket 3 camera at eye-level to simulate the viewpoint of a visually impaired user navigating indoor environments. The recordings covered multiple building types including academic hallways, lounges, and transition zones with varying lighting, object density, and layouts.

Each video was approximately 5–10 minutes in length, 1080p in resolution and included continuous navigation through indoor scenes. We ensured natural transitions and diverse perspectives by varying the walking direction, speed, and head movements. From each video, we extracted frames at 1 frame per second (1 FPS) to balance temporal coverage and annotation feasibility. This resulted in about 1500 of image frames across all videos, which is enough for training a reranker. For each frame, we manually constructed questions that reflect realistic recall needs for VI users. These questions were inspired by common navigational concerns, such as: "Did I pass a staircase near the elevator?" or "Did I see a sign labeled 'exit' after turning left?" Each question is designed to reference temporal memory and spatial reasoning, requiring understanding beyond a single frame. To train and evaluate the MemoryNet module, we created frame-ranking samples for each question to simulate visual memory retrieval. For each question: We selected five candidate

Table 1: Comparison of Different Prompting Strategies for VQA Using BERTScore, Precision, Recall, and F1-score. The table evaluates the impact of various reasoning and caption-based prompts on answer accuracy, highlighting the effectiveness of structured reasoning and contextual grounding in improving VQA performance.

TYPE	PRECISION	RECALL	F1 SCORE
Ground truth	0.45	0.5	0.47
Rationale Before Answer	0.71	0.6	0.65
Rationale After Answer	0.62	0.58	0.59
Rationale Few Shot	0.49	0.64	0.56
Think Step By Step	0.48	0.59	0.63
Caption-Based Prompting	**0.75**	**0.63**	**0.68**

frames from prior portions of the navigation video, representing plausible matches to the queried scene. We use a FAISS-based [33] visual retrieval approach to select top-5 candidate frames for each query image, based on nearest-neighbor search in the feature embedding space. This allows efficient pre-selection of visually similar frames for subsequent relevance annotation and MemoryNet training. Each candidate frame was manually annotated with a binary label, 1 if the frame contained visual evidence relevant to the question, otherwise we label it as 0.

4.2 Overall Results

We evalidated the MemoryNet's effectiveness with FAISS-like baseline and the modified BLIP2's effectiveness with BLIP2 as the baseline.

4.2.1 Top k Recall Evaluation. First, we evaluated MemoryNet over FAISS-like baseline by using top-k recall. The FAISS-like baseline retrieves frames based on visual similarity between the current scene and stored candidates, without considering the semantic relevance of the user's question. In contrast, MemoryNet is a trained neural network that learns to predict relevance by jointly modeling the question and visual content, enabling more accurate retrieval especially in complex scenarios. Top-k recall measures the proportion of queries for which at least one correct result is found among the top-k retrieved candidates. For example, top-1 recall indicates the query where the correct answer is ranked as the most relevant, while top-3 recall reflects whether the correct answer is among the top three predictions. This metric is particularly suitable for our task because the system is intended to assist users in memory recall in the buildings, where presenting one or a few highly relevant visual frames is often sufficient to trigger recognition or provide an answer. We adopt top-k recall as our primary metric because it directly reflects the usability of the system from the user's perspective. In real-world navigation or assistive scenarios, it is acceptable and often desirable to present multiple potentially relevant options rather than relying on a single prediction. Thus, measuring how often the system ranks at least one relevant candidate within the top-k positions gives us a more realistic picture of its practical effectiveness. Especially in our indoor memory recall setting, a missed retrieval in the top-1 but a hit in the top-3 may still be perfectly functional for the user. We show the top 1 and top 3 recall for MemoryNet and Faiss-like baseline separately with different number of training samples as shown in Figure 3.

In this figure, we present a comparative analysis of the top-1 and top-3 recall scores for two different retrieval methods—MemoryNet and a FAISS-like baseline—evaluated across varying numbers of training samples (from 300 to 1500). The purpose of this experiment is to understand how model performance scales with training data size and how each method copes with increasing retrieval complexity.

The FAISS-like baseline initially outperforms MemoryNet in both top-1 and top-3 recall when training data is limited (e.g., 300 samples). This is expected because FAISS is based only on visual vector similarity without learning from how the image relates to the question, which can be highly effective in smaller-scale settings. However, as the dataset scales up beyond 800 samples, MemoryNet begins to surpass FAISS in both top-1 and top-3 recall. This shift suggests that MemoryNet benefits significantly from additional supervision and data-driven learning, allowing it to learn more nuanced patterns of relevance beyond simple vector proximity. Specifically, the top-3 recall of MemoryNet continues to improve and exceeds 0.94 as the number of training samples increases to 1500, whereas the top-3 recall of the FAISS-like method steadily declines, highlighting the scalability limitations of unsupervised retrieval methods in larger and more diverse data scenarios. Similarly, top-1 recall for MemoryNet improves gradually, while FAISS-like performance consistently drops, reinforcing the idea that MemoryNet adapts better to complex data distributions when trained with more examples. This is because the baseline relies solely on visual similarity, which becomes less effective as the number of visually similar but semantically irrelevant frames increases. In contrast, MemoryNet learns to align visual content with the question, making it more robust to complex data distributions. The error bars show how much the results vary when we run the experiment multiple times. This helps us make sure that the trends we see are stable and not just random.

Overall, this analysis underscores the advantage of using a trainable model like MemoryNet for scalable and high-precision visual memory retrieval, especially when sufficient labeled data is available.

To evaluate our system's utility in smart building environments, we categorized test samples into semantic zones (e.g., work area, bookshelf zone, conference zone) and measured top-1 recall accuracy. As shown in Figure 4, retrieval accuracy varies with spatial context: while work areas and conference zones yield high top-1 recall due to distinctive object layouts (e.g., iMacs, large tables),

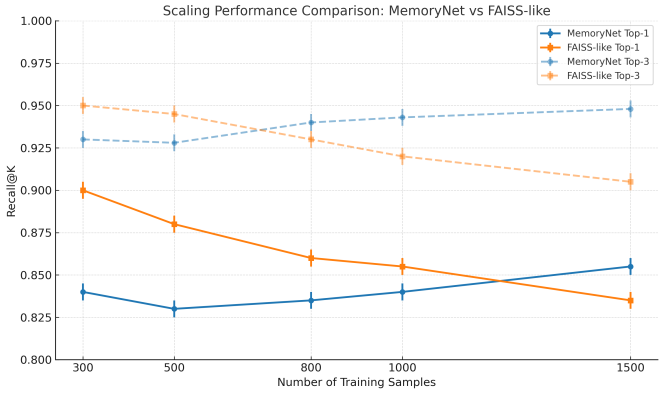


Figure 3: Top-1 and Top-3 recall of MemoryNet and FAISS-like baselines across varying numbers of training samples. MemoryNet achieves more stable and higher Top-3 recall as the candidate set grows, demonstrating its robustness and scalability. In contrast, FAISS-like retrieval degrades notably with larger sets.

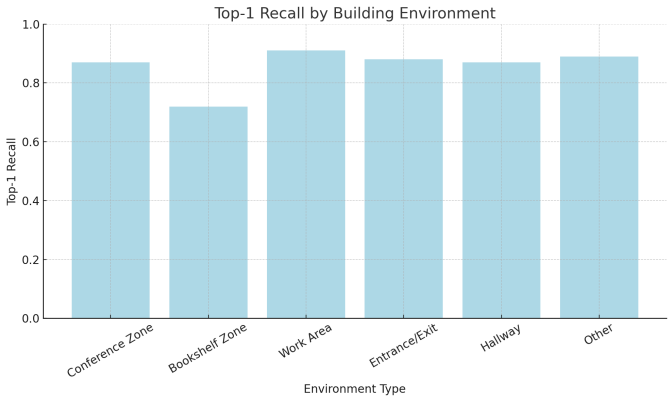


Figure 4: The performance (Top 1 Recall) of MemoryNet on different building environment types

bookshelf zone poses greater challenge due to visually repetitive scenes like bookshelves with same color.

This analysis demonstrates that our MemoryNet is effective in diverse types of building environment and provides insights into deploying such systems in complex indoor environments. For smart buildings equipped with egocentric vision or wearable cameras, PathRecall offers an interpretable memory interface that adapts to spatial semantic cues.

4.2.2 Prompt Selection. We first evaluate different types of prompt that may influence the accuracy of VQA. The metric that we use is BERTScore [34], a widely adopted semantic similarity measure that leverages contextual embeddings from pre-trained transformer models to compare generated answers with reference ground-truth answers. Unlike traditional n-gram-based metrics, such as ROUGE and BLEU, BERTScore captures semantic closeness and contextual alignment, making it particularly effective for evaluating open-ended VQA tasks. Its purpose is to determine the most effective prompting strategies for improving model reasoning and answer accuracy in VQA tasks. The prompting strategies we examine are:

Rationale Before Answer – In this prompt format, the model is encouraged to generate a rationale (explanation or reasoning) before providing the final answer. This approach aims to guide the model toward structured reasoning, improving the quality of the answers by ensuring logical consistency before reaching a conclusion.

Rationale After Answer – The answer is generated first, followed by an explanatory rationale. This approach assesses whether the explanation post-answer influences the correctness and justification of responses, providing insights into post-hoc reasoning in VQA.

Rationale Few-Shot – This method provides a few-shot learning setup in which multiple examples with rationales are included in the prompt before the model generates an answer. We apply this structured reasoning pattern to our model. The purpose is to improve generalization and accuracy of the answers.

Think Step by Step – Inspired by Chain-of-Thought (CoT) [35] prompting, this method explicitly instructs the model to break down the reasoning process into intermediate steps before arriving at an answer. By forcing a stepwise logical progression, we hypothesize this prompt format leads to more robust and interpretable answers.

Caption-Based Prompting – In this method, an image caption is provided as additional context alongside the question before generating an answer. This evaluates whether scene descriptions can help improve accuracy by grounding the model in a structured understanding of the images before answering the question.

We incorporate these templates written in Jinja2 [36] into BLIP2 and investigate the performance in Table 1.

4.2.3 Ablation Study. To evaluate the impact of incorporating question-based prompts on visual question answering (VQA) accuracy, we performed an ablation study that compared the performance under two conditions: (1) using a standard caption prompt and (2) using a question-based caption prompt. The propose of the analysis is to determine whether explicitly integrating the question context into the prompt formulation leads to measurable improvements in response accuracy and relevance. In our experiments, we applied standard prompts and question-aware prompts to the input before feeding it into the VQA model. The standard prompt consists of a generic instruction that does not incorporate specific questions, while the question-aware prompt dynamically adapts based on the content of the question, providing additional context to guide the answer. We also use BLIP-2 to generate the standard caption for the images. The metric is soft scoring, which is used in the AOKVQA dataset to account for multiple valid answers. The quantitative result is shown in Table. 2

Table 2: VQA Accuracy with Standard Caption vs. Question-Aware Caption

Prompt Type	VQA Accuracy(%)
Standard Caption	60.2
Question-Aware Caption	63.9

For qualitative results, we first show some samples taken in the indoor environment in Table. 3. Each sample comprises a standard caption and a question-aware caption, which allows us to analyze how contextual prompts influence generated descriptions. The standard caption provides a general description of the scene, offering a neutral and objective summary of the visual content. In contrast, the question-aware caption is dynamically adjusted on the basis of the given question, tailoring the description to emphasize the relevant aspects of the scene necessary for answering the query.

By comparing these two caption types, we aim to illustrate how question-guided captioning can enhance the informativeness of visual descriptions, ultimately benefiting Visual Question Answering (VQA) tasks. As seen in the examples, question-aware captions tend to highlight specific objects, spatial relationships, or relevant attributes that align with the question. This contrasts with standard captions, which may include unnecessary details or fail to emphasize the key elements needed for an accurate response.

Furthermore, this qualitative analysis helps demonstrate the advantages of integrating adaptive captioning into VQA systems. In scenarios where fine-grained distinctions are crucial, such as recognizing objects among clutter or understanding spatial arrangements, question-aware captions provide a contextually rich and focused scene representation, thereby improving the model's ability to generate precise and relevant answers, especially in challenging cases where generic descriptions may lead to incomplete reasoning.

In general, qualitative results support the hypothesis that question-guided captions significantly enhance VQA accuracy by refining the way visual information is presented and aligned with user queries.

4.2.4 Qualitative Evaluation of VQA Responses. In addition to quantitative results, we present qualitative comparisons to highlight the effectiveness of our model. We adopt zero-shot learning for testing due to the limited size of our customized dataset. We select pictures which are annotated as label 1 in Table 3 from our custom dataset, focusing on challenging scenarios that require precise indoor navigation, object recognition, and understanding of the surrounding environment. These pictures will be selected from MemoryNet during training process and are passed to VQA models for answers.

We present several representative pictures in Table 4 to illustrate the effectiveness of our model in comparison to other models. Beyond our baseline, We also show a comparison with two mainstream VLM to answer the same questions. Specifically, we analyze two key categories of questions: indoor navigation questions and environmental understanding questions. From the results, we observe that our model consistently provides more accurate and contextually relevant answers compared to the baseline. This improvement can be attributed to two critical components in our approach: question-aware captions and the bi-directional cross-attention module, both of which enhance the model's ability to understand the image in the context of the given question.

For indoor navigation questions, which require spatial awareness and object-location relationships, our model demonstrates superior performance by offering precise directional guidance and a clearer understanding of spatial arrangements. This is particularly beneficial for visually impaired users who rely on detailed descriptions for orientation. Similarly, for environmental understanding questions, where the recognition of multiple objects and their relationships is crucial, our model effectively uses the question-driven context to generate responses that go beyond simple object recognition, providing deeper insights into the scene.

However, when it comes to simple object recognition questions, such as identifying the presence of a single object, our model performs on par with Blip2. This is expected since our approach is built upon the BLIP-2 framework, which already exhibits strong object recognition capabilities. Although our enhancements significantly improve contextual reasoning and spatial understanding, they do not introduce major changes to the model's ability to recognize individual objects in straightforward scenarios.

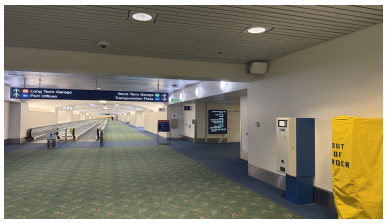
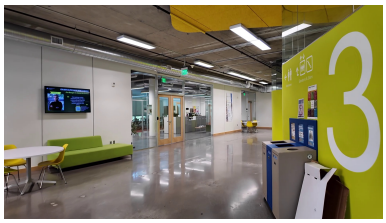
Overall, the results indicate our approach effectively enhances question-specific image comprehension, making it valuable for complex queries that require reasoning beyond direct object detection. These findings further support the practical usability of our model in real-world indoor navigation and assistive scenarios, where users require not only object identification but also question-aware and interactive responses to navigate their environment effectively.

4.2.5 Implementation Details. We trained our model by using a multi-GPU distributed setup on our department high performance computing equipped with two NVIDIA RTX A5000 GPUs (each with 24GB VRAM). We adopt the PyTorch DistributedDataParallel (DDP) training paradigm to maximize GPU utilization and ensure synchronous gradient updates. To accelerate training and reduce memory usage, we use mixed precision training (FP16). This significantly speeds up training without a noticeable decrease in accuracy. We conducted an ablation study by varying the size of the language model in our architecture. Specifically, we replaced the base Vicuna-7B with smaller (3B) and larger (13B) variants, keeping the vision encoder and Q-Former fixed. The results show that increasing LLM size leads to moderate performance gains (e.g., +1.7% accuracy from 7B to 13B), but also significantly increases inference latency and memory footprint. This highlights a trade-off between model capability and deployment cost. At the current stage, our model has not been deployed in a real-world environment. Although training and evaluation were conducted using multi-GPU distributed setups with mixed-precision (fp16) to improve efficiency, actual deployment introduces additional constraints such as latency, memory usage, and model serving frameworks. We plan to explore practical deployment options in future work, such as offloading the vision encoder and Q-Former to a pre-processing stage and using optimized inference frameworks (e.g., vLLM) to serve the language model. It could enable high-throughput, low-latency deployment while maintaining the benefits of multi-modal reasoning.

5 Integration in Smart Building Environments

PathRecall is well-suited for deployment in modern smart buildings as a lightweight, user-centric memory system that enhances spatial awareness and interactive assistance, especially for visually impaired (VI) users. Unlike traditional passive sensing systems (e.g., motion detectors, badge readers), PathRecall allows users to record egocentric videos during indoor navigation and later ask contextual questions about past observations. This enables flexible, retrospective interaction with the environment without requiring dense sensor infrastructure.

Table 3: Table 3: Comparison of Visual Representations with Standard and Question-Aware Captions. The first row shows three MemoryNet-retrieved images, and the following rows compare standard captions with our question-aware captions for each image. Each image is retrieved by MemoryNet based on the associated user question, and the subsequent rows compare the standard caption with a question-aware caption generated using our method.

			
Question	A moment ago, when I stood near the intersection by the out-of-order machine, what direction did the sign point for the Transportation Plaza?	Earlier, when I stood beside the wooden door and looked up at the green exit sign, which direction would I have gone?	When I passed by the recycling area with the large number "3" sign, what kind of seating and signage did I encounter on the left side of the hallway?
Standard Caption	A wide hallway with green carpet and a moving walkway in the center. Overhead lights illuminate directional signs suspended from the ceiling. On the right wall, there is a white kiosk and a machine covered in a yellow "OUT OF ORDER" cloth. A black digital display and multiple hallway entrances are also visible.	A carpeted hallway with a closed wooden office door, a glass window panel, a wall-mounted campus map, and tables with chairs near large windows in the background.	An open indoor hallway with green couches, tables and chairs, a wall-mounted screen, and glass office doors under industrial-style ceilings.
Question Aware Caption	The scene shows an intersection near a yellow-covered out-of-order machine and a white kiosk. Above the walkway, there are directional signs with labels such as “Long Term Garage,” “Port Offices,” and “Transportation Plaza.” The signs include arrows pointing toward different destinations.	A hallway showing an office with a wooden door beneath an overhead EXIT sign, a wall-mounted map, and possible safety equipment such as an emergency phone and fire extinguisher near the corner.	A hallway corner showing a large yellow wall with a bold number “3” sign and a set of recycling and trash bins beneath it, near a lounge area and glass-door offices.

In a smart building setting, users may wear a lightweight camera-equipped device—such as smart glasses, a chest-mounted camera (e.g., GoPro), or even a mobile phone—while traversing indoor spaces such as offices, airports, libraries, or medical centers. The system periodically captures image frames (e.g., one every few seconds) and extracts their visual embeddings using a pretrained image encoder. These frames are stored in a local or cloud-based memory module.

When users later issue natural language questions like "Did I pass the fire extinguisher before entering the meeting room?" or "Where did I see the elevator sign?", the system encodes the query, retrieves relevant frames from the memory module (MemoryNet), and generates question-aware image captions on the fly. The retrieved visual memory and captions are passed to a modified BLIP-2 model to generate the final answer.

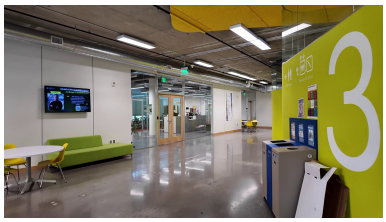
PathRecall differs fundamentally from traditional building sensing systems that rely on continuous data collection and real-time analysis. Traditional passive sensors or camera-based monitoring systems continuously process data streams, consuming energy 24/7, even when no user interaction is needed.Our system is query-driven and event-triggered, meaning that most computation only occurs when a user actively asks a question. The video stream is recorded at a low frame rate (e.g., one frame every few seconds) and preprocessed into lightweight feature embeddings. No captions or heavy visual-language inference is performed until a retrieval event is triggered.

PathRecall can function as an intelligent personal assistant in smart buildings, especially in applications such as (1) Indoor Navigation: Helping VI users recall previous waypoints or signs. (2) Security Auditing: Recalling if sensitive areas were accessed. (3) Cognitive Assistance: Supporting memory recall for elderly users. (4) Workplace Support: Retrieving visual information in large facilities.

6 Conclusion

In this work, we present PathRecall, a memory-augmented VQA system that empowers visually impaired (VI) users to query and

Table 4: Table 4: Qualitative Comparison of VQA Model Outputs Given Identical Retrieved Visual Contexts. Each column shows a question and its corresponding MemoryNet-retrieved image. Rows compare responses from three baseline models (BLIP-2, LLaVA, MiniGPT-4) and our model, all conditioned on the same image-question pair.

			
Question	A moment ago, when I stood near the intersection by the out-of-order machine, what direction did the sign point for the Transportation Plaza?	Earlier, when I stood beside the wooden door and looked up at the green exit sign, which direction would I have gone?	When I passed by the recycling area with the large number "3" sign, what kind of seating and signage did I encounter on the left side of the hallway?
Blip2 [5]	Forward	North	Chairs
Llava-1.5-7B [37]	Follow the signage and identification boards that indicate the location of the Transportation Plaza.	You would probably go ahead or toward the open space.	As you passed by the area with the number "3" sign and recycling bins, on your left you likely noticed a small seating area that includes a round white table surrounded by several yellow chairs.
MiniGPT-4 [38]	The sign at the intersection of the out-of-order machine points to the Transportation Plaza.	Based on the image you provided, the exit sign is pointing to the right, so you would have gone that direction.	When I passed by the recycling area with the large number "3" sign, I encountered a number of chairs and a sign that said "Recycling Area".
Our Model	Right	Left	Lounge Chair

recall past visual experiences in indoor environments. The system integrates three key components: a lightweight MemoryNet that jointly embeds visual scenes and natural-language queries for question-aware memory retrieval, a captioning module that produces query-specific semantic descriptions, and a modified BLIP-2 model that leverages these captions as prompts to enhance final answer generation.

Our evaluations show that our approach significantly improves the accuracy and usability of VQA systems. Our proposed question-aware captioning mechanism effectively extracts key elements from user queries, generating more relevant prompts that improve image comprehension and VQA performance. Through both quantitative and qualitative evaluations, we show that MemoryNet outperforms FAISS-like baselines in top-1 and top-3 recall as the dataset scales, highlighting the benefit of learning-based retrieval over static embedding similarity.

Beyond accessibility, PathRecall also exemplifies how foundation models can be adapted for human-centered analytics and embodied intelligence in the built environment. While it does not rely on traditional sensor inputs, our system complements smart building infrastructure by enabling semantic querying of previously seen spaces, bridging the gap between passive sensing and human memory support. This offers new possibilities for post-hoc inspection, usability evaluation, and assistive navigation without requiring invasive sensor instrumentation.

Despite promising results, one limitation of our study lies in the size of the custom egocentric video dataset, which remains small compared to large-scale VQA corpora. This constraint limits statistical generalization of answer accuracy in indoor scenes. In future work, our goal is to expand the dataset to include a wider range of images and questions, allowing extensive evaluation and insights regarding the effectiveness of the model. We could also explore retrieval accuracy across different question types (spatial, attribute, object recognition) to better understand the strengths and limitations of the system.

References

[1] Inclusive City Maker. Accessibility for customers with vision disabilities in public venues. https://www.inclusivecitymaker.com/accessibility-customers-vision-disabilities-public-venues/, 2023. Accessed: 2025-08-01.

[2] The Architectural Review. Building for the blind. https://www.architectural-review.com/buildings/building-for-the-blind, 2023. Accessed: 2025-08-01.

[3] Dormakaba. 8 examples of accessible spaces for the visually impaired. https://blog.dormakaba.com/8-examples-of-accessible-spaces-for-the-visually-impaired/, 2023. Accessed: 2025-08-01.

[4] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people, 2022.

[5] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions, 2023.

[6] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022.
[7] 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. *CVPR*, 2018.
[8] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *EMNLP*, 2019.
[9] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *CVPR*, 2017.
[10] Bairui Wu, Quan Wang, Rui Yu, Jing Guo, Dongyan Zhao, and Jun Yan. Openvqa: An open platform for visual question answering. *ACM MM*, 2019.
[11] Danna Gurari, Qitao Li, Austin Stangl, Anna Guo, Chenxi Lin, Kristen Grauman, and Jiebo Luo. Captioning images taken by people who are blind. *TPAMI*, 2020.
[12] Ravi Kiran Annam, Danna Gurari, Jiebo Luo, and Kristen Grauman. Vizwiz grand challenge: Answering visual questions from blind people. *CVPR*, 2021.
[13] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and C Lawrence Zitnick. Don't just assume; look and answer: Overcoming priors for visual question answering. *CVPR*, 2018.
[14] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Taking a hint: Leveraging explanations to make vision and language models more grounded. *ICCV*, 2020.
[15] Yevgen Borodin, Jeffrey P Bigham, Gregory Dausch, and I V Ramakrishnan. Web-based touch screen access for visually impaired users. *ASSETS*, 2010.
[16] Heng Cheng and Timothy Luczak. Smart glasses for the visually impaired people: A survey. *Sensors*, 2018.
[17] James M Coughlan and Huiying Shen. Smart technologies for blind navigation: A review. *J. Assistive Tech*, 2018.
[18] Dimitrios Dakopoulos and Nikolaos G Bourbakis. Wearable obstacle avoidance electronic travel aids for blind: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 2010.
[19] Microsoft. Seeing ai - talking camera app for the blind, 2017. https://www.microsoft.com/en-us/ai/seeing-ai.
[20] Be My Eyes. Be my eyes - lend your eyes to the blind, 2015. https://www.bemyeyes.com.
[21] Danna Gurari, Qitao Li, Austin Stangl, Anna Guo, Chenxi Lin, Kristen Grauman, and Jiebo Luo. Image captioning as an assistive technology: Lessons learned. *Journal of Artificial Intelligence Research*, 2020.
[22] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Multi-modal image captioning for the visually impaired. In *Proceedings of the NAACL Student Research Workshop*, 2021.
[23] Lu Yu, Malvina Nikandrou, Jiali Jin, and Verena Rieser. Quality-agnostic image captioning to safely assist people with vision impairment. *arXiv preprint arXiv:2304.14623*, 2023.
[24] Zheng Xu, Yixuan Zhao, Junnan Li, Wayne Xin Zhang, and Tat-Seng Chua. Memvid: Memory-augmented video question answering via cognition-inspired retrieval-augmented generation. *arXiv preprint arXiv:2503.09149*, 2025.
[25] Siyuan Zhang, Chong Wang, Yuting Zhan, and et al. Rewind: Learning to retrieve what to remember for long video question answering. *arXiv preprint arXiv:2411.15556*, 2024.
[26] Jianfeng Zhou, Zekun Li, Jie Chen, Zhiyuan Xu, and Xinchao Liu. Rere: Retrieval-augmented reasoning for visual question answering with natural language explanations. *arXiv preprint arXiv:2408.17006*, 2024.
[27] Kristen Grauman, Andrew Westbury, and et al. Ego4d: Around the world in 3,000 hours of egocentric video. *CVPR*, 2022.
[28] Farid Rahman, Lei Wang, and Michael Brown. Multimodal conversational ai for visually impaired users: Challenges and future directions. *Springer AI Journal*, 2023.
[29] Hui Wang and Jinsoo Lee. Real-time vqa and chatbot integration for enhanced assistive ai. *ACM Transactions on Accessible Computing*, 2023.
[30] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.
[31] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
[32] 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, 2021.
[33] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025.
[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
[36] Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*, 2023.
[37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024.
[38] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009