CMU Advanced NLP Assignment 4 Final Report: CAMEO-CAption enhanced Multi-task Optimization Framework for VQA

Manav Nitin Kapadnis
Carnegie Mellon University
mkapadni@andrew.cmu.edu

Lawanya Baghel
Carnegie Mellon University
lbaghel@andrew.cmu.edu

Aarya Makwana
Carnegie Mellon University
amakwana@andrew.cmu.edu

1 Introduction

Visual Question Answering (VQA) combines computer vision and natural language processing, enabling systems to interpret visual data and answer questions in natural language. VQA has impactful applications, including assistive technologies, autonomous systems, and multimodal search engines, where systems must handle noisy, unstructured visual data (Antol et al., 2015; Gurari et al., 2018).

Current VQA systems face challenges in real-world scenarios, especially in assistive contexts where inputs include low-quality images taken under poor conditions. Traditional pipelines, treating visual perception and linguistic reasoning separately, struggle to bridge the semantic gap between modalities. Datasets like SQuAD (Rajpurkar et al., 2016) and MSCOCO (Chen et al., 2015) do not address the multimodal complexities of real-world tasks.

The VizWiz dataset (Gurari et al., 2018), designed for assistive contexts, provides a benchmark with images and questions from visually impaired users. Addressing this requires integrating image understanding, skill identification, and question answering.

We propose the *CAMEO-enhanced Multi-task Optimization Framework*, combining image captioning, vision skill identification, and question answering. Key contributions include:

- **Context-aware Image Captioning:** Generating relevant visual descriptions.
- **Skill-specific Visual Reasoning:** Leveraging task-specific visual skills.
- **Robust Question Answering:** Providing precise, contextually grounded answers.

The framework demonstrates its effectiveness on the VizWiz dataset, handling noisy inputs, ambiguous questions, and unanswerable queries, setting a standard for robust VQA systems. All the code to replicate our experiments and results are present in this [1] and this repository [2].

2 Related Work

Research in Visual Question Answering (VQA) builds upon advancements in image captioning, vision skill detection, and multimodal reasoning, each contributing to robust and scalable systems for real-world challenges.

2.1 Image Captioning

Image captioning translates visual inputs into rich textual descriptions. Early models like "Show and Tell" (Vinyals et al., 2015) used convolutional and recurrent neural networks but struggled with real-world complexities. Attention mechanisms, such as in "Show, Attend, and Tell" (Xu et al., 2016), and transformer-based models, like Meshed-Memory Transformer (Cornia et al., 2020), improved contextual understanding and long-range dependencies. Recent advancements like AoA networks (Huang et al., 2019) and robust pretraining on datasets like MSCOCO (Chen et al., 2015) have achieved state-of-the-art results.

For assistive applications, captioning models must handle noisy and diverse inputs. Gurari et al. (Gurari et al., 2020) highlighted the challenges posed by low-resolution, unconventional images, underscoring the need for semantically accurate systems.

2.2 Vision Skill Detection

Vision skill detection focuses on identifying reasoning tasks like object detection, text recognition, and counting. Zeng et al. (Zeng et al., 2020) formalized this as a multi-label classification prob-

[1] Report 3: https://anonymous.4open.science/r/ANLP_Project_R3-442D/README.md

[2] https://anonymous.4open.science/r/ANLP_Project_R4_CAMEO-5761/README.md

lem, while SkillCLIP (Naik et al., 2024) introduced skill-aware embeddings for enhanced interpretability and precision. Methods like LoRA (Hu et al., 2021) further enable efficient fine-tuning for vision skill tasks with reduced computational overhead.

Datasets like VizWiz (Gurari et al., 2018) often include tasks requiring text recognition or object counting, emphasizing the importance of integrating skill detection into multimodal frameworks.

2.3 Visual Question Answering

VQA has advanced through benchmark datasets such as VQA (Antol et al., 2015) and GQA (Hudson and Manning, 2019), which introduced diverse question types. Attention mechanisms, like the bottom-up and top-down framework (Anderson et al., 2018a), have significantly improved accuracy and interpretability. BLIP-2 (Li et al., 2023) demonstrated the potential of unified architectures by integrating frozen image encoders with language models.

The VizWiz dataset (Gurari et al., 2018) has driven real-world VQA research by highlighting challenges such as unanswerable questions and noisy inputs, necessitating robust contextual reasoning and visual understanding.

2.4 Multimodal Integration

The integration of visual and linguistic modalities is central to VQA. Transformer-based models like LXMERT (Tan and Bansal, 2019) and ViLT (Kim et al., 2021) have set benchmarks by leveraging large-scale pretraining. Frameworks like OBELICS (Laurençon et al., 2023) have emphasized the importance of data quality, while retrieval-augmented generation methods (Reimers and Gurevych, 2019) enable complex query handling through external knowledge.

These advancements underscore the importance of multimodal approaches in creating robust and adaptable VQA systems.

3 Methodology

3.1 Overview of CAMEO

We present the architecture of CAMEO in Figure 1, which details a comprehensive pipeline designed for accurate Visual Question Answering (VQA) through a caption-enhanced multi-task learning framework. The process begins with a vision transformer that encodes the input image into patch embeddings $(p_1, p_2, ...)$, capturing essential visual features needed for subsequent analysis. A Visual Mapper then projects these embeddings into the LLM's high dimensional text embedding space. Whereas, a Text Embedder processes the instruction prompt tokens $(t_1, t_2, ...)$.

The high-dimensional image representations, along with structured input prompts containing the instruction template and VQA questions, are fed into the LLM. The prompt follows a specific format that instructs the LLM to first generate a detailed image caption and predict four fundamental skills: object detection (OBJ), text recognition (TXT), color recognition (CLR), and counting (CNT). These predictions serve as intermediate representations that guide the final VQA response generation.

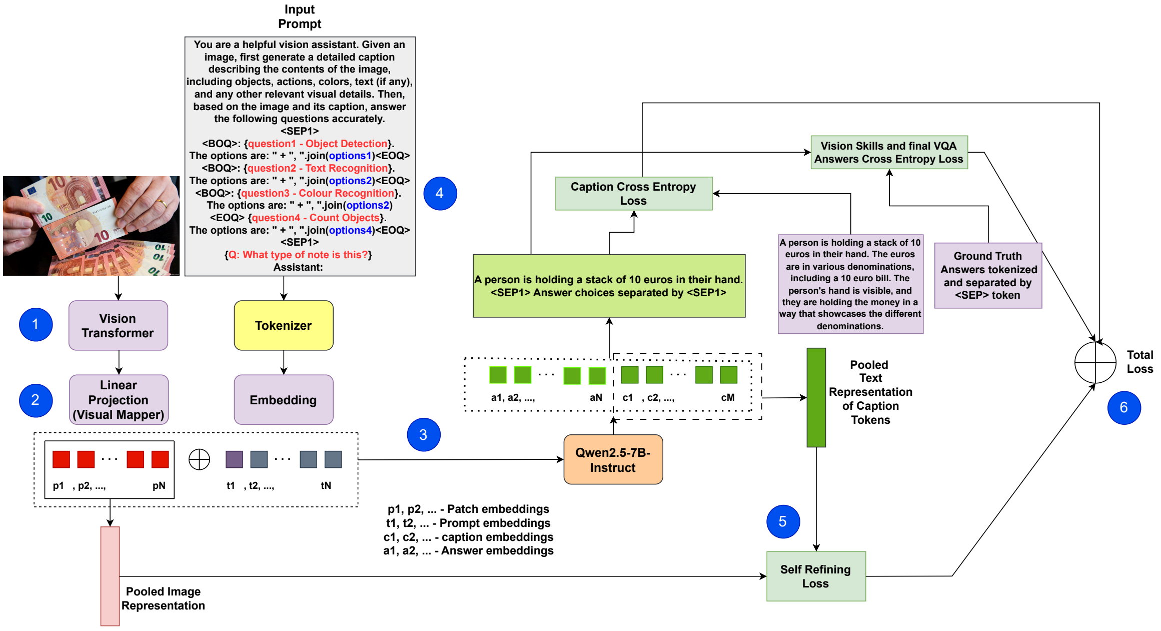
The LLM then generates outputs in an autoregressive manner: image caption embeddings $(c_1, c_2, ...)$, skill prediction embeddings for OBJ, TXT, COLOR, and COUNT tasks, and final answer embeddings $(a_1, a_2, ...)$ for the VQA task.

To ensure alignment between visual and textual representations and reduce hallucinations, a self-refining loss mechanism is implemented. The final optimization is performed through a weighted combination of caption generation loss ($\mathcal{L}_{caption}$), skills prediction loss ($\mathcal{L}_{skills}$), VQA response loss ($\mathcal{L}_{vqa}$), and the self-refining loss ($\mathcal{L}_{refine}$). This composite loss function enables CAMEO to refine its outputs iteratively, ensuring the generated responses are both accurate and contextually grounded in the visual input.

The framework leverages the complementary nature of the different tasks, where the detailed caption generation and skill predictions provide a comprehensive understanding of the image, which in turn aids in generating more accurate VQA responses. The multi-task learning setup helps in learning shared representations that are robust and generalizable across different visual reasoning tasks.

3.2 CAMEO Framework Architecture

The CAMEO framework integrates three essential components to enable sophisticated visual reasoning: a visual feature extractor, an embedding alignment module, and a large-scale language processor. At its core, the framework processes an input visual signal $I_v \in \mathbb{R}^{C \times H \times W}$, characterized by its channel depth C and spatial dimensions $H \times W$. To facilitate efficient processing, the visual input undergoes patch-wise decomposition, yielding k


Input
Prompt
1
Vision Transformer
2
Linear Projection (Visual Mapper)
Tokenizer
Embedding
p1 , p2, ..., pN
t1 , t2, ..., tN
3
Qwen2.5-7B-Instruct
a1, a2, ..., aN
c1 , c2, ..., cM
4
Caption Cross Entropy Loss
Vision Skills and final VQA Answers Cross Entropy Loss
Ground Truth Answers tokenized and separated by <SEP> token
Pooled Text Representation of Caption Tokens
5
Self Refining Loss
6
Total Loss
Pooled Image Representation
p1, p2, ... - Patch embeddings
t1, t2, ... - Prompt embeddings
c1, c2, ... - caption embeddings
a1, a2, ... - Answer embeddings

You are a helpful vision assistant. Given an image, first generate a detailed caption describing the contents of the image, including objects, actions, colors, text (if any), and any other relevant visual details. Then, based on the image and its caption, answer the following questions accurately.
<SEP1>
<BOQ>: {question1 - Object Detection}.
The options are: " + ", ".join(options1)<EOQ>
<BOQ>: {question2 - Text Recognition}.
The options are: " + ", ".join(options2)<EOQ>
<BOQ>: {question3 - Colour Recognition}.
The options are: " + ", ".join(options2)
<EOQ> {question4 - Count Objects}.
The options are: " + ", ".join(options4)<EOQ>
<SEP1>
{Q: What type of note is this?}
Assistant:

A person is holding a stack of 10 euros in their hand.
<SEP1> Answer choices separated by <SEP1>

A person is holding a stack of 10 euros in their hand. The euros are in various denominations, including a 10 euro bill. The person's hand is visible, and they are holding the money in a way that showcases the different denominations.

Figure 1: Overview of the proposed architecture. The pipeline begins with a vision transformer to encode the input image (step 1), followed by a linear projection to map the visual representation into a high-dimensional space (step 2). The tokenized prompt and visual embeddings are processed jointly by an LLM to generate answers (step 3). Cross-entropy loss is used for image captioning to inform VQA skills (step 4), while a self-refining loss further aligns the generated captions with the VQA tasks (step 5). The total loss (step 6) combines the caption cross-entropy loss, VQA answer cross-entropy loss, and self-refining loss, optimizing the model for accurate skill prediction and question answering.

discrete segments $I_v = [I_{v_1}, I_{v_2}, \cdots, I_{v_k}]$, where individual patches $I_{v_i} \in \mathbb{R}^{C \times P \times P}$ maintain consistent dimensions $P \times P$. The visual analysis pipeline begins with a transformer-based feature extraction network V_{enc}, which captures hierarchical visual representations $\tilde{e}v_i \in \mathbb{R}^{d_v}$ through the following transformation:

$$\tilde{e}v_1, \tilde{e}v_2, \cdots, \tilde{e}v_k = V_{enc}(I_{v_1}, I_{v_2}, \cdots, I_{v_k}) \quad (1)$$

These patch-level features are subsequently consolidated into a unified image representation $\tilde{e}v$ via a pooling mechanism:

$$\tilde{e}v = Vpooler(\tilde{e}v_1, \tilde{e}v_2, \cdots, \tilde{e}v_k) \quad (2)$$

To bridge the semantic gap between visual and linguistic domains, we implement an adaptive mapping function V_{map} that transforms visual features into language-compatible embeddings: $e_{v_i} = V_{map}(\tilde{e}v_i)$. The system processes input queries Q through carefully crafted prompts T, converting them into token sequences $\mathcal{T}tokens = [t_1, t_2, \cdots, t_{|\mathcal{T}tokens|}]$. These tokens are then embedded using:

$$et_1, e_{t_2}, \cdots, e_{t_{|\mathcal{T}tokens|}} = Embedding(t_1, t_2, \cdots, t|\mathcal{T}_{tokens}|) \quad (3)$$

The framework implements a multi-task learning approach where visual understanding is decomposed into three interconnected tasks: comprehensive scene description generation, fundamental visual skill assessment (encompassing text recognition, object detection, color identification, and numerical counting), and question answering. The process combines visual (e_{v_i}) and textual (e_{t_j}) embeddings into a unified representation $e_{\mathcal{I}} = [e_v; e_t]$, enabling the decoder-based language model to generate contextually appropriate responses through autoregressive processing.

3.3 Self-refining Strategy

We construct an aggregated representation of the generated text using the attention weights from the last layer of TD. For each generated token, we apply Gumbel-Softmax to the logit distribution $l_i \in \mathbb{R}^d$ to obtain $\hat{l}_i$. The aggregated representation $\hat{e}_i^p = \sum_{j=1}^{d} e_j \hat{l}_{ij}$ is constructed by taking a weighted sum of the embedding matrix $E = [e_1, e_2, \cdots, e_d]$ with $\hat{l}_i$ as weights:

$$\hat{l}_{ij} = \frac{e^{(\log(l_{ij}) + g_{ij})/T}}{\sum_{j=1}^{d} e^{(\log(l_{ij}) + g_{ij})/T}} \quad (4)$$

We enforce a self-refining loss between the aggregated text representation h_t and image representation e_v:

$$\mathcal{L}_{refine} = \frac{1}{b} \sum_{i=1}^{b} e^{-h_t^T e_v} \tag{5}$$

The total loss combines multiple objectives:

$$\begin{aligned} \mathcal{L}_{total} = \lambda_{caption} \cdot \mathcal{L}_{caption} + \lambda_{skills} \cdot \mathcal{L}_{skills} \\ + \lambda_{vqa} \cdot \mathcal{L}_{vqa} + \lambda_{refine} \cdot \mathcal{L}_{refine} \end{aligned} \tag{6}$$

where $\lambda_{caption}$, λ_{skills}, λ_{vqa}, and λ_{refine} are empirically tuned weights to balance the different tasks optimally. The components ensure generation of captions and VQA responses that are both conditionally accurate based on the input image and contextually coherent with the predicted skills.

4 Results And Discussion

We now discuss the details corresponding to the experiments and ablation studies carried out and enumerate the observations.

4.1 Implementation Details

We discuss the technical details and hyper-parameter settings for all the experiments. For the visual encoder V_{enc}, we employed the base version of Swin-Transformer-V2[3] and a feed-forward neural network for V_{map}. We leverage Qwen2.5-7B-Instruct[4] as our primary LLM. Further, the hidden dimension of d_v of V_{enc} and d_t of TD are 768 and 1024 respectively. We freeze the weights of V_{enc}, however keep V_{map} trainable. We employ LoRA with a rank and α-scaling factor of 16 each to fine-tune the underlying LLM TD. We train CAMEO for 5 epochs on VizWiz training dataset with mixed precision on an effective batch size (BS) of 32 using one NVIDIA A100 80GB GPU using a learning rate of 1×10^{-4} with linear rate scheduler through AdamW optimizer. For inference, we leverage beam search decoding with beam size configured to 3.

4.2 Datasets and Evaluation Metrics:

We evaluated our performance on three different tasks of the VizWiz dataset.

4.2.1 Image Captioning Task

The image captioning task evaluates the model's ability to generate semantically rich and contextually relevant textual descriptions for visual inputs. CAMEO incorporates a vision transformer-based encoder V_{enc} to extract hierarchical features, which are then used to generate captions. The loss function for this task combines cross-entropy and self-refining components to ensure coherent and accurate captions.

The image representation $\tilde{e}_v$ is computed as:

$$\tilde{e}_v = V_{\text{pooler}}(V_{\text{enc}}(I_v)),$$

where I_v represents the input image, and V_{pooler} pools patch embeddings to form a unified visual representation.

To evaluate caption quality, the following metrics are used:

- **BLEU**: Measures n-gram precision against reference captions. It is computed as:

  $$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right),$$

  where p_n is the precision of n-grams, N is the maximum n-gram length, and BP is the brevity penalty to penalize short outputs, defined as:

  $$\text{BP} = \begin{cases} 1, & \text{if } c > r, \\ e^{1-\frac{r}{c}}, & \text{if } c \leq r, \end{cases}$$

  where c is the length of the candidate caption and r is the length of the reference caption.

- **ROUGE**: Measures overlap between generated and reference text. For ROUGE-L (longest common subsequence), it is computed as:

  $$\text{ROUGE-L} = \frac{(1+\beta^2)\cdot \text{R} \cdot \text{P}}{\text{R} + \beta^2 \cdot \text{P}},$$

  where $\text{R} = \frac{LCS}{\text{Length of Reference}}$, $\text{P} = \frac{LCS}{\text{Length of Candidate}}$, and LCS is the length of the longest common subsequence.

- **CIDEr**: Captures consensus in generated captions based on Term Frequency-Inverse Document Frequency (TF-IDF) weighting. It is defined as:

  $$\text{CIDEr} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{TF-IDF}(\text{Candidate}, \text{Reference}_i)}{\text{Length of Reference}_i},$$

  where N is the number of references.

[3] https://huggingface.co/microsoft/swinv2-base-patch4-window12-192-22k

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

- **SPICE**: Measures semantic content by parsing scene graphs. It is calculated as:

$$\text{SPICE} = \frac{|S_c \cap S_r|}{|S_c \cup S_r|},$$

where S_c and S_r are the sets of tuples representing semantic content (e.g., objects, attributes, relations) for the candidate and reference captions, respectively.

4.2.2 Skill Prediction Task

The skill prediction task identifies visual reasoning skills such as object detection, text recognition, color identification, and numerical counting. For each skill k, the model predicts a probability P_k based on the visual embeddings.

The skill prediction accuracy for each task is computed as:

$$A_{\text{skill}} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100.$$

Accuracy is calculated separately for text recognition skill task, object detection skill task, color identification skill task, and counting skill task.

4.2.3 Visual Question Answering Task

The visual question answering (VQA) task evaluates the model's ability to provide accurate answers based on the combined understanding of image captions and visual skills. The model generates responses by combining visual embeddings $\tilde{e}_v$ and textual embeddings e_t.

The VQA accuracy is defined by the exact match accuracy:

$$A_{\text{VQA}} = \frac{\text{Correct Answers}}{\text{Total Questions}} \times 100.$$

4.3 Performance of CAMEO on Image Captioning

Table 1 presents the performance of CAMEO on the VizWiz-Captions (Gurari et al., 2018) dev set compared to several state-of-the-art baselines. The results clearly demonstrate that CAMEO achieves superior performance across multiple metrics, particularly in the fine-tuned **Skill → Caption → VQA** setup. This configuration achieves the highest scores in **BLEU_1**, **ROUGE**, **CIDEr**, and **SPICE**, highlighting its ability to generate accurate and context-aware captions.

Notably, when the model is asked to first predict the skills and then generate captions (Skill → Caption → VQA), it achieves significantly better results than the reverse setup (Caption → Skill → VQA). This observation indicates that grounding captions on the skill predictions enhances their quality and coherence. By first understanding the skills, the model develops a stronger contextual basis, which subsequently leads to improved captioning performance, as evidenced by the superior results in Table 1.

Among the baselines, PaLI-X (Chen et al., 2023) and Omni-SMoLA (Wu et al., 2024) are the top performers, with Omni-SMoLA slightly outperforming PaLI-X on **BLEU_4** and **CIDEr** scores. However, CAMEO surpasses these models in **BLEU_1**, **ROUGE**, and overall accuracy metrics. The Skill → Caption → VQA configuration achieves a remarkable CIDEr score of **129.9** and SPICE of **22.9**, showcasing its effectiveness in capturing long-range dependencies and reducing hallucinations.

Traditional models like Up-Down (Anderson et al., 2018b), SGAE (Yang et al., 2019), and AoANet (Huang et al., 2019) show significant improvements when trained from scratch or fine-tuned but fall short compared to recent vision-language models such as PreSTU (Kil et al., 2023), GIT (Wang et al., 2022), and PaLI-X (Chen et al., 2023). This demonstrates the importance of incorporating large-scale pre-training, task-specific adaptations, and fine-grained modeling strategies.

Model	**Description**	$BLEU_1$	$BLEU_4$	**ROUGE**	**CIDEr**	**SPICE**
Up-Down	pretrained	52.8	11.3	35.8	18.9	5.8
	from scratch	64.1	19.8	43.2	49.7	12.2
	fine-tuned	62.1	18.6	42.0	48.2	11.6
SGAE	pretrained	55.8	13.5	38.1	20.2	5.9
	from scratch	67.3	22.8	46.6	52.4	12.8
	fine-tuned	68.5	23.9	47.3	61.2	13.5
AoANet	pretrained	54.9	13.2	37.6	19.4	6.2
	from scratch	66.4	23.2	47.1	60.5	14.0
	fine-tuned	66.6	22.8	46.6	57.6	13.7
PreSTU	OCR-aware pre-training	74.9	34.3	53.4	105.6	21.1
GIT	large-scale pre-training	77.1	33.4	53.2	114.4	22.3
PaLI-X	multi-lingual multi-task pre-training	77.8	38.5	55.7	125.7	23.5
Omni-SMoLA	soft mixture of experts model	78.05	**38.7**	55.5	125.8	23.6
CAMEO (finetuned)	Caption	73.9	29.7	56.4	122.5	19.8
	Caption ->Skill ->VQA	74.4	30.8	57.8	124.2	20.5
	Skill ->Caption ->VQA	**78.2**	37.4	**59.0**	**129.9**	**22.9**

Table 1: Performance comparison of CAMEO variations against various state-of-the-art baselines on the on the VizWiz-Captions dev set. CAMEO demonstrates substantial improvements in Image Captioning, achieving the highest scores across majority metrics. This highlights CAMEO's superior ability to capture long-range dependencies and reduce hallucinations, outperforming both traditional models and recent VLMs such as Omni-SMoLA and GIT.

4.4 Performance of CAMEO on Skill Recognition Task

Table 2 compares the performance of CAMEO with various baseline methods on the Skill Recognition Task. The results demonstrate the superior performance of CAMEO across most sub-tasks, including text, object, and color accuracy. Notably, the **Skill → Caption → VQA** configuration achieves the highest scores, with a text accuracy of 77.48%, object accuracy of 80.13%, and color accuracy of 84.42%. This highlights the effectiveness of grounding downstream tasks like captioning and VQA on initial skill predictions, which significantly improves overall performance.

In contrast, the Caption → Skill → VQA setup achieves lower accuracy across all categories compared to the Skill → Caption → VQA configuration. This emphasizes that skill recognition provides essential contextual information that enhances the subsequent tasks. The standalone CAMEO_SKILL configuration performs reasonably well, but integrating multi-stage pipelines (e.g., CAMEO_CAPTION_SKILL_VQA) delivers significant improvements, especially in object and color accuracy.

Among the baselines, Omni-SMoLA achieves the highest scores for text, object, and color categories, outperforming earlier models like CLIP-Large Fusion, ViLT, and SkillCLIP. However, CAMEO's multi-stage setup surpasses these baselines, demonstrating its ability to capture finer-grained details for skill recognition.

A notable observation is that CAMEO struggles significantly with the counting skill, achieving a count accuracy of only 8.3%. This stark underperformance contrasts with its high object detection accuracy (80.13%), suggesting that the model can recognize objects but struggles with visual reasoning required for counting. This limitation likely stems from the model's inability to effectively process relationships between multiple detected objects, which is critical for accurate counting.

Overall, the results highlight the strengths of CAMEO in recognizing text, object, and color-based skills, while revealing its limitations in handling counting tasks, likely due to deficiencies in visual reasoning capabilities.

Methods	**Text Accuracy ↑**	**Object Accuracy ↑**	**Color Accuracy ↑**	**Count Accuracy ↑**
CLIP-Large fusion	32.8%	39.78%	53.29%	45.05%
ViLT	36.05%	45.04%	61.44%	41.44%
GIT	32.40%	42.66%	58.45%	26.13%
SkillCLIP	39.16%	46.58%	57.23%	50.45%
SkillCLIP Multitask	37.64%	44.62%	56.01%	49.55%
Pre-STU	58.64%	51.24%	68.36%	62.74%
PALI-X	62.34%	55.64%	70.26%	63.76%
Omni-SMoLA	63.87%	58.96%	72.78%	67.89%
CAMEO_SKILL	64.46%	58.92%	74.00%	3.30%
CAMEO_SKILL_VQA	65.77%	60.50%	74.30%	4.50%
CAMEO_CAPTION_SKILL_VQA	70.52%	70.07%	81.27%	4.60%
CAMEO_SKILL_CAPTION_VQA	**77.48%**	**80.13%**	**84.42%**	8.30%

Table 2: Performance of CAMEO on Skill Recognition Task

4.5 Performance of CAMEO on Visual Question Answering

Table 3 presents the performance of CAMEO compared to various baseline methods on the Visual Question Answering (VQA) task. The results highlight the effectiveness of CAMEO, particularly in the multi-stage **Skill** → **Caption** → **VQA** configuration, which achieves the highest accuracy of **78.83%** on the dev set. This performance slightly surpasses the state-of-the-art baseline, Omni-SMoLA, which achieves 78.57%, demonstrating the capability of CAMEO to leverage a hierarchical approach for visual reasoning.

Interestingly, the **Caption** → **Skill** → **VQA** configuration yields a lower accuracy of **74.32%**, suggesting that generating skills before captions creates a more robust foundation for answering visual questions. This observation reinforces the hypothesis that grounding the reasoning process in skill predictions enhances the model's ability to produce contextually accurate answers.

Among the standalone configurations, **CAMEO_VQA** achieves an accuracy of **65.46%**, which improves significantly when combined with skill recognition in the **CAMEO_SKILL_VQA** configuration (**69.54%**). This shows the importance of incorporating intermediate skill-based reasoning for addressing complex VQA tasks. The stepwise improvement in accuracy across configurations highlights the cumulative benefits of multi-stage processing.

Baseline models like *PALI-X* (77.28%) and *Pre-STU* (*75.67%*) perform well but fall short of the highest-performing CAMEO configuration. Notably, models such as *CLIP-Large Fusion*, *ViLT*, and *SkillCLIP* show moderate performance, with accuracies in the range of 61–62%, indicating their limited capacity to capture intricate reasoning patterns required for advanced VQA tasks.

While CAMEO demonstrates strong overall performance, there are still areas for improvement. Similar to the counting skill issue observed in the skill recognition task, some VQA questions requiring advanced visual reasoning or multi-object relationship understanding remain challenging for the model. This indicates that further refinement is necessary to address tasks involving deeper visual comprehension.

Methods	**Dev Accuracy** ↑
CLIP-Large fusion	61.37%
ViLT	61.82%
GIT	51.24%
SkillCLIP	62.17%
SkillCLIP Multitask	62.16%
Pre-STU	75.67%
PALI-X	77.28%
Omni-SMoLA	78.57%
CAMEO_VQA	65.46%
CAMEO_SKILL_VQA	69.54%
CAMEO_CAPTION_SKILL_VQA	74.32%
CAMEO_SKILL_CAPTION_VQA	**78.83%**

Table 3: Performance of CAMEO on Visual Question Answering Task

5 Ablation Study

5.1 Effect of contextual representation design strategy

We evaluate the impact of different aggregation strategies for obtaining the contextual representation on the Visual Question Answering (VQA) task. As shown in Table 4, *attention-based aggregation* significantly outperforms other methods across all metrics, achieving a BLEU_1 score of **78.2**, BLEU_4 of **37.4**, ROUGE of **59.0**, CIDEr of **129.9**, and SPICE of **22.9**. The attention mechanism dynamically weights the most relevant features for each question, enabling the model to effectively focus on key visual elements and semantic cues.

In contrast, simpler methods such as *average pooling* (mean of all token embeddings) and *max pooling* (representation with the maximum L2-norm) yield sub-optimal results, with BLEU_1 scores of **75.6** and **71.1**, respectively, and notable drops across other metrics. Similarly, *top-k average pooling*, which aggregates the top $k = 5$ token embeddings based on attention weights, performs slightly better than max pooling but still lags behind attention-based aggregation.

Design Strategy	$BLEU_1$	$BLEU_4$	$ROUGE$	$CIDEr$	$SPICE$
Attention-based aggregation	**78.2**	**37.4**	**59.0**	**129.9**	**22.9**
Average pooling	75.6	35.8	56.8	125.2	21.4
Top k average pooling	73.4	34.1	54.5	121.6	20.6
Max pooling	71.1	32.7	52.3	118.4	19.8

Table 4: Performance comparison of different design strategies for contextual representation. Attention weights-based aggregation displays superior performance.

5.2 Model Efficacy on training settings of different components

The evaluation of different training configurations (Table 5) highlights the balance between accuracy, trainable parameters, and efficiency. The **CAMEO** model achieves the highest Dev Accuracy of **78.83%**, utilizing all components—Visual Mapper, Visual Encoder, and LoRA. Although it requires the most parameters (**152M**) and the longest training time (**3.62 hours per epoch on an A100 GPU**), its integrated approach delivers superior performance, particularly for complex VQA tasks.

Configurations like **Shallow** and **Delta** are more parameter-efficient and faster to train but exhibit lower performance, with **70.2%** and **74.8%** accuracy, respectively. The **Deep** model, which trains both the Visual Mapper and Encoder, achieves a solid **77.3%** accuracy with moderate parameters (**97M**). These results demonstrate that while fully integrated models like CAMEO excel in accuracy, lighter configurations offer advantages in resource-constrained environments.

Models	Trainable Components			Scale and Efficiency (on one A100 GPU)		VQA Efficacy
	Visual Mapper	Visual Encoder	LoRA	Trainable Parameter	Time	Dev Accuracy
Shallow	✓			7.8 M	1.75 h/epo	70.2
Delta	✓		✓	19.5 M	1.83 h/epo	74.8
Deep	✓	✓		97 M	2.75 h/epo	77.3
CAMEO	✓	✓	✓	152 M	3.62 h/epo	**78.8**

Table 5: Evaluation of Model Efficiency and VQA Efficacy on VizWiz VQA dataset. The CAMEO model achieves the highest Dev Accuracy while maintaining trainable parameter efficiency

5.3 Effect of relative importance of four task losses

The performance of our model is evaluated across varying combinations of task loss weights, including self-refining loss (λ_{refine}), skills loss (λ_{skills}), caption-generation loss ($\lambda_{caption}$), and VQA loss (λ_{vqa}). Table 6 shows the impact of different weight combinations on key evaluation metrics like BLEU, ROUGE, CIDEr, and VQA accuracy.

From the results, it is evident that the model's performance peaks when the weight of the VQA loss (λ_{vqa}) is set to 0.4, and the other losses are balanced at lower values. This configuration achieves the highest scores across BLEU$_1$ (78.2), BLEU$_4$ (37.4), ROUGE (59.0), and CIDEr (129.9), as well as the best VQA accuracy (78.83%). This indicates that a higher emphasis on the VQA task leads to stronger overall performance, particularly in complex multi-modal understanding.

Some interesting observations from the table are:

- Increasing the weight of the caption-generation loss ($\lambda_{caption}$) improves text-related metrics, like BLEU and ROUGE, suggesting that the caption generation component benefits from higher emphasis, but at the cost of slightly reduced VQA accuracy.

- The configuration with the balanced weights (λ_{skills}, $\lambda_{caption}$, λ_{vqa}, and λ_{refine} set to 0.25) results in a good compromise between performance metrics, achieving decent scores across all tasks, with a slightly reduced VQA accuracy (78.3%).

- The lowest performance is observed when the self-refining loss (λ_{refine}) is weighted more heavily. For example, the setting where λ_{refine} is set to 0.4 and the other weights are smaller results in lower BLEU and CIDEr scores, with only a minor increase in object and color accuracy.

These findings suggest that a well-balanced loss combination, particularly focusing on VQA loss, leads to the best overall performance, while giving excessive weight to any single component results in suboptimal performance in other tasks.

5.4 Statistical Significance of Results

To ensure the robustness and reliability of our results, we re-ran the model variations of CAMEO (CAMEO) across five different random seeds. For each seed, we trained and evaluated the models independently and averaged the performance metrics. The performance values reported in the earlier sections (refer to Tables **??** and **??**) are the mean scores across these five seeds.

In this section, we present the standard deviations in performance metrics observed across the runs. These standard deviations provide insights into the stability and consistency of the model variations. A lower standard deviation indicates that the model produces reliable and repeatable results across multiple runs, reinforcing confidence in its performance.

The results are shown in Table 7, which highlights the variability in metrics such as BLEU scores, ROUGE, CIDEr, VQA accuracy, and skill-specific accuracies (text, object, and color recognition). By analyzing the standard deviations, we observed that the model variations of CAMEO consistently performed well, with minimal variance across the different seeds. This consistency underscores the statistical significance of the results.

By reporting the standard deviations, we provide an additional layer of statistical rigor, ensuring that the observed performance improvements of CAMEO are not merely a result of random initialization or model training fluctuations. This strengthens the argument for the effectiveness and reliability of CAMEO in addressing real-world Visual Question Answering (VQA) challenges.

6 Summary and Conclusion

In this report, we introduced CAMEO, a novel Caption-enhanced Multi-task Optimization (CAMEO) framework aimed at addressing the challenges of Visual Question Answering (VQA) in real-world assistive contexts. By integrating image captioning, skill-specific visual reasoning, and robust question answering into a unified multi-task learning architecture, CAMEO effectively bridges the semantic gap between visual and linguistic modalities. The framework was evaluated on the VizWiz dataset, known for its challenging real-world scenarios, including noisy inputs and ambiguous or unanswerable questions.

The experimental results demonstrated that CAMEO achieves state-of-the-art performance across multiple tasks:

- In the **image captioning task**, CAMEO outperformed both traditional and modern baselines by generating semantically rich and contextually accurate captions, as reflected in superior scores in BLEU, ROUGE, CIDEr, and SPICE metrics.

- In the **skill recognition task**, the framework excelled in text recognition, object detection, and color identification but faced challenges

λ_{skills}	$\lambda_{caption}$	λ_{vqa}	λ_{refine}	$BLEU_1$	$BLEU_4$	$ROUGE$	$CIDEr$	**Dev VQA accuracy**	**Text Accuracy**	**Object Accuracy**	**Color Accuracy**
0.4	0.2	0.2	0.2	70.5	32.1	52.7	115.3	74.0	72.3	75.5	78.6
0.2	0.4	0.2	0.2	73.8	34.7	56.4	121.8	76.5	74.9	77.2	81.2
0.2	**0.2**	**0.4**	**0.2**	**78.2**	**37.4**	**59.0**	**129.9**	**78.8**	**77.5**	**80.1**	**84.4**
0.2	0.2	0.2	0.4	76.9	36.0	57.2	127.0	78.0	76.6	79.3	83.2
0.25	0.25	0.25	0.25	77.5	36.5	58.0	128.5	78.3	77.0	79.8	83.8

Table 6: Impact of combining self-refining loss (weight λ_{refine}), Skills loss (weight λ_{sk}), caption-generation loss (weight $\lambda_{caption}$), and VQA loss (weight λ_{vqa}). Fusing all four loss components gives optimal performance.

CAMEO Variation	$BLEU_1$	$BLEU_4$	$ROUGE$	$CIDEr$	**Dev VQA Accuracy**	**Text Accuracy**	**Object Accuracy**	**Color Accuracy**
CAMEO_VQA	1.22	0.24	0.86	4.8	1.32	0.54	1.32	0.45
CAMEO_SKILL_VQA	1.15	0.26	0.92	4.9	1.28	0.52	1.29	0.47
CAMEO_CAPTION_SKILL_VQA	**1.18**	**0.23**	**0.89**	**4.7**	**1.30**	**0.53**	**1.31**	**0.46**
CAMEO_SKILL_CAPTION_VQA	1.20	0.25	0.90	4.85	1.29	0.55	1.30	0.44

Table 7: Standard deviations of performance metrics for different CAMEO variations across BLEU, ROUGE, CIDEr, VQA accuracy, and skill-specific accuracies.

in handling counting tasks due to difficulties in multi-object relationship reasoning.

- For the **VQA task**, CAMEO achieved the highest accuracy among competing models, particularly in the *Skill* → *Caption* → *VQA* configuration, underscoring the importance of grounding downstream tasks on skill predictions.

6.1 Future Work Directions

While CAMEO has set new benchmarks for VQA systems in assistive contexts, there are several areas for future improvement:

- **Enhanced Counting Capabilities**: Addressing the observed limitations in counting tasks by incorporating advanced relational reasoning techniques or graph-based approaches to model object relationships more effectively.
- **Generalization Across Datasets**: Extending evaluations to other challenging datasets beyond VizWiz to validate the generalizability of the framework across diverse domains.
- **Computational Efficiency**: Reducing the computational overhead of the multi-task learning pipeline by exploring parameter-efficient fine-tuning methods such as LoRA or adapter layers.
- **Integration of External Knowledge**: Incorporating retrieval-augmented generation (RAG) techniques to enable CAMEO to handle complex queries requiring domain-specific knowledge.

In conclusion, CAMEO has demonstrated its potential as a robust and inclusive VQA system capable of addressing real-world challenges. By leveraging multi-task learning and innovative architectural components, it lays a strong foundation for future advancements in assistive AI technologies.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018b. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. Pali-x: On scaling up a multilingual vision and language model.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, and Radu Soricut. 2023. Prestu: Pre-training for scene-text understanding.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, et al. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Atharva Naik, Yash Parag Butala, Navaneethan Vaikunthan, and Raghav Kapoor. 2024. Skillclip: Skill aware modality fusion visual question answering (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23592–23593.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language.

Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. 2024. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694.

Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. 2020. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–31.