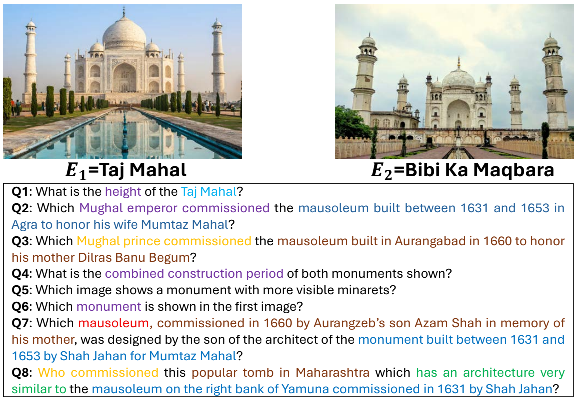
Learning to Ask: Multi-Decoder Fine-Tuning for Multi-Hop Visual Question Generation with External Knowledge

Arpan Phukan[1], **Manish Gupta**[2], **Asif Ekbal**[1]
[1]IIT Patna, India [2]Microsoft, India
arpan_2121cs3@iitp.ac.in, gmanish@microsoft.com, asif@iitp.ac.in

Abstract

Multi-hop visual question generation (VQG) seeks to create coherent, fluent, and contextually rich questions by integrating knowledge from a structured knowledge graph (KG) with information inferred from image pairs, where at least one reasoning step involves a visual relationship between the images. Traditional supervised QG methods which rely on token-level alignment with fixed gold labels struggle to capture diverse valid question formulations. We propose **M3RQG** (**M**ultimodal **M**ulti-hop **M**ulti-decoder **R**etrieval-augmented **Q**uestion **G**eneration), a model-agnostic framework that integrates multimodal inputs (images, KG facts) with a multi-decoder architecture to optimize for multiple labels per sample to design multi-hop questions. M3RQG addresses these challenges: (1) generating meaningful visually-grounded questions given a pair of images, (2) generating rich questions that require multi-hop reasoning across images and KG facts, and (3) integrating diverse question labels during fine-tuning. We extend the WebQA dataset with multi-hop questions generated by GPT-4V and Gemini, resulting in two complementary silver labels per sample. Our approach integrates retrieval-augmented generation (RAG) for accessing external knowledge, a PPO objective with ROUGE-based rewards to prioritize structural correctness, and a named entity overlap loss to improve factual accuracy. Experiments across BART, Phi-3.5, and LLaVA backbones demonstrate significant improvements in fluency, reasoning depth, and relevance. We release our code and dataset to facilitate future research[1].



Q1: What is the height of the Taj Mahal?
Q2: Which Mughal emperor commissioned the mausoleum built between 1631 and 1653 in Agra to honor his wife Mumtaz Mahal?
Q3: Which Mughal prince commissioned the mausoleum built in Aurangabad in 1660 to honor his mother Dilras Banu Begum?
Q4: What is the combined construction period of both monuments shown?
Q5: Which image shows a monument with more visible minarets?
Q6: Which monument is shown in the first image?
Q7: Which mausoleum, commissioned in 1660 by Aurangzeb's son Azam Shah in memory of his mother, was designed by the son of the architect of the monument built between 1631 and 1653 by Shah Jahan for Mumtaz Mahal?
Q8: Who commissioned this popular tomb in Maharashtra which has an architecture very similar to the mausoleum on the right bank of Yamuna commissioned in 1631 by Shah Jahan?

Figure 1: Questions generated using a pair of images and Wikipedia facts. Only Q8 makes use of KG facts about both entities and visual information linking the two images. See Fig. 2 for color keys.

1 Introduction

Multimodal Visual Question Generation (MVQG) has emerged as a pivotal task in multimodal natural language processing, aiming to generate contextually relevant and engaging questions grounded in visual inputs such as images, thereby enhancing machine understanding and interaction capabilities across diverse domains like education, healthcare, and document analysis. It enhances learning experiences in education by prompting students to engage actively with visual content (Krishna et al., 2015) and creating interactive learning materials, quizzes (Huang et al., 2014), clarification questions (Kumar and Black, 2020), and study aids for students (Wang et al., 2018). In accessibility applications, MVQG can assist visually impaired individuals by providing textual descriptions of visual scenes (Stangl et al., 2021), developing automatic tutoring systems (Kumar and Black, 2020; Gala et al., 2021), improving the performance of Question Answering models (Sun et al., 2023), and enabling chatbots to accept multimodal inputs (Zhang et al., 2024). MVQG enhances clinical decision-making and medical education by generating diagnostic, personalized, and instructional questions from multimodal patient data like scans, records, and videos. Lastly, generating thought-provoking questions can stimulate critical thinking and play a vital role in formulating research questions, surveys, and interview protocols.

[1] https://github.com/thePhukan/M3RQG

Findings of the Association for Computational Linguistics: EACL 2026, pages 3899–3918
March 24-29, 2026 ©2026 Association for Computational Linguistics

Despite its importance, MVQG is rather underexplored. While existing MVQG approaches have made significant strides (Chang et al., 2022; Talmor et al., 2021; Yang et al., 2023; Li et al., 2022; Hannan et al., 2020; Hwang et al., 2024; Li et al., 2022), they often fall short in capturing complex relationships within visual scenes. Straightforward single-image-based questions (like Q1, Q2, Q3, Q6 in Fig. 1), while useful for basic comprehension, lack the depth required to engage users in critical thinking or real-world problem-solving. VQG using evidence from multiple images (Chang et al., 2022) generates questions (like Q4 and Q5 in Fig. 1) requiring visual reasoning over both images but no factual knowledge. Multi-hop QG (Yang et al., 2018b) generates questions (like Q7 in Fig. 1) requiring reasoning over multiple documents, but has been studied for text only.

In this paper, we focus on the multi-hop MVQG problem of generating questions given a pair of images and related facts from a knowledge graph (KG) like Wikipedia. Unlike previous MVQG methods, our approach leverages the complementary and contrasting semantics across both images, identifying meaningful visual relationships (e.g., object co-occurrence, stylistic similarities, temporal or spatial links) and augmenting with external factual knowledge from the KG to generate questions (like Q8 in Fig. 1) that require joint reasoning over both images and the KG. As shown in Fig. 2, given 2 images about entities E_1 and E_2, a KG helps obtain linked attributes $\{A_i^1\}_{i=1}^n$ and $\{A_i^2\}_{i=1}^m$. Multi-hop visual questions must contain value of $\geqslant 1$ attribute from $\{A_i^1\}_{i=1}^n$, value of $\geqslant 1$ attribute from $\{A_i^2\}_{i=1}^m$, and $\geqslant 1$ visual link between the 2 images. It typically asks for the value of an attribute (not mentioned in the question) of E_1 or E_2 or the entities themselves. Although we define (and study) such questions from the perspective of just 2 images corresponding to 2 entities, the definition can be easily extended to an arbitrary combination of hops across several images and KG links.

Fig. 1 shows the questions generated using various QG methods for a sample image pair about entities E_1="Taj Mahal" and E_2="Bibi Ka Maqbara." As discussed earlier, questions Q1-Q7 are either about just one image, or are purely text based (and ignore images). Question Q8 satisfies all the required constraints: (1) It uses knowledge from KG about E_2: "popular tomb in Maharashtra." (2) It uses knowledge from KG about E_1: "mausoleum on the right bank of Yamuna commissioned in 1631 by Shah Jahan." (3) It uses the visual link "has an architecture very similar to" based on visual content of the 2 images. (4) It does not directly refer to the entities in the images, or the images themselves. (5) The answer "Azam Shah" cannot be obtained without using the multi-hop reasoning since there are many tombs in Maharashtra. This level of specificity and contextual richness enables more informative and engaging question generation, making it especially valuable for downstream tasks like educational content creation.

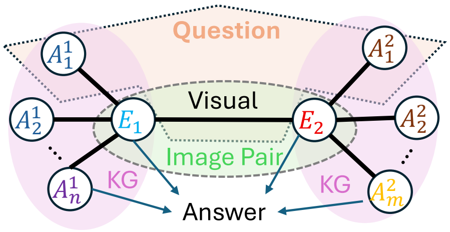


Figure 2: Definition for multi-hop visual questions obtained using an image pair and KG facts.

A multi-hop MVQG system must be able to generate meaningful visually-grounded compositional questions by synthesizing visual cues (that link the two images) with textual context (image captions) and external knowledge (related to the two entities from a KG). Furthermore, one can create several questions given the same input by varying the attributes used in the question, visual links extracted from the images or the attribute queried for. Fortunately, WebQA (Chang et al., 2022) provides questions that involve information across two images. To enhance diversity, we combine (silver) questions generated from GPT-4V and Gemini with the original (gold) question from WebQA (Chang et al., 2022) for each sample to produce questions that balance dataset alignment and creative explorations that provoke curiosity.

We adopt a multimodal encoder combined a Transformer-based multi-decoder architecture, M3RQG, for multi-hop MVQG. We extract relevant facts corresponding to input entities from Wikipedia using Retrieval Augmented Generation (RAG) (Lewis et al., 2020b; Gao et al., 2023). By integrating information from both visual and textual modalities along with RAG documents, the encoder can utilize rich contextual cues. A multi-decoder setup with shared weights enables M3RQG to optimize with respect to multiple labels together. The utilization of the Proximal Policy Optimization (PPO) (Schulman et al., 2017a) with ROUGE-based rewards and Named Entity (NE) overlap (between the RAG documents and the generated question) enhances the model's ability to capture facts, long-range dependencies, and intricate relationships within visual scenes, thus elevating the quality and diversity of generated questions and minimizing model hallucinations.

Overall, we make the following contributions. (1) We propose a novel task of multi-hop MVQG given a pair of images and access to a KG. (2) We extend multi-image dependent instances of WebQA (Chang et al., 2022) with high-quality generated questions from GPT-4V and Gemini. (3) We propose a novel model-independent approach, M3RQG, that exploits multiple labeled questions to generate multimodal multi-hop information seeking questions. (4) We propose a comprehensive training paradigm that integrates PPO with ROUGE-based rewards and Named Entity (NE) overlap scores. (5) We fine-tune models like BART (Lewis et al., 2020a), Phi (Abdin et al., 2024), and LLaVA (Liu et al., 2023) using the proposed approach and demonstrate significant gains over the non-PPO trained models.

2 Related Work

Text-based Question Generation. Early QG methods include template/rule-based approaches (Mazidi and Nielsen, 2014) and sequence-to-sequence models using BiLSTMs/transformers (Du et al., 2017; Wang et al., 2020). Pre-trained models improved performance via transfer learning (Dong et al., 2019; Xiao et al., 2021). Graph-based models leveraged semantic structures (Chen et al., 2023; Pan et al., 2020a), while generative models (VAEs, GANs) enhanced diversity (Wang et al., 2019; Bao et al., 2018). Recent work has explored multi-source reasoning (Su et al., 2020). Key aspects include: (1) QG with Target Answer: Answer-agnostic (Chen et al., 2018) vs. answer-aware methods (span-based (Rajpurkar et al., 2016) or abstract (Bajaj et al., 2016)), (2) QG with input context: Ranges from document (Pan et al., 2020a) to keyword levels (Pan et al., 2020b), and (3) Different question formats: Includes standalone, dialogue-style (Reddy et al., 2019), multilingual (Mitra et al., 2021) and multiple-choice (Gupta and Gupta, 2022). Pan et al. (Pan et al., 2019) provide an extensive survey of recent QG methods.

Visual Question Generation. VQG integrates visual and textual understanding, pioneered by Mostafazadeh et al. (2016); Patil and Patwardhan (2020). Question types include: (1) visually grounded (VQA (Antol et al., 2015), CircuitVQA (Mehta et al., 2024)), (2) commonsense-driven (FVQA (Wang et al., 2017)), and (3) knowledge-based (KVQA (Shah et al., 2019)). Approaches encompass encoder-decoder architectures (Mostafazadeh et al., 2016), compositional models like GNNs (He and Wang, 2023), reinforcement learning (Yang et al., 2018a), and bilinear pooling (Mutan (Ben-Younes et al., 2017)).

Multi-Hop Question Generation. Text-based multi-hop question generation aims to create questions that require reasoning over multiple supporting facts or passages, promoting deeper compositional understanding and information synthesis (Yang et al., 2018b; Qi et al., 2020; Geva et al., 2021; Trivedi et al., 2022). Liao et al. (2023) finetune BART for fluency, while Hwang et al. (2024) use a Transformer-RNN hybrid to incrementally build question complexity. Cross-media integration (WebQA (Chang et al., 2022)) and knowledge graphs (Kumar et al., 2019) enable scalable multi-hop reasoning (Phukan et al., 2025). Key focuses include explainability (Hwang et al., 2024) and adaptive difficulty control (Kumar et al., 2019).

3 Dataset

We leverage the WebQA dataset (Chang et al., 2022)[2], a publicly available comprehensive collection designed to support multimodal question-answering. WebQA comprises image-based and text-based queries as well as image-based and text-based sources of knowledge. Each instance in the dataset is a tuple of (Knowledge Sources, Questions, and Answers), where knowledge sources can be images with captions or textual snippets. The dataset includes 34K training QA pairs, with an additional 5K pairs for development and 7.5K for testing. The dataset was curated through a meticulous crowd-sourcing process through Amazon Mechanical Turk, involving multiple stages of quality checks. It is particularly useful for tasks that require integrating and reasoning over information from both text and images, simulating real-world web search experiences.

For our experiments, we filter the dataset to drop instances requiring no or only a single image as multimodal input. This ensured that our model had access to at least two images for multi-hop MVQG.

[2] https://github.com/WebQnA/WebQA/

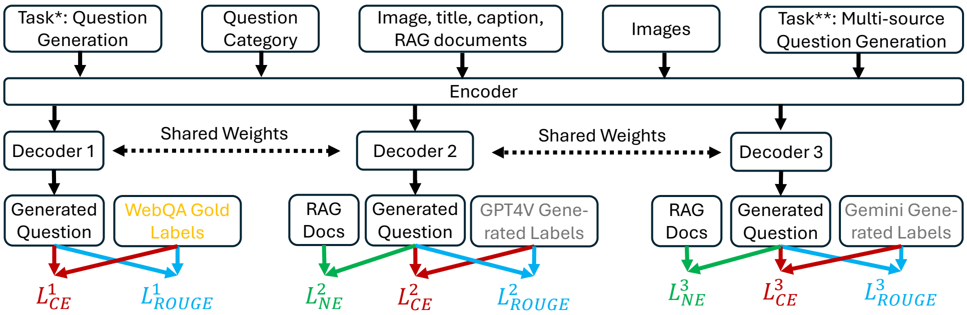


Figure 3: M3RQG Architecture. *Prompt (for decoder 1): *"Generate an english question that meaningfully connects or compares the images and/or passages."*, **Prompt (for decoders 2 and 3): *"You have two images and text passages. Create one multi-hop question that meaningfully connects or compares the images and/or passages."*.

This left us with a train set of 7489, a validation set of 1116, and a test set of 833 samples. The dataset contains gold-standard questions with an average length of 18.02 words, compared to 39.55 words for combined titles and captions (Table 6 in Appendix E). It covers 75 unique topics (e.g., birds, cars, public art) and 6 question types, including but not limited to Yes/No, color, and others.

We extend this dataset by gathering questions for the same samples from GPT-4V and Gemini. We prompt GPT-4V and Gemini to generate questions that integrate additional context. This corresponds to an increased number of discrete reasoning steps that connect facts from various input signals making the generated multi-hop questions more engaging to the user. We quantify this via human and GPT-4o evaluation ensuring our evaluation is fair and unbiased. Note that WebQA's gold questions are not designed to be multi-hop, unlike those generated by GPT-4V and Gemini, which often establish deeper relationships between images and passages.

4 M3RQG Methodology

Fig. 3 shows the overall architecture of the proposed M3RQG method. We leverage a single-encoder multi-decoder Transformer-based architecture. We also utilize RAG for including external knowledge and silver labeled questions from GPT and Gemini. Finally, we utilize cross-entropy loss, NE overlap score and ROUGE-based reward scores for robust training.

4.1 Model Architecture

We experiment with both encoder-decoder and decoder-only architectures. Fig 3 shows an architecture with the encoder-decoder style where we have a single multi-modal encoder but multiple decoders. The model includes one decoder for each label (question) provider: Decoder 1 uses WebQA labels, Decoders 2 and 3 use GPT-4V and Gemini generated labels, respectively. The decoders share weights. To ensure effective training, the multi-decoder framework was designed such that the first decoder guides the model in generating a question that is answerable using information across both the images. Meanwhile, the second and third decoders ensure that the model leverages the additional RAG context provided as input, and hence generate multi-hop KG-aware questions. The model gains a more comprehensive understanding of question generation by distributing the learning process across multiple decoders. Integrating multi-hop reasoning through GPT-4V and Gemini labels enhances the model's ability to generate complex, contextually relevant questions.

4.2 Retrieval Augmented Generation

To enhance the contextual understanding of our model and improve the quality of its generated outputs, we employed a RAG (Lewis et al., 2020b; Gao et al., 2023) approach. Using the title text of the images as queries, we retrieved 10 related Wikipedia pages via the Google Web Search API[3]. This ensures the retrieval of comprehensive and domain-relevant knowledge directly related to the visual context of the input. The detailed cleaning, preprocessing and summarization steps are discussed in Appendix D.

By integrating external information from Wikipedia, the approach significantly reduces the risk of producing incorrect or irrelevant outputs. This contextual framework enhances the model's capacity to establish meaningful connections between disparate elements within the input images.

[3] https://developers.google.com/custom-search/v1/overview

4.3 Prompt Design for Multi-Hop Question Generation

Using WebQA Gold Labels: The WebQA dataset lacks questions utilizing any external KG information. Thus, we prompted the model to generate questions based solely on the available inputs (Fig. 3, Decoder 1). This ensured that the generated questions closely adhered to the gold labels in WebQA and enable the model to learn VQG from multiple images. We leverage the following prompt: *"Generate an english question that meaningfully connects or compares the images and/or passages."*

Utilizing GPT-4V and Gemini Labels: The labels generated by GPT-4V and Gemini incorporated additional external context, enabling more complex reasoning. To utilize these enriched labels, we designed prompts to guide the model in generating multi-hop questions that meaningfully connect or compare images and/or passages along with KG linkages (Figure 3, Decoders 2 and 3). This approach allowed the model to utilize the richer KG context provided by RAG, enabling the generation of complex multi-image multi-hop questions. Prompt: *"You have two images and text passages. Create one multi-hop question that meaningfully connects or compares the images and/or passages."*

4.4 Training

We train the model using standard cross-entropy loss. Further, we also define two rewards using NE-overlap and ROUGE scores, and use them with Proximal Policy Optimization (PPO) (Schulman et al., 2017b).

Supervised Cross-Entropy Loss: Given the model's output logits $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ for three different decoders, the cross-entropy loss is computed as:

$$\mathcal{L}_{CE} = \frac{1}{3}\sum_{i=1}^{3} \alpha_i \cdot \mathcal{L}_{CE}^{i} \tag{1}$$

where $\sum_{i=1}^{3} \alpha_i = 1$ and each cross-entropy term is computed as:

$$\mathcal{L}_{CE}^{i} = -\sum_{t=1}^{T} y_t^i \log p_\theta(\hat{y}_t^i|\mathbf{x}) \tag{2}$$

where y_t^i is 1 for the ground-truth token at timestep t for decoder i (else 0), and $p_\theta(\hat{y}_t^i)$ is the predicted probability at timestep t for token $\hat{y}$ from the i-th decoder.

NE Overlap Score: To ensure that the generated questions effectively incorporate key entities from the external knowledge retrieved by the RAG module, we introduce a NE overlap-based guidance. This function guides M3RQG to utilize as many relevant entities as possible from the retrieved documents when constructing multi-hop questions, improving their informativeness and contextual depth.

We employ a pre-trained BERT-based model[4] for entity extraction, utilizing the standard BIO (Beginning-Inside-Outside) tagging scheme to identify and classify entities. The model can identify O, B-MISC, I-MISC, B-PER, I-PER, B-ORG, I-ORG, B-LOC and I-LOC entities. To balance entity inclusion with relevance, we ensure RAG retrieves only the most relevant knowledge item (based on the title text in the WebQA dataset). We compute the NE overlap between the generated question and the text retrieved by RAG. The NE Overlap-based score penalizes the model when it fails to include entities from the retrieved knowledge, guiding it to generate questions that are structurally coherent and enriched with relevant contextual details. We calculate the NE overlap loss for decoder i as follows:

$$\mathcal{L}_{NE}^{i} = 1 - \frac{|E_{\text{pred}}^{i} \cap E_{\text{RAG\_docs}}|}{|E_{\text{pred}}^{i} \cup E_{\text{RAG\_docs}}|} \tag{3}$$

where E_{pred} and $E_{\text{RAG\_docs}}$ denote the sets of entities extracted from predicted and RAG texts, respectively. We compute NE overlap loss for decoders 2 and 3 and average them out to compute the total NE overlap loss.

$$\mathcal{L}_{NE} = \frac{1}{2}(\beta_1 \cdot \mathcal{L}_{NE}^{2} + \beta_2 \cdot \mathcal{L}_{NE}^{3}) \tag{4}$$

where β_1 and β_2 add up to 1.

ROUGE-based reward: Cross-entropy loss enforces rigid token-level matching, failing to capture semantic equivalence between valid phrasings (e.g., *Is there a cat on the mat?"* vs. *Is there a mat beneath the feline?"*). Hence, we also use ROUGE-L F1 rewards to optimize for structural similarity beyond just token overlap. This rewards semantic alignment while allowing natural syntactic variations. We use the ROUGE-L F1 scores for both the generated sequence and the ground truth sequence to define a reinforcement learning based loss $\mathcal{L}_{ROUGE}^{i}$ (for each decoder i) as discussed in

[4] https://huggingface.co/dslim/bert-base-NER

detail in (Paulus et al., 2018) for a summarization task.

Since our framework leverages three distinct labels (WebQA, GPT-4V, Gemini), the final ROUGE-based loss is computed as follows.

$$\mathcal{L}_{ROUGE} = \frac{1}{3}\sum_{i=1}^{3} \gamma_i \cdot \mathcal{L}^{i}_{ROUGE} \tag{5}$$

where $\mathcal{L}^{i}_{ROUGE}$ is the ROUGE-based losses for decoder i and $\sum_{i=1}^{3} \gamma_i = 1$.

We employ PPO (Schulman et al., 2017b) in our work. PPO learns a policy that balances diversity across label styles with relevance to inputs. This ensures the model has enough flexibility in generation rather than strictly adhering to one "correct" output. PPO enables the model to learn an optimal policy that balances diversity and relevance, guiding it toward producing high-quality, structurally sound, and semantically meaningful multi-hop questions.

Aggregate Loss: Similar to (Paulus et al., 2018), we perform overall training using a mixed learning objective function defined as follows.

$$\mathcal{L}_{\text{final}} = \phi \cdot \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{NE} + \omega \cdot \mathcal{L}_{ROUGE} \tag{6}$$

where $\{\phi, \lambda, \omega\} \in [0, 1]$.

To ensure that $\mathcal{L}_{\text{final}}$ remains positive and nontrivial, we scale the Reinforcement Learning (RL) loss dynamically:

$$\mathcal{L}_{\text{final}} = \max\left(\mathcal{L}_{\text{final}}, 0.1 \cdot \mathcal{L}_{CE}\right) \tag{7}$$

This formulation ensures stable training while balancing supervised and reinforcement learning signals effectively.

5 Experimental Setup

We ran our experiments on an NVIDIA V100 32GB GPU, and to accommodate varying computational demands; we adopted the following batch sizes - for larger models such as Phi 3.5[5] (Abdin et al., 2024), we used a batch size of 1, while for smaller models like BART (Lewis et al., 2020a), we set the batch size to 2 to optimize memory utilization. The training spanned around three days. We set the maximum input token length to 1500, ensuring that longer contexts could be processed while preventing excessive memory consumption. Additionally, image inputs were resized within a controlled range, with a minimum pixel setting of $256 \times 28 \times 28$ and a maximum pixel setting of $1024 \times 28 \times 28$. This ensures each image is encoded using between 256–1024 tokens. The number 28 is derived from the Qwen2-VL's patch configuration: it uses a spatial patch size of 14 combined with a temporal patch size of 2 (14 × 2 = 28). This preprocessing strategy was crucial for handling high-resolution inputs in models such as Qwen2-VL[6] (Wang et al., 2024a), and SmolVLM[7] (Marafioti et al., 2025). We set each α_i to 1/3, each β_i to 1/2 and each γ_i to 1/3. ϕ, λ, ω were all set to 1/3.

6 Results and Analysis

To evaluate the efficacy of our proposed M3RQG framework, we conducted experiments across diverse model architectures, including encoder-decoder models like BART and decoder-only models such as Phi 3.5 and LLaVA.

Recognizing the limitations of traditional automated evaluation metrics, such as BERTScore (Zhang et al., 2019), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and Distinct (Li et al., 2016), in capturing the quality of multi-hop question generation, especially when models are trained using multiple labels, we also incorporate a reference-free metric, RQUGE (Mohammadshahi et al., 2023). To maintain consistency and fairness in evaluation, we compared generated questions against the original gold annotations from the WebQA dataset.

Table 1 presents the primary results for the multimodal multi-hop question generation task, structured into four distinct blocks A-D. More detailed results are shown in Table 4 in Appendix B.

Block A: Baseline models evaluated in a zero-shot setting.

Block B: Experiments on a recently released baseline model (BART-Large variant), finetuned under two configurations: (1) finetune: Vanilla finetuning as proposed in the original paper (Phukan et al., 2024). (2) M3RQG: finetuning with the complete M3RQG framework.

Analyzing Block B and referencing Table 4 in the Appendix B, we observe that integrating a multi-decoder setup introduces deviations from the original dataset labels. Notably, the BART-Large model with a multi-decoder [MD] configuration often generated generic questions, such as "What are

[5] https://huggingface.co/microsoft/Phi-3.5-vision-instruct

[6] https://huggingface.co/Qwen/Qwen2-VL-7B

[7] https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct

	Model	RQUGE	BLEU-1	CIDEr	METEOR	Distinct-1	Distinct-2	BERT-Score	ROUGE-L
A	DeepSeek R1 (DeepSeek-AI, 2025)	2.12	13.06	0.097	24.52	15.30	53.30	48.70	16.12
	Qwen2-VL (Wang et al., 2024a)	1.75	12.83	0.274	15.64	31.21	68.90	48.15	14.84
	ConVQG (Mi et al., 2024)	1.92	3.00	0.172	14.89	18.81	42.97	46.42	7.34
	VisRAG (Yu et al., 2024)	1.95	16.30	0.455	21.30	30.21	70.60	51.62	15.59
	SmolVLM[7]	2.37	06.92	0.182	12.52	42.11	59.93	42.20	10.38
B	BART-Large Variant (Phukan et al., 2024)_finetune	1.80	25.35	0.896	36.89	20.78	65.61	60.38	23.86
	BART-Large Variant (Phukan et al., 2024) +M3RQG	2.41	**27.35**	**1.610**	**44.65**	30.00	73.54	**69.53**	**29.98**
C	LLaVA (Liu et al., 2023)_finetune	1.91	12.74	0.597	18.52	**43.91**	74.00	51.48	13.14
	LLaVA (Liu et al., 2023) + M3RQG	2.86	12.63	0.399	16.95	32.31	**76.92**	52.10	13.63
D	Phi 3.5-V (Abdin et al., 2024)_finetune	2.00	9.80	0.363	13.81	18.42	42.24	33.43	9.69
	Phi 3.5-V (Abdin et al., 2024) + M3RQG	**3.04**	17.91	0.400	23.00	21.08	64.73	51.62	16.68

Table 1: Results on automatic evaluation metrics. RQUGE values for Phi 3.5-V + M3RQG are statistically significant over the DeepSeek R1 baseline (two-tailed t-test, p = $1.01 \times 10^{-12} < 0.05$) signifying that M3RQG indeed helps with Multi-Hop Multimodal Question Generation. Further details are in Table 4 in Appendix B.

Models	Fluency	Multi-Hop Reasoning	Relevance to One Image	Relevance to Both Images	Engagingness
LLaVA (Liu et al., 2023) + M3RQG$_{D2}$	2.60	0.40	1.78	0.78	0.86
LLaVA (Liu et al., 2023) + M3RQG [MD]	1.52	0.00	1.18	0.22	0.26
LLaVA (Liu et al., 2023) + M3RQG [MD + NE]	1.78	0.30	1.46	0.52	0.62
LLaVA (Liu et al., 2023) + M3RQG	2.74	1.26	1.56	1.36	1.42
BART-Large Variant (Phukan et al., 2024) + M3RQG$_{D2}$	2.86	**2.18**	1.30	1.82	1.72
BART-Large Variant (Phukan et al., 2024) + M3RQG [MD]	2.80	1.72	**1.80**	1.66	1.68
BART-Large Variant (Phukan et al., 2024) + M3RQG [MD+ NE]	2.92	1.72	1.78	1.66	1.70
BART-Large Variant (Phukan et al., 2024) + M3RQG	**2.96**	1.96	1.64	2.06	1.74
Phi 3.5-V (Abdin et al., 2024) + M3RQG$_{D2}$	2.86	2.10	1.76	2.40	1.92
Phi 3.5-V (Abdin et al., 2024) + M3RQG [MD]	2.84	2.06	1.62	2.22	1.94
Phi 3.5-V (Abdin et al., 2024) + M3RQG [MD+ NE]	2.44	1.82	1.44	1.94	1.62
Phi 3.5-V (Abdin et al., 2024) + M3RQG	2.88	2.14	1.68	**2.42**	**2.08**

Table 2: GPT-4o evaluation results. Here, MD: Multi-decoder, NE: NE Overlap Score and D_2 is Decoder 2.

the differences between the two paintings?," highlighting a lack of specificity and engagement due to insufficient incorporation of key entities. The introduction of NE overlap scores in the training process [MD+NE] addresses this shortcoming. Comparing [MD] and [MD+NE] rows in Block B of Table 4, there's a marked improvement in the relevance and specificity of the generated questions, underscoring the importance of entity alignment in multi-hop question generation. The complete integration of the M3RQG framework, as depicted in Row 2 of Block B in Table 4 demonstrates the model's enhanced ability to synthesize diverse perspectives into coherent, fact-based multi-hop questions. This comprehensive approach significantly elevates the quality of question generation in the BART architecture. To further validate the versatility and robustness of M3RQG, we extended our experiments to other architectures.

Block C shows results of experiments with the LLaVA architecture, showcasing the impact of M3RQG finetuning.

Block D shows experiments with the Phi 3.5 architecture, highlighting its performance under the M3RQG framework.

Complementing these evaluations, Table 2 presents assessments conducted by GPT-4o based on criteria detailed in Section 7, and among all models augmented with M3RQG, Phi 3.5 consistently outperformed others in terms of relevance to both images and engaginess, demonstrating superior reasoning depth, and image-relevance in generated questions. On the other hand, BART-Large variant performs well on fluency as well as multi-hop reasoning. One can also observe the effectiveness of the M3RQG framework across diverse model architectures and its potential in advancing multi-hop question generation tasks.

7 Human Evaluation

To assess the quality of generated questions, we conducted a human evaluation study focusing on five dimensions: (1) **Fluency:** Evaluates grammatical correctness and naturalness, ranging from 0 (Poor) to 3 (Excellent). (2) **Multi-Hop Reasoning:** Assesses the complexity of reasoning required, from 0 (Single-hop) to 3 (Advanced multi-hop). (3) **Relevance to One Image:** Measures alignment with a single image's content, from 0 (Irrelevant) to 3 (Highly relevant). (4) **Relevance to Both Images:** Evaluates the question's pertinence to both images, from 0 (Irrelevant) to 3 (Highly relevant). (5) **Engagingness:** Assesses the question's ability to provoke interest, from 0 (Not engaging) to 3 (Highly engaging). Detailed human evaluation steps are in Appendix F.

Two annotators with expertise in Natural Lan-

Model	Fluency	Multi-Hop Reasoning	Relevance to One Image	Relevance to Both Images	Engagingness
LLaVA (Liu et al., 2023) + M3RQG$_{D2}$	2.78	0.42	1.42	0.44	0.72
LLaVA (Liu et al., 2023) + M3RQG[MD]	2.48	0.00	0.82	0.68	0.54
LLaVA (Liu et al., 2023) + M3RQG[MD + NE]	2.70	0.20	1.28	0.40	0.44
LLaVA (Liu et al., 2023) + M3RQG	**2.95**	**1.64**	**1.90**	**1.59**	**1.82**
BART-Large Variant (Phukan et al., 2024) + M3RQG$_{D2}$	2.94	2.28	2.00	2.94	2.16
BART-Large Variant (Phukan et al., 2024) + M3RQG[MD]	2.94	1.68	1.86	2.20	1.84
BART-Large Variant (Phukan et al., 2024) + M3RQG[MD+ NE]	2.98	1.60	1.94	2.16	2.20
BART-Large Variant (Phukan et al., 2024) + M3RQG	**2.97**	**1.86**	**1.88**	**2.50**	**2.37**
Phi 3.5-V (Abdin et al., 2024) + M3RQG$_{D2}$	2.96	1.84	1.66	1.82	2.02
Phi 3.5-V (Abdin et al., 2024) + M3RQG[MD]	2.92	2.02	1.74	2.18	2.18
Phi 3.5-V (Abdin et al., 2024) + M3RQG[MD+ NE]	2.46	1.74	1.36	1.88	1.78
Phi 3.5-V (Abdin et al., 2024) + M3RQG	**2.93**	**2.83**	**2.07**	**2.59**	**2.75**

Table 3: Human evaluation results. Here, MD: Multi-decoder, NE: NE Overlap Score and D_2 is Decoder 2.

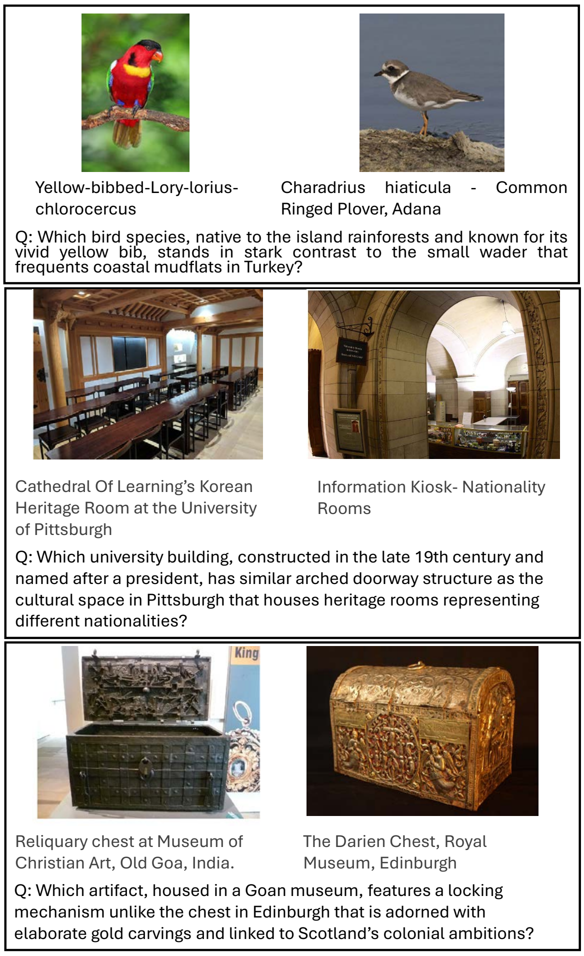

Yellow-bibbed-Lory-lorius-chlorocercus
Charadrius hiaticula - Common Ringed Plover, Adana
Q: Which bird species, native to the island rainforests and known for its vivid yellow bib, stands in stark contrast to the small wader that frequents coastal mudflats in Turkey?
Cathedral Of Learning's Korean Heritage Room at the University of Pittsburgh
Information Kiosk- Nationality Rooms
Q: Which university building, constructed in the late 19th century and named after a president, has similar arched doorway structure as the cultural space in Pittsburgh that houses heritage rooms representing different nationalities?
Reliquary chest at Museum of Christian Art, Old Goa, India.
The Darien Chest, Royal Museum, Edinburgh
Q: Which artifact, housed in a Goan museum, features a locking mechanism unlike the chest in Edinburgh that is adorned with elaborate gold carvings and linked to Scotland's colonial ambitions?
King

Figure 4: Few case studies showcasing the effectiveness of the proposed M3RQG method.

guage Processing tasks independently evaluated 100 randomly selected outputs. They were provided with detailed guidelines and examples for each scoring level to ensure consistency. Inter-annotator agreement measured using Cohen's kappa are shown in Table 7 in Appendix F.

Human evaluation (Table 3) and GPT-4o assessments (Table 2) demonstrate that Phi 3.5-V (Abdin et al., 2024) + M3RQG is the best model for the multi-hop VQG task. Further, Fig. 4 shows a few examples of questions generated by our best model.

However, some limitations were also observed, which we highlight in Table 11 in Appendix I. The first example in the table highlighted a repetition issue in the LLaVA + M3RQG model's output, where the word *"counterpart"* was redundantly used, affecting the question's clarity. In the second example, Phi 3.5-V + M3RQG produced an incoherent output, whereas other models successfully generated relevant multi-hop questions. In the third example, the BART + Large variant (Phukan et al., 2024) + M3RQG model failed to recognize that both images depicted the same entity, resulting in a redundant and nonsensical question. These inconsistencies can be attributed to the uniform prompt and input structure employed across different models in our experiments. While a standard prompt structure facilitates comparative analysis, future work could further explore model-specific prompt tuning to enhance performance.

8 Conclusion

We introduced M3RQG, a model-agnostic framework designed to enhance multimodal, multi-hop question generation by leveraging multiple labels and retrieval-augmented knowledge. We extend the WebQA dataset by adding RAG documents and additional multi-hop questions generated by GPT-4V and Gemini. By integrating a multi-decoder architecture with reinforcement learning via PPO and incorporating a Named Entity overlap score, M3RQG effectively balances alignment with dataset annotations and the generation of complex, information-rich multi-hop questions. Our experiments across diverse architectures, including BART, Phi 3.5, and LLaVA, demonstrate that M3RQG significantly improves generated questions' fluency, reasoning depth, and relevance.

Limitations

While our research advances on Multi-Hop Question Generation with External Knowledge from images, certain limitations must be acknowledged. Our experiments had limited the generalizability across languages as our findings were based only on English. Since online content consumption is inherently multilingual, extending our methodology to support a broader linguistic scope is essential. Incorporating multilingual capabilities would enhance the adaptability of our model, improving its applicability and accessibility for diverse user populations. Furthermore, a key limitation of this study lies in its exclusive reliance on a single dataset for model development and evaluation. While the WebQA dataset provides a robust foundation for multimodal multi-hop question generation, it may affect the generalizability of our findings across diverse datasets. Future research directions should prioritize validation across additional benchmark datasets encompassing broader linguistic, cultural, and contextual variations.

In this paper, we proposed a method which only generates questions but no answers. Although the focus of this paper is on QG and not QA, we acknowledge it as a limitation, and plan to work on gathering answers to these questions as future work, using large visual language models or crowdsourcing. That said, in educational and research contexts, well-posed questions often matter more than answers. Generating questions encourages learners or users to explore, infer, and reason, which is central to active learning, and even scientific inquiry. Further, high-quality, diverse questions can be used to augment QA datasets. In tutoring or accessibility applications, questions can guide attention or prompt reflection.

Another limitation observed is the occasional instability in question generation, attributed to the uniform prompt structure applied across models with varying pre-training formats. Future work could focus on developing model-specific prompt tuning strategies to better align with each architecture's unique characteristics. Exploring the integration of more advanced retrieval mechanisms and dynamic context selection could also enhance the quality and relevance of the generated questions.

Ethical Considerations

This research upholds the highest standards of ethical practice in data handling and dissemination. All human-subject interactions, including annotation tasks and the use of user-generated content, were conducted in compliance with institutional and international ethical guidelines. Before active participation, individuals provided informed consent after being fully apprised of the study's objectives and intended outcomes. Additionally, the research received formal approval from our Institutional Review Board (IRB), which ensures alignment with ethical norms for participant rights and data privacy. Transparency and reproducibility are central to our work. We provide comprehensive documentation of methodologies, data preprocessing protocols, model architectures, and hyperparameters in a publicly accessible repository. Access to the dataset will be restricted to approved researchers who agree to terms limiting use to non-commercial, academic purposes. Furthermore, annotations and evaluations performed by experts were compensated following institutional fair-wage policies, ensuring equitable recognition of their labor.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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*, pages 2425–2433.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Junwei Bao, Yeyun Gong, Nan Duan, Ming Zhou, and Tiejun Zhao. 2018. Question generation with doubly adversarial nets. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2230–2239.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2023. Toward subgraph-guided knowledge graph question generation with graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Ritu Gala, Revathi Vijayaraghavan, Valmik Nikam, and Arvind Kiwelekar. 2021. Real-Time Cognitive Evaluation of Online Learners through Automatically Generated Questions . In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pages 53–58, Los Alamitos, CA, USA. IEEE Computer Society.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 9, pages 346–361.

Pranay Gupta and Manish Gupta. 2022. Newskvqa: Knowledge-aware news video question answering. In *Pacific-asia conference on knowledge discovery and data mining*, pages 3–15. Springer.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886.

Xuehai He and Xin Eric Wang. 2023. Multimodal graph transformer for multimodal question answering. *arXiv preprint arXiv:2305.00581*.

Yi-Ting Huang, Ya-Min Tseng, Yeali S. Sun, and Meng Chang Chen. 2014. Tedquiz: Automatic quiz generation for ted talks video clips to assess listening comprehension. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 350–354.

Seonjeong Hwang, Yunsu Kim, and Gary Geunbae Lee. 2024. Explainable multi-hop question generation: An end-to-end approach without intermediate question labeling. *arXiv preprint arXiv:2404.00571*.

Amrith Krishna, Plaban Bhowmick, Krishnendu Ghosh, Archana Sahu, and Subhayan Roy. 2015. Automatic generation and insertion of assessment items in online video courses. In *Companion Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15 Companion, page 1–4, New York, NY, USA. Association for Computing Machinery.

Vaibhav Kumar and Alan W Black. 2020. ClarQ: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, pages 382–398. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. Mmcoqa: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231.

TongXin Liao, Bin Xu, YiKe Han, Shuai Li, and Shuo Zhang. 2023. Brqg: A bart-based retouching framework for multi-hop question generation. In *International Conference on Advanced Data Mining and Applications*, pages 165–179. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summariation Branches Out, Post-Conference Workshop of ACL 2004*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Karen Mazidi and Rodney Nielsen. 2014. Linguistic considerations in automatic question generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 321–326.

Rahul Mehta, Bhavyajeet Singh, Vasudeva Varma, and Manish Gupta. 2024. Circuitvqa: A visual question answering dataset for electrical circuit images. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 440–460. Springer.

Li Mi, Syrielle Montariol, Javiera Castillo Navarro, Xianjie Dai, Antoine Bosselut, and Devis Tuia. 2024. Convqg: Contrastive visual question generation with multimodal guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4207–4215.

Rajarshee Mitra, Rhea Jain, Aditya Srikanth Veerubhotla, and Manish Gupta. 2021. Zero-shot multilingual interrogative question generation for" people also ask" at bing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3414–3422.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. Rquge: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020a. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475.

Youcheng Pan, Baotian Hu, Qingcai Chen, Yang Xiang, and Xiaolong Wang. 2020b. Learning to generate diverse questions from keywords. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8224–8228. IEEE.

Anupam Pandey, Deepjyoti Bodo, Arpan Phukan, and Asif Ekbal. 2025. The quest for visual understanding: A journey through the evolution of visual question answering. *arXiv preprint arXiv:2501.07109*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Charulata Patil and Manasi Patwardhan. 2020. Visual question generation: The state of the art. *ACM Computing Surveys (CSUR)*, 53(3):1–22.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Arpan Phukan, Manish Gupta, and Asif Ekbal. 2024. ECIS-VQG: Generation of entity-centric information-seeking questions from videos. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14411–14436, Miami, Florida, USA. Association for Computational Linguistics.

Arpan Phukan, Anupam Pandey, Deepjyoti Bodo, and Asif Ekbal. 2025. Videochain: A transformer-based framework for multi-hop video question generation. *arXiv preprint arXiv:2511.08348*.

Peng Qi, Xiao Yang, Yiming Zhang, Danqi Chen, and Christopher D Manning. 2020. Answering complex open-domain questions through multi-hop dense retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3071–3080.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017a. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017b. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA. Association for Computing Machinery.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. *Proceedings of Findings of the Association for Computational Linguistics (EMNLP)*.

Yuxuan Sun, Kai Zhang, and Yu Su. 2023. Multimodal question answering for unified information extraction. *arXiv preprint arXiv:2310.03017*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

Harsh Trivedi, Vidhisha Balachandran, Daniel Khashabi, and Dan Roth. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-hop question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Bingning Wang, Xiaochuan Wang, Ting Tao, Qi Zhang, and Jingfang Xu. 2020. Neural question generation with answer pivot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9138–9145.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Weichao Wang, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Answer-guided and semantic coherent question generation in open-domain conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5066–5076.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024b. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, page 288–304, Berlin, Heidelberg. Springer-Verlag.

Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. 2018. Qg-net: a data-driven question generation model for educational content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, L@S '18, New York, NY, USA. Association for Computing Machinery.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gen: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3997–4003.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018a. Visual curiosity: Learning to ask questions to learn visual recognition. In *Conference on Robot Learning*, pages 63–80. PMLR.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. 2023. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5223–5234.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018b. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Lichao Zhang, Jia Yu, Shuai Zhang, Long Li, Yangyang Zhong, Guanbao Liang, Yuming Yan, Qing Ma, Fangsheng Weng, Fayu Pan, and 1 others. 2024. Unveiling the impact of multi-modal interactions on user engagement: A comprehensive evaluation in ai-driven conversations. *arXiv preprint arXiv:2406.15000*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Overview of Appendix Sections

A Model Nomenclature

This section clarifies the meaning of the following nomenclature:

- MD (Multi-decoder only): No NER Overlap Score or PPO
- MD + NE (Multi-decoder and NER Overlap Score): No PPO
- M3RQG (Multi-decoder + NER Overlap Score + PPO): The full configuration
- Di: i^{th} decoder with NER Overlap Score + PPO

The iterative inclusion of MD, NE, and PPO facilitates the isolation of their contributions.

B Full Results

Our baseline selection is structured to provide a comprehensive and fair evaluation of our proposed M3RQG framework. We establish three distinct categories to ensure meaningful comparison. First, we include state-of-the-art general-purpose multimodal LMs (GPT-4V, Qwen2-VL, DeepSeek, etc.) to represent strong zero-shot or lightly prompted systems widely used in practice, serving as a benchmark for generic capability. Second, we incorporate existing models explicitly adapted for multimodal QA or QG (e.g., VisRAG, MuRAG-finetuned) to situate our work within the landscape of prior research designed for multimodal reasoning with retrieval. Third, we employ ablation baselines using our core backbones without the M3RQG objective (e.g., BART, LLaVA, Phi-3.5 with standard fine-tuning), isolating the specific contribution of our proposed method. This organization ensures we compare against strong generic models, task-specific systems, and controlled ablations, rather than a narrow set of baselines. Furthermore, to quantitatively justify our use of silver data, we report Grounded Question Formability Scores (Table 9) and Answerability Scores (Table 10), which directly compare the quality of our generated silver labels against human-annotated gold standards, demonstrating their viability for training and evaluation.

Table 4 shows the full results table. Here, MD: Multi-decoder, NE: NE Overlap Score and D_i is the i^{th} Decoder.

	Model	RQUGE	BLEU-1	CIDEr	METEOR	Distinct-1	Distinct-2	BERT-Score	ROUGE-L
A	DeepSeek R1 (DeepSeek-AI, 2025)	2.12	13.06	0.097	24.52	15.30	53.30	48.70	16.12
	Qwen2-VL (Wang et al., 2024a)	1.75	12.83	0.274	15.64	31.21	68.90	48.15	14.84
	ConVQG (Mi et al., 2024)	1.92	3.00	0.172	14.89	18.81	42.97	46.42	7.34
	VisRAG (Yu et al., 2024)	1.95	16.30	0.455	21.30	30.21	70.60	51.62	15.59
	SmolVLM[7]	2.37	6.92	0.182	12.52	42.11	59.93	42.20	10.38
	MuRAG_finetuned (Chen et al., 2022)	1.72	20.13	0.888	23.98	31.83	73.38	56.62	20.96
B	BART-Large Variant (Phukan et al., 2024)_zeroshot	1.71	10.67	0.027	13.96	12.93	42.38	35.90	12.34
	BART-Large Variant (Phukan et al., 2024)_finetuned	1.80	25.35	0.896	36.89	20.78	65.61	60.38	23.86
	BART-Large Variant (Phukan et al., 2024) [MD]	1.50	25.15	1.032	31.40	26.21	69.25	58.98	23.39
	BART-Large Variant (Phukan et al., 2024) [MD+NE]	1.94	25.15	1.087	31.71	26.73	69.80	59.19	23.64
	BART-Large Variant (Phukan et al., 2024) + M3RQG$_{D1}$	1.82	25.46	1.264	35.57	31.69	74.14	63.36	24.95
	BART-Large Variant (Phukan et al., 2024) + M3RQG$_{D3}$	1.91	12.91	0.385	15.50	35.48	75.58	49.79	14.29
	BART-Large Variant (Phukan et al., 2024) + M3RQG$_{D2}$	2.00	26.96	1.112	37.00	21.87	65.67	61.36	25.63
	BART-Large Variant (Phukan et al., 2024) +M3RQG	2.41	27.35	1.610	44.65	30.00	73.54	69.53	29.98
C	LLaVA (Liu et al., 2023)_zeroshot	1.74	10.15	0.387	15.31	29.08	64.79	43.206	10.10
	LLaVA (Liu et al., 2023)_finetune	1.91	12.74	0.597	18.52	43.91	74.00	51.48	13.14
	LLaVA (Liu et al., 2023) [MD]	1.87	01.00	0.012	04.93	50.20	84.86	41.30	01.19
	LLaVA (Liu et al., 2023) [MD+NE]	1.88	01.59	0.027	07.06	42.91	83.54	43.87	03.45
	LLaVA (Liu et al., 2023) + M3RQG$_{D1}$	2.03	09.21	0.180	12.06	37.79	79.20	47.04	11.48
	LLaVA (Liu et al., 2023) + M3RQG$_{D3}$	1.88	08.25	0.167	11.90	35.83	76.11	45.44	10.98
	LLaVA (Liu et al., 2023) + M3RQG$_{D2}$	2.41	09.83	0.187	12.92	38.21	80.14	48.03	12.19
	LLaVA (Liu et al., 2023) + M3RQG	2.86	12.63	0.399	16.95	32.31	76.92	52.10	13.63
D	Phi 3.5-V (Abdin et al., 2024)_zeroshot	1.92	11.56	0.168	14.63	14.75	45.30	36.60	11.20
	Phi 3.5-V (Abdin et al., 2024)_finetune	2.00	9.80	0.363	13.81	18.42	42.24	33.43	09.69
	Phi 3.5-V (Abdin et al., 2024) [MD]	1.82	08.32	0.204	11.49	20.65	48.05	31.21	09.66
	Phi 3.5-V (Abdin et al., 2024) [MD+NE]	2.02	16.32	0.282	22.70	18.38	60.53	51.16	15.89
	Phi 3.5-V (Abdin et al., 2024) + M3RQG$_{D1}$	1.85	11.43	0.225	17.29	30.64	71.09	48.13	11.52
	Phi 3.5-V (Abdin et al., 2024) + M3RQG$_{D3}$	2.03	08.92	0.215	14.16	37.19	75.47	46.93	09.22
	Phi 3.5-V (Abdin et al., 2024) + M3RQG$_{D2}$	2.34	17.35	0.324	23.75	18.49	61.82	52.11	16.45
	Phi 3.5-V(Abdin et al., 2024) + M3RQG	3.04	17.91	0.400	23.00	21.08	64.73	51.62	16.68

Table 4: Results. Here, MD: Multi-decoder, NE: NE Overlap Score and D_i is the i^{th} Decoder.

C Gathering Additional Labeled Questions

To generate multi-hop labeled questions for our dataset, we utilized the capabilities of state-of-the-art language models, specifically GPT-4V and Gemini. These models were chosen for their robust understanding of natural language, ability to interpret complex prompts, and adaptability to various input types. By leveraging these models, we were able to obtain high-quality labels that effectively capture semantic and contextual relationships between the two input images. The process involved designing prompts tailored to elicit specific responses from the models. These prompts included examples and contextual information to guide the models toward generating accurate labels that align closely with our intended objective of generating an image based multi-hop question. Refer to Table 5.

D Handling External Information

(1) Cleaning and Preprocessing the Text: To maintain the quality and relevance of the retrieved content, we applied the following cleaning and preprocessing process.

- Removed Special Characters: Stripped newlines, symbols (e.g., [edit], [source]), and other extraneous markers.
- Eliminated Wikipedia-Specific Headers and Footers: Excluded standard Wikipedia elements such as navigation links, “See also,” “References”,“Further reading,” and “External links.”
- Discarded Menus and Sidebars: Removed metadata like “edit,” “view history,” “related changes,” “permanent link,” and similar sidebar elements. This step ensured that only meaningful textual content remained for downstream tasks.

(2) Summarization with Phi3 Mini: Due to the input length constraints of smaller transformer models like BART (1024 for BART-Large), we summarized the cleaned Wikipedia pages to fit their token limits. We employed the lightweight Phi-3-mini-3.8B [8] model for this purpose, using the prompt: *Prompt: “Give me a concise summary of the important points on the given Wikipedia/Wikimedia Commons page in 100 words only.”* This summarization process retained essential information while discarding unnecessary details.

(3) Contextual Integration into Multimodal Large Language Model (MLLM): The summarized outputs from Phi3 Mini were integrated as additional context for the MLLM. This augmented the model’s understanding of the visual and textual inputs, enabling it to generate more informed and

[8] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

contextually relevant outputs.

E Detailed Dataset Statistics

Table 6 shows detailed statistics of our dataset.

F Detailed Human Evaluation

To assess the quality of generation, we conducted a human evaluation based on five key dimensions: *Fluency*, *Multi-Hop Reasoning*, *Relevance to One Image*, *Image Relevance to Both Images*, and *Engagingness*. Each dimension was rated on a 4-point scale (0–3), where higher scores indicate better performance. Below, we detail the evaluation criteria and methodology.

Fluency: This metric evaluates the grammatical accuracy and naturalness of the language used in the generated outputs. Scores range from 0 (Poor) for outputs with grammatical errors and awkward phrasing to 3 (Excellent) for fluent, error-free, and natural language. Examples: "*Is what the color of hat man wears in both pictures.*" (Poor:- 0); "*What is the color of hats men wear in both pictures?*" (Fair:- 1); "*What is the color of the hats the men wear in both pictures?*" (Good:- 2); "What color are the hats worn by the men in both pictures?" (Excellent:- 3).

Multi-Hop Reasoning: This metric evaluates the number of reasoning steps and the complexity required to answer a question, ranging from 0 (Single-hop reasoning), where the answer can be derived from one entity itself, to 3 (Advanced multi-hop reasoning), which requires chaining multiple facts across different entities and modalities. Examples: Single-hop (0): "Who commissioned the Taj Mahal?" (Answerable from one fact: Shah Jahan commissioned it.) Simple multi-hop (1): "Which monument was commissioned by the son of the emperor who built the Taj Mahal?" (Requires two facts: Shah Jahan built Taj Mahal → his son Azam Shah commissioned Bibi Ka Maqbara.) Intermediate multi-hop (2): Uses 1 entity and the images. Advanced multi-hop (3): Uses both images and both entities.

Relevance to One Image: This criterion measures how well the generated question aligns with the content of a single image. Scores range from 0 (Irrelevant) for outputs unrelated to the image to 3 (Highly relevant) for outputs directly addressing the image's content. Examples:"*What is the population of New York City?*" (Irrelevant:- 0); "*Is there any green color in the first image?*" (Slightly relevant:- 1); "*What kind of hats are people wearing in the first image?*" (Mostly relevant:- 2); "*What color are the hats worn by people in the first image?*" (Highly relevant:- 3).

Relevance to Both Images: Similar to the previous dimension, this criterion evaluates the relevance of the output to both images in a multi-image context. A score of 0 (Irrelevant) indicates no connection to the images, while 3 (Highly relevant) indicates a clear and meaningful relationship to the content of both images. Examples: "*What is the capital of Canada?*" (Irrelevant:- 0); "*Do both images have people in them?*" (Slightly relevant:- 1); "*Are there people with hats in both images?*" (Mostly relevant:- 2); "*Compare the styles of hats worn by people in both images.*" (Highly relevant:- 3).

Engagingness: This metric evaluates the overall interest and engagement level of the output. Scores range from 0 (Not engaging) for dull or uninteresting questions to 3 (Highly engaging) for thought-provoking and captivating outputs. Examples: "*Are there hats in the pictures?*" (Not engaging:- 0); "*What colors are the hats in the pictures?*" (Slightly engaging:- 1); "*How do the styles of hats in the pictures differ?*" (Moderately engaging:- 2); "*What do the hats in the images reveal about the social context and time period of the scenes depicted?*" (Highly engaging:- 3).

We employed two annotators with Masters in Computer Science and proficiency in the English language to rate 100 randomly sampled outputs (5 random instances from 10 of the most frequently occurring categories in WebQA) from the M3RQG augmented architectures and baselines. Annotators were provided with a detailed metric and example outputs for each scoring level to ensure consistency. Each output was independently rated across the five dimensions, and inter-annotator agreement was calculated using Cohen's kappa. Refer Table 7.

G GPT-4o as an Evaluator

To ensure scalable and consistent evaluation of generated questions, we employed GPT-4o as an automated human-like evaluator. This section details the prompt design used to guide GPT-4o's assessments.

Prompt: *You are an expert human evaluator with expertise in English. Your task is to analyze and assess the quality of a Generated Question based on specific criteria:*

Fluency

Prompt and Input	Model	Output
Generate a question using the images and linking the following two passages: **Passage 1:** The National Museum of the American Indian is a museum in Washington, DC, that focuses on the history and culture of Native Americans in the United States. **Passage 2:** Robinson Hall is a building on the campus of Brown University in Providence, Rhode Island, designed by Walker & Gould in the High Victorian Gothic style. It was built in 1875-1878 and was named after Ezekiel Robinson, the president of the university at the time.	GPT-4V	Which university building, constructed in the late 19th century in the High Victorian Gothic style, contrasts in purpose with the museum in Washington, DC, dedicated to preserving Native American history and culture?
	Gemini	What state is home to a university building, constructed between 1875 and 1878 in the High Victorian Gothic style, that shares a common regional geography with the primary location of the National Museum of the American Indian?

Table 5: A sample demonstration of the prompt and input structure for GPT-4V and Gemini, designed to generate multi-hop questions.

Aspect	Values
Avg. # words in GPT-4V Generated Questions in Train set	31.25
Avg. # words in GPT-4V Generated Questions in Validation set	31.50
Avg. # words in GPT-4V Generated Questions in Test set	34.61
Avg. # words in GPT-4V Generated Questions Overall	31.29
Avg. # words in Gemini Generated Questions in Train set	14.98
Avg. # words in Gemini Generated Questions in Validation set	14.77
Avg. # words in Gemini Generated Questions in Test set	15.85
Avg. # words in Gemini Generated Questions Overall	14.88
Avg. # words in WebQA Gold Labels in Train set	17.97
Avg. # words in WebQA Gold Labels in Validation set	18.05
Avg. # words in WebQA Gold Labels in Test set	20.76
Avg. # words in WebQA Gold Labels Overall	18.00
Avg. # words in Title and Caption in Train set	39.50
Avg. # words in Title and Caption in Validation set	40.07
Avg. # words in Title and Caption in Test set	44.52
Avg. # words in Title and Caption Overall	39.55
# unique Topics in Train set	75
# unique Topics in Validation set	49
# unique Topics in Test set	19
# unique Topics Overall	75
# unique Categories in Train set	6
# unique Categories in Validation set	6
# unique Categories in Test set	6
# unique Categories Overall	6

Table 6: Dataset Statistics

Score 0: Poor (Grammatical errors and awkward phrasing) Example: "Is what the color of hat man wears in both pictures." Score 1: Fair (Some grammatical errors but understandable) Example: "What is the color of hats men wear in both pictures?" Score 2: Good (Grammatically correct with minor issues) Example: "What is the color of the hats the men wear in both pictures?" Score 3: Excellent (Fluent and natural language with no errors) Example: "What color are the hats worn by the men in both pictures?"

Multi-Hop Reasoning
Score 0: Single-hop (Only needs one entity for the answer) Example: "Who commissioned the Taj Mahal?" Score 1: Simple multi-hop (Multi-hop within KG only without leveraging image link) Example: "Which monument was commissioned by the son of the emperor who built the Taj Mahal?" Score 2: Intermediate multi-hop (Uses 1 entity and the images.) Example: "Which Mughal emperor commissioned the monument that appears larger in these two images, compared to the one built in Aurangabad?" Score 3: Advanced multi-hop (Uses both images and both entities) Example: "Who commissioned this popular tomb in Maharashtra which has an architecture very similar to the mausoleum on the right bank of Yamuna commissioned in 1631 by Shah Jahan?"

Relevance to One Image
Score 0: Irrelevant (Does not relate to either image, assume the images are of a cat and a dog) Example: "What is the population of New York City?" Score 1: Slightly relevant (Relates to one image partially) Example: "Is there any green color in the first image?" Score 2: Mostly relevant (Relates well to one image) Example: "Are there people with hats in the first image?" Score 3: Highly relevant (Directly relates to one image) Example: "What color are the hats worn by people in the first image?"

Relevance to Both Images
Score 0: Irrelevant (Does not relate to both images, assume the images are of a cat and a dog) Example: "What is the capital of Canada?" Score 1: Slightly relevant (Partially relates to both images) Example: "Do both images have people in them?" Score 2: Mostly relevant (Relates well to both images) Example: "Are there people with hats in both images?" Score 3: Highly relevant (Directly relates to both images) Example: "Compare the styles of hats worn by people in both images."

Engagingness
Score 0: Not engaging (Boring or uninteresting) Example: "Are there hats in the pictures?" Score 1: Slightly engaging (Mildly interesting) Example: "What colors are the hats in the pictures?" Score 2: Moderately engaging (Interesting and engaging) Example: "How do the styles of hats in the pictures differ?" Score 3: Highly engaging (Very interesting and captivating) Example: "What do the hats in the images reveal about the social context and time period of the scenes depicted?"

Question: **Context:** **Images:**

Do Not Generate a question yourself. Also, provide

Models	Fluency	Multi-Hop Reasoning	Relevance to One Image	Relevance to Both Images	Engagingness
LLaVA (Liu et al., 2023) +M3RQG	0.928	0.783	0.795	0.874	0.858
BART-Large Variant (Phukan et al., 2024) + M3RQG	0.926	0.837	0.793	0.783	0.861
Phi 3.5-V (Abdin et al., 2024) + M3RQG	0.916	0.861	0.798	0.861	0.874

Table 7: Human Evaluation Kappa Score

a short reasoning for the scores. Output Structure: Fluency: value, Multi-Hop Reasoning: value, Relevance to the images: value, Relevance to One Image: value, Relevance to Both Images: value, Engagingness: value, Reasoning: text

H Multi-Hop Reasoning Performance: Silver vs. Gold Questions

To show that silver labels still meaningfully shape the model's behavior, we perform an LLM-based formability and answerability check. From Tables 9 and 10, we observe that silver labels are more KG-aware and better grounded in both images and retrieved text compared to WebQA gold questions. Additionally, from human and GPT evaluation (Tables 2 and 3), we see that models trained with silver labels produce questions that are more multi-hop-centric, as opposed to those trained only on WebQA gold questions.

I Error Analysis

However, some limitations were also observed, which we highlight in Table 11. In the third example, the Phukan et al. (2024) + M3RQG model generated the question *"What are the differences between the Sherlock Holmes Museum and Sherlock Holmes Museum?"* failing to recognize that both images depicted the same entity, resulting in a redundant and nonsensical query. Additionally, in the second example, Phi 3.5-V + M3RQG produced an incoherent output, whereas other models successfully generated relevant multi-hop questions. The first example in the table highlighted a repetition issue in the LLaVA + M3RQG model's output, where the word *"counterpart"* was redundantly used, affecting the question's clarity. These inconsistencies can be attributed to the uniform prompt and input structure employed across different models in our experiments. Given that each model has been pre-trained with varying input formats, a standardized prompt may not optimally align with all architectures, leading to subpar generation in some instances. While a standard prompt structure facilitates comparative analysis, future work could further explore model-specific prompt tuning to enhance performance.

Models	**HA**	**MHER**	**GR**
LLava + M3RQG	2%	20%	1%
Bart Large + M3RQG	5%	12%	1%
Phi 3.5 V + M3RQG	2%	12%	1%

Table 8: Error type distribution observed in our models

To better understand the limitations of our generated questions, we conducted a manual error analysis on 100 samples, categorizing failures into three distinct types: **Hallucination (HA)**, where the question contains a fabricated entity or fact; **Multi-hop Error (MHER)**, where relevant entities are mentioned but the connecting relation is incorrect or trivial; and **Grammatical error (GR)**. Among all erroneous questions, Table 8 highlights the distribution of error types observed in our model outputs.

J Hallucination Quantification

We assessed hallucinations in our questions generated using automated consistency checks. Specifically, each label in our dataset was evaluated by GPT-5 (similar to the "Auto-Eval" used in HaloQuest (Wang et al., 2024b)) to flag any unsupported claims. This section details the prompt design used to guide GPT-5's assessments. **Prompt:** *You are a strict evaluator. Determine whether the candidate QUESTION could be *faithfully formed* from the provided inputs ('Image A', 'Image B', and the accompanying TEXT).*

Definition of 'formable':

- Every required entity, attribute, or relation referenced by the question is visually or textually grounded in the inputs.
- The question does not hallucinate objects, attributes, counts, or relations absent from the inputs.
- For yes/no or comparative questions, the premises of the question must be supported.

Output JSON with:

- score: in [0,1] where 1=clearly formable, 0=clearly not.

- support: object with boolean flags: image1, image2, text (True if that source contains key evidence).

- verdict: "formable" | "not_formable" | "uncertain".

- rationale: 1-2 sentences explaining your decision.

Labels	Grounded Question Formability Score
Webqa Gold	0.648
Gemini Silver	0.696
GPT 4V Silver	0.804

Table 9: GPT-5 Grounded Question Formability Scores for our gold and generated questions.

Labels	Answerability Score
Webqa Gold question	0.630
Gemini generated questions	0.632
GPT 4V generated questions	0.689

Table 10: GPT-5 Answerability Scores for our gold and generated questions.

- *images: 1-line descriptions of the images.*

The formability scores in Table 9 provide quantitative evidence for the factual grounding of our generated questions. Their significantly higher scores compared to the WebQA gold labels suggest that they are more consistently derived from the provided images and text.

K Downstream QA Utility

We conducted an answerability check on our generated questions via automated evaluation with GPT-5. Similar to the approach in Appendix J, we adapted the “Auto-Eval” method from HaloQuest (Wang et al., 2024b) to judge whether a generated label was answerable from its given context. The prompt designed for this task is described below.

Prompt: *You are a STRICT ANSWERABILITY JUDGE. You will NOT answer the question. You will only decide if the QUESTION can be answered using ONLY the provided inputs: Image A, Image B, and TEXT. Assume no outside/world knowledge.*

Definition of ‘answerable’: A question is answerable if and only if all required entities, attributes, relations, counts, and temporal/spatial conditions are directly grounded in the inputs (visually or textually), and there is sufficient, non-contradictory evidence to compute a single, unambiguous answer.

Core checks (run ALL):

1) ENTITY GROUNDING — Every mentioned object/person/place/concept appears or is named in the inputs.

2) ATTRIBUTE GROUNDING — Stated/asked attributes (color, pose, emotion, state, role, etc.) are supported.

3) RELATION & COMPARISON — Relations (left/right/behind, ownership, part-of) and comparisons (more/less, bigger/smaller, A vs B) are verifiable across Image A, Image B, and/or TEXT.

4) COUNTING/NUMERIC — Counts, ordinals, magnitudes, or measurements are explicitly inferable (not guessed).

5) TEMPORAL/CAUSAL — Before/after/while, changes across A to B, or causes are evidenced (e.g., sequence in TEXT or visual changes across images).

6) COREFERENCE — Pronouns or references (“it”, “they”, “the man”) map unambiguously to grounded entities.

7) OCR/READING — If reading is required (signs, labels, numbers), content is legible and present.

8) AMBIGUITY/CONFLICT — No crucial occlusion, low resolution, or conflicting evidence. If uncertain, prefer “uncertain”.

Disqualifiers (immediately NOT_ANSWERABLE):

- Requires world knowledge, unstated assumptions, or external context (e.g., “Why is this famous?”, “What city is this?” without text evidence).
- Asks for opinions/preferences or hypotheticals not tied to evidence.
- Mentions entities/attributes not present.
- Requires future prediction or intent without evidence.

Scoring rubric: Return a scalar in [0,1]:

- 1.0 = Clearly answerable: all checks pass; evidence is sufficient and unambiguous.
- 0.5 = Uncertain: partial evidence, ambiguous grounding, low legibility, or missing one key element.
- 0.0 = Not answerable: fails any Disqualifier or lacks required evidence.

Support flags (evidence sources):

- image1: True if Image A contains key evidence.
- image2: True if Image B contains key evidence.
- text: True if TEXT contains key evidence.

Evidence pointers: When possible, cite brief phrases from TEXT or short image descriptors (e.g., “A: road sign ‘SLOW”’, “TEXT: ’the red car overtakes the blue car”’). Do NOT invent content.

Mapping score (verdict):

- score $\geqslant 0.85 \Rightarrow$ “answerable”
- $0.35 \leqslant$ score $< 0.85 \Rightarrow$ “uncertain”
- score $< 0.35 \Rightarrow$ “not_answerable”

Special cases:

- Yes/No & Comparative: premises must be explicitly supported for BOTH sides (e.g., A vs B) or by TEXT; otherwise “uncertain” or “not_answerable”.
- Negation (“not”, “none”, “never”): require clear evidence of absence (wide enough view or explicit TEXT).
- Superlatives (“biggest”, “first”): require closed set & clear ordering across the relevant scope.

Guid	WebQA Question	GPT-4V Question	Gemini Question Label	BART-Large Variant (Phukan et al., 2024)+M3RQG	**Phi 3.5+**M3RQG	LLaVA+M3RQG
d5c642ca0dba11ecb1e81171463288e9 	Which bird has a longer beak, an Asian Fairy-bluebird or a Common Hoopoe?	Considering the geographical locations mentioned in the descriptions of the Common Hoopoe and the Asian Fairy-bluebird, what are the similarities in their habitats and how do these locations contribute to the biodiversity of India?	What is the difference between the habitat of the Common Hoopoe and the Asian Fairy-bluebird?	What is the difference between the Common Hoopoe (Upupa epops) and the Great Hornbill?	Which of the two images is of a bird from the family Bucerotidae?	How does the appearance of a blue bird in an image compare to its fair **counterpart counterpart**?
d5bedf120dba11ecb1e81171463288e9 	Do the storefronts on both sides of the Sherlock Holmes Museum both have red boards around around the windows?	What is the address of the Sherlock Holmes Museum as depicted in both images, and in which city is this museum located?	What is the significance of 221B Baker Street to the Sherlock Holmes stories?	What are the differences between the **Sherlock Holmes museum and Sherlock Holmes Museum**?	What two different years did *Flickr user Elliott Brown post images* related to the Sherlock Holmes Museum on Baker Street, London, and what are the key visual elements captured in these images?	How does the exterior design of The Sherlock Holmes Museum differ from its interior?

Table 11: Examples of erroneous questions. Note: The errors are highlighted in bold

Output format (STRICT): Return a SINGLE LINE of JSON matching this schema exactly:

“score”: 0.0 to 1.0,

“verdict”: “answerable” | “uncertain” | “not_answerable”,

“support”: “image1”: bool, “image2”: bool, “text”: bool,

“reasons”: [one or more of “missing_entity”, “missing_attribute”, “missing_relation”, “insufficient_evidence”, “requires_world_knowledge”, “ambiguous_coreference”, “illegible_text”, “occlusion_or_low_resolution”, “conflicting_evidence”, “temporal_unknown”, “count_uncertain”],

“evidence”: [short strings pointing to grounded clues],

“images”: [“image1_desc”: “1 line neutral description of Image A”, “image2_desc”: “1 line neutral description of Image B”],

“rationale”: “Concise 1–2 sentences summarizing why the score/verdict was assigned; mention the decisive checks.”

Now evaluate the candidate QUESTION with this policy.

As shown in Table 10, our generated questions achieve higher answerability scores than the WebQA gold labels. This suggests that the generated questions are not only more directly grounded in the visual and textual inputs but also that the evidence supporting them is sufficient and non-contradictory, a prerequisite for generating unambiguous answers (Pandey et al., 2025). This property is crucial for improving the performance and reliability of downstream tasks.

L Frequently Asked Questions (FAQs)

⁎ How does your method mitigate hallucinations in generated questions?

⇒ We integrate **Retrieval-Augmented Generation (RAG)** to anchor questions in external knowledge (e.g., Wikipedia summaries) and employ **NE overlap loss** to ensure key entities from retrieved documents are preserved. This facilitates factually grounded outputs.

⁎ Why use multiple decoders instead of a single decoder with multi-task learning?

⇒ A single decoder struggles to balance diverse objectives (e.g., single-hop vs. multi-hop generation). Our **multi-decoder architecture** optimizes decoders for distinct label types such as WebQA (for generation structure), GPT-4V, and Gemini (for multi-hop question generation with external knowledge), enabling the model to learn diverse generation styles and integrating multiple perspectives, enhancing robustness, generalization, and adaptability to different contextual complexities.

⁎ How does PPO improve question quality compared to standard cross-entropy loss?

⇒ Cross-entropy loss penalizes token-level deviations, which may discourage valid paraphrases. PPO with ROUGE rewards optimizes for structural similarity and semantic equivalence, allowing flexibility in phrasing while preserving meaning (e.g., “The cat is on the mat” vs. “A feline rests on the mat”).

⁎ Can your framework handle non-English languages or low-resource domains?

⇒ Currently, our experiments are limited to English due to dataset constraints. However, the architecture is language-agnostic; extending it to multilingual settings would require training on parallel corpora and multilingual RAG retrievers (a direction for future work).

⁎ Are the generated questions from GPT-4V/Gemini always better than the dataset's gold labels?

⇒ Not necessarily. While GPT-4V/Gemini enriches questions with external knowledge, WebQA's gold labels ensure task alignment. Our multi-decoder setup balances both: one decoder preserves dataset fidelity, while others enhance complexity via generated questions.

⁎ What is the Named Entity (NE) overlap loss, and what is its role?

⇒ NE overlap loss measures how effectively the generated questions incorporate relevant entities from the external knowledge retrieved. It helps ground generated questions in the context of retrieved knowledge, improving informativeness and relevance.

⁎ Do the questions generated by GPT-4V and Gemini contain any errors or inconsistencies?

⇒ Yes, as with any LLM-generated content, GPT-4V and Gemini questions can contain occasional errors or inconsistencies (e.g., overly specific attributes or world knowledge not fully recoverable from the provided images and retrieved passages). However, from the analysis highlighted in Tables 9 and 10, we observe that while errors do exist, the silver labels are of sufficient quality to provide useful multi-hop, KG-aware supervision, and that any residual noise is mitigated by jointly training with human-written WebQA gold questions.

⁎ Since the method itself uses ROUGE as the RL reward, did you consider the issue of reward hacking?

⇒ We agree that directly optimising a metric via RL can, in principle, lead to reward hacking. In our setup, however, ROUGE-L is only one component of a multi-objective training signal (Equation 6) where cross-entropy (CE) and named-entity overlap (NE) focus on faithfulness to the training labels and grounding in retrieved text, respectively. Empirically, the benefits of our method extend beyond ROUGE-like metrics. We also observe consistent improvements in reference-free metrics (RQUGE), as well as in both human and GPT-4-based evaluations of fluency, multi-hop reasoning, and engagingness.