Diagram-Driven Course Questions Generation

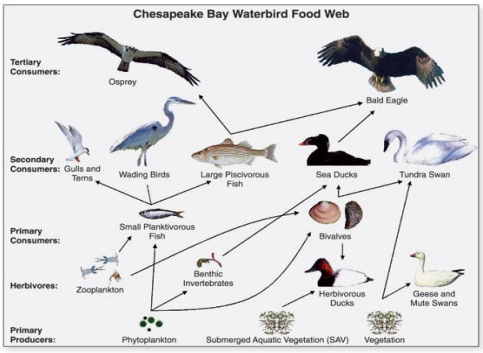
Xinyu Zhang[1,2], **Lingling Zhang**[1,2*], **Yanrui Wu**[1,2], **Muye Huang**[1,2], **Wenjun Wu**[1,2], **Bo Li**,[1,2] **Shaowei Wang**,[1,3] **Basura Fernando**,[4,5] **Jun Liu**[1,3]
[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, China
[4]IHPC, Agency for Science, Technology and Research, Singapore
[5]College of Computing and Data Science, Nanyang Technological University, Singapore
zhang1393869716@stu.xjtu.edu.cn, {zhanglling,liukeen}@xjtu.edu.cn

Abstract

Visual Question Generation (VQG) research focuses predominantly on natural images while neglecting the diagram, which is a critical component in educational materials. To meet the needs of pedagogical assessment, we propose the Diagram-Driven Course Questions Generation (DDCQG) task and construct DiagramQG, a comprehensive dataset with 15,720 diagrams and 25,798 questions across 37 subjects and 371 courses. Our approach employs *course* and *input* text constraints to generate course-relevant questions about specific diagram elements. We reveal three challenges of DDCQG: domain-specific knowledge requirements across courses, long-tail distribution in course coverage, and high information density in diagrams. To address these, we propose the Hierarchical Knowledge Integration framework (HKI-DDCQG), which utilizes trainable CLIP for identifying relevant diagram patches, leverages frozen vision-language models for knowledge extraction, and generates questions with trainable T5. Experiments demonstrate that HKI-DDCQG outperforms existing models on DiagramQG while maintaining strong generalizability across natural image datasets, establishing a strong baseline for DDCQG.

1 Introduction

Visual Question Generation (VQG) represents a pivotal and promising research domain with significant educational applications (Xie et al., 2025; Luo et al., 2024). While VQG focuses on automatically generating questions from visual inputs, current research predominantly addresses natural images while neglecting diagrams (Wang et al., 2024; Zhang et al., 2025; Wang et al., 2025a), a fundamental component of educational materials. This critical gap impedes the advancement of deep learning technologies in educational contexts.



Question: If the number of Bivalves suddenly decreases, what will happen to the number of Geese and Mute Swans?

Text constrained:
(1) Answer: Decrease
(2) Answer & Answer type: Decrease, Relationship
(3) Region: the bounding box of Bivalves
(4) Question type: Relationship
(5) Knowledge triple: (Bivalves, impact, [mask])
(6) Input & Course: Bivalves & Ecological interactions

Figure 1: Comparison of different text constraints for an example in DiagramQG.

The diagram plays a crucial role in pedagogical assessment by facilitating structured information presentation and evaluating students' comprehension of knowledge (Cook, 2006). Therefore, we propose the **Diagram-Driven Course Questions Generation** (DDCQG) task, which aims to encourage models to generate questions based on diagrams for specific courses. These questions are essential for evaluating students' abilities to explain, analyze, and apply course knowledge based on the diagrams (Lambertus et al., 2008).

Furthermore, current VQG research extensively employs various text constraints—including answers (Xie et al., 2025; Mi et al., 2024), answer types (Krishna et al., 2019), image regions (Uehara et al., 2018), question types (Fan et al., 2018), and knowledge triples (Uehara and Harada, 2023). However, these constraints face significant limi-

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5996–6011
November 4-9, 2025

tations when applied to DDCQG task. Research demonstrates that existing approaches frequently produce context-independent questions (Liu et al., 2024a), fail to align with intended assessment objectives (Uehara and Harada, 2023), or merely generate superficial expansions lacking pedagogical depth (Mi et al., 2024), highlighting the need of specific text constraint for DDCQG task.

To address these issues, we present DiagramQG, a comprehensive dataset comprising 15,720 diagrams and 25,798 questions spanning 6 disciplines, 37 subjects, and 371 courses. We propose a novel text constraint with *course* and *input*, as shown in Figure 1, where *course* constraint ensure question relevance to specific educational contexts, and *input* constraint guide question generation around targeted diagram elements. Through analysis of DiagramQG, we identify three fundamental challenges of DDCQG: **domain-specific knowledge requirement**, with models needing to possess specialized knowledge across various disciplines unlike existing VQG datasets that rely primarily on general common sense; **long-tail distribution phenomenon**, where course coverage ranges from abundant to limited, challenging models to generalize across both well-represented and underrepresented courses; and **high object information density**, as diagrams contain concentrated visual information that complicates content interpretation and risks overlooking critical details.

To address these challenges, we propose the Hierarchical Knowledge Integration framework for DDCQG task (HKI-DDCQG) as a strong baseline. This framework employs a trainable CLIP to identify relevant multi-scale diagram patches, utilizes vision-language models like BLIP (Li et al., 2022) and Qwen2.5-VL (Bai et al., 2025) for knowledge extraction, and implements T5 for filtering extracted knowledge to generate the question based on text constraints. The framework then integrates these filtered insights with text constraints and multi-scale diagram patches for question generation. Notably, this frame freezes the VLM's parameters and trains only the remaining parts, thus improving scalability and cost-efficiency.

We evaluate HKI-DDCQG against existing VQG and vision-language models on our DiagramQG dataset and validate its generalizability through experiments on four natural image VQG datasets. Our primary contributions include the following:

- We construct the DiagramQG dataset, incorporating course and input constraints to guide the generation of Diagram-Driven Course Questions that align with educational requirements.
- We propose the HKI-DDCQG, which leverages frozen-parameter vision-language models to enable cost-effective diagram-driven course question generation. The framework demonstrates excellent performance across both DiagramQG and various natural image VQG datasets.
- We conduct comprehensive evaluations of mainstream VQG models, diverse vision-language models, and our HKI-DDCQG on the novel DiagramQG dataset, establishing new benchmarks for diagram-based question generation.

2 Related Work

2.1 Visual Question Generation

VQG has evolved from rule-based approaches to sophisticated neural architectures. While unconstrained approaches often generate generic questions lacking image specificity (Bi et al., 2022), constrained methods have shown success through various strategies, including answer guidance (Xu et al., 2020; Liu et al., 2024c), knowledge enhancement (Xie et al., 2022; Chen et al., 2023), cross-topic models (Liu et al., 2024b), and contrastive learning (Mi et al., 2024). There are also some studies (Xie et al., 2025) on the generation of visual problems with controllable difficulty. However, challenges persist in balancing question diversity with effective visual information utilization, necessitating innovative constraint mechanisms.

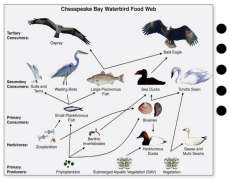
2.2 Diagram Question Answering

Diagram Question Answering (DQA) often demands enhanced reasoning capabilities and domain knowledge. Prior research has focused on improving diagram comprehension through pre-training methods (Gómez-Pérez and Ortega, 2020; Ma et al., 2022; Xu et al., 2023) and utilizing Large Language Models via advanced prompting techniques (Lu et al., 2022; Zhang et al., 2024b; Yao et al., 2024a; Wang et al., 2024; Huang et al., 2025b,a). The effective integration of visual diagrammatic information with background knowledge remains absolutely crucial for improving DQA performance.

3 DiagramQG Dataset

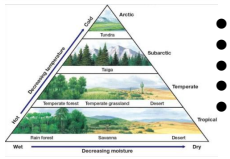
3.1 Data Collection

The data collection process consisted of three distinct phases. First, we gathered diagrams and related questions from existing diagram-related

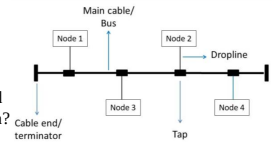


- Discipline: Basic sciences
- Subject: Biology
- Course: Ecological interactions
- Target Text: Bivalves
- Question: If the number of Bivalves suddenly decreases, what will happen to the number of Geese and Mute Swans?

- Discipline: Humanities and social sciences
- Subject: History
- Course : Age of exploration
- Target Text: Columbian Exchange
- Question: Based on the definition of the Columbian Exchange, which arrow could show a part of the Columbian Exchange?



- Discipline: Earth and environmental sciences
- Subject: Climate Science
- Course: Climate zones and biomes
- Target Text: Negative acceleration
- Question: In which biome is tree growth stunted by low temperatures and a short growing season?



- Discipline: Engineering and technology
- Subject: Computer Science
- Course: Network topology
- Target Text: Topology
- Question: What is the type of topology in the graph?

Figure 2: Four different examples of different subjects in DiagramQG dataset.

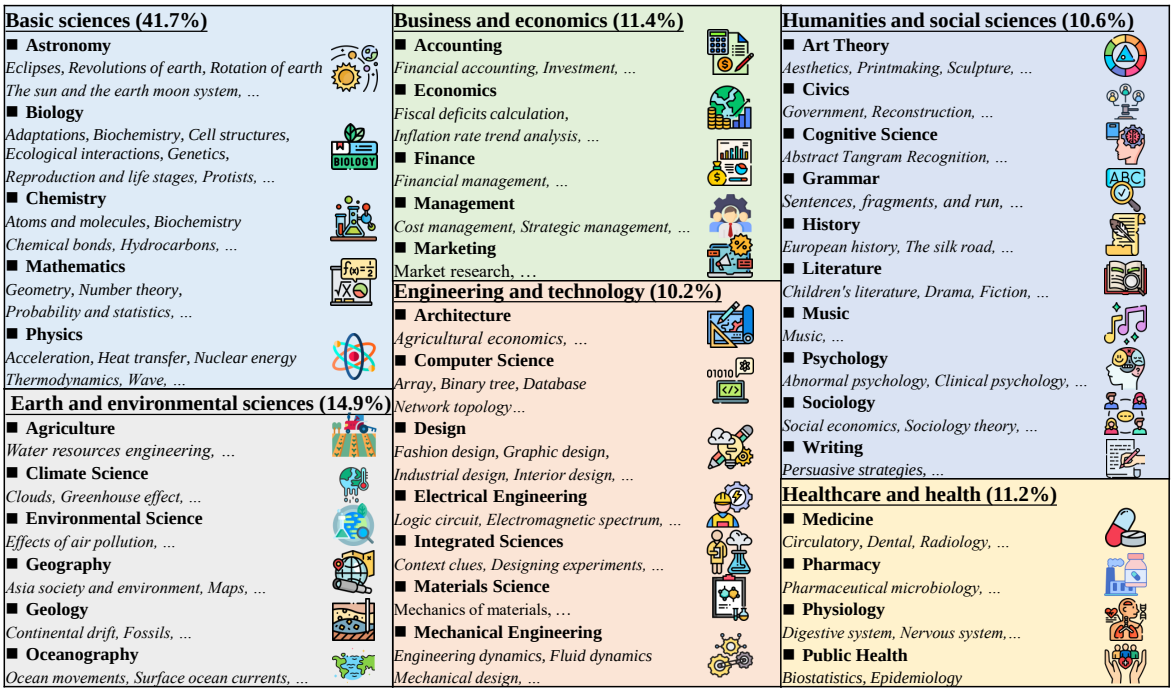


Figure 3: Domain diversity in DiagramQG where each color corresponds to one discipline.

datasets (Kembhavi et al., 2017; Wang et al., 2022; Zhang et al., 2024a; Lu et al., 2022; Yue et al., 2024; Chen et al., 2024), supplemented by diagrams available for academic use from platforms such as Hugging Face and Roboflow. This initial phase yielded a substantial raw dataset containing over 25,000 diagrams and 60,000 questions.

Subsequently, we organized the collected diagrams and questions into six primary disciplines and further categorized them into 37 specific subjects. This structuring process involved mapping questions to their corresponding courses, resulting in the identification of 371 distinct courses.

Finally, highly experienced subject-specific annotators with relevant knowledge backgrounds annotated input text constraints for each diagram-question pair within their specialized domains. A separate group of subject annotators evaluated all samples based on their educational utility using a 100-point scale. Samples scoring below the threshold of 70 are removed, and only the highest-scoring set of text constraints is retained for each diagram-question pair. The resulting DiagramQG dataset contains 25,798 samples associated with 15,720 unique diagrams. Examples from four subjects in the DiagramQG dataset are shown in Figure 2.

3.2 Data Analyse

3.2.1 Domain & Question diversity.

As illustrated in Figure 3, the DiagramQG dataset encompasses 6 disciplines, 37 subjects, and 371 courses across multiple academic domains, with an average of 17.45 words per question. The dataset follows a hierarchical structure, with samples first classified by discipline, then divided into specific subjects (e.g., Biology), and ultimately categorized into courses (e.g., Ecological interactions). For further research, we split DiagramQG into train, val, and test sets with a ratio of 70:5:25. Considering that some courses contain fewer than 5 samples,

Table 1: Comparing characteristics of other datasets and DiagramQG, where Q. and I. mean question and image.

	Num. of Q.	Num. of I.	Object on I.	Images	Text Constraint	Subjects
VQAv2.0	1.1M	20k	3.5	natural	answer	N/A
FVQA	5,826	2k	2.9	natural	answer	N/A
VQG-COCO	25,000	5k	3.3	natural	image caption	N/A
K-VQG	16,098	13K	2.7	natural	knowledge triple	N/A
TQA	26,260	3,455	7.2	diagram	-	10
CSDQA	3,494	1,294	5.2	diagram	-	1
ADE	3.945	3.945	7.4	diagram	-	8
ScienceQA	21,208	10,332	4.3	diagram & natural	-	12
MMMU	11,550	11,264	4.5	diagram & natural	-	30
M3CoT	11,459	11,293	5.4	diagram & natural	-	17
DiagramQG	25,798	15,720	10.5	diagram	input, course	37

we prioritized allocating these courses to the test set to ensure a comprehensive evaluation.

As illustrated in Figure 4, the ratio of questions and diagrams for each course exhibits a pronounced long-tail distribution. This asymmetric distribution indicates that the DiagramQG dataset covers a wide range of courses, some courses have notably limited samples. However, this is a common characteristic in educational scenarios that accurately reflects typical resource allocation patterns in real-world educational environments.

3.2.2 Comparisons to Other Datasets

Table 1 presents a comprehensive comparison between DiagramQG and existing VQG-related and diagram-related datasets, highlighting DiagramQG's distinctive characteristics. DiagramQG encompasses 25,798 questions and 15,720 unique diagrams, substantially surpassing the scale of existing multidisciplinary datasets (such as M3CoT), including those that incorporate both natural images and diagrams. Unlike conventional VQG datasets that primarily focus on image caption and common-sense reasoning, DiagramQG is specifically engineered for educational applications.

DiagramQG demonstrates three significant advantages. First, it exhibits unprecedented educational breadth, spanning 37 distinct academic courses, which exceeds the domain coverage of existing datasets. Second, the dataset introduces a novel constraint that integrates both input phrases and course-specific contextual information, transcending traditional constraint formats. Third, each diagram contains an average of 10.5 objects, representing significantly higher complexity compared to existing datasets in the field. This unique combination of comprehensive course coverage, sophisticated constraint mechanisms, and enhanced

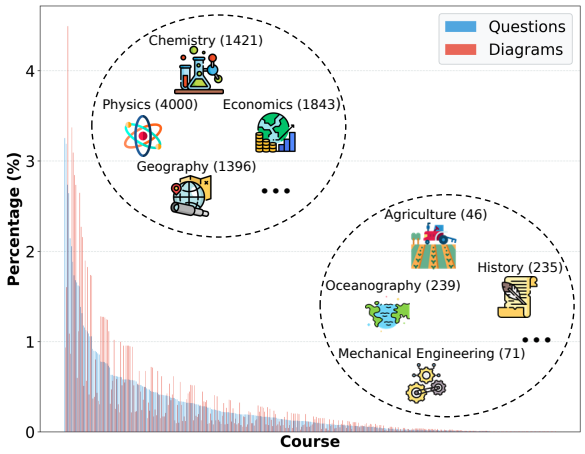


Figure 4: Distribution of diagrams, questions ratios across different courses in DiagramQG.

diagram complexity establishes DiagramQG as the first extensive dataset specifically designed for diagram question generation across diverse educational domains. The dataset facilitates the development of more robust and versatile question-generation systems for educational applications.

3.2.3 Challenges in DiagramQG Dataset

Our comparative analysis reveals three distinctive challenges in DiagramQG that set it apart from existing visual question generation datasets:

- **Domain-specific knowledge requirement:** Unlike other existing VQG datasets that focus on general common sense, DiagramQG dataset consistently requires models to possess and apply different courses across various subjects to generate meaningful and practical questions.
- **Long-tail distribution phenomenon:** The inherent complex long-tail distribution in DiagramQG dataset, where course samples range from abundant to limited, challenges the generalization and performance of models across both sample-rich

and sample-limited courses.

- **High object information density:** The significantly high density of object information in the diagrams complicates content interpretation and risks overlooking critical details, demanding models to possess capabilities in capturing and processing complex visual information.

4 Baseline

4.1 Problem Definition

The Diagram-Driven Course Questions Generation (DDCQG) task aims to generate a pedagogically appropriate question q given a diagram d, an input text t, and a course text c specifically. The generated questions are used to effectively assess students' understanding of the specified course c. This task can be formulated as a conditional generation problem, represented as $p(q|d,t,c)$, where a multimodal model coherently maps visual and textual information into a joint embedding space before decoding questions that satisfy both the text constraint and the course assessment requirement.

4.2 Architecture

We propose the Hierarchical Knowledge Integration (HKI-DDCQG) framework as a baseline for the DDCQG task. This framework is designed to be compatible with any vision-language foundation model and implements a three-stage pipeline for question generation: **HierKnowExtract**, **KnowSelect**, and **KnowFusionQG**, as shown in Figure 5. The **HierKnowExtract** stage extracts knowledge K_s from multi-scale diagram patches P_s using vision-language models with frozen parameters. The **KnowSelect** stage selects the m most relevant knowledge sentences $K_{(t,c)}$ based on text t and course c. The **KnowFusionQG** stage integrates t, c, P_s, and $K_{(t,c)}$ to generate the final question $\hat{q}$.

4.2.1 HierKnowExtract

The **HierKnowExtract** stage obtains diagram patches P_s of different scales related to the input & course and uses vision-language models to extract the knowledge K_s contained in all patches. This process begins with a hierarchical decomposition of diagram d into ordered patches P_d across an n-layered pyramid structure, followed by visual encoding using the CLIP Image Encoder (Radford et al., 2021) to get F^l, as formulated in Equation 1.

$$\begin{cases} F_d = \{F^l\}_{l=1}^{n} \\ P_d = \{p_{i,j}^l \mid l \in [1,n], i,j \in [1,l]\} \\ F^l = \{f_{i,j}^l = \text{CLIP}(p_{i,j}^l) \mid p_{i,j}^l \in P_d\} \\ p_{i,j}^l = d\left[\frac{i-1}{l}H : \frac{i}{l}H, \frac{j-1}{l}W : \frac{j}{l}W\right] \end{cases} \tag{1}$$

where H and W denote the height and width of the input diagram, respectively.

For textual components, the T5 Encoder (Raffel et al., 2020) is employed to process both the input text t and course text c independently. We introduce a learnable linear transformation W_h to ensure seamless compatibility between the CLIP Image Encoder (Radford et al., 2021) and T5 Encoder (Raffel et al., 2020) feature spaces. This effectively facilitates unified similarity computation and subsequent operations in the **KnowFusionQG** phase. The process ultimately culminates in the selection of the most semantically relevant patch from each hierarchical layer to form the patch set P_s, as shown in Equation 2.

$$\begin{cases} e_t = \text{T5}_{\text{enc}}(t), \quad e_c = \text{T5}_{\text{enc}}(c) \\ s_{i,j}^l = \text{sim}(W_h f_{i,j}^l, e_t) + \text{sim}(W_h f_{i,j}^l, e_c) \\ P_s = \{p_{i^*,j^*}^l \mid (i^*,j^*) = \underset{i,j}{\text{argmax}}\, s_{i,j}^l, l \in [1,n]\} \end{cases} \tag{2}$$

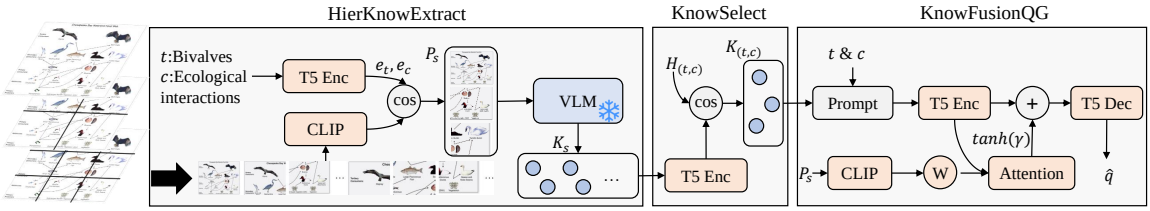
The selected patches P_s, input text t, and course text c are carefully fed into a large-scale vision-language model (VLM) like Qwen2.5-VL (Bai et al., 2025) or BLIP (Li et al., 2022) to obtain diverse knowledge which is related with patch, input and course. The resulting knowledge set K_s is systematically generated according to Equation 3.

$$\begin{cases} K_s = \{k^l \mid l \in [1,n]\} \\ k^l = \text{VLM}(p_{i^*,j^*}^l, t, c), \quad \forall p_{i^*,j^*}^l \in P_s \end{cases} \tag{3}$$

where k^l is the text paragraph retrieved by the VLM, containing several knowledge sentences. To optimize the learning process, the CLIP Image Encoder (Radford et al., 2021), T5 Encoder (Raffel et al., 2020), and the linear transformation W_h employed in this phase share parameters with their counterparts in the third phase, where gradient updates are effectively propagated.

4.2.2 KnowSelect

The **KnowSelect** stage selects the m most relevant knowledge sentences $K_{(t,c)}$ from the extensive knowledge set K_s based on input text t and



(a) Pyramid Layout (b) Framework

Figure 5: (a) Process the diagram into a pyramid structure of patches. (b) Our DiagramQA baseline, HKI-DDCQG, consists of three distinct stages: HierKnowExtract, KnowSelect, and KnowFusionQG. In this framework, orange modules indicate trainable parameters, blue modules represent fixed parameters, and gray modules denote components without learnable parameters.

course text c. The process begins with encoding the knowledge set K_s and the t, c text constraints using the T5 Encoder (Raffel et al., 2020), as formulated in the following Equation 4.

$$\begin{cases} H_K = \text{T5}_{\text{enc}}(K_s) \\ H_{t,c} = \text{T5}_{\text{enc}}(\text{Prompt}(t, c)) \end{cases} \tag{4}$$

where $\text{Prompt}(t, c)$ combines t and c into a prompt following the template: *Given the input text t, identify key knowledge related to the course c.*

For the semantic relevance between the knowledge set ($K_{(t,c)}$) and text constraints (t and c), we employ a scaled dot-product attention mechanism. This computes attention scores between the knowledge tokens and the text constraint, followed by knowledge selection, as formulated in Equation 5:

$$\begin{cases} A = \text{softmax}\left(\dfrac{H_K H_{t,c}^T}{\sqrt{d_k}}\right) \\ K_{(t,c)} = \text{top-m}(AH_K) \end{cases} \tag{5}$$

where d_k denotes the dimensionality of the hidden representations. The top-m() operator selects the m most semantically relevant knowledge sentences based on attention scores, creating a knowledge set $K_{(t,c)}$ that best matches the text constraints.

4.2.3 KnowFusionQG

The **KnowFusionQG** stage integrates the selected diagram patches P_s, input text t, course text c, and knowledge set $K_{(t,c)}$ to generate the diagram-driven course question through a multimodal fusion mechanism. This process begins with encoding the visual and textual inputs into language and vision representations, as shown in Equation 6.

$$\begin{cases} H_t = \text{T5}_{\text{enc}}(\text{Prompt}[t; c; K_{(t,c)}]) \\ H_v = W_h \cdot \text{CLIP}(P_s) \end{cases} \tag{6}$$

where W_h is also used in **HierKnowExtract** phase. This $\text{Prompt}[t; c; K_{(t,c)}]$ synthesizes t, c and $K_{(t,c)}$ into a coherent prompt following the template: *Generate the question including input t to assess course c with the knowledge $K_{(t,c)}$.* To capture the intricate relationships between textual and visual representations, a cross-modal attention mechanism followed by a gated fusion network is employed, as formulated in Equation 7.

$$\begin{cases} H_v^{attn} = \text{Softmax}(\dfrac{H_t H_v^T}{\sqrt{d_k}}) H_v \\ \lambda = W_t H_t + W_v H_v^{attn} \\ H_{fuse} = H_t + \tanh(\lambda) \cdot H_v^{attn} \end{cases} \tag{7}$$

Finally, the fused output H_{fuse} is fed into the T5 decoder (Raffel et al., 2020) to predict the input question $\hat{q}$. This integration of visual and text information through cross-modal attention and gated fusion enables the model to generate questions that are contextually relevant and conceptually focused.

5 Experiments

5.1 Implementation Details

The experimental framework uses T5-Base and T5-Large architectures along with the CLIP (ViT-B/32) model. Optimization is performed with the AdamW optimizer, applying a learning rate of 1e-5 for CLIP and 5e-5 for others, running through 10 epochs of fine-tuning. Key parameters include a maximum input sequence length of 256, an output sequence length of 64, and a batch size of 32, with experiments conducted on one NVIDIA A800 80G GPU. The DiagramQG is split into training, validation, and testing sets with a ratio of 70:5:25, ensuring no overlap of diagrams and questions across these sets. Additionally, each course is represented in both validation and testing sets, despite some courses having relatively few samples.

5.2 Baselines

VQG approaches can be classified into two main categories: *Fine-tuned* and *In-Context Learning*

Table 2: Results on DiagramQG, where GPT-4o (T) and Claude-3.5-Sonnet (T) are meant to generate questions using only text, where, Q-3B, Q-7B, T5-B, T5-L mean, Qwen2.5-VL-3B, Qwen2.5-VL-7B, T5-Base and T5-Large.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Bert-Score	METEOR	CIDEr	ROUGE-L	Fleur
In-Context Learning									
InternVL2.5 (26B) (Wang et al., 2025b)	32.52	19.78	13.14	9.52	92.37	18.51	0.42	28.10	60.31
InternVL2.5 (38B) (Wang et al., 2025b)	33.34	21.15	14.68	10.68	93.76	20.13	0.44	29.66	62.13
MiniCPM-V2(8B) (Yao et al., 2024b)	28.37	15.34	9.66	6.31	88.15	14.26	0.33	25.69	55.14
DeepSeek-VL(7B) (Lu et al., 2024)	28.27	17.13	10.68	7.27	89.88	17.85	0.35	27.17	53.75
Qwen2.5-VL(3B) (Bai et al., 2025)	31.44	18.75	11.64	7.36	88.76	15.51	0.34	27.15	58.95
Qwen2.5-VL(7B) (Bai et al., 2025)	33.51	21.06	13.93	10.52	90.53	18.11	0.51	29.00	61.34
Qwen2.5-VL(72B) (Bai et al., 2025)	41.91	29.45	22.16	17.56	96.86	23.21	0.80	36.57	64.15
GLM4-V pro	37.31	23.81	16.34	12.85	95.66	19.68	0.70	32.07	63.12
Claude-3.5-Sonnet (T)	27.21	15.93	10.00	6.62	92.75	16.10	0.51	24.35	22.76
GPT-4o (T)	23.18	12.78	8.04	4.74	94.82	13.60	0.41	24.22	21.25
Claude-3.5-Sonnet	48.98	40.66	34.68	30.07	98.21	34.62	2.05	45.23	67.69
GPT-4o	51.15	43.08	37.16	32.33	98.94	35.92	2.20	49.60	65.29
Fine-tuned									
K-VQG (Uehara and Harada, 2023)	20.54	15.31	11.98	9.58	73.66	24.76	0.41	26.29	54.25
Patch-TRM (Lu et al., 2021)	21.13	15.90	12.49	10.12	74.46	25.91	0.42	28.09	55.42
ConVQG (BLIP) (Mi et al., 2024)	28.65	21.12	16.38	13.06	74.49	26.93	0.50	28.77	55.75
KC-VQG (BLIP) (Liu et al., 2024a)	30.41	23.95	19.29	15.19	79.62	28.46	0.93	31.55	56.51
KC-VQG (Q-3B) (Liu et al., 2024a)	33.15	25.92	20.92	17.10	83.32	30.57	1.19	34.57	59.76
Qwen2.5-VL-3B (Lora)	50.75	45.37	39.36	35.52	86.04	41.72	3.05	48.21	65.78
Qwen2.5-VL-7B (Lora)	53.64	48.04	41.50	37.56	88.57	44.12	3.22	50.89	68.12
Ours									
HKI-DDCQG(BLIP, T5-B)	44.05	36.31	31.32	27.63	84.02	37.55	2.62	42.65	58.87
HKI-DDCQG(Q-3B, T5-B)	54.55	48.04	43.42	40.90	88.86	47.61	3.74	54.17	64.75
HKI-DDCQG(Q-7B, T5-B)	56.69	49.91	45.07	41.56	89.81	48.22	3.78	55.45	67.84
HKI-DDCQG(BLIP, T5-L)	46.48	39.34	34.54	30.94	85.18	39.50	2.80	45.82	59.25
HKI-DDCQG(Q-3B, T5-L)	57.54	51.12	46.24	44.16	90.11	50.09	4.01	58.18	65.20
HKI-DDCQG(Q-7B, T5-L)	59.63	53.06	48.00	44.85	90.98	50.73	4.04	59.57	69.31

Table 3: Evaluation results on K-VQG dataset.

Dataset	Model	BLEU-4	CIDEr	METEOR
K-VQG	IM-VQG	11.44	17.07	0.26
	K-VQG	18.84	22.79	1.31
	ConVQG	20.01	22.66	1.53
	HKI-DDCQG$_B$	25.34	26.52	1.82
	HKI-DDCQG$_Q$	32.12	30.12	2.08

methods. *Fine-tuned* methods use models like BERT (Devlin, 2018), T5 (Raffel et al., 2020), or GPT-2 (Radford et al., 2019) for question generation, with examples including IM-VQG (Krishna et al., 2019), K-VQG (Uehara and Harada, 2023), Patch-TRM (Lu et al., 2021), ConVQG (Mi et al., 2024), LV2-Net (Liu et al., 2024c), and KC-VQG (Liu et al., 2024a). *In-Context Learning* methods utilize large vision-language models to extract relevant information from diagrams and generate questions using three reference questions. This includes open-source models like Qwen2.5-VL (Bai et al., 2025), MiniCPM-V (Yao et al., 2024b), DeepSeek-VL (Lu et al., 2024), and InternVL2.5(Wang et al., 2025b), as well as closed-source models such as GLM4-V pro, Claude-3.5 Sonnect, and GPT-4o.

5.3 Evaluation metrics

We evaluate our model using established language generation metrics, including BLEU (Papineni et al., 2002), Bert Score, METEOR (Denkowski and Lavie, 2014), CIDEr (Vedantam et al., 2015), and ROUGE (Lin, 2004). We evaluate model performance using the *pycocoevalcap* package to measure alignment between generated and ground truth questions. At the same time, in order to evaluate the relevance of generated questions to diagrams, we introduce the FLEUR evaluation index (Lee et al., 2024) to supplement the evaluation metrics.

5.4 Results

We comprehensively evaluate various models on our DiagramQG dataset, as shown in Table 2. Among the in-context learning methods that utilize Vision-Language Models (VLMs), models with larger parameter counts generally perform significantly better, indicating their abundant knowledge across different subjects. Interestingly, Qwen2.5-VL 3B achieves impressive and competitive results despite having fewer parameters, making it a suitable base model for further comprehensive experiments. Tests conducted with GPT-4o and Claude-3.5-Sonnet using only text inputs result in reduced generation performance, highlighting the critical importance of incorporating diagrams in the question generation process. Among the fine-tuned methods, KVQG and Patch-TRM do not leverage VLMs for additional knowledge, while ConVQG and KC-VQG do. The latter models demonstrate superior performance, with KC-VQG showing im-

Table 4: Results on three VQG datasets.

Dataset	Model	BLEU-1	CIDEr	METEOR
VQG-COCO	MDN	36.0	23.4	0.51
	MC-BMN	40.7	22.6	0.50
	ConVQG	50.2	26.4	0.56
	HKI-DDCQG$_B$	55.6	43.4	0.74
	HKI-DDCQG$_Q$	58.7	47.6	1.08
OK-VQA	IM-VQG	36.47	30.25	0.15
	KVQG	27.18	55.38	0.13
	LV2-Net	29.90	92.17	0.15
	HKI-DDCQG$_B$	31.51	112.61	0.20
	HKI-DDCQG$_Q$	33.82	128.31	0.24
A-OKVQA	IM-VQG	39.30	22.11	0.12
	KVQG	30.56	40.97	0.13
	LV2-Net	32.11	60.06	0.14
	HKI-DDCQG$_B$	33.21	78.42	0.18
	HKI-DDCQG$_Q$	35.31	93.56	0.22

Table 5: Results on DiagramQG with Qwen2.5-VL 3B & T5-Base, where H.K.E, K.S and K.F mean HierKnowExtract, KnowSelect and KnowFusionQG.

H.K.E	K.S	K.F	BLEU-4	METEOR	CIDEr
✓	✓	✓	40.90	47.61	3.74
✓	✓		38.85	45.42	3.48
✓		✓	35.57	44.02	2.95
✓			34.04	42.52	2.82
		✓	14.32	25.64	0.48

proved results as VLM parameters scale up. By observing the Fleur metrics, we find that T5-Base and T5-Large do not show significant improvement in image information understanding, indicating that VLM's contribution to the correlation between the image and the question is more obvious than T5. HKI-DDCQG framework achieves the best performance across all quantitative evaluation metrics, regardless of whether it is based on T5-Base or T5-Large, particularly when using the larger-scale Qwen2.5-VL (Bai et al., 2025) model. These impressive results convincingly validate HKI-DDCQG as an effective baseline for future research by demonstrating the value of incorporating subject knowledge in the DDCQG task.

5.5 Results on Other Dataset

We evaluate HKI-DDCQG's generalization capabilities on four natural image VQG datasets: VQG-COCO, OK-VQA, A-OKVQA, and K-VQG (Tables 3 and 4). Our model consistently outperforms existing approaches, particularly in CIDEr and METEOR metrics, indicating superior semantic comprehension and contextual coherence.

5.6 Ablation experiment

Since **KnowSelect** inherently builds upon **HierKnowExtract**, we cannot evaluate it in isolation. Ablation results (Table 5) demonstrate that knowledge incorporation systematically enhances question generation quality. **KnowSelect** shows greater impact than **KnowFusionQG**, indicating its effectiveness in filtering and retaining relevant information from **HierKnowExtract**'s output.

Table 6: Results on DiagramQG under different settings. We employ HKI-DDCQG (Qwen2.5-VL 3B & T5-Base) to obtain the results.

Setting	BLEU-4	METEOR	CIDEr
n=1,m=4	35.45	41.70	2.98
n=2,m=4	38.13	46.01	3.28
n=4,m=4	40.63	47.25	3.72
n=3,m=1	37.14	44.32	3.35
n=3,m=2	38.85	45.63	3.50
n=3,m=4	41.16	47.74	3.77
n=3,m=5	41.41	48.12	3.78
n=3,m=3	40.90	47.61	3.74

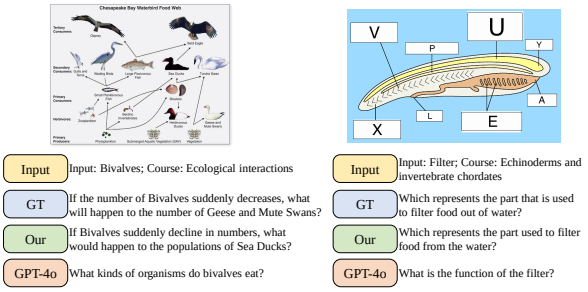


Figure 6: Comparison between HKI-DDCQG and GPT-4o using two examples from the DiagramQG dataset.

5.7 Parameter Sensitive Study

We examine two parameters: the number of pyramid layers (n) for diagram division and the number of knowledge sentences (m) selected for question generation. Results in Table 6 show optimal performance at $n = 3$, with finer divisions adding noise in knowledge extraction. However, performance continues to improve after $m = 3$, although the marginal gains come at the cost of increased input length and computational overhead.

5.8 Case Study

Figure 6 compares HKI-DDCQG with GPT-4o on DiagramQG examples. While GPT-4o generates fluent questions, HKI-DDCQG demonstrates superior capability in producing course-relevant questions that better assess student understanding. More cases are provided in supplementary materials.

6 Conclusion

We present DiagramQG, the first diagram-specific question generation dataset, and HKI-DDCQG, a framework leveraging hierarchical knowledge integration for diagram-driven course question gen-

eration. Experimental results demonstrate HKI-DDCQG's superior performance over existing VQG and vision-language models, advancing intelligent education (Misra et al., 2018) through automated, personalized question generation.

7 Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (No. 62137002, 62293550, 62293553, 62293554, 62450005, 62437002, 62477036, 62477037, 62176209, 62192781, 62306229), 'LENOVO-XJTU' Intelligent Industry Joint Laboratory Project, the Shaanxi Provincial Social Science Foundation Project (No. 2024P041), the Natural Science Basic Research Program of Shaanxi (No. 2023-JC-YB-593), the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001), the Youth Innovation Team of Shaanxi Universities 'Multi-modal Data Mining and Fusion', Project of China Knowledge Centre for Engineering Science and Technology, and the Youth AI Talents Fund of China Association of Automation (Grant No. HBRC-JKYZD-2024-311).

8 Limitation

Our DiagramQG dataset and HKI-DDCQG framework, while advancing educational question generation, face two main limitations. First, our approach generates questions without multiple-choice options. Adding the capability to generate both questions and plausible distractors would enhance the system's educational value, as multiple-choice questions effectively assess student understanding across knowledge levels. However, DiagramQG also establishes the groundwork for future research in the simultaneous generation of diagram-based questions and distractor options. Second, the long-tail distribution across 371 courses poses challenges for uniform performance. Models may show lower effectiveness for underrepresented courses with fewer training examples. However, this distribution reflects real-world educational patterns, and our framework maintains strong generalization across diverse domains. Despite these constraints, DiagramQG significantly advances educational AI research, providing a robust foundation for diagram-driven course question generation.

9 Ethical Statement

In developing DiagramQG, we prioritized ethical considerations throughout the entire process. Our data collection strictly adhered to copyright guidelines, utilizing only publicly available educational resources with appropriate permissions or fair use provisions. All data comply with CC BY-SA 4.0, CC BY-NC-SA, and MIT licenses and have been reviewed to ensure educational value and accuracy. We will publicly release complete datasets under appropriate licenses, both reducing unnecessary carbon footprint and optimizing processing pipelines to lower computational overhead. We fully recognize the broad impacts of automated question-generation systems. Our work aims to assist educators, with these systems designed to complement. To address representation bias, we constructed a diverse dataset spanning 6 disciplines, 37 subjects, and 371 courses. While natural imbalances exist in educational content availability, our dataset provides broader coverage than previous work, enabling contextually appropriate question generation across various educational domains. As an important resource driving AI capabilities in educational reasoning, DiagramQG maintains high standards for data quality and ethical considerations. In all experiments, we strictly comply with all licensing requirements for models and data.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chao Bi, Shuhui Wang, Zhe Xue, Shengbo Chen, and Qingming Huang. 2022. Inferential visual question generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4164–4174.

Jiali Chen, Zhenjun Guo, Jiayuan Xie, Yi Cai, and Qing Li. 2023. Deconfounded visual question generation with causal inference. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5132–5142.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221.

Michelle Patrick Cook. 2006. Visual representations in science education: The influence of prior knowledge

and cognitive load theory on instructional design principles. *Science education*, 90(6):1073–1091.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054.

José Manuel Gómez-Pérez and Raúl Ortega. 2020. Isaaq-mastering textbook questions with pre-trained transformers and bottom-up and top-down attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5469–5479.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025a. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3680–3688.

Muye Huang, Lingling Zhang, Han Lai, Wenjun Wu, Xinyu Zhang, and Jun Liu. 2025b. Vprochart: Answering chart question through visual perception alignment agent and programmatic solution reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3689–3696.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2008–2018.

Amanda Jane Lambertus and 1 others. 2008. Students' understanding of the function concept: Concept images and concept definitions.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hongfei Liu, Guohua Wang, Jiayuan Xie, Jiali Chen, Wenhao Fang, and Yi Cai. 2024a. Knowledge-guided cross-topic visual question generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9854–9864, Torino, Italia. ELRA and ICCL.

Hongfei Liu, Guohua Wang, Jiayuan Xie, Jiali Chen, Wenhao Fang, and Yi Cai. 2024b. Knowledge-guided cross-topic visual question generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9854–9864.

Xumeng Liu, Wenya Guo, Ying Zhang, Xubo Liu, Yu Zhao, Shenglong Yu, and Xiaojie Yuan. 2024c. Look before you leap: Dual logical verification for knowledge-based visual question generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10802–10812, Torino, Italia. ELRA and ICCL.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Jie Ma, Qi Chai, Jingyue Huang, Jun Liu, Yang You, and Qinghua Zheng. 2022. Weakly supervised learning for textbook question answering. *IEEE Transactions on Image Processing*, 31:7378–7388.

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.

Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens Van Der Maaten. 2018. Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Kohei Uehara and Tatsuya Harada. 2023. K-vqg: Knowledge-aware visual question generation for common-sense acquisition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4401–4409.

Kohei Uehara, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Visual question generation for class acquisition of unknown objects. In *Proceedings of the European conference on computer vision (ECCV)*, pages 481–496.

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.

Shaowei Wang, Lingling Zhang, Xuan Luo, Yi Yang, Xin Hu, Tao Qin, and Jun Liu. 2022. Computer science diagram understanding with topology parsing. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–20.

Shaowei Wang, Lingling Zhang, Wenjun Wu, Tao Qin, Xinyu Zhang, and Jun Liu. 2025a. Alignment-guided self-supervised learning for diagram question answering. *IEEE Transactions on Multimedia*, 27:2141–2154.

Shaowei Wang, Lingling Zhang, Longji Zhu, Tao Qin, Kim-Hui Yap, Xinyu Zhang, and Jun Liu. 2024. Cogdqa: Chain-of-guiding learning with large language models for diagram question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, and 1 others. 2025b. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Jiayuan Xie, Mengqiu Cheng, Xinting Zhang, Yi Cai, Guimin Hu, Mengying Xie, and Qing Li. 2025. Explicitly guided difficulty-controllable visual question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25552–25560.

Jiayuan Xie, Wenhao Fang, Yi Cai, Qingbao Huang, and Qing Li. 2022. Knowledge-based visual question generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7547–7558.

Fangzhi Xu, Qika Lin, Jun Liu, Lingling Zhang, Tianzhe Zhao, Qi Chai, Yudai Pan, Yi Huang, and Qianying Wang. 2023. Moca: Incorporating domain pretraining and cross attention for textbook question answering. *Pattern Recognition*, 140:109588.

Xing Xu, Tan Wang, Yang Yang, Alan Hanjalic, and Heng Tao Shen. 2020. Radial graph convolutional network for visual question generation. *IEEE transactions on neural networks and learning systems*, 32(4):1654–1667.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024b. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. 2025. PhysReason:

A comprehensive benchmark towards physics-based reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16593–16615, Vienna, Austria. Association for Computational Linguistics.

Xinyu Zhang, Lingling Zhang, Xin Hu, Jun Liu, Shaowei Wang, and Qianying Wang. 2024a. Alignment relation is what you need for diagram parsing. *IEEE Transactions on Image Processing*.

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, and 1 others. 2024b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

10 Appendix

In this supplementary material, we provide more details on our implementation and experiments as follows:

- Section A: More details on implementation;
- Section B: More details on baselines;
- Section C: More details on DiagramQG;
- Section D: More case studies on Diagram.

A. More details on implementation

In our experimental setup, we developed a framework based on the T5 architecture, utilizing both its Base and Large variants respectively, alongside CLIP (ViT-B/32) for visual understanding capabilities. The framework's trainable parameters, excluding pre-trained components, were initialized following a normal distribution ($\mu = 0$, $\sigma = 0.02$). For optimization, we implemented the AdamW optimizer with a dual learning rate strategy: a conservative $1e-5$ for CLIP components to maintain visual feature integrity, and a higher $5e{-}5$ for remaining components. The training process spanned 20 epochs, incorporating linear learning rate warmup during the initial 2 epochs followed by cosine decay. Our implementation featured specific configurations including 256 tokens for maximum input sequence length, 32 tokens for output sequence length, a batch size of 32, 4 gradient accumulation steps, 0.01 weight decay, and 0.1 dropout rate, with FP16 mixed precision training enabled. All experimental procedures were executed on a hardware setup consisting of two NVIDIA A100 80G GPUs, utilizing CUDA 11.8 and PyTorch 1.13.1.

B. More details on baselines

Visual Question Generation (VQG) approaches can be classified into two main categories: fine-tuning-based methods and large vision-language models.

For **Fine-tuning-based methods**, using the same data split approach as our method, the dataset is divided into training, validation, and test sets in a ratio of 70:5:25. Models are trained on the training set, validated on the validation set, and final results are reported on the test set.

K-VQG (Uehara and Harada, 2023) integrates UNITER, a multi-modal transformer, to encode both visual features (extracted via Faster R-CNN)

and masked knowledge triplets. The model processes visual region features with positional information and combines them with tokenized knowledge triplets through a BART-based decoder, generating questions

Patch-TRM (Lu et al., 2021) employs a hierarchical patch-based transformer that processes diagrams by decomposing them into meaningful patches through a pyramid layout. The model combines ResNet and vision Transformer for visual processing, using attention mechanisms to fuse visual and textual features, enabling effective capture of both local details and global relationships for diagram-focused question generation.

ConVQG (Mi et al., 2024) introduces a contrastive learning framework with dual modality-specific objectives for visual question generation. By contrasting image-specific and text-guided features, the model generates diverse questions that are both image-specific and controllable through flexible textual constraints such as answers, captions, or knowledge triplets, enabling precise control over question content while maintaining visual grounding.

KC-VQG (Liu et al., 2024a) presents a knowledge-guided framework that combines topic-aware visual attention and Large Language Model (LLM) generated knowledge for question generation. The model integrates three components: a topic-aware visual feature extractor, a knowledge extractor with discriminative filtering, and a GPT-2 based decoder, enabling it to generate questions that incorporate both explicit visual information and implicit commonsense knowledge about specified topics.

LV2-Net (Liu et al., 2024c) integrates logical verification into both knowledge acquisition and question generation processes, hence its designation as LV^2-Net. By performing a dual logical structure check—examining the relationships between visual content (V), attributes (A), knowledge (K), ground-truth answers, and the generated questions (Q) at two distinct stages within the knowledge-based visual question generation (KB-VQG) pipeline—LV2-Net is capable of producing a diverse range of insightful knowledge-driven visual questions.

For **large vision-language models** (VLMs), whether open-source or closed-source, we selected three reference questions for each course. These reference questions, along with target and course textual constraints, are incorporated into the prompt during the VLM's question generation process to obtain final outputs.

Open-Source Large vision-language models:

Qwen2.5VL (Bai et al., 2025) is a multimodal model with key enhancements in document parsing, object grounding, and video understanding. It boasts powerful omnidocument parsing capabilities, excelling in diverse document types and languages, including complex elements like tables, charts, and formulas. The model offers precise object grounding with support for various coordinate formats, enabling advanced spatial reasoning. Its ultra-long video understanding is enhanced by dynamic resolution in the temporal dimension, allowing for comprehension of hours-long videos and fine-grained event localization. Furthermore, Qwen2.5VL features improved agent functionality for computer and mobile devices through enhanced grounding, reasoning, and decision-making. Architecturally, it incorporates dynamic FPS sampling and temporal mRoPE for video understanding, along with a streamlined vision encoder utilizing window attention, SwiGLU, and RMSNorm. Available in 3B, 7B, 32B, and 72B parameter versions, its streaming architecture ensures efficient processing of various inputs and strong performance on diagram-related tasks.

MiniCPM-V (Yao et al., 2024b) presents a compact yet effective vision-language model that combines SigLip-400M visual encoder with Qwen2-7B language model, totaling 8B parameters. The model offers efficient multi-image and video understanding capabilities while maintaining competitive performance on vision-language tasks through its streamlined architecture and parameter-efficient design.

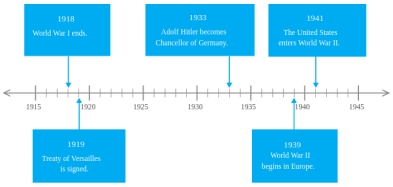
DeepSeek-VL (Lu et al., 2024) introduces a balanced approach to vision-language modeling that builds upon the DeepSeek language model series. The model emphasizes maintaining strong language capabilities while developing visual understanding, featuring high-resolution processing capabilities and a carefully curated training strategy. Through its systematic scaling methodology from 1B to 7B parameters, it achieves competitive performance in practical applications while maintaining efficiency in multi-modal processing.

InternVL 2.5 (Wang et al., 2025b) is an advanced open-source Multimodal Large Language Model (MLLM) series, building on InternVL 2.0 with enhanced training, testing, and data quality. It rivals top commercial models like GPT-4o and

Input: Input text: hybrid structure
Course text: Network topology

GT: Please judge whether it is a hybrid structure.

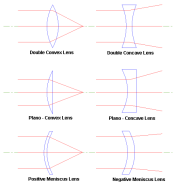
Pred.: What is the difference between the wired and wireless structure?



Input: Input text: World War II
Course text: 20th century american history

GT: How many years passed between the signing of the Treaty of Versailles and the beginning of World War II in Europe?

Pred.: What happened after Theory of Versailles?

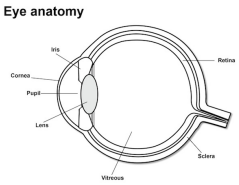


Input: Input text: converge
Course text: Optics

GT: Which of these causes light to converge?

Pred.: What causes the meniscus to curvature?

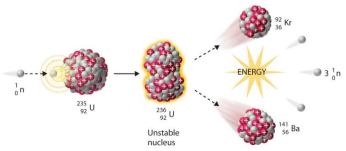
Figure 7: Three different cases (Text Constraint Lost) in DiagramQG dataset.



Input: Input text: Bivalves
Course text: Vision and the eye

GT: Which part controls the diameter and size of pupil?

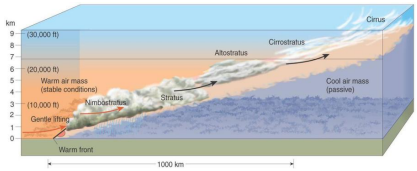
Pred.: What is the function of the pupil?



Input: Input text: energy
Course text: Nuclear energy

GT: Which has more energy Kr or Ba?

Pred.: During this process, what energy is emitted from the nucleus?



Input: Input text: Stratus
Course text: Clouds

GT: In terms of altitude, what are clouds that is in between Stratus and Cirrostratus?

Pred.: What is the top layer of the Stratus?

Figure 8: Three different cases (Diagram Interpretation Bias) in DiagramQG dataset.

Claude-3.5-Sonnet, notably being the first open-source MLLM to exceed 70% on the MMMU benchmark. Key innovations include a Progressive Scaling Strategy for efficient vision encoder and LLM alignment, improved training with Random JPEG Compression and Loss Reweighting, and Well-structured Data Organization for higher quality and training efficiency. InternVL 2.5 aims to advance the open-source multimodal AI community.

Closed-Source Large vision-language models:

GLM4-V is a large-scale multimodal language model developed by ZhiPu, featuring outstanding visual understanding and language generation capabilities. It employs a unified pre-training framework, capable of handling multiple modalities such as text, images, and audio. GLM4-V demonstrates strong performance in tasks like visual question answering, image description, and cross-modal reasoning, and can quickly adapt to new scenarios through few-shot learning. The model supports both Chinese and English, excelling in multilingual understanding and generation.

Claude-3.5-Sonnet is a next-generation multimodal assistant developed by Anthropic, exhibiting exceptional performance in both visual and language understanding. It utilizes an innovative neural network architecture that allows for deep comprehension of image content and complex reasoning. The model has strict controls in terms of safety and ethics, capable of identifying and filtering inappropriate content. A notable feature is its strong contextual understanding and coherent conversational abilities.

GPT-4o is the latest large language model developed by OpenAI, possessing powerful multimodal understanding and generation capabilities. It can process image and text inputs and perform complex reasoning and knowledge integration. This model showcases remarkable few-shot learning abilities, quickly mastering new tasks with only a few examples. GPT-4o also excels in creativity, generating high-quality text, code, and other creative content.

C. More details on DiagramQG

The DiagramQG dataset is a comprehensive collection of questions covering 6 disciplines, 37 subjects, and 371 courses, consisting of 25,798 questions and 15,720 diagrams. This dataset aims to encourage models to generate questions that assess students' course understanding by leveraging the provided input & course text constraints and diagrams, as shown in Figure 2. Accomplishing this task requires two key capabilities from computational models. First, they must demonstrate

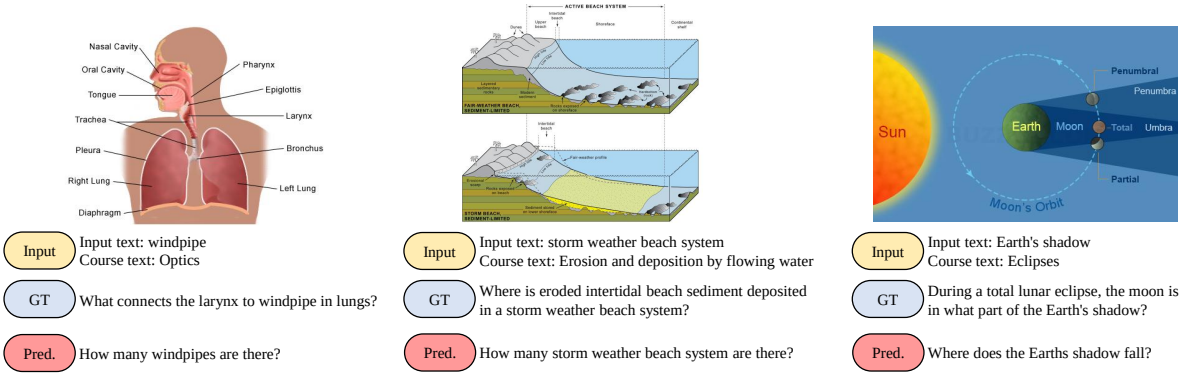


Figure 9: Three different cases (Course Ambiguity) in DiagramQG dataset.

Statistic	Number
Total Diagram	15,720
Total Question	25,798
Total Discipline	6
Total Subject	37
Total Course	371
Train Diagram	11,817
Train Question	17,880
TotTrainal Discipline	6
Train Subject	37
Train Course	351
Val Diagram	1,151
Val Question	1,104
Val Discipline	6
Val Subject	33
Val Course	310
Test Diagram	6,767
Test Question	5,565
Test Discipline	6
Test Subject	37
Test Course	371

Table 7: Main statistics in DiagramQG

sophisticated visual comprehension abilities. Second, they need comprehensive domain knowledge across multiple academic disciplines. Following established methodological practices in machine learning, we implemented a stratified partitioning of the dataset, allocating 70% for training, 5% for validation, and 25% for testing purposes. The comprehensive statistical distribution across these partitions, along with the aggregate dataset metrics, is presented in Table 7. It can be seen that some images have intersections, but this does not affect the experimental results, because its course and question, and can effectively detect overfitting, because overfitting will cause the generated questions to be inconsistent with the current requirements

D More case studies on Diagram

Through extensive examination of the DDCQG (Diagram-based Question Generation) task, we have identified and analyzed three critical bottlenecks: **Text Constraint Lost**, **Diagram Interpretation Bias**, and **Course Ambiguity**. Detailed case studies illustrating these challenges are presented in Figures 7, 8, and 9, respectively.

The first challenge, Text Constraint Lost (Figure 7), manifests when the generated questions fail to maintain fidelity to the target text, resulting in contextually inconsistent question generation. Our analysis suggests that this phenomenon stems from limitations in the model's instruction processing capabilities. Specifically, the neural architecture appears to inadequately preserve and integrate the target text constraints during the generation pipeline, leading to divergent outputs that, while potentially coherent, deviate from the intended textual context.

The second challenge, Diagram Interpretation Bias (Figure 8), reveals a notable gap between the model-generated questions and ground truth questions in terms of course application depth. While the generated questions demonstrate basic assessment of students' course comprehension, they often exhibit simplified or superficial understanding of course concepts. To investigate this phenomenon, we conducted a visualization analysis of the knowledge retrieval process during question generation, as illustrated in Figure 8. Our findings reveal a significant limitation: the retrieved background knowledge corpus lacks crucial content related to key concepts such as "filter blood," thereby constraining the model's ability to generate sophisticated,

course-relevant questions.

The third challenge, Course Ambiguity (Figure 9), represents a important limitation in current DDCQG systems. In these cases, the generated questions demonstrate minimal engagement with course-specific content, instead defaulting to surface-level expansions of the target text. This suggests a deeper architectural limitation in connecting diagram elements with relevant course concepts and pedagogical objectives.

11 Details of human annotators

For data annotation and evaluation, we engaged six graduate students (including both PhD and Master's students) from engineering disciplines who are also co-authors of this paper. All annotators possessed strong backgrounds in different subjects, making them well-qualified for this task. Since the annotators were co-authors actively involved in the research, no formal recruitment process or compensation was required, and they were fully aware of how the data would be used in the study. The annotation process focused solely on content evaluation and did not involve collecting any personal identifying information or expose annotators to any risks. As this research involved co-authors analyzing academic content rather than external human subjects, it was determined to be exempt from formal ethics review board approval. The annotation work was conducted as part of regular academic research activities within our institution. No protected or sensitive demographic information was collected or used in this research.

12 Details of Ai Assistants In Research Or Writing

We used Claude-3.5-Sonnet, o1, o3-mini-high, and Deepseek-R1 to help us write code and polish the paper.