Beyond the Textual: Generating Coherent Visual Options for MCQs

Wanqiang Wang[1], **Longzhu He**[2*], **Wei Zheng**[1]

[1]Beijing Normal University, Beijing, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
{wwq2001, 05166}@mail.bnu.edu.cn, helongzhu@bupt.edu.cn

Abstract

Multiple-choice questions (MCQs) play a crucial role in fostering deep thinking and knowledge integration in education. However, previous research has primarily focused on generating MCQs with textual options, but it largely overlooks the visual options. Moreover, generating high-quality distractors remains a major challenge due to the high cost and limited scalability of manual authoring. To tackle these problems, we propose a Cross-modal Options Synthesis (**CmOS**), a novel framework for generating educational MCQs with visual options. Our framework integrates Multimodal Chain-of-Thought (MCoT) reasoning process and Retrieval-Augmented Generation (RAG) to produce semantically plausible and visually similar answer and distractors. It also includes a discrimination module to identify content suitable for visual options. Experimental results on test tasks demonstrate the superiority of **CmOS** in content discrimination, question generation and visual option generation over existing methods across various subjects and educational levels.

1 Introduction

In the field of education, multiple-choice questions (MCQs) play a crucial role in promoting deep thinking and knowledge integration. Prior studies have shown that well-written MCQs can support learner engagement in higher levels of cognitive reasoning such as application or synthesis of knowledge (Davis, 2009; Zaidi et al., 2018). Among the components of MCQs, the difficulty and relevance of distractors are key indicators of question quality (Gierl et al., 2017; Kumar et al., 2023). However, systematically crafting quality assessment questions and distractors in education is a crucial yet time-consuming, expertise-driven undertaking that calls for innovative solutions (Indran et al., 2024).

With the rapid advancement of large language models (LLMs) that demonstrate remarkable capabilities in scientific question answering, researchers have begun to explore their potential in the automatic generation of educational questions to alleviate human labor, including objective and open-ended formats (Cao and Wang, 2021; Rodriguez-Torrealba et al., 2022). Prior to this study, research on MCQs generation using LLMs primarily focused on two areas: generating questions and textual options based on textual input (Cao and Wang, 2021; Le Berre et al., 2022), and incorporating multimodal inputs to generate questions and textual options enriched with visual information (Yeh et al., 2022; Wang and Baraniuk, 2023; Luo et al., 2024). However, the task of generating MCQs with visual options remains largely unexplored. According to Mayer's cognitive theory of multimedia learning (Mayer, 2005), visual stimuli play an indispensable role in promoting learners' cognitive engagement by facilitating dual-channel processing, reducing extraneous cognitive load, and enhancing the integration of new information with prior knowledge.

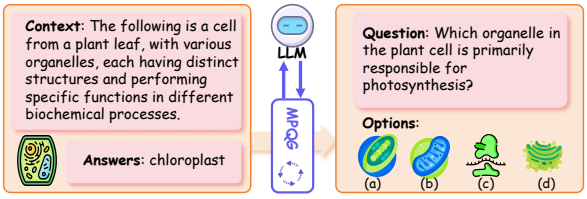


Figure 1: An example of the **CmOS** Framework, showing how it generates a visual MCQ to identify the part responsible for photosynthesis, using content that includes an image, an answer, and the context about plant cell.

There are three major challenges in generating MCQs with visual options. Firstly, not all MCQ options are suitable for visual representation (Butler, 2018). For example, mathematical computation problems typically rely on precise numerical or symbolic expressions which is difficult to convey through static images. Secondly, MCQs with visual option require explicit scaffolding for visual

*Corresponding author. Email: helongzhu@bupt.edu.cn

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6918–6935
November 4-9, 2025

analysis; otherwise, students may incur unnecessary cognitive overload (Kim and Hannafin, 2011). Lastly, current Text-to-Image (T2I) models face key challenges in visual option generation: (i) The inherent variability in visual option generation often leads to inaccurate or unrealistic outputs. (ii) The educational domain lacks specialized image repositories to meet the precise visual needs of academic content. As shown in Figure 1, the goal of our work is to generate an appropriate question and visual options from an input consisting of an image, context, and the answer, via processing pipeline.

To address these challenges, we propose a novel framework named **C**ross-**m**odal **O**ptions **S**ynthesis (**CmOS**), integrating Multimodal Chain-of-Thought (MCoT) reasoning with Retrieval-Augmented Generation (RAG) to generate MCQs with visual options. Specifically, the framework employs an Multimodal Large Language Model (MLLM) to encode multimodal content and embeds it into a four-stage MCoT architecture that separates content discrimination, question and reason generation, alternative pairs screening, and visual options generation. To improve the quality of visual options, we leverage RAG to retrieve similar images from an external educational image database as templates for generation, and then the MLLM and the T2I model are required to optimize based on the templates.

The practical and impactful contributions of this work in MCQs can be summarized as follows:

- We first explore how to generate high-quality MCQs with visual options, addressing a key gap in current MCQs generation research.
- We propose a framework named **CmOS**, which generates questions through multiple questions screening, and produces plausible visual options by providing templates and tuning.
- Experimental results on three tasks indicate that **CmOS** achieves superior performance over existing methods, effectively harnessing the capabilities of both MLLMs and T2I models.

2 Related Works

2.1 Automatic Question Generation

Automatic Question Generation (AQG) is a technology that addresses the high costs and inefficiencies of manually creating educational questions (Brown et al., 2005). Early studies in AQG primarily focus on textual question generation, comprising two categories of generation methods. The first is the Level of Understanding approach, encompassing syntax-based (Afzal et al., 2011; Afzal and Mitkov, 2014) and semantic-based methods (Ai et al., 2015; Afzal and Mitkov, 2014). The second is the Procedure of Transformation approach, including template-based (Kusuma et al., 2018, 2022), rule-based (Singhal and Henz, 2014; Singhal et al., 2015), and statistical methods (Kumar et al., 2015; Uto et al., 2023). However, these methods rely heavily on text as both input and output, exhibiting limited capability in processing multimodal information. In terms of question types, prior research has mainly addressed open-ended and MCQs (Kurdi et al., 2020; Mulla and Gharpure, 2023), with the latter receiving increasing attention due to their standardized format and ease of automated assessment (Touissi et al., 2022; Al Shuraiqi et al., 2024; Newton and Xiromeriti, 2024). With the swift progress of LLMs and MLLMs, researchers have begun to explore their application in generating high-quality educational questions (Mulla and Gharpure, 2023; Yadav et al., 2023; Newton and Xiromeriti, 2024; Al Shuraiqi et al., 2024). Prompt engineering and fine-tuning have been employed to enhance the effectiveness of these models in question generation, with growing interest in their potential for multimodal MCQs generation (Zhao et al., 2022; Wang and Baraniuk, 2023; Luo et al., 2024). However, current studies on multimodal MCQs generation predominantly focus on aligning images with question context, while significantly overlooking the integration of visual information into option generation progress.

2.2 Chain-of-Thought Prompt

Chain-of-Thought (CoT) prompt is an important technique aimed at enhancing the multi-step reasoning capabilities of LLMs. Initial work by Wei et al. (2022) showed that few-shot CoT could significantly improve performance on arithmetic and commonsense tasks. Later, researchers introduced zero-shot variants, such as appending simple phrases like "Let's think step by step". This approach activated reasoning in LLMs without requiring additional examples (Kojima et al., 2022). To minimize the need for manually crafted demonstrations, methods like Auto-CoT were developed (Zhang et al., 2022). These ways leverage LLMs to generate reasoning examples with minimal supervision. Further reliability improvements came from self-consistency decoding, which samples multiple reasoning paths and selects the most frequent

answer (Wang et al., 2022). Beyond text-based tasks, researchers have extended CoT to multimodal content and proposed a method named Multimodal Chain-of-Thought (MCoT) (Zhang et al., 2023). MCoT integrates textual and visual reasoning through several implementation strategies. These include two-stage pipelines that separate rationale generation from answer prediction (Zhang et al., 2023), dual-guidance approaches that disentangle visual and textual reasoning (Jia et al., 2024), and methods that interleave image regions with textual steps (Mitra et al., 2024; Hu et al., 2024). These designs enable MLLMs to perform well across various multimodal benchmarks. More related work is discussed in Appendix A.

3 Method

Problem Definition Given an input consisting of $\mathcal{C} = (T, I, A)$, where T represents the text and I represents the image, and A denotes the answer, the task is to generate an output that includes a high-quality question Q, a visual option A' corresponding to the answer, and multiple visual distractors D_s, each associated with a textual distractor.

3.1 Model Architecture

As illustrated in Figure 2, our **CmOS** consists of four distinct stages: (i) evaluating content convertibility, (ii) generating alternative questions and reasons, (iii) selecting the optimal pair, and (iv) generating option descriptions and visual options. To address the complexity of content discrimination, question and reason generation, and visual option generation, we introduce the MCoT prompt. This method enables MLLMs to identify suitable content for conversion, generate questions and reasons towards the answer, and guide the generation of visual options, based on dynamically exemplars. Specifically, in stage (a), we retrieve relevant multimodal exemplars from an external repository, which includes the original content (text and image) along with convertibility judgments and reasons. They are used to construct dynamic MCoT prompts, enhancing the model's accuracy in content discrimination. In stage (b), we provide three reference processes for each question-reason pair, directing the MLLM to emulate the exemplar reasons. In stage (c), the Optimal Question-Reason Matching (OQRM) module selects the optimal question-reason pair based on internal and external consistency. In stage (d), the option generator produces textual options and their corresponding visual descriptions. Based on these descriptions, relevant images are dynamically retrieved from an external image database, serving as reference templates for generating visual options. Iterative evaluation and tuning using a Text-to-Image (T2I) model further improve the quality of the visual options.

3.2 Exemplars Construction and Retrieval

We construct exemplars using Qwen2.5-VL-72B (Bai et al., 2023), a MLLM with strong visual understanding and instruction-following capabilities. The exemplars include two parts: (i) the original MCQ's context, image, and answer; and (ii) the judgment and reason about its convertibility. We extract 482 questions from the ScienceQA test dataset as the foundation for exemplars construction.

After constructing the exemplars dataset $\mathcal{D}_{\mathcal{E}}$, we introduce an exemplar retrieval mechanism to assist the discriminator in accurately and efficiently determining the convertibility of the given content. Specifically, we adopt the latest FARE encoder (Schlarmann et al., 2024), which enhances robustness over CLIP. The FARE consists of a visual encoder $\phi : I \to \mathbb{R}^D$ and a text encoder $\psi : T \to \mathbb{R}^D$. For a given instance S to be converted, we separately encode its corresponding context t, answer a, and image i as $\mathcal{V}_t$, $\mathcal{V}_a$, and $\mathcal{V}_i$. Let $\mathcal{I} = \{1, 2, \ldots, N\}$ be the index set of exemplars. Denote the encoded vectors of the j-th exemplar's text, answer, and image as $\mathcal{V}_t^j$, $\mathcal{V}_a^j$, and $\mathcal{V}_i^j$, respectively. We compute cross-modality similarities:

$$\mathrm{Sim}_m^j = \frac{\mathcal{V}_m^j \cdot \mathcal{V}_m}{\|\mathcal{V}_m^j\|_2 \, \|\mathcal{V}_m\|_2}, \quad m \in \{t, a, i\}. \quad (1)$$

Select the exemplar by the maximum similarity:

$$j^\star = \arg\max_{j \in \mathcal{I}} \ \max\left(\mathrm{Sim}_t^j, \mathrm{Sim}_a^j, \mathrm{Sim}_i^j\right). \quad (2)$$

The exemplar with index $j^\star$ is concatenated with instance S before being fed into the discriminator.

3.3 Optimal Question-Reason Match

After being processed by the question generator and reason generator with MCoT, several question-reason pairs are produced. To determine which pair is most suitable for generating multiple visual options, we introduce the Optimal Question-Reason Match module (OQRM), which can calculates a Total Match Score (TMS) for each question-reason pair to effectively identify the optimal candidate.

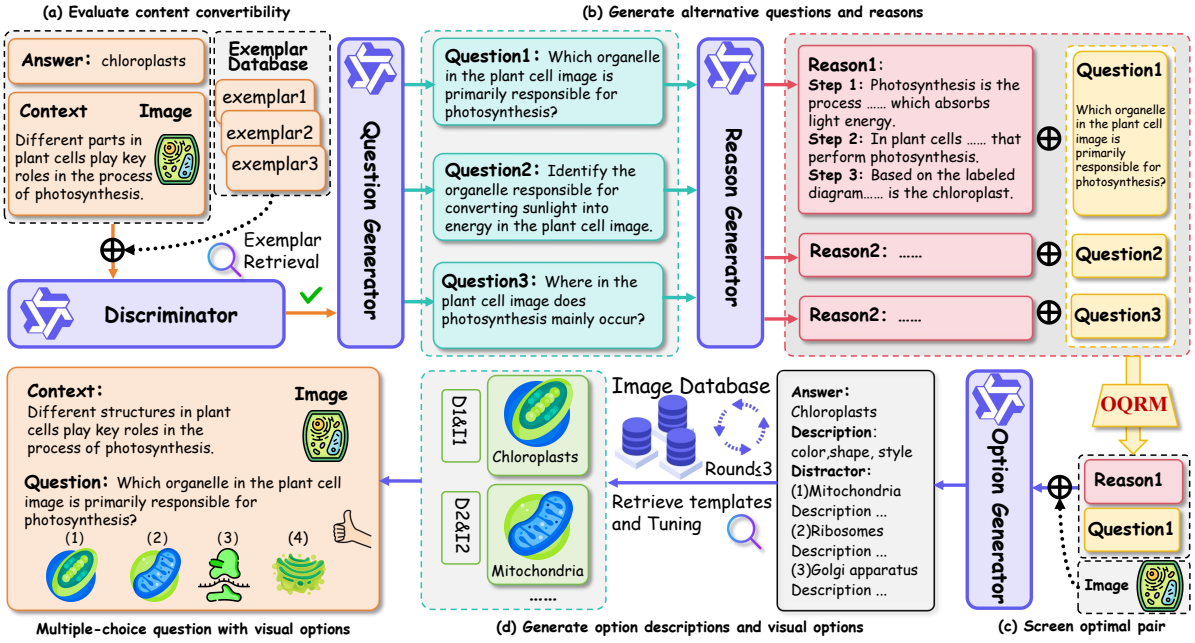


Figure 2: Overview of multimodal educational questions and visual options generation: (a) Evaluate content convertibility: we concatenate the best retrieved exemplar with the instances for content discrimination; (b) Generate alternative questions and reasons: we use prompts to require MLLM to generate diverse questions and reasons; (c) Screen optimal pair: we select the optimal question-reason pair based on internal and external consistency; and (d) Generate option descriptions and visual options: we generate visual options using templates and tuning methods.

The visual encoder ϕ encodes the image i from the instance S to obtain a high-dimensional vector representation $\mathcal{V}_i$. The text encoder ψ processes three textual components: the newly generated question q, the reason r, and the phrase "a photo of A" p, yielding vectors $\mathcal{V}q$, $\mathcal{V}r$, and $\mathcal{V}p$, respectively. The TMS for the k-th ($k \in 1, 2, \ldots, m$) question-reason pair (q_k, r_k) is calculated as the weighted sum of two similarity scores: internal consistency $\mathcal{C}int_k$ and external consistency $\mathcal{C}ext_k$. $\mathcal{C}int_k$ measures the coherence of the question-reason pair within its embedding space $\mathbb{R}^D$, identifying the pair closest to the center $(\mathcal{C}_Q, \mathcal{C}_R)$ to ensure coherence. $\mathcal{C}ext_k$ evaluates the similarity between the pair and the original content. This approach aligns with the self-consistency method proposed by (Wang et al., 2022), which selects the most consistent reason path to mitigate hallucinations and avoid generating irrelevant distractors.

$$\mathcal{C}_{int_k} = \frac{\mathcal{V}_{q_k} \cdot \mathcal{C}_Q}{\|\mathcal{V}_{q_k}\|_2 \|\mathcal{C}_Q\|_2} + \frac{\mathcal{V}_{r_k} \cdot \mathcal{C}_R}{\|\mathcal{V}_{r_k}\|_2 \|\mathcal{C}_R\|_2} \quad (3)$$

where $\mathcal{C}_Q = \frac{1}{m} \sum_{k=1}^{m} \mathcal{V}q_k$, $\mathcal{C}_R = \frac{1}{m} \sum_{k=1}^{m} \mathcal{V}r_k$.

$$\mathcal{C}_{ext_k} = \frac{\mathcal{V}_i \cdot \mathcal{V}_{r_k}}{\|\mathcal{V}_i\|_2 \|\mathcal{V}_{r_k}\|_2} + \frac{\mathcal{V}_{q_k} \cdot \mathcal{V}_p}{\|\mathcal{V}_{q_k}\|_2 \|\mathcal{V}_p\|_2} \quad (4)$$

Finally, considering the hyperparameter α to optimally balance $\mathcal{C}int_k$ and $\mathcal{C}ext_k$, we select the question-reason pair (q^*, r^*) with the highest TMS:

$$(q^*, r^*) = \arg \max_{(q_k, r_k)} \sum (\alpha \mathcal{C}_{int_k} + \mathcal{C}_{ext_k}) \quad (5)$$

3.4 Visual Options Generation

We require the Option Generator to produce t options and their visual descriptive information (including a correct option and $t-1$ distractor options) based on the optimal question-reason pair. Based on the RAG, we propose an adaptive method to generate visual options. This method integrates the excellent text-image alignment ability of the MLLM $\mathcal{G}$ with the generation and enhancement capabilities of the Text-to-Image (T2I) model $\mathcal{P}$.

We construct an image database $\mathcal{D}$ by collecting images i_j from the ScienceQA and generating corresponding captions c_j ($j = 1, 2, \cdots, s$) using MLLM. Given an option description d_i ($i = 1, 2, \cdots, t$), we compute the total similarity between d_i and each image-caption pair as following:

$$Sim_{ij} = \beta \underbrace{\frac{\phi(i_j) \cdot \psi(d_i)}{\|\phi(i_j)\|_2 \|\psi(d_i)\|_2}}_{\text{image-description}} + \underbrace{\frac{\psi(c_j) \cdot \psi(d_i)}{\|\psi(c_j)\|_2 \|\psi(d_i)\|_2}}_{\text{caption-description}} \quad (6)$$

For each option description d_i, we retrieve the

image i_j from $\mathcal{D}$ with the highest Sim_{ij}^{total} as template. Using description d_i and templatei_j, the image generator $\mathcal{P}$ produces a visual image, which is then evaluated by $\mathcal{G}$ to obtain a similarity score Sim_k. If Sim_k meets or exceeds the threshold $\sigma = 0.8$, the image is accepted. Otherwise, $\mathcal{G}$ provides suggestions S to $\mathcal{P}$ for iterative tuning of the generated image up to three rounds.

4 Experiment

4.1 Experimental Setups

Dataset To evaluate our framework's performance in content discrimination, question and visual option generation, we constructed our test datasets based on the ScienceQA benchmark test set $\mathcal{D}_O$(Lu et al., 2022). Each record in $\mathcal{D}_O$ contains the context, question, options, answer, grade, subject, and so on. We randomly sampled 482 instances from $\mathcal{D}_O$ and added "convertible" and "reason" to form the exemplar set $\mathcal{D}_E$. In this paper, *convertible* denotes whether the original content contains core entities or concepts that can be clearly visualized and transformed into image-option questions while maintaining or enhancing cognitive demand. The remaining data comprised $\mathcal{D}_C$, input to the content discrimination module. Each record in $\mathcal{D}_C$ includes the context, image, and answer. From $\mathcal{D}_C$, 812 (40%) convertible instances were selected as $\mathcal{D}_Q$ for downstream tasks. Three human annotators evaluated the convertibility of $\mathcal{D}_C$, and 51.6% of the instances were labeled "TRUE" in the convertible column. They also created new questions for $\mathcal{D}_Q$. A summary of the four datasets can be found in Appendix B and Figure 8.

Metrics We adopt automatic and human evaluation to assess the performance of **CmOS**. Specifically, different tasks are evaluated using tailored metrics. For content discrimination, we use **Accuracy** to measure the its ability to identify convertible content. For question generation, we utilize **BLEU-4** (Papineni et al., 2002) and **ROUGE-L** (Lin, 2004) for question generation, both of which have been widely used in AQG works. For visual option, we adopt the Structural Similarity Index (**SSIM**) (Sara et al., 2019) and the **CLIP-T** (Li et al., 2024) as evaluation metrics. SSIM evaluates the perceptual quality of visual options by jointly modeling luminance, contrast, and structural consistency. CLIP-T quantifies the semantic alignment between generated visual option and the corresponding description. Given the limitations of automatic evaluation in capturing human perception, we further incorporate human evaluation. The specific criteria are presented in the human evaluation section.

Baselines For content discrimination and question generation, we compare our **CmOS** with state-of-the-art (SOTA) methods, including **VL-T5** (Yeh et al., 2022), **MultiQG-Ti** (Wang and Baraniuk, 2023), **Multimodal-CoT** (Zhang et al., 2023), **Chain-of-Exemplar** (Luo et al., 2024) (All these baselines, as well as our **CmOS**, adopt QWEN2.5-VL-7B-INSTRUCT (Bai et al., 2023) as the backbone), and **CHATGPT** under both zero-shot and in-context learning settings (with up to three exemplars in prompts). For visual option generation, due to the limited prior work on educational distractor visuals, we evaluate off-the-shelf model APIs (**FLUX-SCHNELL**, **DALLE-3**, **STABLE DIFFUSION-XL**, **WANX2.1-PLUS**) using identical option descriptions. Our method adopts QWEN2.1-TURBO as the backbone. Baselines and other experimental details are provided in Appendix C.

4.2 Evaluation on Content Discrimination

First, we evaluate the discrimination accuracy of **CmOS** and baselines in determining content convertibility. As shown in Figure 3, **CmOS**, which retrieves similar exemplars from a small pool, significantly improves the accuracy and achieves an average of 88.2%, outperforming all baselines. Notably, although CHATGPT shows lower accuracy in the zero-shot, few-shot prompt significantly enhances its score. Particularly, its average accuracy under the 3 shot becomes comparable to CoE.

In terms of subject domains, all baseline models exhibit the lowest discrimination accuracy on questions from the Natural Sciences (NAT), while achieving the highest accuracy on those from the Language Sciences (LAN). Furthermore, questions accompanied by images (IMG) tend to yield lower discrimination accuracy compared to text-only questions (TXT), indicating potential challenges in multimodal reason. In addition, questions designed for students in higher grades are generally more difficult to discriminate accurately than those intended for lower grades, suggesting increased complexity in advanced educational content.

4.3 Evaluation on Question Generation

Automatic Evaluation Table 1 presents the performance of **CmOS** with $\alpha = 0.6$, compared to SOTA models in terms of BLEU-4 and ROUGE-L.

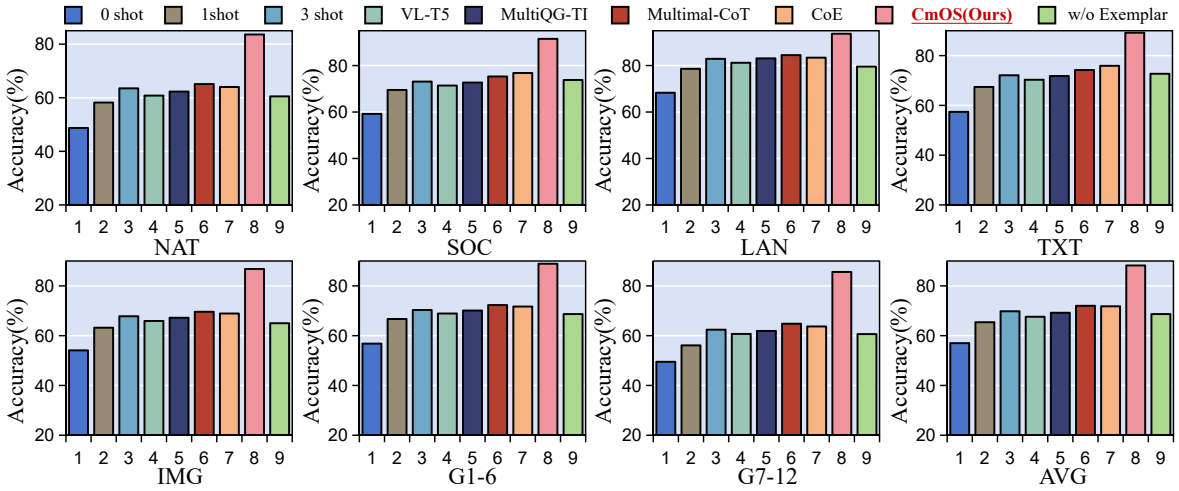


Figure 3: Automatic evaluation results of content discrimination. In terms of accuracy, regardless of subject, modality, or grade, our framework outperforms all baselines, with an average accuracy of **88.2**.

The results show that **CmOS** significantly outperforms all baselines on both metrics, regardless of subject area, modality, or grade level. Among them, MultiQG-TI and Multimodal-CoT, which leverage MLLMs fine-tuned with CoT prompting, slightly outperform VL-T5, a pretrained language model enhanced with visual understanding. Compared to MultiQG-TI and Multimodal-CoT, CoE achieves better performance by integrating exemplar-based CoT reasoning. The table also reports CHATGPT's performance under zero-shot and few-shot. Although CHATGPT benefits from more in-context examples, it remains substantially behind our **CmOS** on the multimodal question generation.

Furthermore, across the three subjects, all baselines consistently achieve the highest performance in Social Science (SOC) and the lowest in Language Science (LAN). These baselines also exhibit improved performance on image-paired questions (IMG) compared to text-only ones (TXT). In contrast, **CmOS** demonstrates stable performance across different subjects and grade levels, highlighting the general ability of the framework in educational content. However, BLEU-4 and ROUGE-L primarily measure surface-level lexical overlap between generated and reference questions, failing to capture semantic relevance. To address this limitation, we incorporate the METEOR metric (Banerjee and Lavie, 2005), which accounts for semantic matching, and also report the results in Appendix D.

Human Evaluation In addition to automatic evaluation, we conduct human evaluation to further assess the quality of generated questions. We randomly sample 50 questions from different methods and recruit three annotators to rate each question on a 1-5 scale across four criteria: (1) **Fluency** (Song and Zhao, 2016), assessing the naturalness and readability of the question; (2) **Grammaticality** (Heilman, 2011), measuring syntactic correctness; (3) **Complexity** (Rodriguez-Torrealba et al., 2022), evaluating the cognitive or linguistic challenge posed by the question; and (4) **Relevance** (Chughtai et al., 2022), measuring how well the question matches the background content. Detailed guidelines are provided in Appendix H.

As shown in Table 2, although the ground-truth questions achieve the highest scores across all metrics, **CmOS** outperforms all baselines and obtains the closest performance to the ground-truth, with an average score of 4.58. These results demonstrate that our framework can generate high-quality educational questions that are fluent, correct, engaging, and relevant to the content. Moreover, both Multimodal-CoT and CoE significantly outperform VL-T5, highlighting the effectiveness of MCoT reasoning. Although CHATGPT trails **CmOS** in complexity and relevance, it surpasses VL-T5 in fluency and grammaticality, indicating its strong ability to generate well-formed natural language context.

4.4 Evaluation on Visual Option Generation

Automatic Evaluation Table 3 presents automatic evaluation results for visual option generation with $\beta = 1.4$. **CmOS** significantly outperforms four baselines in SSIM and CLIP-T by integrating MLLM with T2I model. These results suggest that **CmOS** effectively improves both the structural similarity among visual options and their seman-

Method B-4↑ / R-L↑	Subject NAT	 SOC	 LAN	Modality TXT	 IMG	Grade G1-6	 G7-12	AVG
0 shot	9.2/33.2	5.0/17.2	4.5/25.3	3.5/26.2	8.5/33.6	6.7/30.1	4.2/28.1	4.9/28.3
1 shot	19.5/37.2	19.6/35.8	10.4/26.2	18.8/35.8	19.6/38.6	18.4/35.8	21.7/38.6	19.3/36.6
3 shot	28.6/46.9	29.0/51.6	27.4/53.6	27.3/47.9	29.8/48.8	30.0/48.5	27.1/48.6	29.1/48.5
VL-T5	39.8/55.6	37.3/45.0	34.3/46.1	36.1/50.2	40.7/52.7	38.1/50.9	39.6/54.9	39.1/51.5
MultiQG-TI	43.2/58.7	39.8/47.2	37.3/49.0	38.9/52.7	43.5/55.9	41.0/53.7	42.0/58.1	42.3/55.0
Multimodal-CoT	47.6/63.3	44.4/53.2	42.9/53.8	44.0/57.4	48.3/60.9	47.0/58.9	47.7/63.0	46.6/59.8
CoE	55.7/72.3	52.7/61.5	46.9/57.3	51.6/66.7	56.0/69.9	54.2/67.7	54.8/72.1	54.7/69.9
CmOS	**76.8**/**77.5**	**81.1**/**79.2**	**50.5**/**63.2**	**75.6**/**76.7**	**78.6**/**78.4**	**79.3**/**78.4**	**70.1**/**74.5**	**75.5**/**77.2**
w /o Discriminator	71.6/71.7	77.2/74.2	49.5/60.6	72.3/71.3	73.9/72.5	74.3/73.9	69.8/71.7	72.9/72.1
w /o OQRM	44.8/66.5	41.3/55.9	37.6/57.2	39.1/61.1	43.7/60.6	42.1/59.8	42.6/65.9	42.4/62.3

Table 1: Automatic evaluation results of question generation. ↑: higher is better.

Method	Flu.↑	Gra.↑	Com.↑	Rel.↑	AVG↑
ChatGPT4	4.65	4.71	3.18	3.24	3.94
VL-T5	4.31	4.56	3.49	3.55	3.97
MultiQG-TI	4.61	4.65	3.99	4.12	4.34
Multimodal-CoT	4.65	4.68	4.04	4.25	4.41
CoE	4.65	4.72	4.25	4.29	4.48
CmOS	**4.69**	**4.78**	**4.37**	**4.49**	**4.58**
Groundtruth	4.72	4.82	4.59	4.57	4.68

Table 2: Human evaluation results of question generation. ↑: higher is better.

tic alignment with descriptions. Benefiting from strong semantic understanding and stylistic generalization, WANX2.1-PLUS slightly outperforms the other three baselines in most categories. In contrast, no substantial differences are observed among the remaining baselines. The hyperparameter analysis results for α and β are provided in Appendix E.

To further evaluate performance across disciplines, we analyze the results for three academic subjects. Remarkably, **CmOS** achieves the best performance in the social sciences (SOC), but performs worst in language sciences (LAN), mirroring the trends observed in question generation performance. Moreover, for image-equipped (IMG) questions, **CmOS** obtains the highest scores in both SSIM and CLIP-T, while for text-only (TXT) questions, its performance degrades significantly on both metrics. The findings demonstrate that visual input enhances the semantic fidelity and contextual plausibility of generated descriptions. Additionally, the consistently strong performance of our **CmOS** framework across subjects and grade levels further demonstrates its robust generalization ability.

Human Evaluation Similarly, we invited three qualified annotators to subjectively evaluate the visual options generated by different methods. Using a 5-point Likert scale, they rated each visual option across three important dimensions: (1) **Plausibility** (Luo et al., 2024), assessing coherence with the background content and question context; (2) **Distractibility** (Gierl et al., 2017), measuring the cognitive burden posed by distractor visual options; and (3) **Engagement** (Gierl et al., 2017), reflecting how much the visual options attract learners' attention. Guidelines are detailed in Appendix H.

Table 4 presents a summary of the human evaluation results. Generally, **CmOS** surpasses other T2I models in terms of plausibility and distractibility. This indicates that the visual options it generates are more semantically consistent and possess a higher level of misleadingness, sufficient to test human judgment. However, even though image templates and tuning was carried out to enhance engagement, the improvement is relatively limited. **CmOS** performs slightly worse than DALLE-3 but marginally better than STABLEDIFFUSION-XL. It is worth noting that the average scores across all methods are relatively low. Specifically, **CmOS** only achieves a score of 3.55 out of 5. This situation highlights that creating visual content with pedagogical effectiveness still poses a large challenge.

4.5 Ablation Study

We perform ablation studies to investigate the effects of the proposed approaches in terms of similar exemplar retrieval, optimal question-reason match, template and tuning, as presented in Figure 3, Table 1 and Table 3. There are several notable findings.

Finding1: Figure 3 illustrates that removing the retrieval of similar exemplars forces the model to rely solely on abstract judgment criteria when assessing whether input content can be converted into visual-option questions. This change causes accuracy to drop by 19.5%, suggesting that concrete reason exemplars contribute more effectively to

Method SSIM↑ / CLIP-T↑	Subject NAT	 SOC	 LAN	Modality TXT	 IMG	Grade G1-6	 G7-12	AVG
Flux.schnell	39.0/28.9	41.5/29.1	29.7/29.4	37.6/29.4	42.2/28.2	39.2/29.2	38.8/28.5	39.1/29.0
DALLE-3	40.2/30.1	43.8/29.3	30.6/29.0	38.4/30.2	43.1/29.0	42.4/29.9	39.5/29.3	40.2/29.7
StableDiffusion-XL	40.6/27.3	43.9/28.5	31.4/27.4	39.4/28.4	43.5/27.1	41.2/28.1	40.8/27.5	40.7/27.9
Wanx2.1-plus	41.8/31.6	44.9/30.2	32.1/28.5	40.6/30.4	44.7/32.3	42.1/31.2	40.8/30.1	41.5/30.8
CmOS(Wanx2.1-turbo)	**59.0**/40.6	**61.2**/42.8	**49.7**/**37.6**	**58.3**/38.4	**62.1**/42.7	**59.7**/40.7	**59.1**/39.5	**59.5**/40.2
w / o Discriminator	58.2/29.3	59.4/30.9	49.1/25.6	54.8/26.9	60.3/31.1	57.9/29.3	56.8/27.7	57.5/28.8
w / o Reasoning	50.8/38.0	58.8/40.3	42.0/36.3	50.9/37.6	54.0/41.3	52.5/39.6	51.7/38.0	52.3/39.1
w / o Template	48.0/**41.5**	50.7/**44.6**	38.8/36.9	47.4/**40.1**	50.8/**43.3**	48.9/**41.6**	47.2/**40.3**	48.3/**41.1**
w / o Tuning	54.8/31.7	57.7/35.2	44.8/29.5	53.9/30.5	58.1/35.4	55.4/33.5	54.7/31.4	55.1/32.7

Table 3: Automatic evaluation results of visual option generation. ↑: higher is better.

Method	Plaus.↑	Distra.↑	Enga.↑	AVG↑
Flux-schnell	3.07	2.99	3.90	3.32
DELLE-3	3.13	2.86	**4.20**	3.41
StableDiffusion-XL	3.24	2.82	4.13	3.40
Wanx2.1-plus	3.14	2.75	4.11	3.33
CmOS(Wanx2.1-turbo)	**3.51**	**3.09**	4.17	**3.55**

Table 4: Human evaluation results of visual option generation. ↑: higher is better.

content discrimination than predefined standards.

Finding2: In question generation, as shown in Table 1, removing the discriminator leads to BLEU-4 and ROUGE-L drops of 2.6 and 5.1, reflecting weaker semantic alignment and greater inconsistency. This indicates that unconvertible inputs disrupt downstream processing, thereby diminishing coherence and overall quality. For visual option generation, as shown in Table 3, SSIM remains largely unchanged, whereas CLIP-T decreases by 11.4, the largest drop among ablations. These results indicate that unsuitable inputs hinder the T2I model's ability to align images with text.

Finding3: Table 1 clearly reveals that excluding the OQRM module significantly weakens question generation, with BLEU-4 plummeting by 33.1 and ROUGE-L dropping by 7.3. This sharp decline underscores OQRM's importance in identifying question-reason pairs with strong lexical and semantic fidelity to human-written references.

Finding4: As shown in Table 3, when the reason generator is removed and distractors are produced solely from the question and answer, both SSIM and CLIP-T scores decline to varying extents. Specifically, SSIM decreases by 7.2, while CLIP-T shows a slight drop of 1.1. This indicates that the inclusion of reasons primarily enhances the visual similarity among options, with comparatively smaller effects on image-text alignment.

Finding5: As shown in Table 3, removing the template results in an 11.2 drop in SSIM, while CLIP-T scores unexpectedly improve. This divergence suggests that templates enhance visual consistency across options but may impose constraints that hinder semantic alignment with textual descriptions. Further ablation reveals that disabling tuning by MLLM and T2I model consistently leads to notable declines in both SSIM (-4.4) and CLIP-T (-12.5), reinforcing the importance of iterative tuning strategies for improving intra-option visual similarity and cross-modal coherence.

4.6 Case Study

To qualitatively assess the three important modules in CmOS, Figures 4 and 5 present representative examples from ScienceQA, which include the context, an image, and an answer. Without computing Total Match Scores (TMS) across multiple candidate question-reason pairs to select the one with the highest alignment, the generator tends to generate the most probable question, which often leads to suboptimal performance in question generation.

To further examine the effect of the template and tuning modules on visual option generation, a comparative example are shown in Figure 5. When the template module is incorporated, the generated visual options exhibit improved semantic plausibility and consistency with object properties. In contrast, removing both the template and tuning modules results in outputs that deviate from commonsense expectations and display stylistic inconsistency. Additionally, the example illustrates that the tuning process enables iterative refinement: even if the initial visual option is inadequate, tuning can adjust and improve it toward the desired quality. These findings suggest that the template module ensures baseline plausibility, while the tuning supports optimization, and their combination contributes to overall improvements in visual option quality.

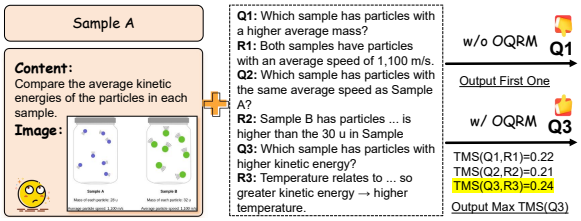


Figure 4: Case in terms of the OQRM module.

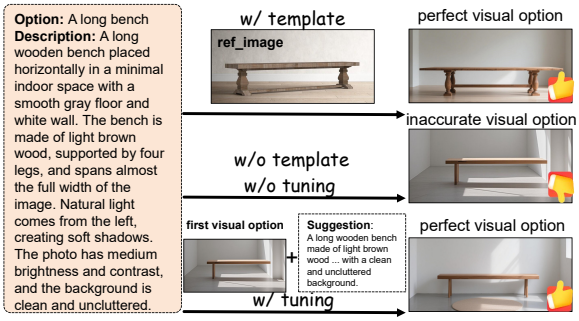


Figure 5: Case regarding the Template and Tuning.

5 Conclusion

In this paper, we present a novel framework called **C**ross-**m**odal **O**ptions **S**ynthesis (**CmOS**), which combines retrieved similar exemplars and Multimodal Chain-of-Thought (MCoT) reasoning to generate educational multiple-choice questions and visual options from multimodal input. Specifically, we utilize MLLMs to encode multimodal contexts and incorporate them into a three-stage Multimodal-CoT framework, namely content discrimination, question generation, and visual option generation. Meanwhile, we introduce a similar exemplar retrieval module to guide the discrimination. What's more, we use OQRM module to select optimal question and reasoning progress. Finally, we leverage template-based and slight tuning strategies to generate educational visual options. Our experimental results on three test sets demonstrate that **CmOS** outperforms all existing methods and models, achieving new state-of-the-art performance. More importantly, our study highlights the potential of visual-option-based multiple-choice question generation to enhance multimodal teaching resources, support personalized learning, and foster deeper understanding in educational settings.

Limitations

Despite its promising performance, our proposed framework still has two limitations.

Exemplar Resource On one hand, similar to CoE framework, our similar exemplar-based strategy for enhancing content discrimination accuracy has an inherent limitation: the dependency on a fixed pool of exemplars. When the input content falls outside the distribution of datasets like ScienceQA, the retrieved exemplars may exhibit low semantic similarity, leading to degraded discrimination performance. To mitigate this, future work could explore adaptive retrieval mechanisms or augment the exemplar pool with more diverse, domain-general instances, possibly using synthetic data or continual learning techniques to improve generalization beyond the source domain.

Visual Option Quality On the other hand, despite improvements in text-image alignment through image templates and lightweight tuning, the generated visual options still suffer from limited visual detail, occasional content hallucinations, and inconsistent style. Human evaluation (Table 4) shows relatively low scores in plausibility and distractibility, indicating that some generated visual options are semantically weak, visually indistinct, or pedagogically uninformative, thereby limiting their effectiveness in supporting deep cognitive processing or engaging prior knowledge. These issues stem from the use of general-purpose image generation models that are not optimized for educational applications, often resulting in irrelevant, ambiguous, or stylistically incoherent outputs. Future work could fine-tune generation models on curated educational datasets and apply task-specific constraints or multimodal alignment objectives to improve clarity, visual contrast, semantic accuracy, and pedagogical value.

Ethics Statement

We comply with institutional ethical guidelines throughout this study. No private or non-public data was used. For human annotation (Sections 4.3 and 4.4), six annotators were recruited from the schools of education at local universities via public advertisements, with clear disclosure of compensation terms. All annotators were senior undergraduate or graduate students in education-related programs who participated on a part-time basis. Each annotator was compensated at a rate of 55 CNY per hour, which exceeds the local minimum hourly wage for part-time employment in 2025 (23.5 CNY/hour). The annotation process did not involve any personally sensitive information.

References

Naveed Afzal and Ruslan Mitkov. 2014. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing*, 18:1269–1281.

Naveed Afzal, Ruslan Mitkov, and Atefeh Farzindar. 2011. Unsupervised relation extraction using dependency trees for automatic generation of multiple-choice questions. In *Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada, May 25-27, 2011. Proceedings 24*, pages 32–43. Springer.

Renlong Ai, Sebastian Krause, Walter Kasper, Feiyu Xu, and Hans Uszkoreit. 2015. Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 26–33.

Somaiya Al Shuraiqi, Abdulrahman Aal Abdulsalam, Ken Masters, Hamza Zidoum, and Adhari AlZaabi. 2024. Automatic generation of medical case-based multiple-choice questions (mcqs): a review of methodologies, applications, evaluation, and future directions. *Big Data and Cognitive Computing*, 8(10):139.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826.

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education*, pages 327–339. Springer.

Andrew C Butler. 2018. Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, 7(3):323–331.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. *arXiv preprint arXiv:2107.00152*.

Rudeema Chughtai, Farooque Azam, Muhammad Waseem Anwar, Wasi Haider But, and Muhammad Umar Farooq. 2022. A lecture centric automated distractor generation for post-graduate software engineering courses. In *2022 International Conference on Frontiers of Information Technology (FIT)*, pages 100–105. IEEE.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Barbara Gross Davis. 2009. *Tools for teaching*. John Wiley & Sons.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. *Preprint*, arXiv:2103.10360.

Fares Fawzi, Sarang Balan, Mutlu Cukurova, Emine Yilmaz, and Sahan Bulathwela. 2024. Towards human-like educational question generation with small language models. In *International Conference on Artificial Intelligence in Education*, pages 295–303. Springer.

Yifan Gao, Piji Li, Irwin King, and Michael R Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. *arXiv preprint arXiv:1906.06893*.

Mark J Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of educational research*, 87(6):1082–1116.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. Chaincqg: Flow-aware conversational question generation. *arXiv preprint arXiv:2102.02864*.

Michael Heilman. 2011. *Automatic factual question generation from text*. Carnegie Mellon University.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.

Inthrani Raja Indran, Priya Paranthaman, Neelima Gupta, and Nurulhuda Mustafa. 2024. Twelve tips to leverage ai for efficient and effective medical question generation: a guide for educators using chat gpt. *Medical Teacher*, 46(8):1021–1026.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. 2024. Dcot: Dual chain-of-thought prompting for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Minchi C Kim and Michael J Hannafin. 2011. Scaffolding problem solving in technology-enhanced learning environments (teles): Bridging research and theory with practice. *Computers & Education*, 56(2):403–417.

Myo-Kyoung Kim, Rajul A Patel, James A Uchizono, and Lynn Beck. 2012. Incorporation of bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *American journal of pharmaceutical education*, 76(6):114.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Devang Kulshreshtha, Muhammad Shayan, Robert Belfer, Siva Reddy, Iulian Vlad Serban, and Ekaterina Kochmar. 2022. Few-shot question generation for personalized feedback in intelligent tutoring systems. In *PAIS 2022*, pages 17–30. IOS Press.

Archana Praveen Kumar, Ashalatha Nayak, Manjula Shenoy, Shashank Goyal, and 1 others. 2023. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225:120022.

Girish Kumar, Rafael E Banchs, and Luis Fernando D'Haro. 2015. Automatic fill-the-blank question generator for student self-assessment. In *2015 IEEE Frontiers in Education Conference (FIE)*, pages 1–3. IEEE.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Selvia Ferdiana Kusuma, Daniel Oranova Siahaan, and Chastine Fatichah. 2022. Automatic question generation with various difficulty levels based on knowledge ontology using a query template. *Knowledge-Based Systems*, 249:108906.

Selvia Ferdiana Kusuma and 1 others. 2018. Automatic question generation for 5w-1h open domain of indonesian questions by using syntactical template-based features from academic textbooks. *Journal of Theoretical and Applied Information Technology (JATIT)*, 96(12):3908–3923.

Salima Lamsiyah, Abdelkader El Mahdaouy, Aria Nourbakhsh, and Christoph Schommer. 2024. Fine-tuning a large language model with reinforcement learning for educational question generation. In *International Conference on Artificial Intelligence in Education*, pages 424–438. Springer.

Guillaume Le Berre, Christophe Cerisara, Philippe Langlais, and Guy Lapalme. 2022. Unsupervised multiple-choice question generation for out-of-domain q&a fine-tuning. In *60th annual meeting of the association for computational linguistics*, volume 2, pages 732–738. Association for Computational Linguistics.

Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. 2024. UNIMO-G: Unified image generation through multimodal conditional diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6173–6188, Bangkok, Thailand. Association for Computational Linguistics.

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

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.

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. ACL.

Richard E Mayer. 2005. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 41(1):31–48.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Philip Newton and Maira Xiromeriti. 2024. Chatgpt performance on multiple choice question examinations in higher education. a pragmatic scoping review. *Assessment & Evaluation in Higher Education*, 49(6):781–798.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Ricardo Rodriguez-Torrealba, Eva Garcia-Lopez, and Antonio Garcia-Cabot. 2022. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258.

Umme Sara, Morium Akter, and Mohammad Shorif Uddin. 2019. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. 2024. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*.

Rahul Singhal and Martin Henz. 2014. Automated generation of region based geometric questions. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 838–845. IEEE.

Rahul Singhal, Martin Henz, Shubham Goyal, and FE Browder. 2015. A framework for automated generation of questions across formal domains. In *the 17th international conference on artificial intelligence in education*, pages 776–780.

Linfeng Song and Lin Zhao. 2016. Question generation from a knowledge base with web exploration. *arXiv preprint arXiv:1610.03807*.

Will Thalheimer. 2014. Learning benefits of questions. Technical report, Work-Learning Research. Version 2.0.

Toyin Tofade, Jamie Elsner, and Stuart T Haines. 2013. Best practice strategies for effective use of questions as a teaching tool. *American journal of pharmaceutical education*, 77(7):155.

Youness Touissi, Ghita Hjiej, Abderrazak Hajjioui, Azeddine Ibrahimi, and Maryam Fourtassi. 2022. Does developing multiple-choice questions improve medical students' learning? a systematic review. *Medical Education Online*, 27(1):2005505.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 119–129.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zichao Wang and Richard Baraniuk. 2023. Multiqg-ti: Towards question generation from multi-modal sources. *arXiv preprint arXiv:2307.04643*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, and 1 others. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.

Gautam Yadav, Ying-Jui Tseng, and Xiaolin Ni. 2023. Contextualizing problems to student interests at scale in intelligent tutoring system using large language models. *arXiv preprint arXiv:2306.00190*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Min-Hsuan Yeh, Vicent Chen, Ting-Hao'Kenneth' Haung, and Lun-Wei Ku. 2022. Multi-vqg: Generating engaging questions for multiple images. *arXiv preprint arXiv:2211.07441*.

Nikki L Bibler Zaidi, Karri L Grob, Seetha M Monrad, Joshua B Kurtz, Andrew Tai, Asra Z Ahmed, Larry D Gruppen, and Sally A Santen. 2018. Pushing critical thinking skills with multiple-choice questions: does bloom's taxonomy work? *Academic Medicine*, 93(6):856–859.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*.

A Related Works

Educational question is an indispensable component of instructional resources, serving multiple functions including assessment, guidance, feedback, and the promotion of active learning (Tofade et al., 2013; Thalheimer, 2014). However, manually authoring educational questions is a complex and resource-intensive task that requires professional training, domain knowledge, and instructional experience (Davis, 2009; Kim et al., 2012). To address the high cost and inefficiency associated with manual question generation, Automatic Question Generation (AQG) technologies have emerged

as a promising solution (Brown et al., 2005), and have been widely applied in dialogue systems (Gao et al., 2019; Gu et al., 2021; Bulathwela et al., 2023) and intelligent tutoring systems (Kulshreshtha et al., 2022; Xu et al., 2022; Yadav et al., 2023), becoming a prominent research focus within the field of Artificial Intelligence in Education to support personalized learning (Bulathwela et al., 2023; Fawzi et al., 2024; Lamsiyah et al., 2024).

B Dataset Details

Figure 8 shows four datasets distribution across 3 subjects (NAT, SOC, LAN), 2 modalities (IMG, TXT), and 2 grade ranges (G1-6, G7-12). Question types: NAT = natural science, SOC = social science, LAN = language science, TXT = only containing text, IMG = containing image, G1-6 = grades 1-6, G7-12 = grades 7-12. For $\mathcal{D}_O$, NAT has $2,252$, SOC $1,100$, and LAN 889; IMG has $2,224$ and TXT $2,017$; G1-6 has $2,723$ and G7-12 has $1,518$. For $\mathcal{D}_E$, NAT has 257, SOC 128, and LAN 97; IMG has 228 and TXT 254; G1-6 has 309 and G7-12 has 173. For $\mathcal{D}_C$, NAT has $1,990$, SOC 969, and LAN 787; IMG has $1,795$ and TXT $1,964$; G1-6 has $2,723$ and G7-12 has $1,036$. For $\mathcal{D}_Q$, NAT has 576, SOC 236; IMG has 571 and TXT 241; G1-6 has 568 and G7-12 has 244.

C Implementation Details

To ensure fair and reproducible evaluation of visual option generation baselines, we report the detailed settings of all off-the-shelf generative models used in our experiments. As part of our pipeline, we employ Qwen2.5-VL-7B-Instruct (Bai et al., 2023) for content discrimination, question generation, and the production of visual option descriptions. Moreover, we uniformly set the decoding parameters to `top-p` = 0.8 and `temperature` = 0.7. To generate corresponding images, we use Wanx2.1-turbo as our main image synthesis backbone. Additionally, we utilize the robust Contrastive Language–Image Pretraining model FARE (Schlarmann et al., 2024) for retrieval and evaluation. Below, we present the configurations of these models and the prompting details for MCoT.

Models and Platforms

(1) FLUX-SCHNELL

- `Access`: Alibaba Bailian platform[1]
- `guidance_scale`: 3.5
- `num_inference_steps`: 50
- `image_size`: 1024×1024
- `seed`: 42

(2) DALLE-3

- `Access`: OpenAI official API[2]
- `model`: dall-e-3
- `num_inference_steps`: N/A
- `image_size`: 1024×1024
- `seed`: N/A

(3) STABLEDIFFUSION-XL

- `Access`: Alibaba Bailian platform
- `guidance_scale`: 10
- `num_inference_steps`: 50
- `image_size`: 1024×1024
- `seed`: N/A

(4) WANX2.1-PLUS

- `Access`: Alibaba Bailian platform
- `guidance_scale`: N/A
- `num_inference_steps`: N/A
- `image_size`: 1024×1024
- `seed`: 42
- `negative_prompt` = N/A

(5) WANX2.1-TURBO

- `Access`: Alibaba Bailian platform
- `guidance_scale`: N/A
- `num_inference_steps`: N/A
- `image_size`: 1024×1024
- `seed`: 42
- `negative_prompt` = N/A

[1] https://bailian.console.aliyun.com

[2] https://platform.openai.com

Prompts

The first red-shaded panel presents the prompt used to guide QWEN2.5-VL-72B for exemplar construction. The second blue-highlighted section shows the prompts employed to instruct **CmOS**, VL-T5, MultiQG-TI, MultiModal-CoT, and CoE in content discrimination, question and reason generation, textual option generation, visual description generation, and visual option optimization. Note that the last three tasks are exclusive to **CmOS**. The third yellow-marked panel provides the prompts used to instruct CHATGPT for content discrimination and question generation under both zero-shot and few-shot settings (CD = Content Discrimination, QG = Question Generation).

Box 1: Prompt input for exemplar construction

Context: ...
Image: ...
Answer: ...
Please analyze whether the above content can be transformed into a multiple-choice question with images as options, based on the following three dimensions:
(1) Whether the answer itself is suitable for visual transformation;
(2) Whether the key entities in the context are suitable for visual transformation;
(3) Which form of transformation (if any) provides greater educational value, or whether neither form is suitable or meaningful in an educational context.

Reasoning: ...
Convertible: ...

Box 2: Prompt input for CmOS and baselines

Content Discrimination Prompt Input
Context: ...
Image: ...
Answer: ...
Refer to the following exemplar to determine whether the original content can be converted into a question format with visual options and give the reason.
Exemplar
Context: ...
Image: ...
Answer: ...
Reason: ...
Judgment: ...

Question Generation Prompt Input
Context: ...
Image: ...
Answer: ...
Judgment: ...
Refer to the following three exemplars. Generate a new question suitable for visual options basing on the original content, provide the corresponding answer and reason.
Exemplar1
Context: ...
Image: ...
Answer: ...
Question: ...
Reason: ...
Exemplar2
Exemplar3

Option Generation Prompt Input
Context: ...
Question: ...
Answer: ...
Reason: ...
Refer to the following exemplar content to generate multiple options and description that are related to the answer and have a certain degree of interference.
Exemplar
Context: ...
Question: ...
Answer: ...
Reason: ...
Options: (a) ...; (b) ...; (c) ...; (d) ...

Visual Option Generation Prompt Input
Option: a picture of ×××.
Description: Color; Green; Shape.
Please refer to the following reference image, Generate a corresponding image according to the visual description of this option.

Optimization Prompt Input
Visual Option: ...
Description: Color; Style; Shape.
Please calculate the similarity between the given visual option and the descriptive text, and provide optimization suggestions.

Reference Image: ...

Box 3: Zero-shot and few-shot settings for ChatGPT

0 shot CD and QG
Context: ...
Answer: ...
Image: ...
Determine whether this content can be converted into a visual option question. If it is convertible, generate a question based on the corresponding content.

1 shot CD and QG
Context: ...
Answer: ...
Image: ...
Refer to the exemplar, determine whether this content can be converted into a visual option question. If it is convertible, generate a question based on the corresponding content.
Exemplar: ...

3 shot CD and QG
Context: ...
Answer: ...
Image: ...
Refer to these 3 exemplars, determine whether this content can be converted into a visual option question. If it is convertible, generate a question based on the corresponding content.
Exemplar 1: ...
Exemplar 2: ...
Exemplar 3: ...

D METEOR

Unlike BLEU-4 and ROUGE-L, which focus on lexical overlap, METEOR (MTR) also captures semantic and content-level similarities. Table 7 shows that **CmOS** consistently outperforms SOTA methods in question generation under the METEOR metric. While its score in the language sciences (LAN) is lower than in natural (NAT) and social sciences (SOC), it still significantly exceeds other methods. This suggests that **CmOS** generates semantically aligned questions, aided by its multi-question filtering strategy.

Moreover, MultiQG-TI and Multimodal-CoT that are fine-tuned with CoT prompting, achieve better performance than VL-T5. Although VL-T5 benefits from stronger visual understanding, it lags behind in semantic coherence during question generation. In contrast, CoE leverages Chain of Exemplar reasoning to further enhance its generation quality, outperforming both MultiQG-TI and Multimodal-CoT. Additionally, the table includes results for ChatGPT under zero-shot and few-shot settings. While ChatGPT benefits from increased contextual exemplars and achieves METEOR scores that approach those of some baselines, it still falls considerably short of **CmOS** in multimodal question generation, highlighting its limitations in complex reasoning and visual-semantic integration.

Similarly, we evaluated the performance of **CmOS** after removing the Optimal Question-Reason Match (OQRM) module. A noticeable performance drop was observed, with the average METEOR score decreasing by 17.3 points. This result highlights the importance of OQRM in enhancing the semantic alignment between generated questions and questions authored by humans.

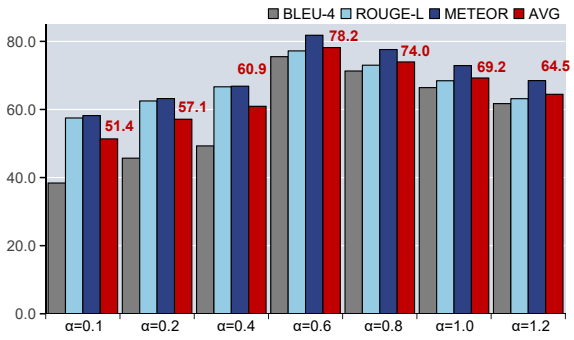


Figure 6: The overall performance of question generation with varying α.

E Hyperparameter Analysis

Question Generation

To determine the optimal hyperparameter α for selecting the best question-reason pair, we systematically examined the BLEU and ROUGE-L scores across varying values of α from 0.1 to 1.2 in steps of 0.1 or 0.2. What's more, we sampled 300 instances from D_c where the value of "convertible" is true, ensuring no overlap with the dataset D_Q. As shown in Figure 6, both scores exhibit a trend of first increasing and then decreasing. Notably, BLEU-4, ROUGE-L and METEOR reach their peak values when $\alpha = 0.6$. The average of the

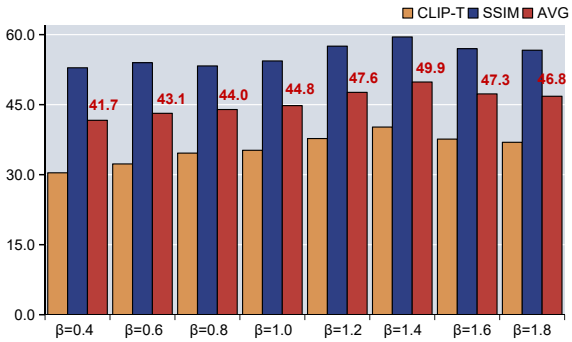


Figure 7: The overall performance of question generation with varying β.

two scores also reaches its highest value of 78.2 at this point. Therefore, we select $\alpha = 0.6$ as the optimal value.

Visual Option Generation

To determine the optimal hyperparameter β for balancing the influence of the image itself and its caption during template retrieval, we systematically evaluated the changes in structural similarity (SSIM) and text-image similarity (CLIP-T) scores for the generated visual options across different values of β. As shown in the Figure 7, β was gradually increased from 0.4 to 1.8 in increments of 0.2. The results indicate that both SSIM and CLIP-T scores exhibit a trend of initially increasing and then decreasing as β increases. Specifically, when $\beta = 1.4$, the SSIM score reaches its peak at 59.3, and the CLIP-T score also achieves its highest value of 40.4. Therefore, we select $\beta = 1.4$ as the optimal value within the range of our experimental settings for the subsequent visual option generation task.

F Analysis of Question Diversity

We adopt Distinct-n scores to evaluate the diversity of questions generated. Specifically, this metric calculates the number of unique n-grams at the corpus level, where higher values indicate greater diversity. We consider values of n ranging from 1 to 4. As shown in the table 5, overall performance improves with increasing n. Both CHATGPT's 0-shot and few-shot settings exhibit relatively high question diversity. Aside from CHATGPT, **CmOS** only falls slightly behind CoE on Distinct-1, while it surpasses the baseline methods on other three metrics. When compared to the Groundtruth, we find that both CoE and **CmOS** achieve scores close to human-generated questions, suggesting that these two methods better approximate the style and distribution of human-authored question diversity.

Method	Distinct-1	Distinct-2	Distinct-3	Distinct-4
0shot	19.68	38.32	49.89	58.81
1shot	18.46	34.16	44.49	53.32
3shot	17.21	30.80	39.95	47.93
VL-T5	12.11	23.95	30.93	37.74
MultiQG-TI	13.39	24.54	31.63	38.02
Multimodal-CoT	15.92	23.31	36.20	40.47
CoE	17.20	29.47	38.18	45.74
CmOS	16.85	34.65	40.72	47.22
Grondtruth	17.41	31.56	40.46	47.97

Table 5: Distinct-n results of different methods.

G Analysis of Different Base Models

To analyse the generality of **CmOS**, we conduct an experiment to utilize other base models in place of QWEN2.5-VL-7B-INSTRUCT as the backbone for question and visual option generation, including LLAMA3.2-11B-VISION (Grattafiori et al., 2024), LLAVA (Lu et al., 2022), INSTRUCTBLIP (Dai et al., 2023), MPLUG-OWL (Ye et al., 2024), and VISUALGLM-6B (Du et al., 2022). Note that we employ the same prompt for all base models to ensure fairness in the comparison. As summarized in Table 9, QWEN2.5-VL-7B-INSTRUCT outperforms all the rest base models, showcasing its high applicability and suitability in our framework. Generally, while there are slight difference in performance among the 6 base models, they consistently demonstrate superior performance in both question and visual option generation, which further confirms the effectiveness and versatility of our **CmOS** framework.

Method	QG			VOG	
	B-4 ↑	R-L ↑	MTR ↑	SSIM ↑	CLIP-T ↑
Qwen-VL	**75.50**	**77.20**	**81.80**	**59.5**	**40.2**
LlaMA	74.74	76.51	80.50	57.5	39.0
LLaVA	73.62	75.13	80.69	57.7	38.3
InstructBLIP	72.10	75.61	76.41	55.9	38.7
mPLUG-Owl	71.23	72.66	76.10	56.4	38.6
VisualGLM	58.63	60.61	64.06	47.3	33.1

Table 6: Detailed performance of **CmOS** with different base models. (MTR=METEOR) ↑: higher is better.

H Guideline of Human Evaluation

Table H presents the evaluation form used to guide human annotators, consisting of three sections: case details, question evaluation, and visual option evaluation.

Method MTR↑	Subject NAT	SOC	LAN	Modality TXT	IMG	Grade G1-6	G7-12	AVG
0-shot	32.8	29.2	24.0	29.2	32.1	31.2	30.4	30.8
1-shot	46.9	43.1	33.7	45.3	46.2	44.2	48.7	45.5
3-shot	51.2	49.4	42.3	51.2	50.3	51.0	50.4	50.8
VL-T5	54.4	48.1	48.3	54.5	51.2	52.8	54.5	53.1
MultiQG-TI	59.4	48.7	50.5	57.2	54.5	56.8	55.3	56.0
Multimodal-CoT	65.1	57.9	56.4	62.1	59.1	60.7	62.1	61.4
CoE	72.3	68.3	63.5	70.1	67.8	68.5	70.0	69.1
CmOS	**80.9**	**84.1**	**72.6**	**81.2**	**82.3**	**83.3**	**78.4**	**81.8**
w/o OQRM	62.4	64.9	45.6	62.4	64.6	65.4	60.3	63.5
w/o Discriminator	71.7	74.2	63.6	71.3	72.5	73.9	69.7	72.2

Table 7: Additional automatic evaluation results of question generation. (MTR=METEOR) ↑: higher is better.

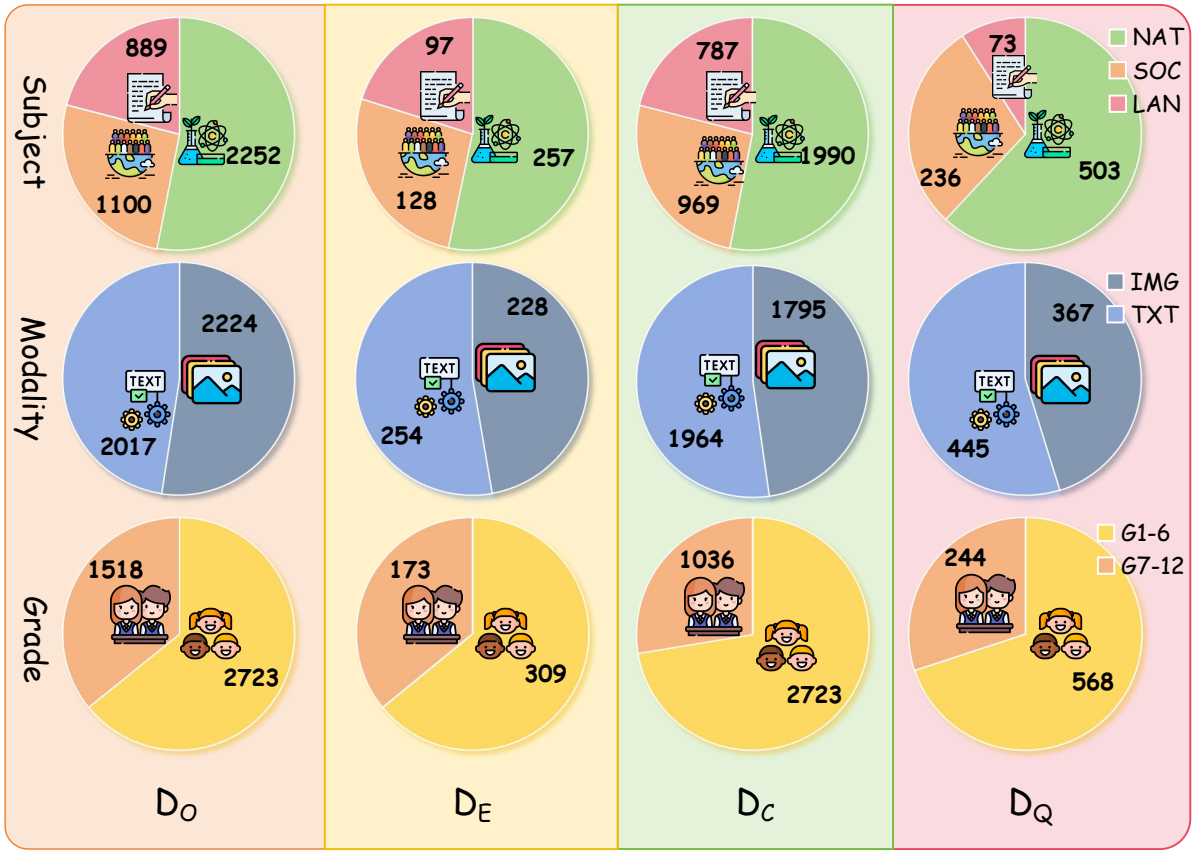


Figure 8: Dataset statistics of ScienceQA test benchmark and our test sets. Question types: NAT = natural science, SOC = social science, LAN = language science, TXT = containing text context, IMG = containing image context, G1-6 = grades 1-6, G7-12 = grades 7-12.

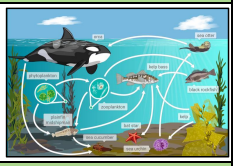
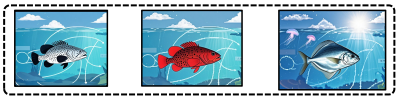
<table>
<tr><th colspan="2">Guideline of Generation Quality Evaluation</th></tr>
<tr><td colspan="2">This study aims to evaluate the quality of the question and visual options. Each case provides a context, image, answer and groundtruth. You need to assess the generated question and visual options from the following aspects.</td></tr>
<tr><th colspan="2">Case</th></tr>
<tr><td colspan="2">Context: Below is a food web from an ocean ecosystem in Monterey Bay, off the coast of California.The arrows in a food web represent how matter moves between organisms in an ecosystem.
Answer: black rockfish.
Groundtruth Question: Which of these organisms contains matter that was once part of the phytoplankton?
  </td></tr>
<tr><th colspan="2">Question Evaluation</th></tr>
<tr><td colspan="2">Fluency: whether the questions are natural and easy to read and understand for students of corresponding grade.</td></tr>
<tr><td>Options</td><td>1. Very disfluent 2.Disfluent 3. Neutral 4. Fluent 5. Very fluent</td></tr>
<tr><td>Examples</td><td>1. "Which of the following organisms is the primary consumer in this food web?" — This question is grammatically correct, natural-sounding, and easy to understand.
2. "Which of organism is the primary consumer in this food web is?" — This question contains grammatical errors and awkward phrasing, making it moderately readable.</td></tr>
<tr><td colspan="2">Grammaticality: whether the question is syntactically correct and follows standard grammar rules.</td></tr>
<tr><td>Options</td><td>1. Very ungrammatical 2.Ungrammatical 3. Neutral 4. Grammatical 5. Very grammatical</td></tr>
<tr><td>Examples</td><td>1. "Which of the following organisms is the primary consumer in this food web?" — Very grammatical
2. "Which of organism is the primary consumer in this food web is?" — Somewhat ungrammatical</td></tr>
<tr><td colspan="2">Complexity: whether the question poses an appropriate level of cognitive challenge suitable for the students.</td></tr>
<tr><td>Options</td><td>1. Very simple 2. Simple 3. Neutral 4. Complex 5. Very complex</td></tr>
<tr><td>Examples</td><td>1. "Which organism preys on golden algae in this food web?" is fairly thought-provoking and necessitates some reasoning effort to answer.
2. "What is the largest fish in this picture?" shows completely unchallenging to answer the question.</td></tr>
<tr><td colspan="2">Relevance: how well the question aligns with and reflects the background content or topic it is intended to assess.</td></tr>
<tr><td>Options</td><td>1. Very irrelevant 2. Irrelevant 3. Neutral 4. Relevant 5. Very relevant</td></tr>
<tr><td>Examples</td><td>1. "Which organism preys on golden algae in this food web?" — Refers to specific relationships in the food web, though not directly aligned with the given answer.
2. "Which of these organisms contains matter that was once part of the phytoplankton?" — Directly connected to the movement of matter in the food web and aligned with both context and answer</td></tr>
<tr><th colspan="2">Visual Option Evaluation</th></tr>
<tr><td colspan="2">Plausibility: whether the option is coherent and consistent with the background content and question context.</td></tr>
<tr><td>Options</td><td>1. Very implausible 2. Implausible 3. Neutral 4. Plausible 5. Very plausible</td></tr>
<tr><td>Examples</td><td>  The image shows a black rockfish, matching the food chain and the answer — highly relevant.
  The image shows a seagull, which is not part of the food chain in this question, so the relevance is low.</td></tr>
<tr><td colspan="2">Distractibility: whether the visual options pose a meaningful cognitive challenge and potential confusion.</td></tr>
<tr><td>Options</td><td>1. Very simple 2. Simple 3. Neutral 4. Distracting 5. Very distracting</td></tr>
<tr><td>Examples</td><td> 
The three images have similar backgrounds and object outlines, showing high distractibility.
 
The three images differ in background and outlines, making them easy to tell apart with low distractibility.</td></tr>
<tr><td colspan="2">Engagement: how much the visual options are appealing, interesting, and likely to capture the attention of learners.</td></tr>
<tr><td>Options</td><td>1. Very unengaging 2. Unengaging 3. Neutral 4. Engaging 5. Very engaging</td></tr>
<tr><td>Examples</td><td> 
They are colorful, have attractive backgrounds, and clear objects, making them quite engaging.
 
The three images have missing backgrounds and are black-and-white, resulting in very low engagement.</td></tr>
</table>

Figure 9: Guideline of human evaluation for question and visual option generation quality.