Context Selection and Rewriting for Video-based Educational Question Generation

Mengxia Yu, Bang Nguyen, Olivia Zino, Meng Jiang

University of Notre Dame, Notre Dame, USA
{myu2,bnguyen5,ozino,mjiang2}@nd.edu

Abstract

Educational question generation (EQG) is a crucial component of intelligent educational systems, significantly aiding self-assessment, active learning, and personalized education. While EQG systems have emerged, existing datasets typically rely on predefined, carefully edited texts, failing to represent real-world classroom content, including lecture speech with a set of complementary slides. To bridge this gap, we collect a dataset of educational questions based on videos from real-world lectures. On this realistic dataset, we find that current methods for EQG struggle to accurately generate questions from educational videos, particularly in aligning with specific timestamps and target answers. Common challenges include selecting informative contexts from extensive transcripts and ensuring generated questions meaningfully incorporate the target answer. To address the challenges, we introduce a novel framework utilizing large language models (LLMs) for dynamically selecting and rewriting contexts based on target timestamps and answers in lecture videos. First, our framework selects contexts from both lecture transcripts and video keyframes based on answer relevance and temporal proximity. Then, we integrate the contexts selected from both modalities and rewrite them into answer-containing knowledge statements, to enhance the logical connection between the contexts and the desired answer. Quantitative evaluation and human evaluation show that our approach improves the quality and relevance of the generated questions.

Code and Data — https://github.com/mengxiayu/COSER

Introduction

In-class quizzes are often used to engage students and assess their understanding of lecture content. The quizzes typically feature multiple-choice questions (MCQs) due to their objectivity, efficiency, and scalability. The growing prevalence of online learning has further increased demand for high-quality MCQs. However, manually producing such questions is a resource-intensive process that requires substantial time, expertise, and effort. Specialists must be involved to ensure that the questions are accurate, relevant, of appropriate difficulty, and cover the target spectrum of knowledge and skills (Mucciaccia et al. 2025). Automatic Question Generation (QG) systems offer promising solutions to alleviate this burden on educators. While existing QG techniques have shown progress in generating questions from structured textual content, extending these capabilities to lecture videos presents unique challenges. Unlike well-organized written materials, lecture videos contain unstructured, lengthy, and often noisy speech transcripts. Generating high-quality MCQs from such content requires precise alignment with target answers and specific timestamps in the lecture. Key challenges include: (1) selecting the most informative and relevant context from lengthy lecture transcripts and complementary slides; (2) ensuring generated questions appropriately use the context and meaningfully incorporate the given answer.

Existing educational question generation (EQG) methods attempt to identify relevant contexts on which the generated questions are grounded (Ghanem et al. 2022). Simplistic extraction heuristics result in either overly broad, irrelevant contexts, or overly narrow contexts that lack sufficient information, both of which negatively affect the quality of the generated questions (Noorbakhsh et al. 2025). Recent studies on long-context EQG (Wang et al. 2023; Ding, Hong, and Yao 2024) performed context selection with supervised training, which relied on annotated training data and supervised finetuning. Critically, existing EQG datasets (Chen et al. 2018; Gong, Pan, and Hu 2022; Hadifar et al. 2023a; Xu et al. 2022) are mostly built from high-quality, predefined contexts (e.g. manually labeled segments from textbooks or carefully edited academic resources), or manually corrected transcripts of lecture videos. This setup does not realistically reflect typical educational settings. First, lecture speech in real-world classroom includes filler words, disfluencies, and auto-transcription errors. Second, the information conveyed in lecture speech is not as concise or structured as in textbooks. Third, the language used in the lecture could be informal, designed to elaborate on the formal knowledge provided in a set of slides. The mismatch between datasets based on idealized texts and real lecture speech poses significant challenges for the practical development and evaluation of EQG technology. To address this gap, we construct a dataset for video-based EQG that reflects a more realistic setting. Our dataset includes audio recordings of live lectures in college classrooms and the associated video recordings of their screens. We collect multiple-choice quiz questions by having educators watch the videos, pause them to create questions as

desired, and record the associated timestamps. This dataset serves as a testbed for video-based EQG.

Our new dataset reveals a clear need for approaches that construct concise and relevant contexts from lecture videos for EQG. Large language models (LLMs) demonstrate the potential to address this challenge through their strong language modeling and long-context capabilities. That said, this paper introduces a novel LLM-based EQG framework specifically designed to: (1) Dynamically select context segments given a specific timestamp and guided by a desired answer, (2) Rewrite context segments to ensure clarity, conciseness, and explicitly incorporate the answer text, and (3) Integrate multi-modal information from both textual data (e.g., audio-transcribed segments) and visual information (e.g., video frames of slides). By effectively selecting and rewriting context, the framework improves the quality, specificity, and educational alignment of generated question stems.

In summary, this paper makes three key contributions:

- A new dataset supporting in-class video-based EQG, which consists of lecture recordings and educator created timestamp-based MCQs.
- A novel framework named **COSER**, which integrates **Co**ntext **Se**lection and **R**ewriting explicitly tailored for answer and timestamp-aware EQG.
- A more reliable reference-based metric NLI score for question generation.

Related Work

Educational Question Generation with LLM

The rapid growth of large language models has shed light on automatic question generation to facilitate education. Wang et al. (2022) first explored EQG with ChatGPT under different basic settings, and found more in-context demonstrations for few-shot generation to be more effective. Lee et al. (2024) leveraged few-shot prompting with LLMs for EQG. MCQGen (Hang, Wei Tan, and Yu 2024) created answer-agnostic MCQs with chain-of-thought and self-refined strategies. Agrawal et al. (2024) constructed knowledge graphs from educational contexts, which were then used to design prompts for LLMs to generate questions for interactive learning. Maity, Deroy, and Sarkar (2025) investigated in-context learning strategies to generate questions better aligned with Bloom's revised taxonomy. However, none of the studies has explored improving noisy long context for video-based EQG.

Context Modeling for Question Generation

Regardless of their long-context capabilities, LLMs are prone to be distracted by irrelevant context (Pan et al. 2024; Wu et al. 2024; Shi et al. 2023). In answer-aware question generation, identifying answer-relevant contexts is important. Previous studies (Sun et al. 2018; Liu et al. 2019) predicted "clue" words based on their proximity to the answer, which performed well in generating questions from short contexts. Ding, Hong, and Yao (2024) trained a BART-based model to identify salient sentences as an auxiliary task for QG. Li and Zhang (2024) leveraged an LLM to identify answer-containing sentences as key points when generating answer plans for QG. Xia et al. (2023) trained a model to plan content at fine-grained level of phrases and coarse-grained level of sentences, thus obtaining answer-relevant summaries of the context. The approach showed the effectiveness of incorporating answer spans with contexts in multi-hop QG. Hadifar et al. (2023b) ranked document segments by relevance and diversity. Savaal (Noorbakhsh et al. 2025) was a multi-stage QG framework that retrieved and summarized information to improve conciseness and relevancy of contexts. Unlike previous approaches that require annotated data for fine-tuning, our study focuses on context selection and rewriting for zero-shot QG with LLMs.

		LLM-Frontier			DL-Intro		
		Mean	Min	Max	Mean	Min	Max
Transcript	# words	3,447.6	1,133	10,109	10,075.7	7,571	13,145
	# seg.	161.8	40	508	377.8	207	540
Keyframe	# frames	22.5	1	69	70.6	39	94
MCQ	# choices	4.0	4	4	3.9	3	4
Question	# words	12.9	3	37	14.4	3	63
Answer	# words	7.0	1	40	9.2	1	29
Distractor	# words	5.9	1	40	8.3	1	31

Table 1: Statistics of our AIRC dataset. **Question** stem is the target output of QG.

Dataset: AIRC

Overview

Existing EQG datasets such as LearningQ (Chen et al. 2018), KhanQ (Gong, Pan, and Hu 2022), and EduQG (Hadifar et al. 2023a) primarily deal with short contexts. FairytaleQA (Xu et al. 2022) includes long contexts from books, but not lecture transcripts.

We present **AIRC** (short for Artificial Intelligence in Real Classroom), a dataset for video-based EQG. The dataset consists of two college-level courses collected from real classrooms. One course, **LLM-Frontier**, is a graduate-level course about frontier research on Large Language Models. It consists of 27 research talk-style lectures, covering topics such as instruction tuning, pre-training, and reinforcement learning. The other course, **DL-Intro**, is an undergraduate-level course on deep learning. It consists of 8 one-hour lectures, covering various topics such as graph neural networks, computer vision, and language modeling. Descriptive statistics of our dataset are shown in Tab. 1. Compared to existing EQG datasets in Table 2, our dataset provides full recordings of live lectures in real classrooms, reflecting the practical challenges of EQG with long and noisy context.

Data Collection

We collected educational question data through a systematic process. First, we gathered authentic lecture recordings of the courses. These lectures were captured by the widely used platforms Zoom and YouTube, along with transcripts generated via their automatic captioning features. The video recordings primarily consisted of the instructors' screen-sharing sessions, typically displaying lecture slides.

Dataset	LearningQ			EduQG	FairytaleQA	AIRC (ours)	
	TED-Ed	Khan-Video	Khan-Doc			LLM-Frontier	DL-Intro
Source	e-Learning Platform			e-Library	e-Library	Live Lectures	
Materials	Videos	Videos	Textbooks	Textbooks	Storybooks	Videos	Videos
Level	K12	K12	K12	Undergrad	Children	Graduate	Undergrad
Avg. Words	847.6	1,370.8	1,306.6	12,641.5	2,313.4	3,447.6	10,075.7

Table 2: Compared to existing EQG datasets with educator created questions, our dataset is under realistic settings for in-class quiz question generation.

Next, annotators were instructed to review the lecture videos carefully. We recruited three volunteer annotators, one professor and two graduate students who have served as teaching assistants in related courses. They manually identified and documented timestamps associated with key instructional moments. For each timestamp, annotators created multiple-choice quiz questions designed to assess students' understanding of the associated context. To ensure the answer-aware QG task is optimizable and the collected questions can be used as references, we instructed the annotators to follow these guidelines: (1) ensure that the question is grounded on the lecture content, and (2) avoid generic answers, e.g., "yes","no", "none of the above", etc. In total, we have collected 352 MCQs for LLM-Frontier and 70 for DL-Intro.

Data Postprocessing

For transcripts that do not have punctuation, we run a punctuation restoration model (Guhr et al. 2021). For visual information, we first crop video frames to include only the lecturers' screens, then run a keyframe detection algorithm to extract slide images. Lastly, we associate keyframes with transcript segments based on the timestamps. To obtain textual descriptions of slide images, we prompt an LLM, i.e. GPT-4o-mini (Achiam et al. 2023), to describe each extracted keyframe.

Preliminary

Answer and Timestamp-Aware Educational Question Generation

We formally define the task of answer and timestamp-aware question generation (ATEQG) as follows: Given a lecture video V, consisting of both audio-transcribed speech and visual keyframes of the lecturer's screen, the objective is to generate a meaningful and contextually relevant question conditioned upon a specified timestamp t and a target answer A. The target answer is not necessarily a span from the transcript.

Specifically, the input lecture video comprises transcribed speech segments denoted by $S = [S_1, S_2, ..., S_N]$, where each segment $S_i \in S$ is a sentence. Additionally, the visual modality of the video is represented by a set of visual keyframes $F = [F_1, F_2, ..., F_M]$, each corresponding to visual content captured from the lecturer's screen. Crucially, these keyframes and transcript segments are temporally aligned, such that each keyframe F_j can be associated with one or multiple transcript segments S_i.

For any given timestamp $t \in [0, T]$, where T denotes the total duration of the lecture video, and t maps to a transcript segment S_{i_t}, where $i_t \in [1, N]$. The provided answer span A is represented as a sequence of tokens $A = [a_1, a_2, ..., a_k]$.

The task then is to generate a natural-language question $Q = [q_1, q_2, ..., q_l]$ that is semantically coherent and contextually grounded in the aligned transcript segment S_{i_t}, its associated keyframes, and the specified span A. Mathematically, this task can be framed as modeling the following conditional probability: $P(Q \mid \mathbf{S}, \mathbf{F}, t, A) = P([q_1, q_2, \ldots, q_l] \mid C, A)$, where the context C is an informative representation derived from the transcript segments and keyframes relevant to the question.

Context Construction for ATEQG

We decompose the ATEQG task into two main stages: context construction and question generation. Given a question generation function G_ϕ, which produces a question Q conditioned on a context C and answer span A, we have: $Q = G_\phi(C, A)$. In this formulation, the question generation parameters ϕ remain fixed. Thus, context construction becomes the sole focus of optimization. Formally, the objective is to construct a context that contains precisely the information needed for the fixed generator G_ϕ to successfully produce the target question Q based on that context and the given answer.

Method

In this section, we propose COSER, an LLM-based framework for ATEQG. As shown in Fig. 1, COSER involves two main stages: (1) **Context selection** extracts transcript segments and visual keyframes conditioned on the given timestamp and answer span, and (2) **Context rewriting** revises the extracted context such that it is better suited for QG.

Context Selection

First, we prompt an LLM to select relevant context for creating a quiz question. Given a lecture represented by audio transcript segments $\mathbf{S}$ and visual keyframes $\mathbf{F}$, timestamp t, answer span A, the context selection process aims to identify a continuous subset of segments $S_{a:b} = [S_a, S_{a+1}, ...S_b]$, where $1 \leq a \leq b \leq N$, or keyframes $F_{u:v} = [F_u, F_{u+1}, ..., F_v]$, where $1 \leq u \leq v \leq M$. The extracted context should meet the following requirements: (1) *integrity*: for transcript-based context, the output should be one or more unaltered segments directly extracted from

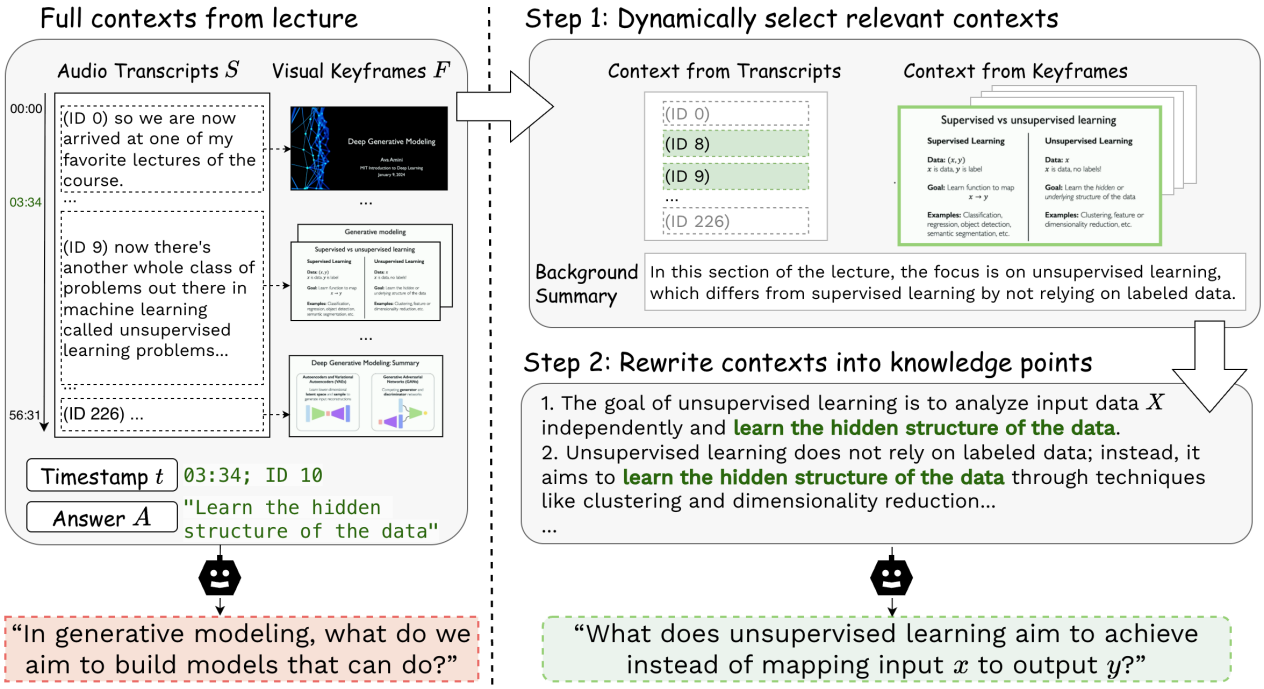

Figure 1: Our proposed framework COSER. Using all lecture content (left) as context results in generated questions that are too general and fail to incorporate keywords. COSER (right), which (1) dynamically selects relevant contexts from both transcripts and keyframes, and (2) integrates and rewrites them into answer-containing knowledge points, yields more specific and relevant question.

the lecture transcripts; (2) *relevancy*: it must be sufficient and concise, clearly providing all information required for creating the quiz question; (3) *temporal proximity*: the selected segments might be near the given timestamp.

Context Rewriting

Next, to incorporate the desired answer, we ask the model to rewrite the extracted context into several concise knowledge points or statements that contain the answer span in them.

Formally, it is expressed as: $Rewrite_\theta\left(S_{p:q}, F_{u:v}, t, A\right) = K_1, ..., K_r$. Each knowledge statement K_i is a sequence of tokens $K_i = [k_{i,1}, k_{i,2}], ..., k_{i,m_i}$ that satisfies three requirements: (1) *explicitness*: each statement K_i explicitly includes the provided answer span $A : A \subseteq K_i, \forall K_i \in \{K_1, ..., K_r\}$; (2) *atomity*: all statements should contain atomic knowledge and eliminate ambiguous references and indirect speech; (3) *multi-granularity*: the set of statements should cover knowledge of different levels of granularity, ranging from conceptual or high-level knowledge to specific details.

The rewriting stage plays a crucial role in bridging the gap between raw lecture transcripts and QG by converting implicit relationships into explicit statements and integrating the desired answers while maintaining natural language flow.

Multi-Modal Integration

To integrate information from both modalities, we first extract context from audio transcripts and visual keyframes separately. Then, we instruct the model to rewrite the transcript-based context into knowledge points, with context from keyframes as complementary information.

Experiments

Evaluation

NLI Score Recent QG studies (Luo et al. 2024; Ding, Hong, and Yao 2024; Wang et al. 2023) adopt traditional natural language generation metrics such as BLEU, ROUGE, and BERTScore. However, these metrics often emphasize surface-level similarity and lexical overlap, which may not accurately reflect semantic fidelity in QG tasks (Mohammadshahi et al. 2023). We aim to identify a metric that can reflect semantic fidelity of candidate question. To this end, we experiment on two benchmarks on paraphrasing questions: Minimal Edited Questions (MEQ) (Zhang et al. 2023) and Quora Question Pair (QQP). These benchmarks consist of positive pairs of questions that are paraphrases of each other, and negative pairs of questions that are not. We investigated the ability of QG evaluation metrics in accurately distinguishing the positive pairs of questions from negative ones. Specifically, we experimented with the following off-the-shelf scoring metrics: (1) Natural Language Inference (NLI) models, including three state-of-the-art DeBERTa-based models; (2) LLM zero-shot and few-shot; (3) Parascore (Shen et al. 2022); (4) ROUGE (Lin 2004); (5) BLEU (Papineni et al.

	GPT-4o-mini			Llama-3.3-70B			Qwen-2.5-72B		
	NLI	R-L	RQG	NLI	R-L	RQG	NLI	R-L	RQG
Transcripts Only									
All	31.25	**28.71**	3.18	24.05	26.84	2.94	23.01	27.30	2.61
Rule-Best	28.75	22.90	2.97	32.21	26.31	3.03	29.77	27.32	2.88
Direct	32.42	25.70	3.68	33.22	27.46	3.85	32.81	27.32	3.63
+ Rewrite	**34.80**	25.10	**4.86**	33.44	27.29	4.98	35.66	28.75	4.88
CoT	31.86	24.83	4.18	35.45	29.31	4.31	33.36	30.19	4.05
+ Rewrite	34.42	24.80	4.85	35.33	27.66	**4.99**	**37.76**	29.43	**4.86**
Keyframes Only									
All	21.58	20.84	2.52	15.48	18.30	2.29	15.46	18.75	2.06
Rule-Best	26.68	24.85	2.90	26.12	25.46	2.78	24.62	23.50	2.48
Direct	28.22	23.27	2.99	28.23	22.98	3.09	28.10	24.19	3.02
+ Rewrite	29.07	23.90	2.82	29.65	25.46	4.83	31.12	25.22	4.85
CoT	26.35	24.17	2.96	28.08	24.74	3.17	26.36	23.45	2.95
+ Rewrite	30.84	25.13	3.13	31.28	26.98	4.85	30.32	25.05	4.84
Multi-Modal									
CombineMM	33.41	27.12	4.06	36.07	**31.04**	4.31	35.78	**30.83**	3.97
+ Rewrite	34.49	27.17	4.80	**36.19**	28.10	4.96	35.57	30.00	4.80

*Notes: R-L denotes Rouge-L F1; RQG denotes RQUGE. Higher means better for all metrics.

Table 3: Results on LLM-Frontier. Rewriting (highlighted in light green) consistently improves NLI and RQG score.

2002); (6) BERTSsore (Zhang* et al. 2020). Results show that NLI models outperform other metrics. NLI score evaluates the logical entailment between the candidate and the reference, thereby capturing deeper semantic relationships. We select the best-performing NLI model `sileod/deberta-v3-large-tasksource-nli` on Hugging Face as our reference-based metric, termed **NLI score**.

We also report a traditional reference-based metric **ROUGE-L**, and a reference-free metric **RQUGE** (Mohammadshahi et al. 2023). RQUGE scores a question candidate from 1 to 5, based on its answerability given the context and target answer. We allow LLM to generate up to 5 questions, and calculate NLI@5, RQUGE@5, and RougeL@5. We evaluate QG under zero-shot setting, using all data from LLM-Frontier and DL-Intro for testing.

Experimental Settings

Context Settings We compare our method against the following context settings: (1) **All**. Use the entire transcript or the complete set of keyframe descriptions as contexts, presented as an ordered list. Each segment is appended to an ordered ID. In the instruction, we also provide the segment ID indicating the timestamp. (2) **Rule-k** . Use the given timestamp to locate relevant content, then apply a fixed-length context window of size k, where k denotes the number of transcript segments or keyframes. We experiment with $k \in \{1, 3, 5, 7, 9, 11\}$. and report the best-performing setting as Rule-Best. For our LLM-based context selection, we explore two strategies: (3) **Direct** Provide the model with straightforward instructions for selecting relevant context. (4) **Chain-of-thought (CoT)** After providing instructions, encourage LLM to explicitly output its reasoning process (Wei et al. 2022). Specifically, we manually define the reasoning process as first listing all relevant segments, then refining its selection on a sentence-by-sentence basis. We also explore a basic combination strategy (**CombineMM**) where we concatenate segments selected by CoT from each modality.

Question Generation and Base Models We use a consistent prompt for QG for all context settings, which has been carefully engineered and optimized under Rule-Best settings. We allow the LLM to generate up to 5 questions. We use GPT-4o-mini (Achiam et al. 2023), Llama-3.3-70B-Instruct-Turbo (Grattafiori et al. 2024), Qwen-2.5-72B-Instruct-Turbo (Yang et al. 2024) as our base LLMs. We use the same LLM for context selection, rewriting, and question generation.

Results and Analysis

Main Results

Results on LLM-Frontier and DL-Intro are shown in Table 3 and Table 4, respectively.

Using all transcripts or keyframes as context is suboptimal for QG. In most cases (except for GPT on LLM-Frontier), using all transcripts or keyframes as context leads to the lowest performance, indicating that excessive or unfiltered context negatively impacts question quality. While the best fixed-length context window (Rule-Best) outperforms full context, our LLM-based CoT selection yields better context. Specifically, direct extraction (Direct), as a basic strategy, outperforms Rule-Best with all models on LLM-Frontier, but underperforms Rule-Best on DL-Intro. However, CoT consistently improves over Direct, yielding higher NLI scores than All and Rule-Best. These results shows the effectiveness of CoT in dynamically selecting relevant context for QG.

	GPT-4o-mini			Llama-3.3-70B			Qwen-2.5-72B		
	NLI	R-L	RQG	NLI	R-L	RQG	NLI	R-L	RQG
Transcripts Only									
All	34.64	29.86	3.05	28.49	**28.69**	2.91	25.45	26.41	2.63
Rule-Best	33.99	28.35	2.97	38.32	28.76	3.25	37.17	26.59	3.14
Direct	31.48	24.04	3.41	35.54	26.21	3.57	36.11	25.93	3.40
+ Rewrite	35.19	26.53	**4.76**	38.82	26.69	**4.95**	36.34	27.71	4.79
CoT	32.78	24.66	3.94	**39.34**	27.07	3.87	33.36	30.19	3.96
+ Rewrite	38.36	27.36	4.70	38.98	27.60	4.94	37.33	29.14	4.80
Keyframes Only									
All	28.34	23.76	2.89	20.89	18.22	2.66	25.04	22.02	2.65
Rule-Best	31.23	24.74	2.47	24.23	24.23	2.73	26.40	22.52	2.53
Direct	29.08	24.94	2.86	36.64	26.04	3.31	36.11	25.93	3.06
+ Rewrite	32.94	22.05	4.60	35.05	24.67	4.93	37.19	26.60	**4.80**
CoT	32.38	25.27	2.96	35.29	25.10	3.32	32.10	24.38	3.07
+ Rewrite	35.49	**29.19**	4.66	36.35	24.99	4.93	36.21	26.61	4.76
Multi-Modal									
CombineMM	37.65	26.79	3.89	38.61	27.22	3.92	37.12	**30.38**	3.79
+ Rewrite	**41.09**	25.32	4.71	37.57	25.92	4.90	**38.56**	29.11	4.79

*Notes: R-L denotes Rouge-L F1; RQG denotes RQUGE. Higher means better for all metrics.

Table 4: Results on DL-Intro. Rewriting (highlighted in light green) consistently improves NLI and RQG score. CombineMM+Rewrite

Rewriting context improves relevancy and answerability. Rewriting improves NLI scores in most (24 out of 30) settings. For instance, CoT with rewriting reaches 38.36, up from 32.78 without rewriting, using GPT on DL-Intro. A similar trend is observed in keyframe-based context. The gains suggest that rewriting helps better align questions with instructional goals. Notably, rewriting consistently improves on RQUGE score. This suggests LLM context rewriting, as the essential component of our framework, is effective in re-organizing extracted transcripts or keyframes into answer-containing knowledge points that serve as concise and relevant context.

Transcripts are more useful context than keyframes, moreover, combining both modalities further improves question quality. Using transcripts alone consistently outperforms keyframe-based contexts across all settings. Moreover, the best performance is achieved when combining both modalities. On DL-Intro, CombineMM reaches an NLI of 37.65 without rewriting and 41.09 with rewriting, marking the highest scores overall. This demonstrates the complementary nature of the lecturer's speech and slides and the effectiveness of the rewriting step in enhancing question quality. While CombineMM does not always achieve the highest scores across all datasets, this may be due to our current, simplistic method of combining the two modalities, which presents an opportunity for future improvement.

Impacts of Rule-Based Context Window

As demonstrated in Figure 2, varying context window sizes has a notable influence on the quality of generated questions, and simply expanding the window often fails to yield consistent improvements. Moreover, providing the entire context in most cases did not lead to the highest performance, suggesting that using all available information can introduce noise or dilute key details for question formation. We also observed that the optimal context window size differed across both datasets and models, indicating that there is no one-size-fits-all approach. These findings highlight the need for adaptive and selective context retrieval strategies, rather than relying on a static, predetermined context window.

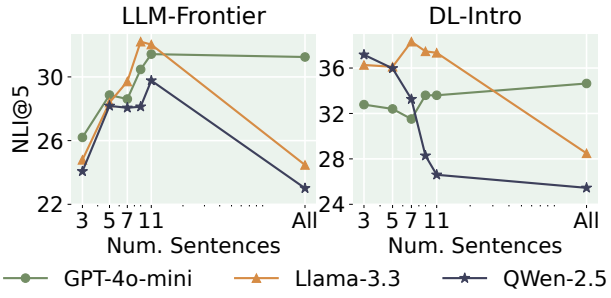

Figure 2: Increasing context window does not always improve question generation.

Human Evaluation

To assess the performance of our question generation (QG) methods, we conduct a qualitative human evaluation. With limited human effort, we focus on evaluating the effectiveness of the Rewrite and Multimodal combination modules. Our selection included the Direct and CoT methods (transcript-only) and the CombineMM method, both with and without the Rewrite module. We randomly sample 50 instances from

Reference Question: How does the ReLU activation function work?
Target Answer: It transforms all negative values into 0, while leaving positive values intact.

Generated Questions with Different Context Settings

All Transcript: What does the nonlinearity operation in a convolutional neural network do to negative values ? (NLI: 15.88)

CoT: What does the rectified linear unit (ReLU) activation function do to negative values in the output? (NLI: 41.32)

CoT+Rewrite: What is the primary transformation performed by the ReLU function ? (NLI: 42.77)

CombineMM: In the context of deep learning, what is the primary role of the ReLU activation function ? (NLI: 40.32)

CombineMM+ Rewrite: How does ReLU affect values during its operation? (NLI: 47.95)

Contexts from CombineMM

Extracted: ***Contextual summary***: In the context of convolutional neural networks, activation functions play a critical role ... ***From Transcript Excerpt***: "a common activation for convolutional neural networks are, rectified linear units or relu, right? think of this as really just deactivating pixels in your feature map that are negative, right? so anything that's positive gets passed on and anything that's negative gets set to zero. " ***From Keyframe Excerpt***: "... Here's a summary of the key points: 1. **Application**: ... 4. **Graph Representation**: The function $g(z) = \max(0, z)$ illustrates how ReLU operates, where values less than zero are set to zero. ..."

Rewritten: 1. Activation functions are crucial in convolutional neural networks as they introduce nonlinearity into the model. 2. The rectified linear unit (ReLU) is a common activation function that is applied after every convolution operation in CNNs. 3. ReLU operates as a pixel-by-pixel transformation, where " It transforms all negative values into 0, while leaving positive values intact. "

Table 5: An example in LLM-Frontier. With extracted and rewritten contexts, generated questions contains more relevant information aligned with instructor's intent and less imprecise or irrelevant information.

ModelSetting	LLM-Frontier			DL-Intro		
	Nat.	Rel.	Ans.	Nat.	Rel.	Ans.
Direct	3.65	3.72	2.79	4.60	4.05	3.77
+ Rewrite	3.87	3.57	3.29	4.67	4.28	3.47
CoT	4.33	3.72	2.97	4.60	4.47	3.92
+ Rewrite	4.35	4.00	3.49	4.58	4.50	4.08
CombineMM	**4.39**	3.99	3.55	4.65	4.37	3.85
+ Rewrite	4.37	**4.07**	**3.91**	**4.68**	**4.67**	**4.45**

Table 6: Human Evaluation Results.

the LLM-Frontier dataset and 40 instances from DL-Intro dataset. We engage 3 human judges, all fluent in English and experienced in education in AI fields. Judges are asked to rate each generated question on a 5-point Likert scale (1=Very Poor, 5=Excellent) in terms of three criteria: Naturalness (Nat.), Relevance (Rel.), and Answerability (Ans.) following previous studies (Zhang et al. 2021; Nguyen et al. 2024).

The evaluation results are summarized in Tab. 6. Our analysis shows that (1) Rewrite module provides a consistent boost to performance, especially in Answerability. With CombineMM, Rewrite module improves Answerability by 0.6 (from 3.85 to 4.45). (2) Our final method, CombineMM + Rewrite, consistently achieves the highest scores across nearly all criteria and datasets. We measured inter-annotator agreement on the ranking of the 6 model settings using Kendall's W (ranging from 0 to 1). For the LLM-Frontier dataset, the coefficients were 0.74 (Naturalness), 0.72 (Relevance), and 0.81 (Answerability), indicating strong agreement. For the DL-Intro dataset, the scores were 0.61, 0.61, and 0.59 respectively, showing moderate agreement. Overall, these values demonstrate a reliable and consistent evaluation.

Case Study

Tab. 5 shows a test case from LLM-Frontier. In this case, based on the target answer, the instructor intended to assess the operation of ReLU activation function. With All Transcript as context, the generated question involves a general concept "nonlinearity" and an irrelevant concept "convolutional neural network". The CoT context selection has correctly identified the relevant concept of ReLU, but still only asks about "negative values" imprecisely (ignoring positive values) in the generated question. The rewriting step and the integration of both modalities have rectified this issue, further aligning the generated question with instructor's intent.

Conclusion

In this work, we address the challenge of constructing appropriate contexts for generating educationally aligned questions. First, we construct a dataset of instructor-written quiz questions based on real-world classroom lectures, highlighting the limitations of existing EQG solutions when applied in realistic settings with long and noisy contexts. To bridge this gap, we propose an LLM-based EQG framework that first extracts relevant context pieces, then rewrites them into answer-containing atomic statements, and finally integrates multi-modal information from transcripts and slides for QG. Experiments with three LLMs demonstrate that COSER produces more relevant and concise contexts, thereby improving relevancy and answerability.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, G.; Pal, K.; Deng, Y.; Liu, H.; and Chen, Y.-C. 2024. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23164–23172.

Chen, G.; Yang, J.; Hauff, C.; and Houben, G.-J. 2018. LearningQ: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Ding, C.; Hong, Y.; and Yao, J. 2024. SGCM: Salience-Guided Context Modeling for Question Generation. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14755–14762. Torino, Italia: ELRA and ICCL.

Ghanem, B.; Lutz Coleman, L.; Rivard Dexter, J.; von der Ohe, S.; and Fyshe, A. 2022. Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2131–2146. Dublin, Ireland: Association for Computational Linguistics.

Gong, H.; Pan, L.; and Hu, H. 2022. KHANQ: A dataset for generating deep questions in education. In *Proceedings of the 29th international conference on computational linguistics*, 5925–5938.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guhr, O.; Schumann, A.-K.; Bahrmann, F.; and B"ohme, H. J. 2021. FullStop: Multilingual Deep Models for Punctuation Prediction. *CEUR Workshop Proceedings*.

Hadifar, A.; Bitew, S. K.; Deleu, J.; Develder, C.; and Demeester, T. 2023a. EduQG: A multi-format multiple-choice dataset for the educational domain. *Ieee Access*, 11: 20885–20896.

Hadifar, A.; Bitew, S. K.; Deleu, J.; Hoste, V.; Develder, C.; and Demeester, T. 2023b. Diverse Content Selection for Educational Question Generation. In Bassignana, E.; Lindemann, M.; and Petit, A., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 123–133. Dubrovnik, Croatia: Association for Computational Linguistics.

Hang, C. N.; Wei Tan, C.; and Yu, P.-D. 2024. MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access*, 12: 102261–102273.

Lee, U.; Jung, H.; Jeon, Y.; Sohn, Y.; Hwang, W.; Moon, J.; and Kim, H. 2024. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29(9): 11483–11515.

Li, K.; and Zhang, Y. 2024. Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 4715–4729. Bangkok, Thailand: Association for Computational Linguistics.

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

Liu, B.; Zhao, M.; Niu, D.; Lai, K.; He, Y.; Wei, H.; and Xu, Y. 2019. Learning to Generate Questions by LearningWhat not to Generate. In *The World Wide Web Conference*, WWW '19, 1106–1118. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Luo, H.; Deng, Y.; Shen, Y.; Ng, S.-K.; and Chua, T.-S. 2024. Chain-of-Exemplar: Enhancing Distractor Generation for Multimodal Educational Question Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7978–7993. Bangkok, Thailand: Association for Computational Linguistics.

Maity, S.; Deroy, A.; and Sarkar, S. 2025. Can large language models meet the challenge of generating school-level questions? *Computers and Education: Artificial Intelligence*, 8: 100370.

Mohammadshahi, A.; Scialom, T.; Yazdani, M.; Yanki, P.; Fan, A.; Henderson, J.; and Saeidi, M. 2023. RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 6845–6867. Toronto, Canada: Association for Computational Linguistics.

Mucciaccia, S. S.; Paixão, T. M.; Mutz, F. W.; Badue, C. S.; de Souza, A. F.; and Oliveira-Santos, T. 2025. Automatic Multiple-Choice Question Generation and Evaluation Systems Based on LLM: A Study Case With University Resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2246–2260.

Nguyen, B.; Yu, M.; Huang, Y.; and Jiang, M. 2024. Reference-based Metrics Disprove Themselves in Question Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13651–13666. Miami, Florida, USA: Association for Computational Linguistics.

Noorbakhsh, K.; Chandler, J.; Karimi, P.; Alizadeh, M.; and Balakrishnan, H. 2025. Savaal: Scalable Concept-Driven Question Generation to Enhance Human Learning. *arXiv preprint arXiv:2502.12477*.

Pan, R.; Cao, B.; Lin, H.; Han, X.; Zheng, J.; Wang, S.; Cai, X.; and Sun, L. 2024. Not All Contexts Are Equal: Teaching LLMs Credibility-aware Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19844–19863. Miami, Florida, USA: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Shen, L.; Liu, L.; Jiang, H.; and Shi, S. 2022. On the Evaluation Metrics for Paraphrase Generation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3178–3190. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E.; Schärli, N.; and Zhou, D. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sun, X.; Liu, J.; Lyu, Y.; He, W.; Ma, Y.; and Wang, S. 2018. Answer-focused and Position-aware Neural Question Generation. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3930–3939. Brussels, Belgium: Association for Computational Linguistics.

Wang, X.; Liu, B.; Tang, S.; and Wu, L. 2023. SkillQG: Learning to Generate Question for Reading Comprehension Assessment. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13833–13850. Toronto, Canada: Association for Computational Linguistics.

Wang, Z.; Valdez, J.; Basu Mallick, D.; and Baraniuk, R. G. 2022. Towards Human-Like Educational Question Generation with Large Language Models. In *Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I*, 153–166. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-11643-8.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, S.; Xie, J.; Chen, J.; Zhu, T.; Zhang, K.; and Xiao, Y. 2024. How Easily do Irrelevant Inputs Skew the Responses of Large Language Models? In *First Conference on Language Modeling*.

Xia, Z.; Gou, Q.; Yu, B.; Yu, H.; Huang, F.; Li, Y.; and Cam-Tu, N. 2023. Improving Question Generation with Multi-level Content Planning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 800–814. Singapore: Association for Computational Linguistics.

Xu, Y.; Wang, D.; Yu, M.; Ritchie, D.; Yao, B.; Wu, T.; Zhang, Z.; Li, T.; Bradford, N.; Sun, B.; Hoang, T.; Sang, Y.; Hou, Y.; Ma, X.; Yang, D.; Peng, N.; Yu, Z.; and Warschauer, M. 2022. Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 447–460. Dublin, Ireland: Association for Computational Linguistics.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhang, R.; Guo, J.; Chen, L.; Fan, Y.; and Cheng, X. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1): 1–43.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, Z.; Yu, W.; Ning, Z.; Ju, M.; and Jiang, M. 2023. Exploring Contrast Consistency of Open-Domain Question Answering Systems on Minimally Edited Questions. *Transactions of the Association for Computational Linguistics*, 11: 1082–1096.