SynMQG: Disentanglement and Mutual-Information Optimization for Synergistic Multi-modal Question Generation

Junjie Zhang[1], Shengyong Ding[2], Shuangyin Li[1]*
[1]School of Computer Science, South China Normal University, Guangzhou, China
[2]Faculty of Data Science, City University of Macau, Macau, China
junjiezhang_cs@m.scnu.edu.cn, syding@cityu.edu.mo, shuangyinli@scnu.edu.cn

Abstract

Synergistic multi-modal question generation aims to generate questions grounded in synergistic semantics distributed across multiple modalities, rather than being specified by isolated information from any single modality alone. In this setting, modality-specific cues and cross-modal shared semantics must be jointly well-organized so that the generated question reflects synergistic multi-modal contribution and captures genuinely synergistic information. However, this task faces two core challenges: *how to disentangle multi-modal information for synergistic information modeling*, and *how to preserve synergistic contribution during question generation*. To address these challenges, we propose **SynMQG**, a synergistic dependency-aware framework derived from a mutual-information-based theoretical analysis, which integrates disentangled representation learning with synergy-aware optimization for synergistic multi-modal question generation. Specifically, SynMQG disentangles the multi-modal context into visual-specific, textual-specific, and shared semantics, thereby explicitly organizing heterogeneous information for synergistic information modeling. Based on these disentangled representations, it further introduces a reasoning chain as an intermediate scaffold to structure cross-modal information before generation. Moreover, SynMQG proposes GRPO with a mutual-information-based synergy reward, which explicitly measures whether the generated question preserves effective contribution of synergistic information, encouraging the model to generate questions that depend on synergistic multi-modal semantics rather than superficial single-modal cues. Experiments on ScienceQA and MultimodalQA show that SynMQG consistently outperforms representative MQG baselines across standard text-generation metrics, MLLM-based evaluation, and human evaluation, while generating questions with stronger synergistic multi-modal grounding and higher multi-modal dependency. The source code is available at https://github.com/JunjieCS/SynMQG.

1 Introduction

Multi-modal question generation (MQG) aims to generate questions from heterogeneous inputs such as text, tables, and images. It has become an important task for benchmark construction, reasoning-oriented data augmentation, and multi-modal content understanding [18]. Compared with conventional text-based question generation, MQG broadens the input space from a single textual source to diverse multi-modal contexts, making it suitable for scenarios where information is naturally presented through multiple forms.

*Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

As MQG is increasingly used in reasoning-oriented scenarios, generating questions from multi-modal inputs is no longer only a matter of input coverage, but also a matter of cross-modal dependency. Many real-world multi-modal reasoning tasks require coordinated understanding across text, tables, and images, rather than being solvable from any single modality alone, as reflected in recent multi-modal reasoning benchmarks [19, 31]. However, in traditional MQG tasks, a generated question may still be largely determined by the textual context alone or by the visual content alone, even when both modalities are available. Moreover, simply providing multiple modalities during generation does not guarantee that the generated question depends on their joint information. This distinction motivates us to move from general MQG to *Synergistic Multi-modal Question Generation*.

Synergistic multi-modal question generation focuses on generating questions that depend on synergistic information across multiple modalities. Here, synergistic information denotes the non-redundant information that cannot be inferred from any individual modality, but emerges only when multiple modalities are jointly considered. Therefore, the generated question should be determined by cross-modal interaction rather than by modality-specific cues alone. This formulation is aligned with the intuition of partial information decomposition (PID), which separates unique, shared, and synergistic information among multiple sources [52]. From this perspective, synergistic MQG extends conventional MQG by emphasizing whether the generated question preserves genuine cross-modal dependency, as illustrated in Figure 1.

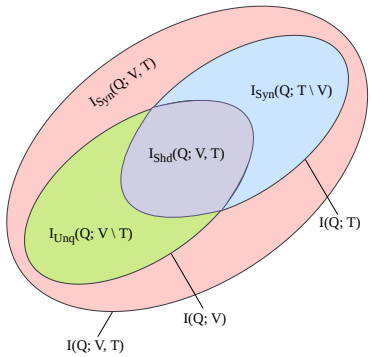

Figure 1: Illustration of synergistic multi-modal question generation. The information used to generate question Q from visual input V and textual input T can be decomposed into shared, modality-unique, and synergistic components.

Despite this need, existing MQG methods remain limited in enforcing such synergistic dependency. MultiQG-TI is among the first works to explicitly formulate question generation from image–text sources, converting visual inputs into textual signals through image captioning and OCR before feeding them into a text-only generator [48]. Chain-of-Exemplar (CoE) further studies multi-modal question generation in multiple-choice settings by combining exemplar retrieval, multi-modal large language models, and chain-of-thought reasoning [32]. Although these methods establish strong baselines, they mainly improve multi-modal *conditioning*, rather than explicitly modeling whether the generated question preserves non-redundant contribution from different modalities. As a result, they may still produce questions that are visually irrelevant, text-dominant, or only loosely connected to a coherent multi-modal reasoning process.

We argue that the challenges of existing MQG methods mainly lie in two aspects when the goal is *synergistic* multi-modal question generation. First, the model needs to *disentangle* heterogeneous multi-modal information so that synergistic information can be properly modeled. Visual inputs often contain fine-grained attributes, spatial relations, and structural cues, whereas textual inputs provide linguistic context, definitions, or explanatory knowledge. When heterogeneous inputs are directly textualized or naively concatenated, these distinct types of information can be blurred into a mixed representation space, making it difficult to preserve the information required for synergistic question grounding. Second, the model needs to *preserve* synergistic contribution during question generation. Standard generation objectives reward similarity to references, but do not explicitly enforce whether the generated question retains dependence on synergistic information, rather than being specified by a single modality alone. Thus, the key issue in synergistic MQG is not only how to combine multiple modalities, but also how to ensure that the generated question preserves such synergistic dependency. These challenges motivate us to design the framework from both the representation and optimization perspectives. To support synergistic information modeling, the model should avoid collapsing heterogeneous multi-modal information into a single mixed representation space, and instead explicitly distinguish *visual-specific*, *textual-specific*, and *shared* semantics. To preserve synergistic contribution during generation, the learning objective should further encourage the generated question to remain dependent on *synergistic* information rather than being specified by a single modality alone.

Thus, to address the above challenges, we propose **SynMQG**, a synergistic dependency-aware framework for synergistic multi-modal question generation. At the representation level, SynMQG disentangles the multi-modal context into modality-specific and cross-modal information, providing a structured basis for synergistic information modeling. At the generation level, SynMQG introduces a reasoning chain as an intermediate structure to organize multi-modal information before question generation. Moreover, SynMQG incorporates a mutual-information-based synergy reward into GRPO to explicitly encourage the generated question to preserve synergistic contribution from multi-modal information. Experiments on ScienceQA and MultimodalQA show that SynMQG consistently outperforms representative MQG baselines across standard text-generation metrics, MLLM-based evaluation, and human evaluation.

Our contributions are summarized as follows:

- We propose SynMQG, a synergistic dependency-aware MQG framework, and provide a theoretical formulation that characterizes question generation through cross-modal cooperative dependency.
- We introduce Synergy-Oriented Semantic Decomposition, a structured representation mechanism that separates modality-specific semantics from shared cross-modal information for synergistic reasoning.
- We develop Dependency-Preserving Policy Optimization, a synergy-guided reinforcement learning strategy that encourages questions grounded in synergistic multi-modal information.

2 Related Work

Question Generation. Question generation (QG) has been widely studied in natural language processing, and existing surveys typically organize prior work according to input sources, including unstructured text, structured knowledge, and visual content [18, 37, 55]. Early neural QG mainly focused on text-based answer-aware question generation for reading comprehension [13]. Later studies improved QG through answer-relevant context modeling, semantic graph reasoning, difficulty control, and reasoning guidance [9, 22, 38]. QG has also been extended to educational, structured, mathematical, and knowledge-grounded scenarios [7, 12, 14, 15, 28, 49]. Although these studies have advanced controllable, reasoning-aware, and structurally grounded QG, most of them still assume that the target question can be generated from a single dominant information source.

Visual Question Generation. Visual question generation (VQG) extends QG to visual inputs. Mostafazadeh et al. [34] introduced the task of generating natural questions about images, and subsequent studies connected VQG with visual question answering [24] or extended it to multi-image scenarios [54]. However, VQG remains primarily vision-centered, where textual evidence is usually absent or not treated as an equally necessary information source. In contrast, recent multi-modal question answering benchmarks show that realistic reasoning often requires coordinated evidence from text, tables, and images [19, 31]. This observation motivates a stronger requirement for multi-modal question generation: generated questions should not merely be related to multi-modal inputs, but should depend on information that emerges through cross-modal interaction.

Multi-modal Question Generation. Direct studies on multi-modal question generation remain limited. MultiQG-TI formulates question generation from image–text inputs by textualizing visual information through captioning and OCR before generation [48]. CoE further uses multi-modal large language models, chain-of-thought reasoning, and contextualized exemplar retrieval for educational question and distractor generation [32]. Although these methods demonstrate important progress, they mainly focus on how to consume multi-modal inputs through textualization, prompting, or staged generation. However, they do not explicitly distinguish whether a generated question relies on visual-only, textual-only, redundant cross-modal shared, or truly synergistic information [52]. This motivates our focus on *synergistic multi-modal question generation*, where questions are expected to rely on the joint contribution of modality-specific and shared cross-modal semantics.

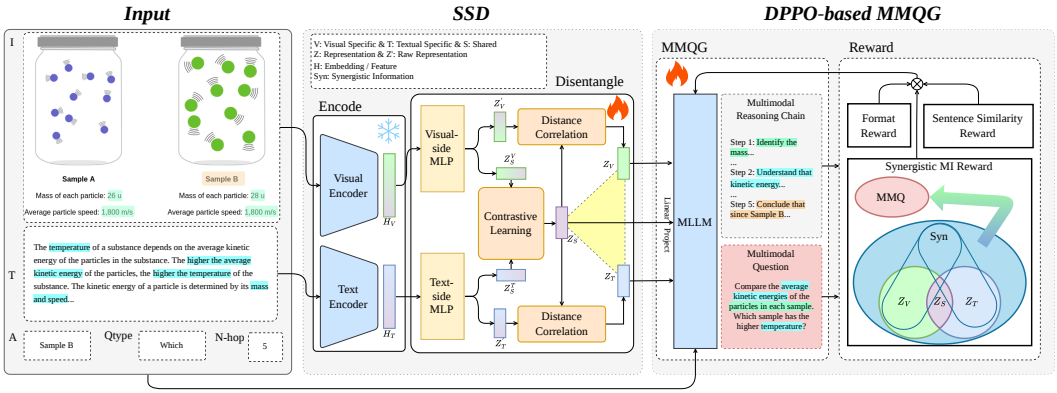

Figure 2: Overview of SynMQG. SSD is learned through contrastive shared alignment and distance-correlation-based semantic separation. The generator is then optimized by DPPO with a synergistic dependency-based reward, encouraging the generated question to depend on cross-modal interaction rather than single-modal cues.

3 Method

3.1 Overview

We propose **SynMQG** for synergistic multi-modal question generation, where the generated question should rely on the joint contribution of visual and textual information rather than on either modality alone. This formulation is inspired by partial information decomposition, which separates the information provided by multiple sources into shared, unique, and synergistic components [5, 51, 52]. Let V and T denote visual and textual information groups, and let $\mathbf{C}$ denote auxiliary generation constraints such as the target answer or question type.

As shown in Figure 2, the key idea of SynMQG is to make synergistic dependency explicitly *representable* and *optimizable*. To make it representable, **Synergy-Oriented Semantic Decomposition** (SSD) learns three semantic variables: visual-specific semantics Z_V, textual-specific semantics Z_T, and shared semantics Z_S. SSD is driven by two information-theoretic goals: maximizing the dependency between visual-side and textual-side shared semantics, and minimizing redundant dependency among private and shared semantic streams. Derived from mutual-information analysis, the semantic decomposition objective is instantiated by contrastive learning for shared alignment and distance correlation for dependency separation [8, 44]. Intuitively, SynMQG first separates what is uniquely provided by the visual modality, what is uniquely provided by the textual modality, and what is commonly grounded by both modalities. To make synergistic dependency optimizable during generation, **Dependency-Preserving Policy Optimization** (DPPO) introduces a synergistic dependency-based reward. It estimates whether the final question obtains additional likelihood support from the joint presence of visual-specific and textual-specific semantics beyond their individual effects. The policy is then optimized with GRPO to prefer questions that are valid, semantically faithful, and dependent on cross-modal interaction [43].

3.2 Synergy-Oriented Semantic Decomposition

Inspired by partial information decomposition [5, 51, 52], the SSD module separates multi-modal semantics into visual-specific Z_V, textual-specific Z_T, and shared Z_S information. For a concrete input instance, encoder features h_v, h_t are first obtained from H_V, H_T, and the intermediate projections z'_v, z'_t, z^v_s, z^t_s in Figure 2 are regularized to produce the corresponding realizations z_v, z_t, z_s. Formally, $z_v = f_v(h_v), \quad z_t = f_t(h_t), \quad z^v_s = f^v_s(h_v), \quad z^t_s = f^t_s(h_t)$, with the shared representation computed as $z_s = \mathrm{Fuse}(z^v_s, z^t_s)$. We formalize this decomposition with a mutual-information objective: the shared semantics should maximize cross-modal common information, while the private semantics should minimize redundant pairwise dependencies:

$$\max_{\psi} \mathcal{J}_{\mathrm{SSD}} = I(Z^V_S; Z^T_S) - I(Z_V; Z_S) - I(Z_T; Z_S) - I(Z_V; Z_T), \tag{1}$$

where ψ denotes the parameters of the SSD module.

Contrastive shared alignment. The shared representation should capture information consistently grounded in both modalities. We therefore align the visual-side shared representation z_s^v and the textual-side shared representation z_s^t with a symmetric InfoNCE objective. For each training example, the visual-side shared representation and the textual-side shared representation from the same sample form a positive pair, while representations from different samples in the mini-batch are used as negatives:

$$\mathcal{L}_{\text{share}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\text{sim}(z_{s,i}^v, z_{s,i}^t)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(z_{s,i}^v, z_{s,j}^t)/\tau)} + \log\frac{\exp(\text{sim}(z_{s,i}^t, z_{s,i}^v)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(z_{s,i}^t, z_{s,j}^v)/\tau)}\right], \quad (2)$$

where B is the mini-batch size, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature parameter. This objective encourages Z_S to preserve cross-modal common grounding by maximizing agreement between shared representations extracted from the two modalities [46, 47]. The lower-bound interpretation is provided in Appendix A.2.

Distance-correlation semantic separation. As shared semantics are aligned, modality-specific semantics should remain separated from shared semantics and avoid cross-modal leakage. Since directly minimizing mutual information among high-dimensional latent variables is unstable, we use distance correlation as a tractable surrogate for the mutual-information-based independence objective:

$$\mathcal{L}_{\text{sep}} = \text{dCor}(z_v, z_s) + \text{dCor}(z_t, z_s) + \text{dCor}(z_v, z_t). \quad (3)$$

Minimizing this term discourages redundant dependence among visual-specific, textual-specific, and shared semantic streams. Since distance correlation vanishes if and only if two variables are independent under mild conditions, it provides an effective surrogate for separating latent semantic factors [21, 27]. The detailed discussion is provided in Appendix A.3. Thus, the final semantic decomposition objective is: $\mathcal{L}_{\text{dis}} = \lambda_{\text{share}}\mathcal{L}_{\text{share}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}$, where λ_{share} and λ_{sep} balance shared alignment and semantic separation.

3.3 Dependency-Preserving Policy Optimization

The decomposition above provides a structured semantic interface, but supervised learning alone does not guarantee that a generated question truly uses the joint visual-textual contribution. A question may still be fluent and relevant while depending only on visual-specific cues, textual-specific cues, or redundant shared information. We therefore introduce a dependency-preserving reward to explicitly optimize synergistic dependency during policy learning.

Conditioned on the decomposed semantics and auxiliary constraints, the generator first produces a reasoning chain r [50] and then generates the final question q:

$$p_\theta(r, q \mid z_v, z_t, z_s, c) = p_\theta(r \mid z_v, z_t, z_s, c)p_\theta(q \mid r, z_v, z_t, z_s, c). \quad (4)$$

During reward computation, we keep the generated reasoning chain fixed and only re-score the final question under different semantic conditioning sets. This avoids changing the intermediate context when measuring whether the same question depends on joint visual-textual semantics.

Synergistic dependency score. We characterize synergy as the additional dependency contributed by jointly using visual-specific and textual-specific semantics under shared grounding:

$$\Delta_{\text{syn}}(Q; Z_V, Z_T \mid Z_S, C) = I(Q; Z_V, Z_T \mid Z_S, C) - I(Q; Z_V \mid Z_S, C) - I(Q; Z_T \mid Z_S, C). \quad (5)$$

This score becomes positive when the joint visual-textual dependency provides additional information beyond the sum of the two single-modality dependencies [25].

Synergistic dependency-based reward. Since the population-level dependency score is not directly optimizable during generation, we estimate a sample-level surrogate by comparing the likelihood of the same generated question under different semantic conditioning sets, as detailed in Appendix A.5. For the g-th generated sample $y^{(g)} = (r^{(g)}, q^{(g)})$, we keep the reasoning chain fixed and define:

$$\begin{aligned}\Delta_{\text{syn}}^{(g)} = \; & \log p_\theta(q^{(g)} \mid r^{(g)}, z_v, z_t, z_s, c) - \log p_\theta(q^{(g)} \mid r^{(g)}, z_v, z_s, c) \\ & - \log p_\theta(q^{(g)} \mid r^{(g)}, z_t, z_s, c) + \log p_\theta(q^{(g)} \mid r^{(g)}, z_s, c),\end{aligned} \quad (6)$$

where the first term evaluates the likelihood of the question when all decomposed semantics are provided. The second and third terms evaluate the likelihood when only one modality-specific semantic variable is provided together with the shared semantics. The last term evaluates the likelihood using only the shared semantics and auxiliary constraints. The log-likelihood of the question can be computed by summing token-level log probabilities:

$$\log p_\theta(q^{(g)} \mid r^{(g)}, z_v, z_t, z_s, c) = \sum_{m=1}^{|q^{(g)}|} \log p_\theta \left(q_m^{(g)} \mid q_{<m}^{(g)}, r^{(g)}, z_v, z_t, z_s, c \right). \tag{7}$$

In practice, each semantic variable can be converted into a semantic prompt. For reduced conditioning sets, the unavailable semantic prompts are masked out, while the remaining prompts are kept unchanged. The synergy score measures how much additional likelihood support the final question receives from jointly using z_v and z_t, beyond their individual effects under the shared semantics z_s. A larger value indicates that the question is more strongly supported by the interaction between visual-specific and textual-specific semantics.

Thus, the synergy reward is defined as $R_{\text{mi}}^{(g)} = \phi\Big(\Delta_{\text{syn}}^{(g)} - \gamma_{\text{syn}}\Big)$, where γ_{syn} is a synergy margin and $\phi(\cdot)$ is a monotone shaping function. The margin prevents the policy from being rewarded for weak or noisy interaction effects, while the shaping function maps the raw synergy score into a stable reward signal. To maintain generation quality without changing the main synergy-oriented objective, we combine the synergy reward with two auxiliary quality constraints: $R^{(g)} = R_{\text{rule}}^{(g)} + R_{\text{sim}}^{(g)} + R_{\text{mi}}^{(g)}$, where $R_{\text{mi}}^{(g)}$ is the primary dependency-preserving reward that encourages the generated question to rely on synergistic multi-modal information. The other two terms serve as auxiliary constraints: $R_{\text{rule}}^{(g)}$ checks whether the output follows the required format and produces a valid question, while $R_{\text{sim}}^{(g)}$ measures semantic agreement with the reference question using an external scorer. These auxiliary rewards are used to prevent degenerate generations and preserve question quality, rather than to replace the synergy objective.

Training objective. SynMQG is trained in two stages [36]. The first stage performs supervised fine-tuning with: $\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{dis}}$, where $\mathcal{L}_{\text{gen}}$ trains reasoning-question generation and $\mathcal{L}_{\text{dis}}$ learns the decomposed semantic space. The second stage initializes from the SFT policy and applies GRPO with the reward $R^{(g)}$. For each input, we sample a group of candidate reasoning-question outputs, compute their structural, semantic, and synergistic dependency rewards, and normalize the rewards within the group to obtain relative advantages. The full GRPO objective is provided in Appendix H.

4 Experiments

4.1 Datasets

We conduct experiments on **ScienceQA** [30] and **MultimodalQA** [45]. These two datasets provide complementary evaluation scenarios for multi-modal question generation: ScienceQA focuses on educational science questions with image–text contexts, while MultimodalQA emphasizes heterogeneous information involving text, tables, and images. Since our goal is to study *synergistic* MQG, we retain samples whose questions require the joint use of multi-modal information. Detailed of data processing are provided in Appendix B.2.

4.2 Baselines

We compare SynMQG with two groups of baselines: task-specific MQG models and general-purpose multi-modal LLMs. All methods follow the same setting: given the multi-modal context and target answer, the model generates a question. The task-specific baselines include VL-T5 [10], MultiQG-TI [48], MMCoT [57] and CoE [32]. We also evaluate GPT-4o [35] and Qwen3-VL [1] under zero-shot and few-shot settings. More details are provided in Appendix B.3.

Table 1: Main results on MultimodalQA and ScienceQA. The best result in each column is shown in **bold**, and the second-best result is underlined.

Dataset (→)	MultimodalQA								ScienceQA							
Method (↓)/Metric (→)	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**MTR**	**BSc**	**BRT**	**B-1**	**B-2**	**B-3**	**B-4**	**R-L**	**MTR**	**BSc**	**BRT**
MLLMs																
ChatGPT 0-shot [35]	17.6	8.5	4.9	3.0	22.4	19.2	61.2	32.3	36.1	22.2	14.6	10.3	35.7	37.2	67.1	48.7
ChatGPT 1-shot [35]	17.4	8.5	4.9	3.0	22.6	18.9	61.1	32.1	38.1	23.7	15.2	10.2	36.0	37.4	68.2	49.3
ChatGPT 3-shot [35]	17.1	8.3	4.8	2.9	22.5	18.7	61.2	32.0	37.7	23.2	14.7	9.9	35.7	37.2	68.0	49.0
Qwen 0-shot [1]	23.3	10.5	5.6	3.1	21.9	20.4	61.1	32.8	30.2	17.9	11.5	7.8	34.3	37.6	66.7	48.3
Qwen 1-shot [1]	23.5	10.9	5.9	3.3	22.4	20.7	61.0	32.9	30.7	18.5	11.8	7.9	34.2	37.7	66.8	48.8
Qwen 3-shot [1]	23.0	10.6	5.7	3.2	22.3	20.7	61.0	32.9	31.0	18.6	11.9	7.9	34.9	37.9	67.1	48.9
MQG Models																
VL-T5 [10]	42.8	26.4	17.7	12.4	35.9	37.9	71.8	45.9	71.3	64.7	60.4	57.2	69.9	73.4	84.5	70.4
MultiQG-TI [48]	42.5	26.9	18.3	13.0	35.0	37.8	71.1	43.9	66.0	58.8	54.5	51.5	66.6	67.4	83.3	71.1
MMCoT [57]	42.1	25.6	17.0	11.9	34.7	37.5	71.0	45.3	75.6	68.0	62.8	59.3	73.3	75.1	87.0	73.0
CoE [32]	38.7	24.2	16.4	11.6	34.0	37.7	70.6	42.5	76.9	69.7	64.6	60.8	75.6	78.4	87.8	75.9
Ours w/o GRPO	42.8	25.8	17.0	11.6	34.3	38.0	70.7	45.1	71.0	63.5	58.2	54.6	73.0	77.2	85.9	71.2
Ours	**45.4**	**29.0**	**20.1**	**14.5**	**37.6**	**41.5**	**72.7**	**47.4**	**79.7**	**72.8**	**68.4**	**65.6**	**77.3**	**80.8**	**89.4**	**77.1**

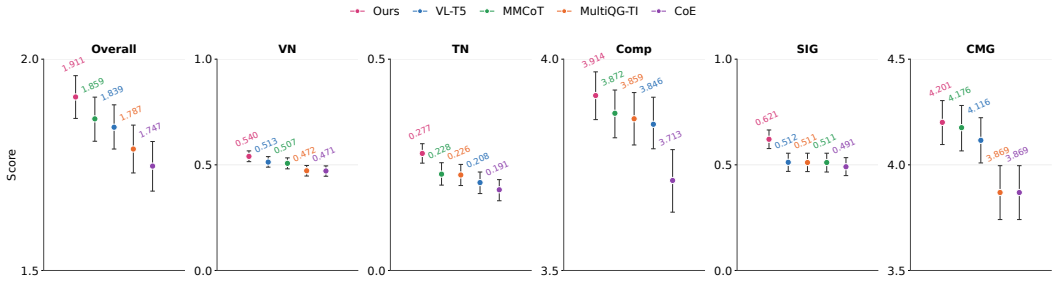

Figure 3: MLLM-based evaluation results on MultimodalQA. The figure compares different methods under the proposed multi-dimensional evaluation protocol.

4.3 Evaluation Metrics

We evaluate generated questions from three perspectives. First, we report standard text-generation metrics, including BLEU (*B*) [39], ROUGE-L (*R-L*) [26], METEOR (*MTR*) [3], BERTScore (*BSc*) [56], and BLEURT (*BRT*) [42]. Second, following QGEval [16], we conduct MLLM-based evaluation from five synergy-oriented dimensions: visual necessity (*VN*), textual necessity (*TN*), visual-textual complementarity (*Comp*), synergistic interaction gain (*SIG*), and cross-modal grounding (*CMG*), which assess whether questions genuinely rely on complementary visual-textual information. Third, following prior work [32], we conduct human evaluation with N-way preference comparisons under multi-modal synergy and clarity. More details are provided in Appendix B.4.

4.4 Main Results

Table 1 reports the main results on MultimodalQA and ScienceQA. SynMQG achieves the best performance across both datasets, demonstrating its effectiveness for synergistic multi-modal question generation. The improvements over prompted MLLMs indicate that general instruction-following ability alone is insufficient for MQG, since this task requires fine-grained grounding and explicit dependency modeling across visual and textual modalities.

Among existing baselines, CoE shows different behavior across the two datasets. Since CoE relies on retrieved training examples as prompts, it benefits more from the relatively regular educational-style patterns in ScienceQA, but becomes less stable on MultimodalQA, where questions involve more diverse contexts and more complex cross-modal reasoning. In contrast, SynMQG does not rely on exemplar matching. It explicitly models visual-specific, textual-specific, and shared semantics, and further optimizes the generation process toward synergistic visual-textual interaction. This enables

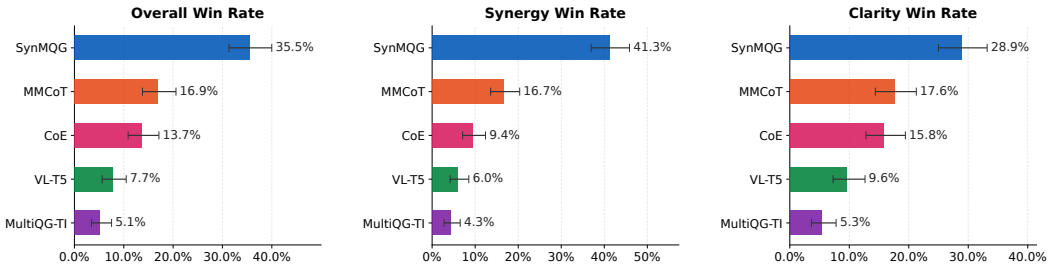

Figure 4: Human preference evaluation on MultimodalQA. The figure compares different methods under human judgments.

Table 2: Inter-annotator agreement and Gemini–human correlation across evaluation dimensions.

Dimension	**Round 1** α	**Round 2** α	**Pearson** r
VN	0.216	0.287	0.431
TN	0.249	0.344	0.408
SIG	0.191	0.397	0.338
CMG	0.427	0.864	0.322
Comp	0.331	0.817	0.192

SynMQG to better preserve modality-specific information while encouraging questions that require coordinated reasoning over both modalities.

4.5 MLLM-based Evaluation

Figure 3 presents the MLLM-based evaluation results on MultimodalQA. SynMQG consistently outperforms all baselines across the evaluated dimensions, achieving the best overall performance. The most pronounced improvement is observed in *SIG*, indicating that the generated questions more effectively capture synergistic interactions between visual and textual information. The strong performance in *VN* and *TN* further suggests that both modalities contribute meaningfully to question generation, rather than allowing either modality to dominate. Meanwhile, the improvements in *Comp* and *CMG* reflect better integration of complementary information and more reliable cross-modal grounding. These findings support the effectiveness of disentangling shared and modality-specific semantics and explicitly optimizing synergistic dependencies, enabling SynMQG to generate questions whose underlying information more strongly depends on the joint contribution of both modalities.

4.6 Human Evaluation

Figure 4 reports the human preference results on MultimodalQA. SynMQG achieves the highest win rate across all criteria, especially under the synergy criterion, indicating that human evaluators more often prefer its questions when judging whether they require joint visual-textual reasoning. This result further supports the effectiveness of explicitly modeling modality-specific and shared semantics and optimizing synergistic dependency. Compared with baselines, SynMQG is better at generating questions that are not only clear and well-formed but also dependent on coordinated multi-modal reasoning. Overall, the human evaluation confirms that SynMQG improves both perceived question quality and synergy-oriented multi-modal dependency.

4.7 Human Evaluation and Evaluator Reliability

To assess the reliability of MLLM-based evaluation, we further examine the agreement among human annotators and the consistency between Gemini and human judgments. Following QGEval [16], three annotators independently evaluated 40 randomly sampled questions from each method, resulting in 200 questions in total. We report Krippendorff's α to measure inter-annotator agreement and Pearson's r to quantify the correlation between Gemini and human evaluations across five dimensions.

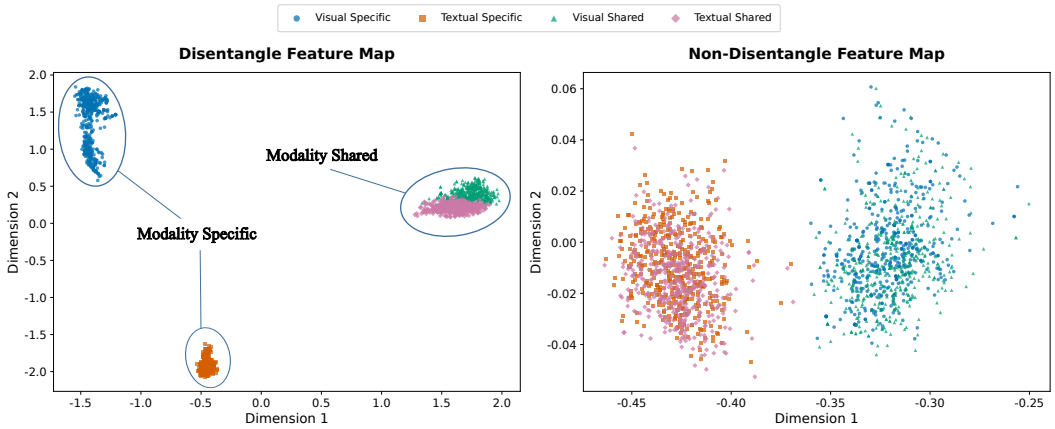

Figure 5: Semantic representation visualization on MultimodalQA. The left subfigure shows the feature space learned with SSD, while the right subfigure shows the feature space without disentanglement.

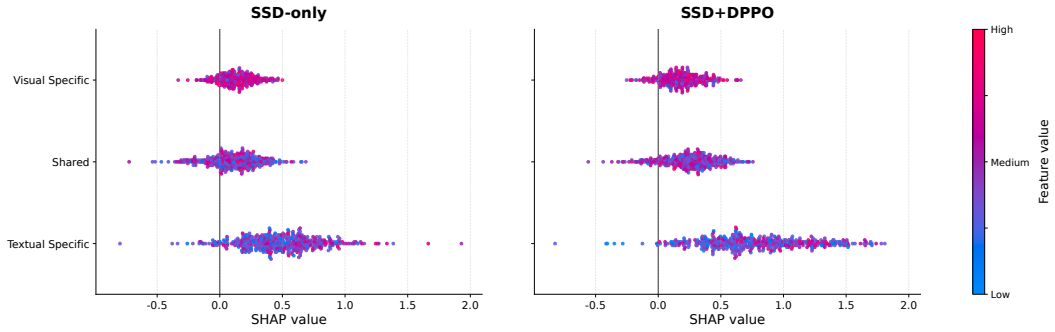

Figure 6: SHAP-based feature contribution analysis on MultimodalQA. The left subfigure reports SSD-only results, while the right subfigure reports SSD+DPPO results.

As shown in Table 2, inter-annotator agreement improves consistently across all dimensions after rubric clarification, with particularly substantial gains in *CMG* and *Comp*. This suggests that clearer evaluation criteria help reduce ambiguity in human assessments of cross-modal relationships. Meanwhile, Gemini exhibits positive correlations with human judgments across all dimensions, with stronger alignment in *VN* and *TN* than in *Comp*. Although the correlations remain modest, these results indicate that Gemini captures aspects of human-perceived multi-modal question quality, supporting its use as a complementary automated evaluator.

4.8 Analysis of SSD

To analyze SSD, we visualize the semantic space and feature contributions on MultimodalQA. As shown in Figure 5, SSD learns a more structured representation space, where visual-specific and textual-specific features are clearly separated, while visual-side shared and textual-side shared features are closely aligned. This indicates that SSD can distinguish modality-specific evidence while preserving shared cross-modal grounding. Figure 6 further shows that SSD+DPPO produces more stable and positively oriented contributions from the disentangled semantic streams than SSD-only training. This suggests that SSD provides meaningful semantic factors, and DPPO further encourages the generator to use them for dependency-aware question generation. Overall, these results confirm that SSD supports both semantic disentanglement and synergistic multi-modal question generation.

Table 3: Ablation study on the MultimodalQA dataset.

Model	B-1	B-2	B-3	B-4	R-L	MTR	BSc	BRT
SynMQG	**45.4**	**29.0**	**20.1**	**14.5**	**37.6**	**41.5**	**72.7**	**47.4**
w/o GRPO	42.8 (↓ 5.73%)	25.8 (↓ 11.03%)	17.0 (↓ 15.42%)	11.6 (↓ 20.00%)	34.3 (↓ 8.78%)	38.0 (↓ 8.43%)	70.7 (↓ 2.75%)	45.1 (↓ 4.85%)
w/o DPPO	38.4 (↓ 15.42%)	22.2 (↓ 23.45%)	14.4 (↓ 28.36%)	9.9 (↓ 31.72%)	30.7 (↓ 18.35%)	33.0 (↓ 20.48%)	68.0 (↓ 6.46%)	42.3 (↓ 10.76%)
w/o SSD	42.1 (↓ 7.27%)	25.6 (↓ 11.72%)	17.0 (↓ 15.42%)	11.9 (↓ 17.93%)	34.7 (↓ 7.71%)	37.5 (↓ 9.64%)	71.0 (↓ 2.34%)	45.3 (↓ 4.43%)

Table 4: Performance comparison of different MLLM backbones on MultimodalQA. The best result for each metric is highlighted in bold.

Backbone	B-1	B-2	B-3	B-4	R-L	MTR	BSc	BRT
Qwen3-VL-2B	45.0	29.1	20.2	14.6	38.0	41.2	72.7	47.0
Qwen2.5-VL-3B	45.8	28.9	19.6	13.8	35.8	40.6	72.5	47.3
LLaVA-1.5-7B	45.3	28.7	19.6	14.0	36.9	41.3	72.2	46.7
InternVL3-8B	43.5	27.6	18.8	13.3	37.0	40.4	72.4	46.1

4.9 Ablation Study

Table 3 presents the ablation results on MultimodalQA. Removing any component consistently degrades performance, confirming that SynMQG benefits from both semantic decomposition and dependency-aware optimization.

w/o SSD. The performance degradation indicates that semantic decomposition is essential for organizing heterogeneous multi-modal information. When visual-specific, textual-specific, and shared semantics are not explicitly structured, complementary evidence from different modalities becomes harder to identify, which weakens synergy-oriented question generation.

w/o DPPO. The substantial drop shows that supervised fine-tuning alone is insufficient to ensure synergistic dependency. Without an explicit dependency-preserving objective, the model may still generate fluent questions, but these questions are less likely to rely on the joint contribution of visual and textual information.

w/o GRPO. The decline suggests that the synergy reward needs an effective policy optimization mechanism. Group-relative optimization helps compare candidate questions within the same input and turns the dependency-aware reward into a more stable training signal, leading to stronger synergy-dependent generation.

4.10 Impact of Different Backbones

As shown in Table 4, SynMQG maintains consistent performance across MLLM backbones with different architectures and parameter scales, ranging from 2B to 8B. Despite variations in generation quality, all four backbones achieve comparable results across lexical and semantic evaluation metrics. Notably, smaller models remain competitive with larger ones, suggesting that the effectiveness of SynMQG does not depend on a specific model capacity. These findings demonstrate the adaptability of SynMQG across diverse MLLM architectures and model sizes, highlighting its potential as a backbone-agnostic framework for multi-modal question generation.

5 Conclusion

In this paper, we proposed SynMQG, a synergistic dependency-aware framework for synergistic multi-modal question generation. SynMQG explicitly decomposes multi-modal inputs into visual-specific, textual-specific, and shared semantics, and further optimizes question generation to preserve synergistic dependency. Through Synergy-Oriented Semantic Decomposition and Dependency-Preserving Policy Optimization, SynMQG encourages generated questions to rely on complementary cross-modal information rather than single-modality shortcuts. Experiments on MultimodalQA and ScienceQA demonstrate that SynMQG consistently improves automatic metrics, LLM-based evaluation, and human preference results. These results suggest that modeling and preserving synergistic dependency is essential for high-quality multi-modal question generation.

Acknowledgments and Disclosure of Funding

This work was supported by the Science and Technology Development Fund, Macao SAR, under the joint research funding program with the Department of Science and Technology of Guangdong Province (Grant No. 0176/2026/AGJ).

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, et al. Qwen3-vl technical report, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, et al. Qwen2.5-vl technical report, 2025.

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[4] Thomas Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions*, (53):370–418, 12 1763. ISSN 0260-7085. doi: 10.1098/rstl.1763.0053.

[5] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. ISSN 1099-4300. doi: 10.3390/e16042161.

[6] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, 16(2):101–133, 2001.

[7] Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. Scalable educational question generation with pre-trained language models. In Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 327–339, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-36272-9.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[9] Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.465.

[10] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation, 2021.

[11] Thomas M. Cover and Joy A. Thomas. *Entropy, Relative Entropy, and Mutual Information*, chapter 2, pages 13–55. John Wiley and Ltd Sons, 2005. ISBN 9780471748823. doi: https://doi.org/10.1002/047174882X.ch2.

[12] Chenhe Dong, Ying Shen, Shiyang Lin, Zhenzhou Lin, and Yang Deng. A unified framework for contextual and factoid question generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):21–34, 2024. doi: 10.1109/TKDE.2023.3280182.

[13] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1123.

[14] Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. A feasibility study of answer-agnostic question generation for education. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1919–1926, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.151.

[15] Hady Elsahar, Christophe Gravier, and Frederique Laforest. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 218–228, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1020.

[16] Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. QGEval: Benchmarking multi-dimensional evaluation for question generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.658.

[17] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control, 2016.

[18] Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. A survey on neural question generation: Methods, applications, and prospects. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8038–8047. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/889. Survey Track.

[19] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7879–7886, Apr. 2020. doi: 10.1609/aaai.v34i05.6294.

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

[21] Ruifan Huang and Haixia Liu. Fairness via independence: A (conditional) distance covariance framework, 2025.

[22] Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. Improving question generation with to the point context. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3216–3226, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1317.

[23] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014.

[24] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6116–6124, 2018. doi: 10.1109/CVPR.2018.00640.

[25] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying and modeling multimodal interactions: An information decomposition framework, 2023.

[26] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[27] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1088–1097, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539302.

[28] Tianqiao Liu, Qiang Fang, Wenbiao Ding, Hang Li, Zhongqin Wu, and Zitao Liu. Mathematical word problem generation from commonsense knowledge graph and equations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4225–4240, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.348.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[30] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc., 2022.

[31] Haohao Luo, Ying Shen, and Yang Deng. Unifying text, tables, and images for multimodal question answering. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.626.

[32] Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.432.

[33] Nicholas Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949. ISSN 01621459, 1537274X.

[34] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1170.

[35] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, et al. Gpt-4o system card, 2024.

[36] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[37] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. Recent advances in neural question generation, 2019.

[38] Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. Semantic graphs for generating deep questions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.135.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135.

[40] Igal Sason. On reverse pinsker inequalities, 2015.

[41] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013.

[42] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704.

[43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[44] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.

[45] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021.

[46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[48] Zichao Wang and Richard Baraniuk. MultiQG-TI: Towards question generation from multi-modal sources. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 682–691, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.55.

[49] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. Towards human-like educational question generation with large language models. In Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 153–166, Cham, 2022. Springer International Publishing. ISBN 978-3-031-11644-5.

[50] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[51] Michael Wibral, Viola Priesemann, Jim W. Kay, Joseph T. Lizier, and William A. Phillips. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain and Cognition*, 112:25–38, 2017. ISSN 0278-2626. doi: https://doi.org/10.1016/j.bandc.2015.09.004. Perspectives on Human Probabilistic Inferences and the 'Bayesian Brain'.

[52] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010.

[53] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[54] Min-Hsuan Yeh, Vincent Chen, Ting-Hao Huang, and Lun-Wei Ku. Multi-VQG: Generating engaging questions for multiple images. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 277–290, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.19.

[55] Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1), September 2021. ISSN 1046-8188. doi: 10.1145/3468889.

[56] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[57] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024.

A Theoretical Analysis

A.1 Formal Objective of Synergistic Multi-modal Question Generation

To formalize synergistic multi-modal question generation, we first represent each input instance as a combination of visual information, textual information, and auxiliary generation constraints:

$$X = (V, T, C), \tag{8}$$

where V denotes the visual modality, T denotes the textual modality, and C denotes auxiliary constraints such as the answer, question type, or other task-specific control variables.

Since the goal of MQG is to generate a natural-language question, we denote the target question as a discrete sequence from the question space:

$$Q \in \mathcal{Q}. \tag{9}$$

To distinguish modality-specific semantics from cross-modal common semantics, we decompose the visual and textual modalities into private and shared latent variables:

$$V \longrightarrow (Z_V, Z_S^V), \qquad T \longrightarrow (Z_T, Z_S^T). \tag{10}$$

Because the two shared variables are extracted from different modalities but should describe the same common semantics, we fuse them into a unified shared representation:

$$Z_S = \mathrm{Fuse}(Z_S^V, Z_S^T). \tag{11}$$

To make the shared space preserve modality-invariant semantics, we encourage the visual-side and textual-side shared variables to have high mutual information, following the multi-view contrastive learning perspective [46]:

$$I(Z_S^V; Z_S^T). \tag{12}$$

To make the private spaces non-redundant, we penalize the dependency between each private variable and the shared variable, as well as the dependency between the two private variables:

$$I(Z_V; Z_S), \qquad I(Z_T; Z_S), \qquad I(Z_V; Z_T). \tag{13}$$

Combining shared alignment and private separation gives the ideal disentanglement objective:

$$\max_{\psi} \ \mathcal{J}_{\mathrm{dis}} = I(Z_S^V; Z_S^T) - I(Z_V; Z_S) - I(Z_T; Z_S) - I(Z_V; Z_T), \tag{14}$$

where ψ denotes the parameters of the semantic decomposition module.

Since Q is discrete while Z_V, Z_T, and Z_S are continuous latent variables, the dependency between a question and a semantic variable should be written as mixed-variable conditional mutual information [11]:

$$I(Q; Z \mid U) = \int_{\mathcal{U}} \int_{\mathcal{Z}} p(z, u) \sum_{q \in \mathcal{Q}} p(q \mid z, u) \log \frac{p(q \mid z, u)}{p(q \mid u)} \, dz \, du. \tag{15}$$

To interpret this dependency as uncertainty reduction, the same quantity can equivalently be written in entropy form:

$$I(Q; Z \mid U) = H(Q \mid U) - H(Q \mid Z, U). \tag{16}$$

This entropy form shows that a semantic variable is useful for question generation when it reduces the uncertainty of Q under the same conditioning context.

A.2 Contrastive Lower Bound for Shared Semantics

To justify the shared-alignment objective, we first simplify the notation by treating the visual-side and text-side shared representations as two paired semantic views:

$$X_S = Z_S^V, \qquad Y_S = Z_S^T. \tag{17}$$

Because the shared representations should be consistent across modalities, we aim to maximize the mutual information between the two paired views [46]:

$$I(X_S; Y_S) = \int_{\mathcal{X}_S} \int_{\mathcal{Y}_S} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dy \, dx. \tag{18}$$

To expose the connection between mutual information and contrastive learning, we rewrite the joint density as a product of the marginal density and the conditional density:

$$p(x, y) = p(x)p(y \mid x). \tag{19}$$

Substituting this factorization into the mutual information formula yields a density-ratio form:

$$I(X_S; Y_S) = \int_{\mathcal{X}_S} \int_{\mathcal{Y}_S} p(x, y) \log \frac{p(y \mid x)}{p(y)} \, dy \, dx. \tag{20}$$

To make the density ratio explicit, we define the ratio between the conditional distribution and the marginal distribution:

$$r(x, y) = \frac{p(y \mid x)}{p(y)}. \tag{21}$$

Using this definition, mutual information becomes the expected log-density ratio between true paired samples and independently sampled negatives:

$$I(X_S; Y_S) = \int_{\mathcal{X}_S} \int_{\mathcal{Y}_S} p(x, y) \log r(x, y) \, dy \, dx. \tag{22}$$

Following the contrastive predictive coding and InfoNCE formulation [47], we convert density-ratio estimation into a B-way contrastive classification problem:

$$y_{1:B} = (y_1, \ldots, y_B). \tag{23}$$

To indicate which candidate is paired with the anchor, we introduce a uniformly distributed positive index:

$$I \in \{1, \ldots, B\}, \qquad p(I = i) = \frac{1}{B}. \tag{24}$$

Conditioned on the positive index $I = i$, the positive candidate is sampled jointly with x, while all other candidates are sampled independently from the marginal distribution:

$$p(x, y_{1:B} \mid I = i) = p(x, y_i) \prod_{j \neq i} p(y_j). \tag{25}$$

To derive the Bayes-optimal classifier for identifying the positive sample, we apply Bayes' rule [4]:

$$p(I = i \mid x, y_{1:B}) = \frac{p(x, y_i) \prod_{j \neq i} p(y_j)}{\sum_{k=1}^{B} p(x, y_k) \prod_{j \neq k} p(y_j)}. \tag{26}$$

To reveal the density-ratio structure inside the posterior, we multiply and divide each term by the common factor $p(x) \prod_{j=1}^{B} p(y_j)$:

$$p(I = i \mid x, y_{1:B}) = \frac{\frac{p(x, y_i)}{p(x)p(y_i)}}{\sum_{k=1}^{B} \frac{p(x, y_k)}{p(x)p(y_k)}}. \tag{27}$$

Using the definition of $r(x, y)$, the Bayes-optimal posterior becomes a normalized density ratio over all candidates:

$$p(I = i \mid x, y_{1:B}) = \frac{r(x, y_i)}{\sum_{k=1}^{B} r(x, y_k)}. \tag{28}$$

Since the true density ratio is unavailable in practice, we approximate its logarithm with a trainable score function [47]:

$$s_\omega(x, y) \approx \log r(x, y). \tag{29}$$

This approximation induces a model posterior over the positive index:

$$q_\omega(I = i \mid x, y_{1:B}) = \frac{\exp s_\omega(x, y_i)}{\sum_{k=1}^{B} \exp s_\omega(x, y_k)}. \tag{30}$$

In implementation, the score is instantiated by a temperature-scaled similarity function:

$$s_\omega(x, y) = \frac{\mathrm{sim}(x, y)}{\tau}. \tag{31}$$

For mini-batch training, the visual-to-text contrastive loss encourages each visual shared representation to identify its paired textual shared representation:

$$\mathcal{L}_{\mathrm{NCE}}^{V\to T} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\mathrm{sim}(z_{S,i}^V, z_{S,i}^T)/\tau\big)}{\sum_{j=1}^{B} \exp\Big(\mathrm{sim}(z_{S,i}^V, z_{S,j}^T)/\tau\Big)}. \tag{32}$$

Symmetrically, the text-to-visual contrastive loss encourages each textual shared representation to identify its paired visual shared representation:

$$\mathcal{L}_{\mathrm{NCE}}^{T\to V} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\mathrm{sim}(z_{S,i}^T, z_{S,i}^V)/\tau\big)}{\sum_{j=1}^{B} \exp\Big(\mathrm{sim}(z_{S,i}^T, z_{S,j}^V)/\tau\Big)}. \tag{33}$$

To avoid directional bias, we define the shared-alignment loss as the average of the two contrastive directions:

$$\mathcal{L}_{\mathrm{share}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{NCE}}^{V\to T} + \mathcal{L}_{\mathrm{NCE}}^{T\to V}\right). \tag{34}$$

To connect the contrastive loss with mutual information, we define the population contrastive cross-entropy:

$$\mathcal{C}_\omega = -\sum_{i=1}^{B} \int p(i, x, y_{1:B}) \log q_\omega(I = i \mid x, y_{1:B})\, dx\, dy_{1:B}. \tag{35}$$

To show that this cross-entropy is minimized by the true posterior, we decompose it into the conditional entropy plus a KL divergence:

$$\mathcal{C}_\omega = H(I \mid X_S, Y_{S,1:B}) + \int p(x, y_{1:B}) D_{\mathrm{KL}}\left(p(I \mid x, y_{1:B}) \parallel q_\omega(I \mid x, y_{1:B})\right)\, dx\, dy_{1:B}. \tag{36}$$

Because KL divergence is nonnegative, the contrastive cross-entropy upper-bounds the conditional entropy of the positive index:

$$\mathcal{C}_\omega \geq H(I \mid X_S, Y_{S,1:B}). \tag{37}$$

Since the positive index is uniformly sampled from B candidates, its entropy is:

$$H(I) = \log B. \tag{38}$$

Subtracting the contrastive cross-entropy from $\log B$ gives a lower bound on the information revealed about the positive index:

$$\log B - \mathcal{C}_\omega \leq H(I) - H(I \mid X_S, Y_{S,1:B}). \tag{39}$$

The right-hand side is exactly the mutual information between the positive index and the observed anchor-candidate set:

$$H(I) - H(I \mid X_S, Y_{S,1:B}) = I(I; X_S, Y_{S,1:B}). \tag{40}$$

Under the standard contrastive sampling construction [47], this positive-index information cannot exceed the mutual information between the paired views:

$$I(I; X_S, Y_{S,1:B}) \leq I(X_S; Y_S). \tag{41}$$

Combining the above inequalities gives the InfoNCE-style lower bound:

$$I(X_S; Y_S) \geq \log B - \mathcal{C}_\omega. \tag{42}$$

Returning to the original shared semantic variables, minimizing the contrastive loss increases a lower-bound surrogate of the shared mutual information:

$$I(Z_S^V; Z_S^T) \geq \log B - \mathcal{C}_\omega. \tag{43}$$

Therefore, the shared-alignment loss provides a tractable optimization signal for preserving cross-modal common semantics.

A.3 Asymptotic Connection from dCor to Mutual Information

We provide a conditional asymptotic justification showing that, under standard regularity assumptions, driving distance correlation to zero also drives mutual information to zero. We present the derivation for one semantic pair, namely the visual-specific representation Z_V and the shared representation Z_S. The same argument applies to (Z_T, Z_S) and (Z_V, Z_T).

Consider a sequence of learned semantic representations $\{(Z_{V,n}, Z_{S,n})\}_{n=1}^{\infty}$. To compare the actual joint distribution with the distribution under independence, we define

$$P_n = P_{Z_{V,n} Z_{S,n}}, \qquad Q_n = P_{Z_{V,n}} P_{Z_{S,n}}. \tag{44}$$

Here, P_n denotes the joint distribution of $Z_{V,n}$ and $Z_{S,n}$, while Q_n denotes the product of their marginal distributions.

Let $p_n(z_V, z_S)$ and $q_n(z_V, z_S)$ be the densities of P_n and Q_n, respectively. Since Q_n corresponds to the independent coupling of the two marginals, its density factorizes as

$$q_n(z_V, z_S) = p_{V,n}(z_V) p_{S,n}(z_S). \tag{45}$$

We assume that these distributions are supported on a common compact set, have finite first moments, and belong to an L_1-compact regular density class. To prevent the KL divergence from being dominated by arbitrarily sharp density ratios, we further assume the following uniform density-ratio bound:

$$0 \leq \frac{p_n(z_V, z_S)}{q_n(z_V, z_S)} \leq M, \qquad M < \infty. \tag{46}$$

Under these assumptions, our goal is to show that convergence of distance correlation to zero implies convergence of mutual information to zero:

$$\mathrm{dCor}(Z_{V,n}, Z_{S,n}) \to 0 \quad \Longrightarrow \quad I(Z_{V,n}; Z_{S,n}) \to 0. \tag{47}$$

We first connect distance correlation with distance covariance. By definition, distance correlation is the normalized form of distance covariance:

$$\mathrm{dCor}^2(Z_{V,n}, Z_{S,n}) = \frac{\mathrm{dCov}^2(Z_{V,n}, Z_{S,n})}{\sqrt{\mathrm{dCov}^2(Z_{V,n}, Z_{V,n})\, \mathrm{dCov}^2(Z_{S,n}, Z_{S,n})}}. \tag{48}$$

Because the representations lie in a common compact set, the self-distance covariances in the denominator are uniformly bounded. Therefore, if the normalized dependence vanishes, the unnormalized distance covariance also vanishes:

$$\mathrm{dCor}(Z_{V,n}, Z_{S,n}) \to 0 \quad \Longrightarrow \quad \mathrm{dCov}(Z_{V,n}, Z_{S,n}) \to 0. \tag{49}$$

Next, we use the independence-characterizing property of distance covariance [44]. Distance covariance is zero if and only if the joint distribution factorizes into the product of its marginals:

$$\mathrm{dCov}(Z_V, Z_S) = 0 \quad \Longleftrightarrow \quad P_{Z_V Z_S} = P_{Z_V} P_{Z_S} \quad \Longleftrightarrow \quad Z_V \perp Z_S. \tag{50}$$

Since the density class is L_1-compact and the distance-covariance functional is continuous on this regular class, convergence of distance covariance to zero implies that the joint distribution approaches the product-marginal distribution in total variation [41]:

$$\|P_n - Q_n\|_{\mathrm{TV}} \to 0. \tag{51}$$

We now connect total variation convergence with mutual information. Mutual information is the KL divergence between the joint distribution and the product-marginal distribution:

$$I(Z_{V,n}; Z_{S,n}) = D_{\mathrm{KL}}(P_n \| Q_n) = \int p_n(z_V, z_S) \log \frac{p_n(z_V, z_S)}{q_n(z_V, z_S)} \, dz_S \, dz_V. \tag{52}$$

To express the KL divergence in terms of the deviation from independence, we define the density ratio between P_n and Q_n:

$$r_n(z_V, z_S) = \frac{p_n(z_V, z_S)}{q_n(z_V, z_S)}. \tag{53}$$

Since both p_n and q_n are probability densities, the signed deviation of the density ratio from one has zero integral under q_n:

$$\int q_n(z_V, z_S)\big(r_n(z_V, z_S) - 1\big) \, dz_S \, dz_V = \int p_n(z_V, z_S) \, dz_S \, dz_V - \int q_n(z_V, z_S) \, dz_S \, dz_V = 0. \tag{54}$$

Using the zero-mean identity above, we can subtract $r_n - 1$ inside the KL integrand without changing its value:

$$D_{\mathrm{KL}}(P_n \| Q_n) = \int q_n(z_V, z_S) \left[r_n(z_V, z_S) \log r_n(z_V, z_S) - r_n(z_V, z_S) + 1 \right] \, dz_S \, dz_V. \tag{55}$$

The bounded density-ratio assumption ensures that the nonlinear KL integrand can be controlled by a linear deviation term, consistent with reverse-Pinsker type arguments for upper-bounding relative entropy by total variation under additional regularity conditions [40]. Thus, we define the finite constant

$$C_M = \sup_{0 \le r \le M,\, r \ne 1} \frac{r \log r - r + 1}{|r - 1|} < \infty. \tag{56}$$

By the definition of C_M, the centered KL integrand is upper-bounded by the absolute deviation from independence:

$$r \log r - r + 1 \le C_M |r - 1|. \tag{57}$$

Substituting this bound into the KL expression gives an upper bound in terms of the L_1 distance between p_n and q_n:

$$\begin{aligned} I(Z_{V,n}; Z_{S,n}) &= D_{\mathrm{KL}}(P_n \| Q_n) \\ &\le C_M \int q_n(z_V, z_S) \left| \frac{p_n(z_V, z_S)}{q_n(z_V, z_S)} - 1 \right| \, dz_S \, dz_V \\ &= C_M \int |p_n(z_V, z_S) - q_n(z_V, z_S)| \, dz_S \, dz_V \\ &= 2 C_M \|P_n - Q_n\|_{\mathrm{TV}}. \end{aligned} \tag{58}$$

Since the total variation distance has been shown to converge to zero, the upper bound above directly implies that the mutual information also converges to zero:

$$I(Z_{V,n}; Z_{S,n}) \to 0. \tag{59}$$

Combining the previous implications gives the full asymptotic chain from distance correlation to mutual information:

$$\begin{aligned} \mathrm{dCor}(Z_{V,n}, Z_{S,n}) \to 0 &\Longrightarrow \mathrm{dCov}(Z_{V,n}, Z_{S,n}) \to 0 \\ &\Longrightarrow \|P_n - Q_n\|_{\mathrm{TV}} \to 0 \\ &\Longrightarrow I(Z_{V,n}; Z_{S,n}) \to 0. \end{aligned} \tag{60}$$

Therefore, under the stated compactness, continuity, and bounded density-ratio conditions, minimizing distance correlation is asymptotically consistent with reducing the mutual information between Z_V and Z_S:

$$\mathrm{dCor}(Z_{V,n}, Z_{S,n}) \to 0 \quad \Longrightarrow \quad I(Z_{V,n}; Z_{S,n}) \to 0. \tag{61}$$

A.4 Cross-Entropy as an Upper Bound on Conditional Entropy

After disentanglement, the decomposed variables should still preserve information useful for question generation, so we group them with the constraint variable:

$$W = (Z_V, Z_T, Z_S, C). \tag{62}$$

To measure the irreducible uncertainty of the target question under the true data distribution, we define the population conditional entropy:

$$H_{p^\star}(Q \mid W) = -\int_{\mathcal{W}} p(w) \sum_{q \in \mathcal{Q}} p^\star(q \mid w) \log p^\star(q \mid w)\, dw. \tag{63}$$

Since the true conditional distribution is unknown, training instead minimizes the model cross-entropy:

$$H(p^\star, p_\theta \mid W) = -\int_{\mathcal{W}} p(w) \sum_{q \in \mathcal{Q}} p^\star(q \mid w) \log p_\theta(q \mid w)\, dw. \tag{64}$$

To show that cross-entropy is a valid surrogate, we compare it with the true conditional entropy:

$$H(p^\star, p_\theta \mid W) - H_{p^\star}(Q \mid W) = \int_{\mathcal{W}} p(w) \sum_{q \in \mathcal{Q}} p^\star(q \mid w) \log \frac{p^\star(q \mid w)}{p_\theta(q \mid w)}\, dw. \tag{65}$$

The difference is an average KL divergence between the true conditional distribution and the model conditional distribution:

$$H(p^\star, p_\theta \mid W) - H_{p^\star}(Q \mid W) = \int_{\mathcal{W}} p(w) D_{\mathrm{KL}}\left(p^\star(\cdot \mid w) \| p_\theta(\cdot \mid w)\right)\, dw. \tag{66}$$

Because KL divergence is nonnegative, model cross-entropy upper-bounds the true conditional entropy:

$$H_{p^\star}(Q \mid W) \le H(p^\star, p_\theta \mid W). \tag{67}$$

Substituting the semantic state $W = (Z_V, Z_T, Z_S, C)$ gives the upper bound used in our generation objective:

$$H_{p^\star}(Q \mid Z_V, Z_T, Z_S, C) \le H(p^\star, p_\theta \mid Z_V, Z_T, Z_S, C). \tag{68}$$

Thus, minimizing generation cross-entropy helps prevent the disentangled representations from becoming separated but uninformative.

A.5 Mutual-Information-based Synergistic Reward

After obtaining informative and separated semantic variables, we derive the synergistic reward from a structured partial information decomposition (PID) view [5, 25, 51, 52]. In our semantic decomposition, Z_V denotes visual-specific semantics, Z_T denotes textual-specific semantics, and Z_S denotes shared cross-modal semantics. Therefore, the visual semantic source can be represented as (Z_V, Z_S), while the textual semantic source can be represented as (Z_T, Z_S).

Under the PID formulation, the information that the two semantic sources jointly provide about the generated question Q can be decomposed as:

$$\begin{aligned} &I(Q; Z_V, Z_T, Z_S \mid C) \\ =&I_\cap\big(Q; (Z_V, Z_S), (Z_T, Z_S) \mid C\big) + I^V_{\mathrm{uniq}} + I^T_{\mathrm{uniq}} + I_{\mathrm{syn}}(Q; Z_V, Z_T \mid Z_S, C), \end{aligned} \tag{69}$$

where $I_\cap$ denotes the redundant information shared by the two semantic sources, I^V_{uniq} and I^T_{uniq} denote the visual-unique and textual-unique information, and I_{syn} denotes the synergistic information.

The single-source dependencies are:

$$I(Q; Z_V, Z_S \mid C) = I_\cap\big(Q; (Z_V, Z_S), (Z_T, Z_S) \mid C\big) + I^V_{\mathrm{uniq}}, \tag{70}$$

$$I(Q; Z_T, Z_S \mid C) = I_\cap\big(Q; (Z_V, Z_S), (Z_T, Z_S) \mid C\big) + I^T_{\text{uniq}}. \tag{71}$$

Therefore, the PID synergistic component can be written as:

$$\begin{aligned} I_{\text{syn}}(Q; Z_V, Z_T \mid Z_S, C) =& I(Q; Z_V, Z_T, Z_S \mid C) - I(Q; Z_V, Z_S \mid C) \\ & - I(Q; Z_T, Z_S \mid C) + I_\cap\big(Q; (Z_V, Z_S), (Z_T, Z_S) \mid C\big). \end{aligned} \tag{72}$$

In our structured semantic decomposition, the redundant cross-modal information is explicitly represented by the shared semantic variable Z_S. Thus, we instantiate the PID redundancy term as:

$$I_\cap\big(Q; (Z_V, Z_S), (Z_T, Z_S) \mid C\big) \triangleq I(Q; Z_S \mid C). \tag{73}$$

Substituting this redundancy instantiation into the PID synergy gives:

$$\begin{aligned} I_{\text{syn}}(Q; Z_V, Z_T \mid Z_S, C) =& I(Q; Z_V, Z_T, Z_S \mid C) - I(Q; Z_V, Z_S \mid C) \\ & - I(Q; Z_T, Z_S \mid C) + I(Q; Z_S \mid C). \end{aligned} \tag{74}$$

Using the chain rule of mutual information, we have:

$$I(Q; Z_V, Z_T, Z_S \mid C) = I(Q; Z_S \mid C) + I(Q; Z_V, Z_T \mid Z_S, C), \tag{75}$$

$$I(Q; Z_V, Z_S \mid C) = I(Q; Z_S \mid C) + I(Q; Z_V \mid Z_S, C), \tag{76}$$

$$I(Q; Z_T, Z_S \mid C) = I(Q; Z_S \mid C) + I(Q; Z_T \mid Z_S, C). \tag{77}$$

Substituting the above three identities cancels the shared redundancy term and yields:

$$\text{Syn}(Q; Z_V, Z_T \mid Z_S, C) = I(Q; Z_V, Z_T \mid Z_S, C) - I(Q; Z_V \mid Z_S, C) - I(Q; Z_T \mid Z_S, C). \tag{78}$$

This formulation shows that, after conditioning on the shared semantic variable Z_S, the remaining synergistic signal corresponds to the joint visual-textual dependency that cannot be explained by either modality-specific semantic stream alone.

Since Q is discrete and the semantic variables are continuous, the joint dependency term is written as a mixed-variable conditional mutual information [11]:

$$\begin{aligned} & I(Q; Z_V, Z_T \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_V} \int_{\mathcal{Z}_T} p(z_v, z_t, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_t, z_s, c) \log \frac{p(q \mid z_v, z_t, z_s, c)}{p(q \mid z_s, c)} \, dz_t \, dz_v \, dz_s \, dc. \end{aligned} \tag{79}$$

The visual-only dependency measures the information contributed by visual-private semantics alone:

$$\begin{aligned} & I(Q; Z_V \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_V} p(z_v, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_s, c) \log \frac{p(q \mid z_v, z_s, c)}{p(q \mid z_s, c)} \, dz_v \, dz_s \, dc. \end{aligned} \tag{80}$$

To subtract this visual-only term from the joint term, we rewrite it under the same joint measure:

$$p(z_v, z_s, c) p(q \mid z_v, z_s, c) = \int_{\mathcal{Z}_T} p(z_v, z_t, z_s, c) p(q \mid z_v, z_t, z_s, c) \, dz_t. \tag{81}$$

Using this identity, the visual-only dependency can be expressed with the same integration variables as the joint dependency:

$$\begin{aligned} & I(Q; Z_V \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_V} \int_{\mathcal{Z}_T} p(z_v, z_t, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_t, z_s, c) \log \frac{p(q \mid z_v, z_s, c)}{p(q \mid z_s, c)} \, dz_t \, dz_v \, dz_s \, dc. \end{aligned} \tag{82}$$

The text-only dependency measures the information contributed by text-private semantics alone:

$$\begin{aligned} & I(Q; Z_T \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_T} p(z_t, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_t, z_s, c) \log \frac{p(q \mid z_t, z_s, c)}{p(q \mid z_s, c)} \, dz_t \, dz_s \, dc. \end{aligned} \tag{83}$$

To subtract the text-only term from the joint term, we similarly rewrite it under the joint measure:

$$p(z_t, z_s, c)p(q \mid z_t, z_s, c) = \int_{\mathcal{Z}_V} p(z_v, z_t, z_s, c)p(q \mid z_v, z_t, z_s, c)\, dz_v. \tag{84}$$

Using this identity, the text-only dependency can also be expressed with the same integration variables:

$$\begin{aligned} &I(Q; Z_T \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_V} \int_{\mathcal{Z}_T} p(z_v, z_t, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_t, z_s, c) \log \frac{p(q \mid z_t, z_s, c)}{p(q \mid z_s, c)}\, dz_t\, dz_v\, dz_s\, dc. \end{aligned} \tag{85}$$

Now that all three terms share the same measure, substituting them into the synergy definition yields a four-term log-likelihood interaction:

$$\begin{aligned} &\mathrm{Syn}(Q; Z_V, Z_T \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_V} \int_{\mathcal{Z}_T} p(z_v, z_t, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_t, z_s, c) \\ &\quad \cdot \Big[\log p(q \mid z_v, z_t, z_s, c) - \log p(q \mid z_v, z_s, c) \\ &\qquad - \log p(q \mid z_t, z_s, c) + \log p(q \mid z_s, c) \Big]\, dz_t\, dz_v\, dz_s\, dc. \end{aligned} \tag{86}$$

For a concrete generated question, we define the pointwise synergistic score as the inner four-term log interaction:

$$\begin{aligned} \delta_{\mathrm{syn}}(q; z_v, z_t, z_s, c) = &\log p(q \mid z_v, z_t, z_s, c) - \log p(q \mid z_v, z_s, c) \\ &- \log p(q \mid z_t, z_s, c) + \log p(q \mid z_s, c). \end{aligned} \tag{87}$$

Equivalently, this pointwise score can be written as a log-ratio between joint support and individual-modal support:

$$\delta_{\mathrm{syn}}(q; z_v, z_t, z_s, c) = \log \frac{p(q \mid z_v, z_t, z_s, c)p(q \mid z_s, c)}{p(q \mid z_v, z_s, c)p(q \mid z_t, z_s, c)}. \tag{88}$$

Using the pointwise score, the population-level synergy becomes the average pointwise interaction under the full conditional question distribution:

$$\begin{aligned} &\mathrm{Syn}(Q; Z_V, Z_T \mid Z_S, C) \\ &= \int_{\mathcal{C}} \int_{\mathcal{Z}_S} \int_{\mathcal{Z}_V} \int_{\mathcal{Z}_T} p(z_v, z_t, z_s, c) \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_t, z_s, c)\delta_{\mathrm{syn}}(q; z_v, z_t, z_s, c)\, dz_t\, dz_v\, dz_s\, dc. \end{aligned} \tag{89}$$

For a fixed input instance, we write the local semantic state as:

$$x = (z_v, z_t, z_s, c). \tag{90}$$

The local synergistic contribution is the expected pointwise score over all possible generated questions:

$$\mathrm{Syn}_x = \sum_{q \in \mathcal{Q}} p(q \mid z_v, z_t, z_s, c)\delta_{\mathrm{syn}}(q; z_v, z_t, z_s, c). \tag{91}$$

Because the question space is a large discrete sequence space, we approximate the expectation by Monte Carlo samples from the current policy:

$$q^{(g)} \sim p_\theta^Q(\cdot \mid z_V, z_T, z_S, C), \qquad g = 1, \ldots, G. \tag{92}$$

The Monte Carlo estimate of the local synergy is the average pointwise score over sampled questions [33]:

$$\widehat{\mathrm{Syn}}_x = \frac{1}{G} \sum_{g=1}^{G} \delta_{\mathrm{syn}}(q^{(g)}; z_v, z_t, z_s, c). \tag{93}$$

To implement the four-term interaction with ablated inputs, we first define the full semantic state:

$$x_{\mathrm{VTS}} = (z_V, z_T, z_S, C). \tag{94}$$

We then define three reduced semantic states by removing text-private, visual-private, or both private components:

$$x_{\mathrm{VS}} = (z_V, \varnothing_T, z_S, C), \qquad x_{\mathrm{TS}} = (\varnothing_V, z_T, z_S, C), \qquad x_{\mathrm{S}} = (\varnothing_V, \varnothing_T, z_S, C). \tag{95}$$

Following common variational and ablation-based approximation practice for intractable conditional distributions [17], the theoretical pointwise score is instantiated by the current model likelihood under the full and reduced semantic states:

$$\begin{aligned}\delta_{\mathrm{syn},\theta}(q^{(g)}; x_{\mathrm{VTS}}, x_{\mathrm{VS}}, x_{\mathrm{TS}}, x_{\mathrm{S}}) = &\log p_\theta^Q(q^{(g)} \mid x_{\mathrm{VTS}}) - \log p_\theta^Q(q^{(g)} \mid x_{\mathrm{VS}}) \\ &- \log p_\theta^Q(q^{(g)} \mid x_{\mathrm{TS}}) + \log p_\theta^Q(q^{(g)} \mid x_{\mathrm{S}}).\end{aligned} \tag{96}$$

Since $q^{(g)}$ is a sequence, its likelihood is computed as the sum of token-level log-likelihoods:

$$\log p_\theta^Q(q^{(g)} \mid x) = \sum_{\ell=1}^{|q^{(g)}|} \log p_\theta\left(q_\ell^{(g)} \mid q_{<\ell}^{(g)}, x\right). \tag{97}$$

For each sampled question, the sample-level synergistic score is defined as the model-based pointwise interaction:

$$\mathrm{Syn}_\theta^{(g)} = \delta_{\mathrm{syn},\theta}(q^{(g)}; x_{\mathrm{VTS}}, x_{\mathrm{VS}}, x_{\mathrm{TS}}, x_{\mathrm{S}}). \tag{98}$$

Averaging over sampled questions gives the empirical synergistic estimator:

$$\widehat{\mathrm{Syn}}_\theta = \frac{1}{G}\sum_{g=1}^{G} \mathrm{Syn}_\theta^{(g)}. \tag{99}$$

To convert the synergistic score into an optimization signal, we enforce a margin on the sampled question-level score:

$$R_{\mathrm{syn}}^{(g)} = \phi\left(\mathrm{Syn}_\theta^{(g)} - \gamma_{\mathrm{syn}}\right), \tag{100}$$

where γ_{syn} is the synergy margin and $\phi(\cdot)$ is a monotonic shaping function.

To balance structural validity, semantic similarity, and synergistic dependency, the final reward is defined as a sum of three reward components:

$$R^{(g)} = R_{\mathrm{rule}}^{(g)} + R_{\mathrm{sim}}^{(g)} + R_{\mathrm{syn}}^{(g)}. \tag{101}$$

The policy optimization objective maximizes the expected reward over inputs and generated questions:

$$\max_\theta \int_{\mathcal{X}} p(x) \sum_{q \in \mathcal{Q}} p_\theta^Q(q \mid x_{\mathrm{VTS}}) R(q, x)\, dx. \tag{102}$$

The reward function for a generated question combines rule-based quality, similarity, and the synergy-margin term:

$$R(q, x) = \lambda_{\mathrm{rule}} R_{\mathrm{rule}}(q) + \lambda_{\mathrm{sim}} R_{\mathrm{sim}}(q) + \lambda_{\mathrm{syn}} \phi\left(\mathrm{Syn}_\theta(q, x) - \gamma_{\mathrm{syn}}\right). \tag{103}$$

Therefore, the reward encourages the generated question to exceed the required synergy margin:

$$\mathrm{Syn}_\theta(q, x) > \gamma_{\mathrm{syn}}. \tag{104}$$

This means that a preferred question should receive strong support from the joint visual-textual semantic state, while its likelihood should decrease when either modality-specific component is removed.

B Experimental Details

B.1 Implementation Details

We use Qwen2.5-VL-7B-Instruct [2] as the backbone model for SynMQG. To ensure a fair comparison, all baselines are implemented with the same backbone. All experiments are conducted with random seed of 42. We use AdamW [29] as the optimizer and train all models with bfloat16 precision on a single NVIDIA A800 80GB GPU. All methods are fine-tuned with LoRA for parameter-efficient training [20].

SynMQG is trained in two stages. In the first stage, we perform supervised fine-tuning to learn grounded reasoning-question generation and the decomposed semantic space. The per-device batch size is set to 8, with gradient accumulation steps of 16. The learning rate is 1×10^{-5}, and the warmup ratio is 0.1. We train the first stage for 30 epochs. The maximum source length is set to 2048, and the maximum target length is set to 512. In the second stage, we initialize the policy from the supervised checkpoint and apply GRPO with the proposed dependency-preserving reward. During this stage, only the LoRA parameters are updated, while the backbone parameters remain frozen. The per-device batch size is set to 4, with gradient accumulation steps of 8. The GRPO sampling group size is set to 8, and the sampling temperature is 0.7. The learning rate is 1×10^{-6}. We train the second stage for 5 epochs, corresponding to more than 1000 policy optimization steps. The GRPO clipping threshold is set to $\eta = 0.2$, and the KL coefficient is set to $\beta = 0.1$.

B.2 Data Processing

We construct the experimental data from ScienceQA [30] and MultimodalQA [45]. Since our goal is synergistic multi-modal question generation, we first filter the original datasets to retain samples that contain both visual and textual information. Samples without visual input or without meaningful textual context are removed. For images without human-written captions, we use Qwen-VL to generate auxiliary image captions. These captions are used only during data processing to help the filtering model understand the visual content. They are not used to replace the original image during model training or evaluation.

After this preliminary filtering, we use GPT-4o to analyze the original question and determine whether answering it requires both modalities. Specifically, GPT-4o is given the image, textual context, generated caption if available, original question, and reference answer. It then produces a reasoning result with a prefix label indicating the modality dependency type. We retain samples whose reasoning prefix is labeled as multi-modal. The detailed filtering pseudo-code is provided in Appendix H. After filtering, we obtain 6080 multi-modal examples from ScienceQA and 4571 multi-modal examples from MultimodalQA. These filtered samples are used for training and evaluation.

B.3 Baseline Details

VL-T5. VL-T5 formulates vision-and-language tasks as text generation problems and unifies different multi-modal tasks under a sequence-to-sequence framework [10]. Since VL-T5 was originally designed as a general vision-language generation model rather than specifically for synergistic MQG, we adapt it by converting MQG into a conditional text generation task. Given the image, textual context, and auxiliary constraints, the model is trained to directly generate the target question. Under our implementation, VL-T5 serves as a direct vision-language generation baseline.

MultiQG-TI. MultiQG-TI is a multi-modal question generation method that generates questions from textual and image-based sources [48]. It is closely related to our task because it directly studies question generation from multi-modal inputs. We adapt MultiQG-TI to our unified backbone by conditioning Qwen2.5-7B-Instruct on the same image-text input and training it with supervised question generation. This baseline represents the standard MQG setting where the model learns to generate questions from multi-modal sources.

MMCoT. MMCoT introduces multimodal chain-of-thought reasoning in language models, where the model first performs intermediate reasoning over multi-modal inputs before producing the final answer [57]. Since MMCoT is originally designed for multi-modal reasoning rather than question generation, we convert it into an MQG baseline by changing the final output from an answer to

a generated question. Specifically, the model is prompted or trained to first produce a reasoning chain based on the image and text, and then generate a question according to the reasoning process. This baseline allows us to examine whether adding a chain-of-thought reasoning process alone can improve MQG.

CoE. Chain-of-Exemplar enhances multimodal educational question generation by using exemplar-based reasoning and generation [32]. We adapt CoE by providing representative exemplars in the input context and asking the model to generate a question for the current multi-modal input. The exemplar-guided setting helps the model imitate question patterns from similar examples. This baseline is used to test whether example-based prompting or training can improve question generation quality.

B.4 Evaluation Metrics

B.4.1 Automatic Metrics

BLEU. BLEU measures the n-gram precision between generated text and reference text, with a brevity penalty to discourage overly short outputs [39]. In MQG, BLEU reflects how much lexical overlap the generated question has with the reference question. However, because different valid questions may use different surface expressions, BLEU mainly captures surface-level similarity and may underestimate semantically correct paraphrases.

ROUGE-L. ROUGE-L is based on the longest common subsequence between the generated question and the reference question [26]. It captures sentence-level overlap while allowing non-contiguous matching. Compared with BLEU, ROUGE-L is more recall-oriented and can reflect whether the generated question preserves important reference content.

METEOR. METEOR evaluates text generation by considering unigram matching, stemming, synonym matching, and word order penalty [3]. Because METEOR allows more flexible lexical matching than exact n-gram overlap, it is useful for evaluating generated questions that are semantically close to the reference but expressed with different words.

BERTScore. BERTScore computes semantic similarity between generated and reference text using contextualized token embeddings [56]. It measures token-level semantic alignment rather than exact lexical overlap. Therefore, BERTScore is more robust to paraphrasing and is suitable for evaluating question generation, where multiple valid formulations may exist.

BLEURT. BLEURT is a learned evaluation metric trained to predict human judgments of text generation quality [42]. Compared with purely lexical metrics, BLEURT can better capture semantic adequacy and fluency. We include BLEURT as a learned reference-based metric to complement BLEU, ROUGE-L, METEOR, and BERTScore.

B.4.2 Dependency-aware Multi-dimensional Evaluation

Although reference-based metrics are useful for measuring similarity to the gold question, they cannot directly determine whether a generated question genuinely depends on both modalities. For example, a question may obtain high lexical similarity while still being answerable from the visual input alone or from the textual input alone. Therefore, we further conduct dependency-aware multi-dimensional evaluation.

Following the idea of multi-dimensional question generation evaluation in QGEval [16], we design five evaluation dimensions for synergistic MQG: Visual Necessity, Textual Necessity, Cross-modal Complementarity, Synergistic Interaction Gain, and Cross-modal Grounding. All dimensions are evaluated using Gemini-2.5-Pro under a unified evaluation protocol. For modality-dependency dimensions, Gemini-2.5-Pro evaluates the answerability of the same generated question under four input conditions: visual-textual input $(V + T)$, visual-only input (V), textual-only input (T), and context-only input (C). For complementarity and grounding dimensions, Gemini-2.5-Pro directly evaluates whether the question is supported by distinct and linked visual-textual evidence. The full evaluation prompts are provided in Appendix J.

To make the evaluation procedure explicit, let s_{V+T}, s_V, s_T, and s_C denote the Gemini-2.5-Pro answerability scores of the same question under the visual-textual, visual-only, textual-only, and context-only settings, respectively. Each answerability score is assigned on a 1–5 scale, where a higher score indicates that the question can be more reliably answered under the corresponding input condition. We compute positive dependency gains using $[x]_+ = \max(x, 0)$. Therefore, the gain-based dimensions, including Visual Necessity, Textual Necessity, and Synergistic Interaction Gain, lie in the range $[0, 4]$. In contrast, Cross-modal Complementarity and Cross-modal Grounding are directly rated on the 1–5 scale.

Visual Necessity. Visual Necessity measures whether the generated question requires visual evidence. It is computed as the answerability gain obtained by adding the visual input to the textual-only condition:

$$\mathrm{VN} = [s_{V+T} - s_T]_+. \tag{105}$$

A high score means that removing the visual input would make the question difficult or impossible to answer correctly. This dimension tests whether the question is genuinely grounded in the visual modality rather than being answerable from text alone.

Textual Necessity. Textual Necessity measures whether the generated question requires textual evidence. It is computed as the answerability gain obtained by adding the textual input to the visual-only condition:

$$\mathrm{TN} = [s_{V+T} - s_V]_+. \tag{106}$$

A high score means that removing the textual context would make the question difficult or impossible to answer correctly. This dimension prevents the model from generating questions that rely only on visual cues while ignoring the text.

Cross-modal Complementarity. Cross-modal Complementarity measures whether the visual and textual inputs provide distinct but mutually necessary evidence for answering the generated question. Different from VN and TN, which are computed from answerability differences, complementarity is directly rated by Gemini-2.5-Pro on a 1–5 scale according to whether the two modalities provide non-redundant evidence and whether both are needed to support the answer:

$$\mathrm{Comp} = s_{\mathrm{comp}}, \tag{107}$$

where $s_{\mathrm{comp}} \in \{1, 2, 3, 4, 5\}$ denotes the direct complementarity score. A high score means that the question requires combining different evidence from the visual and textual inputs, rather than relying on repeated or substitutable information.

Synergistic Interaction Gain. Synergistic Interaction Gain measures whether the complete visual-textual input provides more useful support than the two single-modality conditions considered independently. Inspired by an interaction-style dependency term, we compute it as:

$$\mathrm{SIG} = [s_{V+T} - s_V - s_T + s_C]_+. \tag{108}$$

A high score indicates that the generated question benefits from the interaction between visual and textual evidence, rather than from independent single-modality cues. This dimension is most directly aligned with our goal of synergistic multi-modal question generation.

Cross-modal Grounding. Cross-modal Grounding measures whether the generated question can be linked to concrete evidence in both modalities. Gemini-2.5-Pro is asked to identify the supporting visual evidence, the supporting textual evidence, and their cross-modal relation. Let g_V and g_T denote the visual-grounding and textual-grounding scores on the 1–5 scale. We compute cross-modal grounding as:

$$\mathrm{CMG} = \min(g_V, g_T). \tag{109}$$

A high score means that the question is grounded in specific visual evidence and specific textual evidence, and that the two pieces of evidence are meaningfully connected.

Overall Score. In addition to the five diagnostic dimensions, we report an Overall score to summarize the general quality of each generated question under the dependency-aware evaluation protocol. The Overall score is directly rated by Gemini-2.5-Pro on a 1–5 scale by considering question validity, answerability, multi-modal dependency, and grounding quality:

$$\text{Overall} = s_{\text{overall}}, \tag{110}$$

where $s_{\text{overall}} \in \{1, 2, 3, 4, 5\}$. We report the average score over all evaluated samples. Since Overall is a direct holistic judgment rather than the arithmetic mean of VN, TN, Comp, SIG, and CMG, it is reported as a separate evaluation dimension in the figures.

B.4.3 Human Evaluation

To further assess whether the generated questions satisfy human preferences, we conduct a human evaluation based on N-way preference comparison. We invite 10 domain-aware annotators to participate in the evaluation through a structured questionnaire. For each test instance, evaluators are given the image, the textual context, the reference answer, and multiple candidate questions generated by different methods. Instead of assigning absolute scores to each candidate, evaluators are asked to select the best question among all candidates. This comparison-based setting reduces score calibration bias and directly reflects which method produces more preferable questions under the same multi-modal input.

We evaluate candidate questions from two complementary perspectives: *multi-modal synergy* and *multi-modal clarity*. The first criterion focuses on whether a question genuinely requires joint visual-textual reasoning. A question is preferred under this criterion if it cannot be reliably answered from the image alone or the text alone, but requires combining information from both modalities. The second criterion focuses on the overall quality and well-posedness of the question. A question is preferred under this criterion if it is relevant to the given image-text input, semantically clear, and has a single determinate answer.

For each criterion, evaluators select one winner from the candidate set in the questionnaire. To reduce position bias, we conduct each comparison twice: once using the original candidate order and once using a randomly shuffled order. A candidate is regarded as the final winner for a criterion only if it is selected consistently in both orders; otherwise, the comparison is treated as a tie. The overall winner for an instance is determined by the number of criteria won. If multiple candidates win the same number of criteria, the instance is also counted as a tie. The detailed human evaluation guidelines are provided in Table 11.

We report the win rate of each method under the two criteria, as well as the overall win rate. Following common practice for preference-based evaluation, we also report 95% confidence intervals based on the Wilson score interval [6, 53]. This allows us to measure not only which method is preferred more often, but also the reliability of the observed preference differences.

C Ablation Study on ScienceQA

Table 5 reports the ablation results on ScienceQA. SynMQG achieves the best performance across all metrics, and removing any key component leads to consistent degradation. This confirms that the improvements of SynMQG come from both the semantic decomposition stage and the dependency-preserving optimization stage.

w/o SSD. Without SSD, the performance drops on all metrics, especially on ROUGE-L, METEOR, and BLEURT. This shows that even in the relatively regular educational-style ScienceQA setting, explicitly separating visual-specific, textual-specific, and shared semantics remains beneficial. SSD helps the model organize multi-modal evidence more clearly, which improves the quality and faithfulness of generated questions.

w/o DPPO. The largest degradation is observed when DPPO is removed. BLEU-4 drops by 16.18%, and METEOR and BLEURT also decrease substantially. This indicates that supervised learning alone cannot sufficiently encourage the model to generate questions that depend on the joint contribution of image and text. DPPO provides an explicit dependency-preserving signal, which is crucial for producing synergy-oriented questions.

Table 5: Ablation study on the ScienceQA dataset.

Metric	SynMQG	w/o GRPO	w/o DPPO	w/o SSD
BLEU-1	**79.7**	71.0 (↓ 10.92%)	69.8 (↓ 12.42%)	75.6 (↓ 5.14%)
BLEU-2	**72.8**	63.5 (↓ 12.77%)	61.0 (↓ 16.21%)	68.0 (↓ 6.59%)
BLEU-3	**68.4**	58.2 (↓ 14.91%)	55.1 (↓ 19.44%)	62.8 (↓ 8.19%)
BLEU-4	**65.6**	54.6 (↓ 16.77%)	51.3 (↓ 21.80%)	59.3 (↓ 9.60%)
ROUGE-L	**77.3**	73.0 (↓ 5.56%)	70.1 (↓ 9.31%)	73.3 (↓ 5.17%)
METEOR	**80.8**	77.2 (↓ 4.46%)	73.4 (↓ 9.16%)	75.1 (↓ 7.05%)
BERTScore	**89.4**	85.9 (↓ 3.91%)	84.7 (↓ 5.26%)	87.0 (↓ 2.68%)
BLEURT	**77.1**	71.2 (↓ 7.65%)	69.6 (↓ 9.73%)	73.0 (↓ 5.32%)

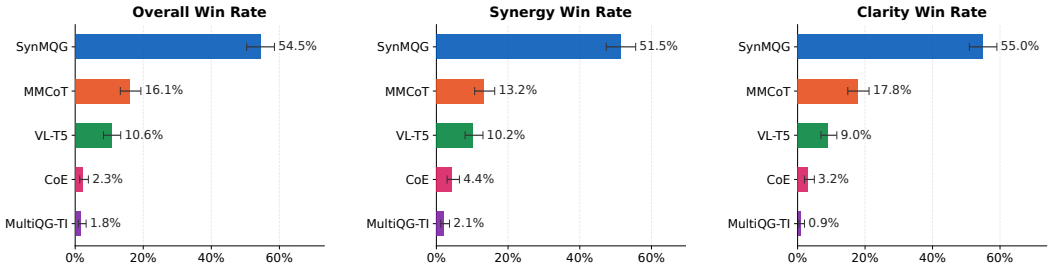

Figure 7: Human preference evaluation on ScienceQA. The figure reports overall, synergy, and clarity win rates of different methods in N-way comparison.

w/o GRPO. Removing GRPO also causes a clear performance drop, though the degradation is smaller than removing DPPO. This suggests that the synergy reward requires an effective policy optimization mechanism to be translated into generation improvements. GRPO helps compare multiple candidate generations within each group and pushes the model toward questions with stronger multi-modal dependency.

Overall, the ablation results on ScienceQA are consistent with those on MultimodalQA. SSD provides the semantic foundation for disentangling multi-modal evidence, while DPPO and GRPO further enhance the model's ability to generate questions that require coordinated visual-textual reasoning.

D Human Evaluation on ScienceQA

Figure 7 reports the human preference results on ScienceQA. SynMQG achieves the highest win rate across all criteria, with a 54.5% overall win rate, a 51.5% synergy win rate, and a 55.0% clarity win rate. This shows that human evaluators consistently prefer SynMQG-generated questions in terms of both synergistic multi-modal dependency and general question quality. The strong synergy win rate indicates that SynMQG is more effective at generating questions that require joint visual-textual reasoning rather than relying on single-modal shortcuts.

Among the baselines, MMCoT obtains the second-best results, suggesting that explicit reasoning chains help improve human-perceived question quality. VL-T5 also performs moderately well, while CoE and MultiQG-TI receive much lower preference rates on ScienceQA. This suggests that exemplar-based or image-to-text generation strategies may produce plausible questions, but they are less effective at generating questions that are both clear and strongly dependent on complementary visual-textual evidence. Overall, the human evaluation further confirms the advantage of SynMQG in producing synergy-oriented and well-formed multi-modal questions.

E SSD Analysis on ScienceQA

Figure 8 visualizes the learned semantic representations on ScienceQA. With SSD, the feature space shows a clear semantic structure: visual-specific and textual-specific features are separated into distinct regions, while visual-shared and textual-shared features are aligned in the shared semantic region. This indicates that SSD can effectively distinguish modality-private evidence from cross-modal shared grounding, even in the relatively regular educational-style ScienceQA setting. In

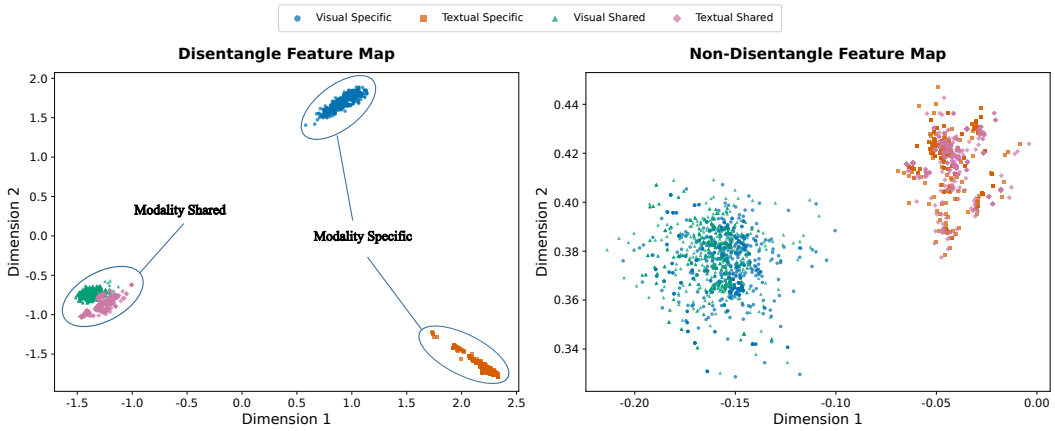

Figure 8: Semantic representation visualization on ScienceQA. The left and right subfigures compare feature distributions with and without SSD, showing that SSD better separates modality-specific semantics and aligns shared semantics.

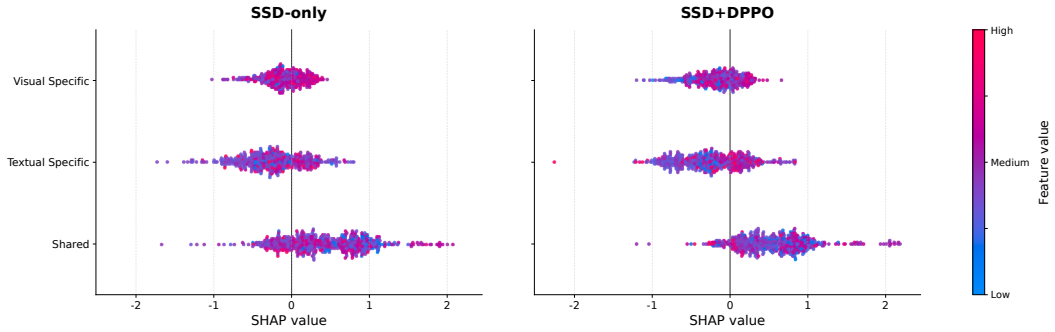

Figure 9: SHAP-based contribution analysis on ScienceQA. The left and right subfigures compare SSD-only and SSD+DPPO, showing that DPPO strengthens the contribution of disentangled semantic streams.

contrast, the non-disentangled representation mainly forms modality-driven clusters, making it harder to separate private evidence from shared information.

Figure 9 further presents the SHAP-based contribution analysis of the disentangled streams. Under SSD-only training, the shared stream already contributes strongly, while visual-specific and textual-specific streams show less stable effects. After introducing DPPO, the contribution distribution becomes more concentrated and remains positively oriented, especially for the shared semantic stream. This suggests that SSD provides meaningful semantic factors for organizing multimodal evidence, while DPPO further encourages the generator to exploit these factors in a dependency-aware manner. Together, the two analyses show that SynMQG benefits from both semantic decomposition and dependency-preserving optimization for synergy-oriented question generation.

F MLLM-based Evaluation on ScienceQA

Figure 10 presents the LLM-based evaluation results on ScienceQA. SynMQG achieves the highest overall score among all methods, showing that it provides the best balance across synergy-oriented evaluation dimensions. In particular, SynMQG obtains the best results on *VN*, *Comp*, and *CMG*, indicating that the generated questions more effectively require visual evidence, exploit complementary visual-textual information, and maintain strong cross-modal grounding.

Compared with MultimodalQA, the performance gap on ScienceQA is relatively smaller. This is mainly because ScienceQA contains more regular educational-style questions, where exemplar-based and reasoning-based methods such as CoE and MMCoT can also benefit from structured question

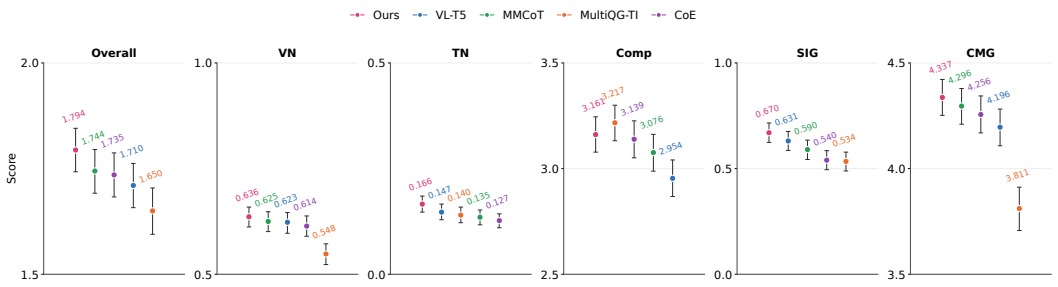

Figure 10: MLLM-based evaluation results on ScienceQA. The figure compares different methods on Overall, visual necessity (VN), textual necessity (TN), complementarity (Comp), synergistic interaction gain (SIG), and cross-modal grounding (CMG). Error bars denote the standard deviation across samples.

Table 6: Diversity comparison on ScienceQA and MultimodalQA. D-n denotes Distinct-n. The best result in each column is shown in **bold**, and the second-best result is underlined.

Method	ScienceQA				MultimodalQA			
	D-1	D-2	D-3	D-4	D-1	D-2	D-3	D-4
ChatGPT 0-shot	0.17	0.33	0.46	0.55	0.30	0.63	0.81	0.90
ChatGPT 1-shot	0.16	0.32	0.43	0.51	0.30	0.62	0.80	0.89
ChatGPT 3-shot	0.17	0.32	0.43	0.51	0.30	0.62	0.80	0.88
Qwen 0-shot	**0.19**	**0.39**	**0.51**	**0.60**	**0.31**	0.66	**0.84**	**0.92**
Qwen 1-shot	0.18	0.35	0.45	0.52	**0.31**	0.65	0.83	0.91
Qwen 3-shot	0.18	0.35	0.45	0.51	**0.31**	**0.67**	**0.84**	**0.92**
VL-T5	0.12	0.23	0.31	0.37	0.20	0.47	0.68	0.81
MultiQG-TI	0.10	0.19	0.25	0.29	0.20	0.47	0.67	0.80
MMCoT	0.12	0.23	0.30	0.35	0.20	0.50	0.71	0.84
CoE	0.10	0.18	0.23	0.27	0.21	0.50	0.71	0.83
SynMQG	0.11	0.21	0.27	0.32	0.21	0.49	0.71	0.83

patterns. For example, CoE performs competitively and achieves the highest scores on *TN* and *SIG*, suggesting that retrieved examples are useful for capturing textual cues and producing interaction-like gains in this dataset. However, SynMQG remains stronger on the overall score and grounding-related dimensions, showing that explicitly modeling visual-specific, textual-specific, and shared semantics still brings consistent advantages. MultiQG-TI performs clearly lower, especially on interaction and grounding dimensions, which further suggests that simply converting visual information into text is insufficient for high-quality synergistic MQG.

G Diversity Analysis

Table 6 reports the lexical diversity of generated questions using Distinct-n [23]. Overall, general-purpose LLMs achieve higher Distinct scores than task-specific MQG models, especially Qwen 0-shot and Qwen 3-shot. This is expected because open-ended prompting tends to produce more varied surface forms, while supervised MQG models are more constrained by dataset-specific question styles and answer-grounded generation patterns.

Among task-specific models, SynMQG maintains a moderate level of diversity. On ScienceQA, its Distinct scores are higher than CoE and MultiQG-TI, but lower than VL-T5 and MMCoT, indicating that the generated questions remain relatively controlled in this educational-style dataset. On MultimodalQA, SynMQG achieves comparable diversity to CoE and MMCoT, with Distinct-3 and Distinct-4 reaching 0.71 and 0.83, respectively. This suggests that SynMQG does not improve synergy by collapsing into repetitive templates; instead, it preserves reasonable lexical diversity while focusing generation on cross-modal dependency.

H Algorithms

This section summarizes the data processing and training procedures for SynMQG. Algorithm 1 describes the construction of multi-modal reasoning-chain data, including caption integration and multi-modal sample filtering. Algorithm 2 presents the two-stage training procedure, where SSD is

Algorithm 1: Data Processing for Synergistic MQG

Input: Raw dataset $\mathcal{D}^{raw} = \{s_i\}_{i=1}^{N}$ with $s_i = (q_i, a_i, t_i, \mathcal{P}_i, c_i)$; image root $\mathcal{I}$. Here, q_i is the original question, a_i is the target answer, t_i is the textual context, $\mathcal{P}_i$ is the set of image paths, and c_i denotes image captions.

Output: Processed dataset $\mathcal{D}^{mm}$ with samples $(q_i, a_i, \mathcal{V}_i, t_i, c_i, y_i, \tau_i, h_i)$, where $\mathcal{V}_i$ is the loaded visual input, y_i is the supervised target, τ_i is the question type, and h_i is the number of reasoning hops.

$\mathcal{D}^{mm} \leftarrow \emptyset$;
foreach $s_i \in \mathcal{D}^{raw}$ **do**
  `// Step 1: parse raw fields`
  $(q_i, a_i, t_i, \mathcal{P}_i, c_i) \leftarrow \mathrm{Parse}(s_i)$;
  `// Step 2: filter samples without valid multimodal input`
  if $|\mathcal{P}_i| = 0$ ***or*** $t_i = \emptyset$ **then**
    continue;
  `// Step 3: normalize textual and visual inputs`
  $t_i \leftarrow \mathrm{MergeTextTable}(s_i)$;
  $c_i \leftarrow \mathrm{ConcatCaption}(c_i)$;
  $\mathcal{V}_i \leftarrow \mathrm{LoadImage}(\mathcal{P}_i, \mathcal{I})$;
  `// Step 4: remove samples whose images cannot be loaded`
  if $|\mathcal{V}_i| = 0$ **then**
    continue;
  `// Step 5: obtain multimodal dependency annotation`
  $(m_i, \tau_i, h_i, r_i) \leftarrow \mathrm{Annotate}(q_i, a_i, t_i, c_i, \mathcal{V}_i)$;
  `// Step 6: retain only truly multimodal samples`
  if $m_i \neq$ `yes` **then**
    continue;
  `// Step 7: build supervised reasoning-question target`
  $y_i \leftarrow [$`Reasoning Chain`; r_i; `Question`; $q_i]$;
  `// Step 8: add processed sample`
  $\mathcal{D}^{mm} \leftarrow \mathcal{D}^{mm} \cup \{(q_i, a_i, \mathcal{V}_i, t_i, c_i, y_i, \tau_i, h_i)\}$;
return $\mathcal{D}^{mm}$;

first trained with supervised generation and disentanglement losses, and DPPO is then applied to optimize synergistic dependency.

I Limitations, Ethical Considerations, and Broader Impact

Limitations. Several limitations remain due to the practical difficulty of optimizing and evaluating synergistic multi-modal dependency at scale. First, SynMQG uses a fixed synergy threshold in the dependency-preserving reward. Although a dynamic threshold would be more flexible, it requires reliable sample-level estimates of multi-modal dependency during training. Such estimates are difficult to obtain without additional evaluators or expensive online calibration, and may further destabilize policy optimization. We therefore adopt a fixed validation-selected threshold and leave adaptive reward calibration for future work.

Second, SSD contains weighting hyperparameters for disentanglement-related losses. These coefficients cannot be arbitrarily tuned: stronger constraints may over-separate useful shared semantics, while weaker constraints may fail to reduce modality entanglement. Since full hyperparameter search for multi-modal LLM training is computationally impractical, we use a stable configuration selected on the validation set. Future work may study automatic loss balancing to reduce manual tuning.

Third, we mainly validate SynMQG on a single backbone. A broader backbone comparison would require retraining the full two-stage pipeline, including SSD training and DPPO optimization, for each model. This is beyond the available computational budget, especially because policy optimization with multi-modal inputs is substantially more expensive than standard supervised fine-tuning. Future work may examine whether the same design transfers to larger or structurally different MLLMs.

Algorithm 2: Two-Stage Training of SynMQG

Input: Processed dataset $\mathcal{D}^{mm} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = (\mathcal{V}_i, t_i, c_i, a_i, \tau_i, h_i)$ denotes the multimodal input, $y_i = (r_i, q_i)$ denotes the reasoning chain and target question; pretrained MLLM policy π_θ; SSD parameters ϕ; group size G; training epochs E_{ssd} and E_{dppo}.

Output: Trained SynMQG parameters (ϕ^*, θ^*).

Stage 1: SSD-based Supervised Training;
Initialize SSD module ϕ and trainable projection/LoRA parameters;
Freeze pretrained backbone parameters when required;
for $e = 1$ **to** E_{ssd} **do**
  foreach *mini-batch* $\mathcal{B} \subset \mathcal{D}^{mm}$ **do**
    $(x, y) \leftarrow \mathcal{B}$;
    $(Z_V, Z_T, Z_S) \leftarrow \text{SSD}_\phi(x)$;
    $\mathbf{x}_{VTS} \leftarrow \text{Project}(Z_V, Z_T, Z_S)$;
    $\hat{y} \leftarrow \pi_\theta(\cdot \mid \mathbf{x}_{VTS})$;
    $\mathcal{L}_{gen} \leftarrow -\sum_{(x,y)\in\mathcal{B}} \sum_{l=1}^{|y|} \log p_\theta(y_l \mid y_{<l}, \mathbf{x}_{VTS})$;
    $\mathcal{L}_{dis} \leftarrow \lambda_{shared}\mathcal{L}_{shared} + \lambda_{dcor}\mathcal{L}_{dcor}$;
    $\mathcal{L}_{ssd} \leftarrow \mathcal{L}_{gen} + \mathcal{L}_{dis}$;
    Update trainable parameters by minimizing $\mathcal{L}_{ssd}$;
  Evaluate on validation set and save the best SSD checkpoint;

Stage 2: DPPO-based Dependency Optimization;
Load the best SSD checkpoint;
Initialize policy π_θ from the Stage-1 checkpoint;
Construct frozen reference policy π_{ref};
for $e = 1$ **to** E_{dppo} **do**
  foreach *mini-batch* $\mathcal{B} \subset \mathcal{D}^{mm}$ **do**
    foreach $x_i \in \mathcal{B}$ **do**
      for $g = 1$ **to** G **do**
        $y_i^{(g)} = (r_i^{(g)}, q_i^{(g)}) \sim \pi_\theta(\cdot \mid x_i)$;
        $R_i^{(g)} \leftarrow \lambda_{\text{rule}} R_{\text{rule}}^{(g)} + \lambda_{\text{sim}} R_{\text{sim}}^{(g)} + \lambda_{\text{mi}} R_{\text{mi}}^{(g)}$;
      $A_i^{(g)} \leftarrow \dfrac{R_i^{(g)} - \text{mean}_g(R_i^{(g)})}{\text{std}_g(R_i^{(g)}) + \epsilon}$;
    $\rho_i^{(g)} \leftarrow \dfrac{\pi_\theta(y_i^{(g)} \mid x_i)}{\pi_{\theta_{\text{old}}}(y_i^{(g)} \mid x_i)}$;
    $\mathcal{L}_{clip} \leftarrow -\dfrac{1}{|\mathcal{B}|G} \sum_{i,g} \min\Big(\rho_i^{(g)} A_i^{(g)}, \text{clip}(\rho_i^{(g)}, 1-\epsilon, 1+\epsilon) A_i^{(g)}\Big)$;
    $\mathcal{L}_{kl} \leftarrow \text{KL}\big(\pi_\theta(\cdot \mid x_i) \,\|\, \pi_{\text{ref}}(\cdot \mid x_i)\big)$;
    $\mathcal{L}_{dppo} \leftarrow \mathcal{L}_{clip} + \beta\mathcal{L}_{kl}$;
    Update trainable policy parameters by minimizing $\mathcal{L}_{dppo}$;
  Evaluate on validation set and save the checkpoint with the best reward;
return (ϕ^*, θ^*);

Ethical Considerations. SynMQG is designed to generate questions from given multi-modal contexts and target answers. Since the model relies on the provided image-text input, potential risks mainly come from biased, noisy, or inappropriate source data. If the input context contains social stereotypes, sensitive attributes, or misleading information, the generated questions may inherit or amplify these issues. To reduce such risks, datasets should be carefully filtered before training and deployment, and generated questions should be reviewed in high-stakes educational or assessment scenarios.

Another concern is hallucination. Although SynMQG encourages grounding in visual and textual evidence, it may still generate questions that over-interpret images or assume information not explicitly provided. This is particularly important when generated questions are used for learning, testing, or decision-support applications. Therefore, SynMQG should be used as an assistance tool rather than a fully autonomous question authoring system, especially in settings where factual correctness and fairness are critical.

Broader Impact. SynMQG can benefit multi-modal education, dataset construction, and evaluation by generating questions that require coordinated reasoning over visual and textual evidence. Compared with questions that can be answered from a single modality, synergy-oriented questions may better assess whether learners or models can integrate complementary information across modalities. This can support the construction of more challenging multimodal benchmarks and more informative educational materials.

At the same time, stronger automatic question generation may reduce the cost of producing large-scale evaluation data, which could accelerate research on multimodal reasoning. However, this also increases the need for careful quality control, because automatically generated questions may be used at scale before being fully verified. We therefore encourage future applications of SynMQG to combine automatic generation with human review, dataset documentation, and transparent reporting of filtering and evaluation protocols.

J Prompt Templates

This appendix provides the prompt templates used in data processing, model training, and evaluation. The placeholders enclosed by brackets are replaced with the corresponding instance-level inputs. Specifically, Table 7 shows the prompt for data processing, Table 8 shows the prompt for model training, Tables 9 and 10 provide the prompts for MLLM-based evaluation, and Table 11 presents the rule for human preference evaluation.

Table 7: Prompt template used for data processing. The prompt is used to construct multimodal reasoning-chain annotations and auxiliary labels.

Prompt Template for Data Processing

Task:
Given the image information, textual context, answer, and original question, construct a multimodal reasoning-chain annotation for the question. The reasoning chain should explain how the answer can be reached by using the provided multi-modal information.
Instructions:
1. Read the image captions, textual context, original question, and reference answer.
2. Generate a step-by-step reasoning chain that connects the visual information and textual information to the answer.
3. Each reasoning step should explicitly indicate the source of evidence using `[Image]` or `[Text]`.
4. Identify the question type and the number of reasoning hops.
5. Use only the given image and text information; do not introduce external knowledge.
Input:
Images: `[IMAGES-BASE64]`
Image Captions: `[CAPTIONS]`
Text Context: `[TEXT]`
Original Question: `[QUESTION]`
Reference Answer: `[ANSWER]`
Output Required:
Return the result in JSON format:

```
{
  "is_multimodal":  "yes/no",
  "question_type":  "<question type>",
  "chain_hops":  <number of reasoning steps>,
  "reasoning_chain":  [
    "Step 1:  [Image/Text] ...",
    "Step 2:  [Image/Text] ..."
  ]
}
```

Table 8: Prompt template used for training. The model is required to generate a reasoning chain followed by a question.

Prompt Template for Training

System Prompt:
Given a known answer, generate a multimodal reasoning chain that leads to this answer, followed by a question that matches the question type.
User Prompt:
Based on the images and text, generate a `[HOP_COUNT]`-step reasoning chain that leads to the given answer, followed by a question.
LANGUAGE: Respond only in English. Do not use any other language.
OUTPUT: Generate only the reasoning steps and question. Do not add any prefix or suffix text.
Images: `[IMAGES]`
Text Context: `[TEXT]`
Answer: `[ANSWER]`
Question Type: `[QUESTION_TYPE]`
Output Format:
Generate exactly `[HOP_COUNT]` steps followed by a question in this format:

```
[Reasoning Chain]
Step 1:  [Image/Text] ...
Step 2:  [Image/Text] ...
...
Step [HOP_COUNT]: [Image/Text] ...
[Question]
Your question here
```

Table 9: Prompt template used for LLM-based answerability evaluation. The modality condition is replaced by visual-only, textual-only, visual-textual, or context-only settings.

Prompt Template for Answerability Evaluation

System Prompt:
You are a strict multimodal question evaluator. Based on the given rubric, rate whether the question can be answered and output your result in JSON format. Output JSON only, no other content.
Task:
You are given `[MODALITY_CONDITION]`. Evaluate whether the question can be answered correctly under this condition.
Text Context: `[TEXT]`
Question to Evaluate: `[QUESTION]`
Choices: `[CHOICES]`
Reference Answer: `[ANSWER]`
Scoring Rubric:
5 – The question can be fully and clearly answered under the given condition.
4 – The question can be mostly answered, with minor uncertainty.
3 – The given condition provides partial information.
2 – The given condition barely helps.
1 – The question cannot be answered under the given condition.
Output Required:
`{"score": <1-5>, "reason": "<brief justification>"}`

Table 10: Prompt template used for complementarity and cross-modal grounding evaluation.

Prompt Template for Complementarity and Grounding Evaluation

System Prompt:
You are a strict multimodal question quality evaluator. Based on the given rubric, rate whether the question is supported by distinct visual and textual evidence. Output JSON only, no other content.
Input:
Image: `[IMAGE]`
Text Context: `[TEXT]`
Question to Evaluate: `[QUESTION]`
Reference Answer: `[ANSWER]`
Complementarity Criterion:
Evaluate whether the image and text each provide distinct, non-redundant pieces of information that are jointly needed to answer the question.
Cross-modal Grounding Criterion:
Evaluate whether the question is grounded in specific visual evidence and specific textual evidence, and whether the two pieces of evidence are meaningfully linked.
Output Required:

```
{
  "score": <1-5>,
  "visual_evidence": "<specific visual evidence>",
  "textual_evidence": "<specific textual evidence>",
  "cross_modal_link": "<relationship between visual and textual evidence>",
  "reason": "<brief justification>"
}
```

Table 11: Guideline template used for human-preference evaluation. Annotators compare anonymized candidate questions according to multimodal synergy and clarity.

Guideline of Human Preference Evaluation
This study aims to evaluate the quality of generated multimodal questions. Each case provides an image, a text context, a reference answer, and several anonymized candidate questions. Please select the best question according to multimodal synergy, clarity, and answerability. The method names are hidden and the order of candidate questions is randomized.
Case

Context: `[TEXT]`
Image: `[IMAGE]`
Reference Answer: `[ANSWER]`
Candidate Questions:
`A: [QUESTION_A]`
`B: [QUESTION_B]`
`C: [QUESTION_C]`
`...`

Question Preference Evaluation
▷ **Multimodal Synergy:** whether the generated question requires jointly understanding the image and the text context. **Guideline:** A high-synergy question cannot be answered from the image alone or the text alone. It should require combining evidence from both modalities to obtain the correct answer. **Examples:** 1. A question that requires matching visual evidence with textual information shows strong multimodal synergy. 2. A question that can be answered only by reading the text shows weak multimodal synergy. 3. A question that can be answered only by inspecting the image shows weak multimodal synergy.
▷ **Clarity:** whether the generated question is easy to understand and semantically clear. **Guideline:** The preferred question should be fluent, concise, and unambiguous. Avoid choosing questions with unclear references, grammatical errors, or confusing wording. **Examples:** 1. A question with a clear subject and explicit reasoning target shows high clarity. 2. A vague question such as "What is shown here?" shows low clarity. 3. A question with multiple possible interpretations should not be preferred.
▷ **Relevance:** whether the generated question is aligned with the given image-text input and reference answer. **Guideline:** The selected question should be directly grounded in the provided context and should correspond to the reference answer. Questions introducing unsupported facts or unrelated entities should not be selected. **Examples:** 1. A question asking about evidence explicitly present in the image-text pair is relevant. 2. A question about information not mentioned or shown in the input is irrelevant.
▷ **Answer Uniqueness:** whether the generated question has a single determinate answer. **Guideline:** The preferred question should be answerable with the given reference answer. Avoid questions that allow multiple valid answers or require external knowledge beyond the provided input. **Examples:** 1. A question whose answer is exactly supported by the reference answer has strong answer uniqueness. 2. A question that can be answered in several different ways is ambiguous and should not be preferred.
Annotation Rule
Please select **the single best candidate question** among A, B, C, etc. If multiple questions have similar quality, choose the one with stronger multimodal synergy. If their synergy is similar, choose the one with better clarity and answer uniqueness. A fluent but unimodal question should not be preferred over a question that requires stronger image-text integration.
Response Format
Winner: `[A/B/C/...]` **Reason:** Briefly explain why the selected question is better.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction match the proposed SynMQG framework and the reported experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical formulation and derivations are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides datasets, preprocessing, implementation details, evaluation settings, and algorithms in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides an open code repository in the abstract and describes the main reproduction details in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the training setup, data processing, optimizer, backbone, and evaluation protocols in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars for MLLM-based evaluation and 95% confidence intervals for human preference evaluation using the Wilson score interval.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute setup in Appendix B, including the GPU type, memory, precision, optimizer, and training configuration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics and discusses ethical considerations in Appendix I.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive impacts and potential risks in Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pretrained models, image generators, or scraped datasets requiring additional safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the existing datasets, models, metrics, and baselines, and the released code follows the corresponding research-use settings.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases new research code and documents the data processing, training, evaluation prompts, and algorithms in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper provides the evaluation criteria and prompt templates for the preference comparison in Appendix B.4.3 and Appendix J.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The study includes human preference judgments over generated questions, but no sensitive personal information, biometric data, or identifiable participant data is collected.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the use of LLMs and MLLMs for data filtering, auxiliary captioning, evaluation, and as the main backbone model.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.