Summary of Changes and Response to Reviewers

We sincerely thank the Area Chair and the reviewers for their thorough and constructive feedback. We have undertaken a significant revision of our manuscript to address every concern raised. We believe the paper is now substantially stronger, with a clearer focus and more robust evidence supporting our contributions.

The most critical changes include:

1. **Human Performance Benchmark:** We have conducted a new human performance evaluation and added the results to Table 1. This directly addresses the major concern about the utility of the visual modality by demonstrating that humans perform significantly better with visual context, even if current models do not.
2. **Clarified Contributions & Focus:** We have restructured the introduction to explicitly enumerate the paper's contributions (lines 72-80) and added a dedicated example (Figure 1) to illustrate the task, resolving ambiguity about the paper's focus.
3. **Enhanced Clarity and Justification:** We have added justifications for our evaluation metrics (Appendix A.3), clarified ethical considerations regarding the source dataset (Ethics Statement), and improved the overall readability by converting some tables to text and ensuring all code/data links are functional and documented.

Below, we provide a point-by-point response to the issues raised by the reviewers.

Response to Meta-Review and Reviewer Concerns

We have grouped the responses by topic to address the overlapping and most critical points raised by the reviewers and the Area Chair.

1. Major Concern: Unclear Contribution of Visual Data (Area Chair CXKo, Reviewer CTb8)

- **Concern:** The performance gain from adding video was marginal, questioning the need for the visual modality. The Area Chair suggested a human evaluation to demonstrate its benefit.
- **Our Response:** We agree this is a crucial point and have made it a central focus of our revision.
    - **Action:** We conducted a new **human evaluation** and added two rows for human performance (with and without visual context) to **Table 1**. The results show a significant performance increase when humans have access to the image, confirming that the visual modality is essential for correctly answering questions in our dataset.
    - **Action:** This result reinforces our paper's main argument: EDUVIDQA presents a valuable and unsolved challenge for current VLMs, which still struggle to effectively integrate and reason over multimodal educational content in the same way humans do. We have strengthened this point in our discussion.

2. Paper Focus, Structure, and Clarity (Area Chair CXKo, Reviewer K99F)

- **Concern:** The paper lacked a clear focus, and the contributions were not immediately apparent. An illustrative example was requested.
- **Our Response:** We have restructured the paper's opening to improve clarity and focus.
    - **Action:** We added a new **Figure 1** at the beginning of the paper to provide a concrete, illustrative example of a QA pair from EDUVIDQA.
    - **Action:** We added a dedicated **"Contributions" paragraph** at the end of the Introduction (lines 72-80) that explicitly lists the paper's primary contributions and points the reader to the relevant sections.

3. Dataset Justification and Validation (Reviewer VdU6, Reviewer K99F, Reviewer CTb8)

- **Concern:** Questions were raised about the statistical validation of annotations, justification for evaluation metrics, dataset legality (LearningQ), and its distinction from other datasets.
- **Our Response:** We have added clarifications throughout the paper to address these points.
    - **Action (Metrics):** We added a new paragraph in **Appendix A.3** detailing why **correctness, coherence, relevance, and visual grounding** are crucial and appropriate metrics for the educational domain.
    - **Action (Ethics):** We have added a note to the paper clarifying the ethical considerations and public availability of the LearningQ dataset at its time of creation, as suggested.
    - **Action (Dataset Distinction):** We added a paragraph in the experiments section (lines 253-258) to articulate how EDUVIDQA differs from existing QA datasets.
    - **Action (Validation):** We have clarified our validation process (using inter-annotator agreement) within the text. The new human evaluation data in Table 1 further substantiates that the dataset's context is sufficient to answer the questions.

4. Minor Issues and Reproducibility (Reviewer VdU6, Reviewer K99F)

- **Concern:** Issues included the cost of GPT prompts, confusing statements about fine-tuning GPT-4, broken links, and lack of code documentation.
- **Our Response:** We have corrected these issues.
    - **Action:** We have reworded the text to remove any ambiguity regarding GPT-4.
    - **Action:** We have corrected all links and added a **README file** to the code repository to improve usability.
    - **Action:** We clarified that while shorter prompts are cheaper, our experiments confirmed they led to a drop in evaluation quality, justifying our current prompt design as a necessary trade-off for accuracy.