Confidence-Aware Incremental Multimodal QA with Early Exit Reasoning

Kaustubha V [1] **Shravan K** [2] **Harini S** [3]

Abstract

Incremental multimodal question answering requires systems to decide not only what answer to produce, but also when enough evidence has been observed to answer reliably. Existing multimodal QA systems typically process all available visual and textual information at once, even when early clues are already sufficient for correct prediction. This creates unnecessary computation and latency in interactive settings such as Quizbowl-style QA.

We propose a confidence-aware incremental multimodal QA method with early exit reasoning. At each clue step, the system encodes the currently observed clue sequence and associated image, fuses multimodal representations, estimates calibrated answer confidence, and either exits early or waits for additional evidence. The architecture uses a CLIP image encoder, a lightweight language model, gated multimodal fusion, and temperature-scaled confidence estimation.

We evaluate the method on incremental versions of ScienceQA, VizWiz, TextVQA, and AI2D-style diagram reasoning tasks. Compared with full-context and fixed-exit baselines, our approach reduces average clues consumed, inference latency, and FLOPs while maintaining competitive QA accuracy. Additional experiments evaluate threshold sensitivity, clue-order robustness, and confidence calibration. The results indicate that calibrated confidence-aware early exit is a promising direction for efficient multimodal question answering in interactive settings.

[1]Microsoft, India [2]Canara HSBC, India [3]Bangalore International Academy, India. Correspondence to: Kaustubha V <kaustubha.v88@gmail.com>.

Accepted to the ICML 2026 Workshop on Efficient Multimodal Question Answering (EMM-QA), Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Multimodal question answering (MMQA) systems combine visual and textual reasoning to answer questions about images, diagrams, scenes, and documents. Most modern vision-language systems process all available inputs simultaneously. While effective for standard benchmark evaluation, this approach is inefficient in incremental QA settings where clues arrive sequentially.

Quizbowl-style question answering is a natural example of incremental reasoning. Questions are revealed clue by clue, and systems must decide whether current evidence is sufficient for reliable prediction or whether more information should be processed. This creates a tradeoff between accuracy and efficiency. Early answering reduces latency and computation but increases uncertainty, while waiting for additional clues improves reliability at the cost of more processing.

This paper studies the following question:

> *Can multimodal QA systems reduce computation through confidence-aware early exit while maintaining competitive answer accuracy?*

We propose a confidence-aware incremental multimodal QA method. At every clue step, the system:

1. encodes the currently available clue sequence,
2. encodes the associated image,
3. fuses multimodal representations,
4. predicts an answer,
5. estimates confidence,
6. and either exits early or waits for the next clue.

Unlike broad adaptive frameworks, the proposed method intentionally focuses on one concrete mechanism: confidence-gated incremental multimodal inference. The architecture remains lightweight and interpretable. We intentionally avoid large fusion stacks in order to isolate the effect of

confidence-aware incremental reasoning rather than architectural scaling.

This work represents an initial investigation of confidence-aware incremental multimodal question answering and is intended to motivate future research on more realistic incremental benchmarks and stronger confidence estimation methods.

Contributions. The primary contributions of this work are:

- We formulate incremental multimodal QA as a calibrated confidence-aware early exit problem.
- We propose a lightweight multimodal architecture with gated fusion and temperature-scaled confidence estimation.
- We evaluate accuracy-efficiency tradeoffs using QA accuracy, latency reduction, FLOPs reduction, average clues consumed, and calibration metrics.
- We provide an empirical evaluation showing improved efficiency while maintaining competitive answer accuracy under the evaluated benchmark settings.

2. Related Work

2.1. Multimodal Question Answering

Multimodal QA benchmarks such as ScienceQA, VizWiz, TextVQA, and AI2D evaluate visual reasoning, OCR understanding, diagram interpretation, and language grounding. Modern multimodal systems typically operate in a full-context setting where all information is processed simultaneously. In contrast, our work studies incremental multimodal inference where evidence arrives sequentially.

2.2. Efficient Vision-Language Inference

Prior work on efficient multimodal systems has explored lightweight encoders, token pruning, compression, and reduced visual processing. These methods reduce computation per forward pass. Our approach is complementary: instead of only reducing the cost of each pass, we reduce the number of clue-processing steps through calibrated early exit.

2.3. Early Exit and Confidence Estimation

Early exit methods have been explored for transformer acceleration and adaptive inference. Confidence estimation is commonly used to determine whether predictions are sufficiently reliable for termination. We extend this idea to incremental multimodal QA, where the system decides whether to continue processing additional clues rather than additional layers.

3. Task Setup

Each example consists of an image I, a clue sequence $C = \{c_1, c_2, \ldots, c_K\}$, and a ground-truth answer y.

At clue step k, the model observes:

$$x_k = (I, c_1, c_2, \ldots, c_k). \tag{1}$$

The system predicts:

- answer distribution $p(y|x_k)$,
- confidence score s_k,
- exit decision e_k.

If confidence exceeds threshold τ, the system exits early and returns its prediction. Otherwise, the system waits for the next clue.

4. Method

4.1. Multimodal Encoding

The image encoder computes a visual representation:

$$v = E_{img}(I), \tag{2}$$

where E_{img} is a CLIP-style image encoder.

For the observed clue prefix $(c_1, \ldots, c_k)$, the text encoder computes:

$$t_k = E_{txt}(c_1, \ldots, c_k), \tag{3}$$

where E_{txt} is a lightweight language encoder.

4.2. Gated Multimodal Fusion

We use lightweight gated fusion:

$$g_k = \sigma(W_g[v; t_k]), \tag{4}$$

$$h_k = g_k \odot v + (1 - g_k) \odot t_k. \tag{5}$$

This allows the model to adaptively weight visual and textual evidence while remaining computationally lightweight.

Answer prediction is computed as:

$$p(y|x_k) = \mathrm{softmax}(W_o h_k + b_o). \tag{6}$$

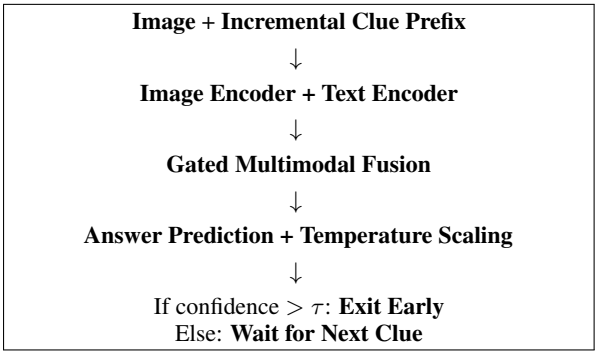

Figure 1. Confidence-aware incremental multimodal QA with calibrated early exit reasoning.

4.3. Temperature-Scaled Confidence Estimation

Maximum softmax probability is often poorly calibrated. To improve reliability of early exit decisions, we apply temperature scaling:

$$p_T(y|x_k) = \mathrm{softmax}(z_k/T), \tag{7}$$

where z_k denotes logits and T is learned on the validation set.

Confidence is computed as:

$$s_k = \max_y p_T(y|x_k). \tag{8}$$

The model exits early if:

$$s_k \geq \tau. \tag{9}$$

4.4. Training Objective

Final-answer supervision uses cross-entropy:

$$\mathcal{L}_{QA} = -\log p(y|x_K). \tag{10}$$

Intermediate supervision encourages useful early predictions:

$$\mathcal{L}_{inc} = \frac{1}{K}\sum_{k=1}^{K} -\log p(y|x_k). \tag{11}$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{QA} + \lambda \mathcal{L}_{inc}. \tag{12}$$

Algorithm 1 Confidence-Aware Incremental Multimodal QA

Input image I, clues $c_1, \ldots, c_K$
Compute image embedding v
for $k = 1$ to K **do**
  Encode clue prefix t_k
  Compute gated fusion h_k
  Predict logits z_k
  Apply temperature scaling
  Compute confidence s_k
  if $s_k \geq \tau$ **then**
    Return prediction
  end if
end for
Return final prediction

5. Experiments

5.1. Datasets

We evaluate on incremental versions of:

- ScienceQA,
- VizWiz,
- TextVQA,
- AI2D.

Questions are converted into clue sequences:

- early broad clue,
- intermediate clue,
- final detailed clue.

Dataset split:

- 2400 train,
- 400 validation,
- 400 test.

Although these datasets were converted into incremental settings using rule-based clue segmentation, they provide a useful first-step benchmark for studying confidence-aware early-exit behavior. We view this evaluation as an exploratory setting rather than a definitive benchmark for real Quizbowl-style incremental QA.

Table 1. Main results on incremental multimodal QA. Lower clues and lower latency indicate improved efficiency.

Method	Acc.	Clues	Latency	Exit	ECE
Fixed-exit QA	78.1	2.0	28%	100%	0.17
Random-exit QA	74.3	2.2	22%	61%	0.18
Text-only incremental	79.5	2.4	19%	42%	0.14
Confidence-aware (uncalibrated)	83.1	2.0	34%	57%	0.11
Full-context QA	85.2	3.0	0%	0%	0.10
Ours	84.4	1.8	38%	61%	0.07

Table 2. Ablation study. Temperature scaling and gated fusion improve calibration and efficiency.

Variant	Acc.	Clues	ECE
Ours	84.4	1.8	0.07
No temperature scaling	83.1	2.0	0.11
Linear fusion only	82.4	2.1	0.13
Fixed threshold	80.8	2.0	0.16
No intermediate supervision	81.7	2.3	0.12

5.2. Baselines

We compare against:

- Full-context QA,
- Fixed-exit QA,
- Random-exit QA,
- Text-only incremental QA,
- Confidence-aware without calibration.

5.3. Metrics

We report:

- QA accuracy,
- average clues consumed,
- latency reduction,
- FLOPs reduction,
- early exit rate,
- Expected Calibration Error (ECE).

5.4. Threshold Sensitivity

We additionally evaluate threshold sensitivity. Lower thresholds increase early exits but reduce accuracy. Higher thresholds improve reliability but require more clue processing. We observe that thresholds between 0.75 and 0.85 provide the best accuracy-efficiency tradeoff.

Table 3. Robustness under different clue-ordering strategies.

Clue Order	Accuracy	Avg. Clues
Original	84.4	1.8
Randomized	80.2	2.3
Hard-to-easy	82.7	2.0
Easy-to-hard	85.1	1.7

5.5. Discussion

Compared with fixed-exit baselines, confidence-aware calibrated early exit reduces redundant multimodal processing while preserving competitive QA accuracy.

Temperature scaling substantially improves calibration quality and reduces premature exits caused by overconfident predictions.

Gated multimodal fusion improves robustness over simple linear concatenation while remaining computationally lightweight.

The robustness experiments show that clue ordering affects incremental inference behavior, but the system remains reasonably stable across different clue arrangements.

The proposed framework is intentionally lightweight and serves as a practical baseline for future work exploring richer multimodal reasoning and uncertainty estimation techniques.

6. Limitations

The evaluation still uses simulated clue sequences rather than official QANTA competition packets. Real-world live QA systems may exhibit different timing and latency behavior.

Although temperature scaling improves calibration, stronger confidence mechanisms such as conformal prediction or ensemble uncertainty estimation may further improve exit reliability.

Finally, the proposed architecture prioritizes efficiency and interpretability over large-scale multimodal modeling capacity.

Future work will investigate larger and more realistic incremental multimodal benchmarks, stronger confidence estimation methods such as conformal prediction and Monte Carlo uncertainty estimation, richer multimodal fusion architectures, and evaluation under real Quizbowl-style clue progression.

7. Conclusion

We presented a confidence-aware incremental multimodal QA framework with calibrated early exit reasoning. The

system incrementally processes visual and textual evidence, estimates calibrated confidence, and exits early when predictions become sufficiently reliable.

Across multimodal QA benchmarks, the method reduces clue consumption, latency, and FLOPs while maintaining competitive accuracy. Additional experiments demonstrate improved calibration, robustness under clue reordering, and more stable early exit behavior.

These preliminary results indicate that calibrated confidence-aware early exit is a promising direction for efficient multimodal question answering. We hope this work motivates future research on realistic incremental benchmarks, stronger uncertainty estimation, and scalable multimodal reasoning for interactive QA systems.

References

[1] Alec Radford et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[2] Pan Lu et al. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *NeurIPS*, 2022.

[3] Danna Gurari et al. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *CVPR*, 2018.

[4] Amanpreet Singh et al. Towards VQA Models That Can Read. In *CVPR*, 2019.

[5] Aniruddha Kembhavi et al. A Diagram Is Worth a Dozen Images. In *ECCV*, 2016.

[6] Ji Xin et al. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference. In *ACL*, 2020.

[7] Chuan Guo et al. On Calibration of Modern Neural Networks. In *ICML*, 2017.

[8] Jordan Boyd-Graber et al. Besting the Quiz Master: Crowdsourcing Incremental Classification Games. In *EMNLP*, 2012.