

Original Article

Classroom VQA: A Vietnamese Visual Question Answering Dataset and a PhoBERT-EVA02 Co-Cross-Attention Baseline for Classroom Learning-Behavior Monitoring

Huy Tran^{1*}, Linh Luong¹

Faculty of Information Technology, Ho Chi Minh City University of Education, Vietnam.

*tranquanghuy@hcmue.edu.vn

Received: 25 May 2026; Revised: 28 June 2026; Accepted: 20 July 2026; Published: 29 August 2026

Abstract - Keeping track of how students behave in class is hard to do consistently, and right now it mostly comes down to a teacher's own judgment, formed by watching the room. Visual Question Answering (VQA) datasets rarely help here: most are English, built around everyday photos, or simply not designed with a classroom in mind. This paper presents Classroom VQA, a Vietnamese dataset built specifically for classroom learning-behavior understanding, comprising 3,000 classroom images drawn from Places 365 and 9,000 question-answer pairs produced and checked through a six-stage annotation pipeline. The questions span four categories – confirmation, object and attribute, counting, and action and state – chosen to cover different kinds of visual reasoning a classroom-monitoring system might need. A PhoBERT-EVA02 baseline is also proposed, using bidirectional Co-Cross-Attention so that Text-to-Image and Image-to-Text signals can inform one another. Tested across four EVA02 backbone sizes, this baseline reaches 70.78–71.44% accuracy, with under one percentage point separating the smallest and largest models. Classroom VQA is compared against six existing VQA datasets, and its question distribution, category-level error patterns, ethical considerations, and limitations are examined in detail. Together, the dataset and baseline give Vietnamese classroom VQA a concrete starting point and, more broadly, a foundation for education-oriented multimodal research.

Keywords - Visual Question Answering dataset, Benchmark construction, Classroom behavior monitoring, Cross-Attention, EVA02, PhoBERT, Vietnamese NLP.

1. Introduction

Classroom monitoring matters because it is one of the more direct windows into teaching quality, yet the way it is usually done – a teacher watching the room and forming an impression – does not scale well and leaves plenty of room for subjective judgment [1, 2]. Computer-vision systems built for classroom analytics have tried to help, but most are trained to spot a fixed list of behaviors: a raised hand, a bowed head, someone reading or writing. That works until the question gets more specific or more open-ended, such as “how many students currently have their hands up” or “does this class still look attentive.” Visual Question Answering (VQA) is a better fit for this kind of need, since the question itself – rather than a fixed label set – determines what the model has to look for in the image [3].



The real obstacle to using VQA here is less about modeling and more about data. The VQA benchmarks that already exist – VQA v2 [4], COCO-QA [5], CLEVR [6] – were built around English questions and either everyday photographs or synthetic scenes, none of which look anything like a classroom. A handful of domain-specific VQA datasets exist outside education – VQA-MED [7] for radiology is one example – but nothing comparable has been built for classroom settings. Vietnamese is further behind still: ViVQA [8] appears to be the only public Vietnamese VQA dataset, and its images come from MS-COCO rather than from anything classroom-related. Three factors compound this gap specifically for Vietnamese classrooms.

First, Vietnamese is tonal and diacritic-heavy, and questions produced by machine translation tend to lose precision or come out sounding stilted – not the way a teacher would actually phrase something. Second, there is no Vietnamese equivalent of the large-scale vision-language pretraining corpora available in English, so multimodal alignment cannot simply be learned from scale alone; a carefully built dataset matters more here than it would for a resource-rich language. Third, Vietnamese classrooms look and behave differently from the generic scenes found in most scene-recognition corpora – different uniforms, different seating arrangements, different teacher-student dynamics – so a model trained only on generic imagery has little real basis for transferring to this setting. Put together, no existing resource offers classroom imagery, behavior-oriented questions, and Vietnamese annotations at the same time, and closing that gap is what motivates this work.

Three questions guided the study. Can a quality-controlled pipeline produce a Vietnamese, classroom-specific VQA dataset whose annotations are actually learnable and internally consistent (RQ1)? Does a fairly ordinary multimodal baseline – a Vietnamese language encoder and a vision encoder tied together through bidirectional cross-attention – perform stably on this dataset regardless of how large the vision backbone is (RQ2)? And which categories of classroom-behavior questions turn out to be hardest, and what does that difficulty reveal about where the dataset and the baseline still fall short (RQ3)?

With that gap in mind, this study introduces Classroom VQA, a Vietnamese dataset built specifically for classroom learning-behavior monitoring. Its construction is documented in enough detail – data collection, annotation, verification, category breakdown, and statistics – that other researchers should be able to reproduce or extend it without much guesswork. To check that the dataset is actually usable in practice, a compact multimodal baseline combining PhoBERT [9], and EVA02 [10] through bidirectional Co-Cross-Attention was also built and tested; its purpose is less about chasing state-of-the-art numbers and more about confirming, empirically, that the dataset supports stable learning.

The rest of the paper follows a fairly standard path. Section 2 covers related work on VQA methods, VQA datasets, and Vietnamese-language resources, and closes by separating out what is genuinely new in this paper from what is adapted or reused. Section 3 walks through how Classroom VQA was built – collection, annotation, verification, statistics, and examples. Section 4 describes the PhoBERT-EVA02 baseline and its cross-attention mechanism. Section 5 reports the experiments, including a breakdown of errors by question category. Section 6 discusses what the results mean, the ethical considerations involved, and where the work is limited. Section 7 closes with a summary and directions for future work.

In summary, this paper contributes the following:

- Classroom VQA, a Vietnamese VQA dataset for classroom learning-behavior monitoring, built from 3,000 curated images and 9,000 question-answer pairs through a six-stage annotation and verification pipeline.
- A PhoBERT-EVA02 baseline that uses bidirectional Co-Cross-Attention so that question semantics and classroom visual evidence can inform each other in both directions.

- A systematic evaluation of the dataset itself – comparisons against existing VQA datasets, an explicit account of what is novel versus adapted, per-category analysis, experiments across backbone sizes, and a discussion of failure cases, ethics, and limitations.

2. Related Work

2.1. Visual Question Answering: Task and Approaches

Antol et al. [3] were the first to formalize VQA as the task of answering a free-form question about an image, and the methods that followed can be grouped into three broad families [11]. Joint-embedding approaches map image and question features into a shared space before classifying an answer; Neural Image QA, which pairs a CNN with an LSTM [12], and mQA [13] are early examples of this.

Attention-based approaches let the model focus on whichever image regions actually matter for a given question, rather than processing the whole image uniformly – the Neuro-Symbolic Concept Learner (NS-CL) [14] and SelRes combined with LXMERT [14] fall into this group, the latter reaching over 72% accuracy on VQA v2. A third family, model-agnostic methods such as RUBi [15] and Learned-Mixin + H [16], tries to reduce language bias without touching the underlying architecture at all. What none of these three families addresses is a more basic question: what dataset is actually available to reason over in the first place. That is the gap this paper is aimed at, specifically for the classroom-monitoring domain.

2.2. VQA Datasets: From General-Domain to Domain-Specific

VQA research has arguably been shaped as much by which datasets became available as by any single algorithmic breakthrough. DAQUAR [17] was an early attempt but stayed small and confined to indoor, object-centric scenes. COCO-QA [5] and VQA v2 [4] scaled things up considerably by drawing on MS-COCO and crowd-sourced questions, and VQA v2 in particular added complementary image pairs specifically to cut down on models guessing from language patterns alone. CLEVR [6] took a different route entirely, generating synthetic 3D scenes so that compositional reasoning could be studied without the noise of real photography.

As general-purpose imagery turned out to be insufficient for some applications, domain-specific datasets followed – VQA-MED [7], for instance, targets radiology images and clinical questions, a reasonable illustration of why a specialized field ends up needing its own dataset rather than a repurposed general one. Education, and classroom-behavior monitoring specifically, has not had an equivalent resource, despite having its own visual vocabulary (postures, gestures, furniture, group activity) and its own kinds of questions (attention, participation, instructional state) that general-purpose datasets simply do not exercise.

2.3. Vietnamese Language Resources and VQA

Vietnamese NLP has come a long way – PhoBERT [9] is a solid example of a pretrained language model – but multimodal Vietnamese resources are still thin on the ground. ViVQA [8] is, as far as can be determined, the only publicly documented Vietnamese VQA dataset, and it was built by adapting an MS-COCO-derived question set into Vietnamese rather than by designing around a particular application; it has no systematic classroom or student-behavior content.

Outside VQA specifically, other Vietnamese education-related NLP work tells a similar story: the UIT-VSFC student-feedback corpus [20], for example, shows that domain-built Vietnamese datasets clearly outperform generic, translated ones on education-related tasks – more evidence that a purpose-built classroom dataset is worth the effort rather than a shortcut through translation. Table 1 lines Classroom VQA up against six representative datasets, and the comparison is fairly blunt: it is the only one on the list that is simultaneously classroom-specific, made of real (non-synthetic) images, and in Vietnamese.

Table 1. Comparison of the proposed Classroom VQA dataset with representative VQA datasets

Dataset	Image source	Images	Q&A pairs	Domain / context
DAQUAR (2014)	NYUDv2	1,449	12,468	Indoor scenes, very small scale
COCO-QA (2015)	COCO	123,287	117,684	Everyday life, auto-generated questions
VQA v2 (2017)	COCO	204,721	1,105,904	Everyday life, reduced language bias
CLEVR (2017)	3D Graphics	100,000	853,554	Logical reasoning, no real images
VQA-MED (2019)	PubMed	4,200	15,292	Medical images (X-ray, CT, MRI)
ViVQA (2021)	COCO	10,328	15,000	Everyday life, Vietnamese
Classroom VQA (proposed)	Places365	3,000	9,000	Learning behavior, classroom, Vietnamese

2.4. Novelty Positioning: Novel versus Adapted Components

Since the reported results reflect both the dataset and the baseline together, it is worth being explicit about which one is doing the work. Table 2 breaks the contribution down into three groups. PhoBERT [9], and EVA02 [10] are used exactly as published, with no architectural changes – no novelty claim is attached to either encoder. The bidirectional Co-Cross-Attention fusion (Section 4.2) follows the general dual-stream cross-attention idea used in LXMERT-style models; what differs here is the specific symmetric Text-to-Image / Image-to-Text formulation, tuned for aligning Vietnamese question semantics with classroom visual evidence – an adaptation, not a new attention mechanism. The genuinely new parts of this work are the dataset itself (the Vietnamese classroom imagery, the four-category taxonomy, and the six-stage verified pipeline described in Section 3) and the finding that a deliberately simple fusion mechanism is enough to establish a stable benchmark, with nothing architecturally elaborate required. That last point matters for interpreting the results: because the baseline was kept intentionally plain, its consistent accuracy across backbone sizes (Section 5.3) says more about the quality of the dataset than about any particular strength of the model.

Table 2. Novel, adapted, and reused components of the proposed dataset and baseline

Component	Status	Description
PhoBERT Text Encoder [9]	Reused (Not Novel)	Off-the-shelf pretrained Vietnamese language encoder, used without architectural modification
EVA02 Vision Encoder [10]	Reused (Not Novel)	Off-the-shelf pretrained Vision Transformer, used without architectural modification
Co-Cross-Attention fusion	Adapted	Dual-stream cross-attention paradigm adapted into a symmetric bidirectional T2I / I2T formulation for this task
Classroom VQA Dataset	Novel	First Vietnamese, classroom-specific VQA dataset with a four-category taxonomy and a documented six-stage verified pipeline
Six-stage Annotation Pipeline	Novel	New protocol designed to yield consistent, quality-controlled classroom-behavior annotations in Vietnamese

3. The Classroom VQA Dataset: Construction, Design, and Characteristics

3.1. Data Collection

Three things mattered when building the image collection: the scenes had to look like real classrooms, they had to support human-centric questions, and they had to work with Vietnamese-language annotation. In practice, that meant looking for real-world classroom photos with clear evidence of teaching or learning – an identifiable teacher and/or student, and enough visual quality to make out actions, postures, or states.

Images were drawn from Places365 [18], a large scene-recognition dataset that happens to include a dedicated classroom category. It is a reasonable starting point: the scenes are real and diverse, and because the dataset was

never built with VQA in mind, it avoids the kind of dataset-specific artifacts that can creep in otherwise. Not every image in the category was usable, though, so a filtering pass removed anything with poor visual quality, ambiguous content, watermarks, or synthetic overlays, keeping only images with clear instructional activity.

After filtering, 3,000 classroom images remained, spanning a reasonable range of classroom layouts, camera angles, and teacher–student interactions – enough variety to support meaningful questions about actions, postures, and states. Because these images come from a public, credited benchmark rather than from footage recorded for this study, no new personal data was collected and no individual student or teacher was identified; the ethical implications of this choice, and of any future deployment, are discussed separately in Section 6.2.

3.2. Data Annotation and Validation

The dataset was put together through six stages (Figure 1), broadly following the annotation approach used for ViVQA [8] and VQA v2 [4] but adapted for the classroom domain and for Vietnamese.

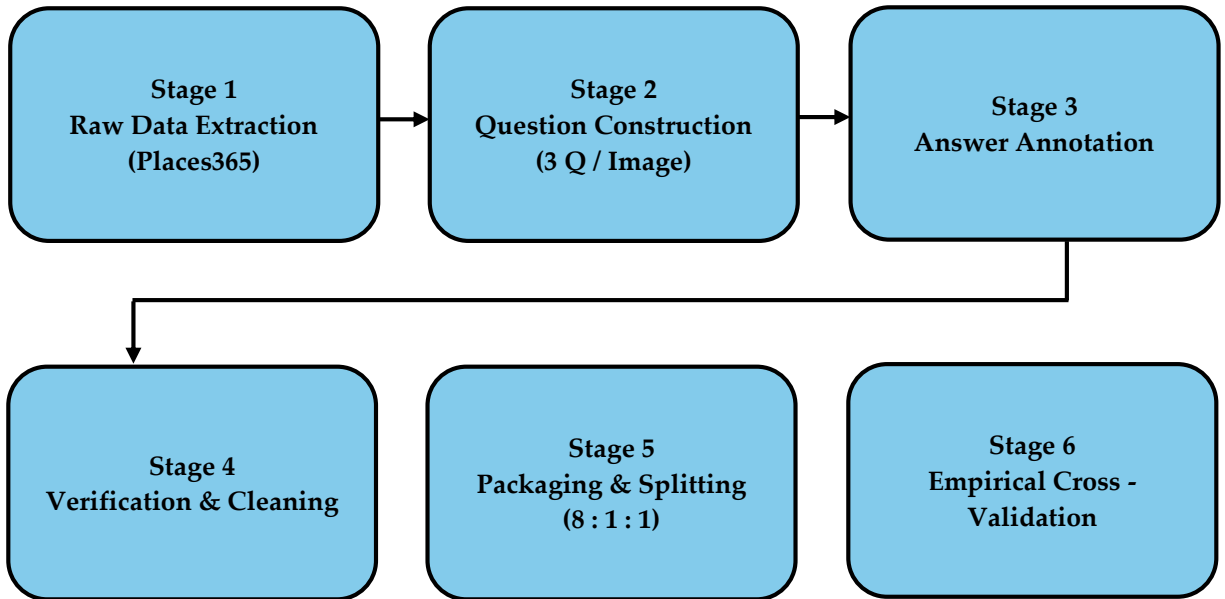


Fig. 1 Six-stage pipeline for constructing the Vietnamese Classroom VQA dataset

Stage 1 - Raw Data Extraction: The Places365 classroom category started with 5,000 images. A filtering pass removed anything low-quality, near-duplicate, or visually ambiguous, leaving 3,000 images that clearly show classroom activity with an identifiable teacher or student.

Stage 2 - Question Construction: Annotators wrote exactly three questions per image, each covering a different angle: object or attribute recognition, counting, or human posture and action. Fixing the number at three was a deliberate choice – with only one question allowed, annotators tend to gravitate toward whatever is most obviously happening in the frame and overlook everything else. Requiring three pushes them to look at secondary details, background elements, and spatial relationships instead. This step produced 9,000 questions across the 3,000 images.

Annotator Team and Guidelines: Annotation was carried out by a team of native Vietnamese speakers working from a written guideline document that laid out the four question categories, examples of acceptable and unacceptable questions, and a rule that every answer had to be grounded in something visible in the image rather than assumed. Before annotating the full set, the team ran through a calibration round on a held-out sample;

disagreements that came up during calibration were discussed and used to tighten the guidelines before the main annotation effort began.

Stage 3 - Answer Annotation: Each question received a short answer, grounded strictly in what could actually be seen in the image rather than in a plausible guess. Figure 2 shows a typical annotation screen: a classroom image, the question “Who is teaching?”, and the corresponding answer, “teacher.”

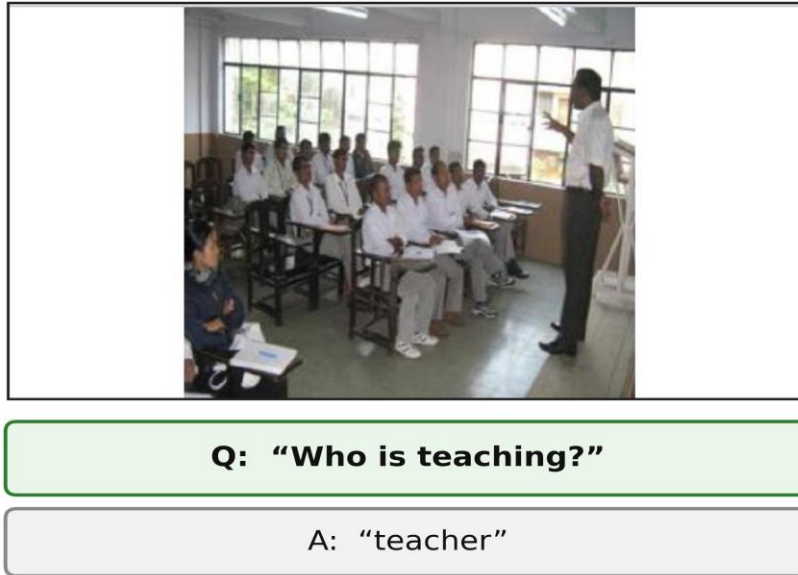


Fig. 2 Representative annotation example: an image paired with the question “Who is teaching?” and the grounded answer “teacher”

Stage 4 - Verification and Cleaning: All 9,000 pairs went through an independent review pass looking for three kinds of problems: questions too vague or ambiguous to answer with confidence, answers that did not actually match the image, and surface issues such as typos or inconsistent abbreviations. Anything flagged was either fixed or dropped, and fixes were re-checked before being accepted.

Stage 5 - Packaging and Splitting: The cleaned dataset was shuffled at the image level, so that all three questions tied to a given image stayed together within the same split, then divided 8:1:1 into training, validation, and test sets.

Stage 6 - Empirical Validation: As a final check, the finished dataset was run through several Vietnamese-capable VQA architectures (Section 4) to see whether it actually supported stable training or whether some annotation issue had slipped through. The results held up consistently (Section 5), which is reasonable evidence that no further cleaning pass was needed.

3.3. Dataset Statistics and Analysis

The verified questions fall into four categories that, together, cover a decent range of visual reasoning (Figure 3). Confirmation questions – simple yes/no checks on whether something is present – make up the largest share, at 55%. Object and attribute questions, asking about identity, color, material, or position, account for 18%. Counting questions, which ask the model to tally instances of something, make up another 18%. Action and state questions – the hardest category, since they call for reading a posture or an ongoing action rather than simply naming an object – come in at 9%.

That split is not accidental. Confirmation questions are relatively easy and give the model a large, low-noise training signal, while the smaller counting and action/state categories concentrate the harder reasoning where it matters most for this application. Including action and state questions specifically is what pushes the dataset past static object recognition and toward genuine behavioral understanding.

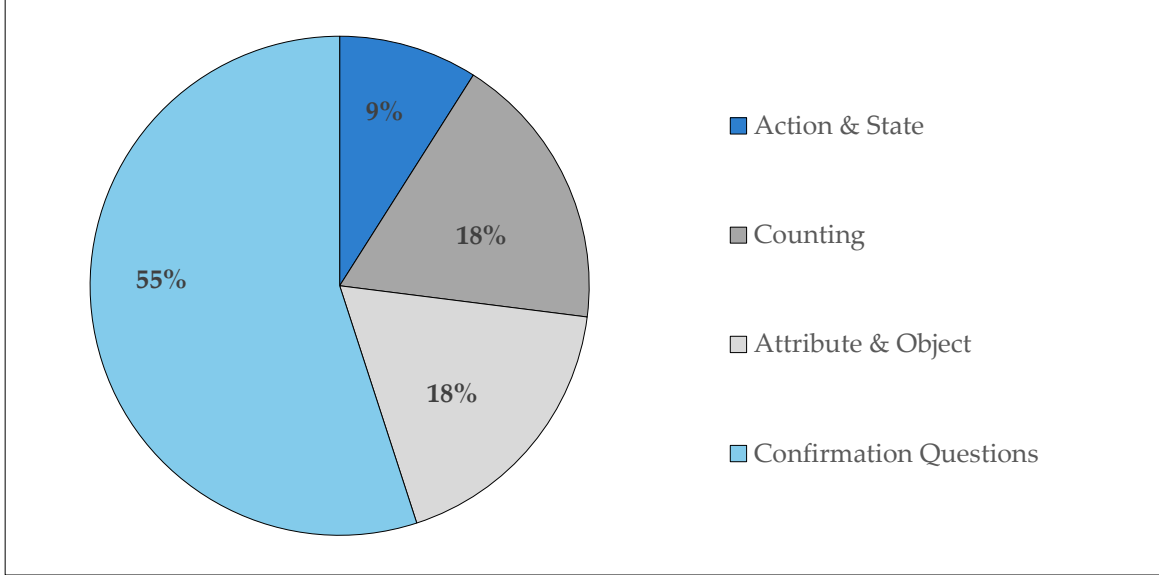


Fig. 3 Distribution of question categories in the Classroom VQA dataset

The finished dataset splits 8:1:1 as described above, giving 2,400/300/300 images and 7,200/900/900 question-answer pairs for training, validation, and test, respectively (Table 3). This exact split is released alongside the dataset so that future work can report directly comparable numbers.

Table 3. Sample distribution across the Classroom VQA data splits

Split	Images	Questions	Ratio (%)
Training	2,400	7,200	80
Validation	300	900	10
Test	300	900	10
Total	3,000	9,000	100

3.4. Illustrative Examples

Figure 4 shows one example from each category: a confirmation question about whether class is in session, an attribute question about a desk's color, a counting question about how many people appear in the frame, and an action question about what the teacher is doing.

4. Validation Model: A PhoBERT-EVA02 Co-Cross-Attention Baseline

Dataset quality cannot really be judged from the construction process alone, so a multimodal baseline was built to test whether Classroom VQA actually supports learning and behaves consistently. The baseline was kept deliberately simple, relying on a well-established fusion approach rather than anything architecturally new, so that the evaluation says something about the dataset rather than about clever model design (see also the novelty discussion in Section 2.4).

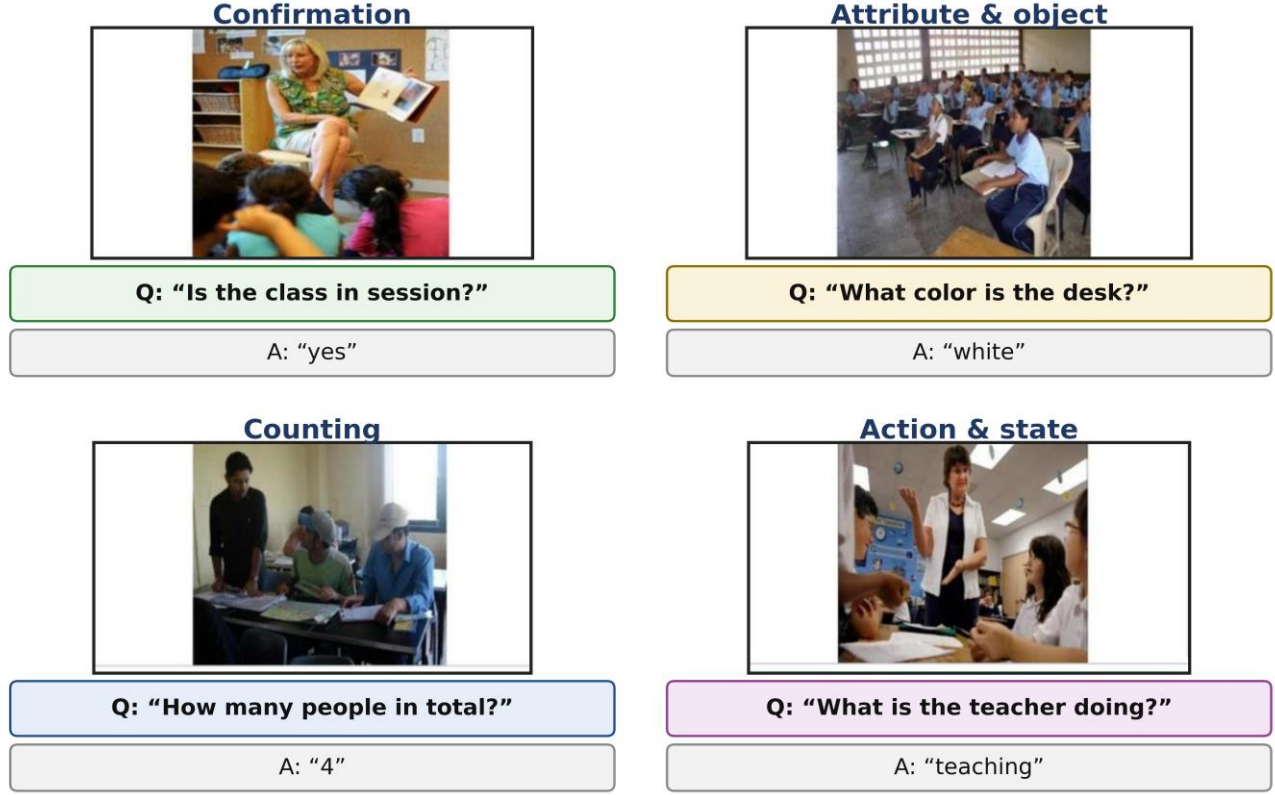


Fig. 4 One example from each of the four question categories: confirmation (“Is the class in session?” → “yes”), attribute (“What color is the desk?” → “white”), counting (“How many people in total?” → “4”), and action/state (“What is the teacher doing?” → “teaching”).

Framed formally, VQA is a multi-class classification or answer-generation problem: given an image (I) and a question (Q), the goal is to pick the answer ($\hat{a}$) from a candidate set (Ω) that maximizes the conditional probability:

$$\hat{a} = \operatorname{argmax}_{a \in \Omega} p(a|I, Q; \theta) \quad (1)$$

Where, θ is the set of learnable parameters.

4.1. Overall Architecture

The baseline uses a two-branch architecture: text and vision are encoded separately, then brought together through cross-modal attention and fusion (Figure 5). The text branch handles the Vietnamese question and its linguistic context; the vision branch extracts visual evidence from the classroom image. The two are then combined and passed through a classifier to predict an answer.

On the text side, each question is first segmented with VnCoreNLP and then encoded with the pretrained PhoBERT-base model [9] – 12 Transformer layers, 12 attention heads, hidden size 768. For a batch of (B) questions with maximum length (L), this produces contextualized representations $F_{text} \in \mathbb{R}^{(B \times L \times 768)}$.

On the vision side, each 224×224 image is split into 256 non-overlapping patches and encoded with EVA02 [10], a Vision Transformer variant that uses SwiGLU activation, Sub-Layer Normalization, and 2D Rotary Position Embedding (2D RoPE). This produces image representations $F_{visual} \in \mathbb{R}^{(B \times 256 \times 768)}$.

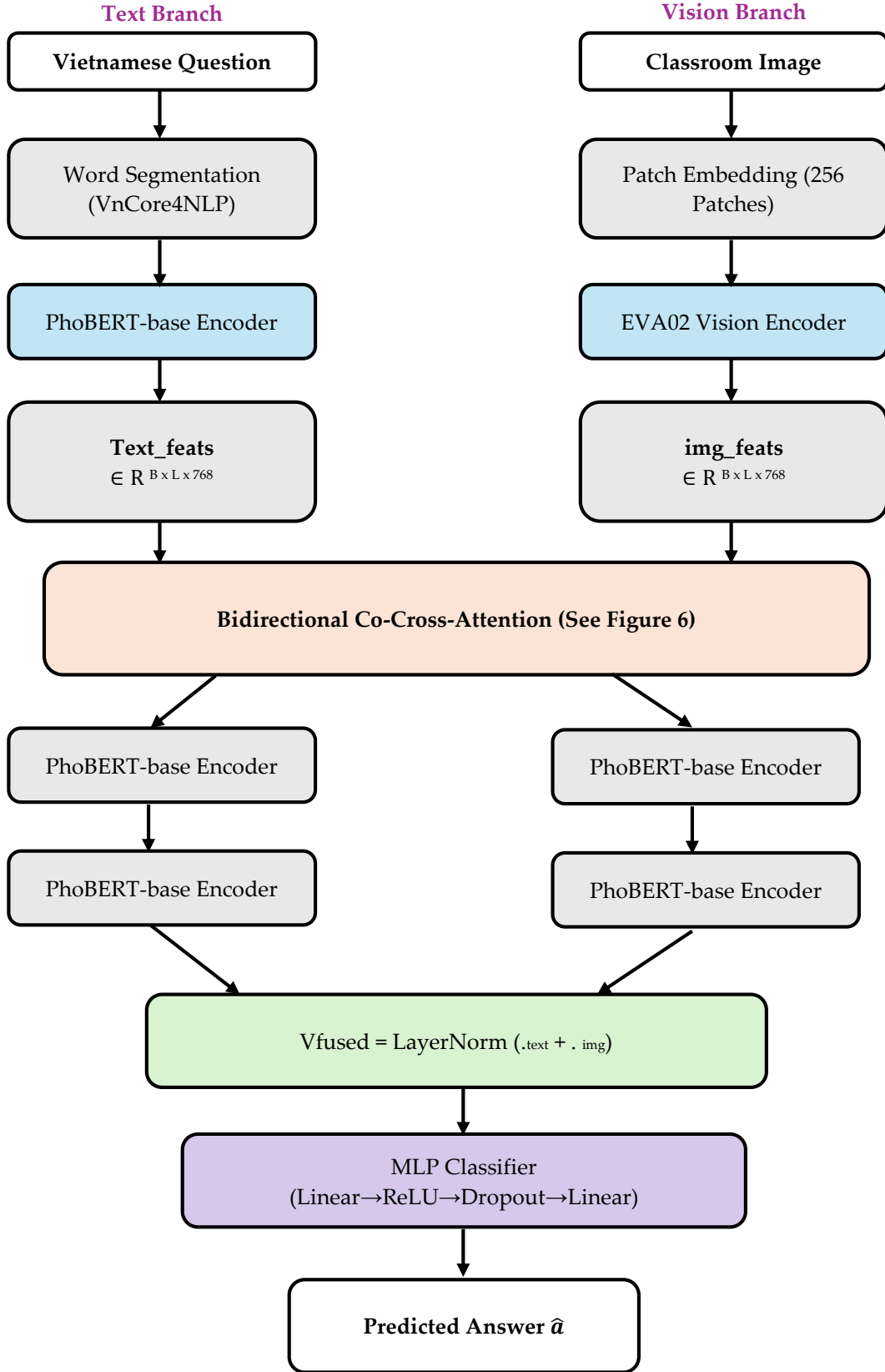


Fig. 5 Overall architecture of the validation model combining PhoBERT, EVA02, and the Co-Cross-Attention mechanism

4.2. Co-Cross-Attention Mechanism

To let the two modalities genuinely inform each other, rather than simply being concatenated or summed, a bidirectional Co-Cross-Attention mechanism is used: each modality attends to the other and pulls in whatever information is relevant. Given a feature matrix $X \in \mathbb{R}^{(n \times d_{\text{model}})}$, the Query, Key, and Value representations are obtained through learned linear projections:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \quad (2)$$

Where, W^Q , W^K , and W^V are the learnable projection matrices. Attention itself follows the usual scaled dot-product form:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V \quad (3)$$

Where, d_k is the dimensionality of the Key representations; the $1/\sqrt{d_k}$ scaling keeps large dot products from pushing the softmax into a near-zero-gradient regime, which would otherwise slow training down.

The cross-modal interaction runs in two directions. In the Text-to-Image (T2I) stream, the text representation acts as the Query while the visual representation supplies the Key and Value, so each question token can attend to whichever image regions are relevant to it:

$$O_{T \rightarrow I} = \text{softmax}(Q_{\text{text}} K_{\text{visual}}^T / \sqrt{d_k}) \cdot V_{\text{visual}} \quad (4)$$

The Image-to-Text (I2T) stream reverses the roles – the visual representation becomes the Query, and text supplies the Key and Value:

$$O_{I \rightarrow T} = \text{softmax}(Q_{\text{visual}} K_{\text{text}}^T / \sqrt{d_k}) \cdot V_{\text{text}} \quad (5)$$

Between the two streams, the mechanism gives a genuinely bidirectional link: T2I lets the question pull in relevant visual evidence, while I2T lets visual features pick up the linguistic context of the question. Both outputs then feed into fusion and classification.

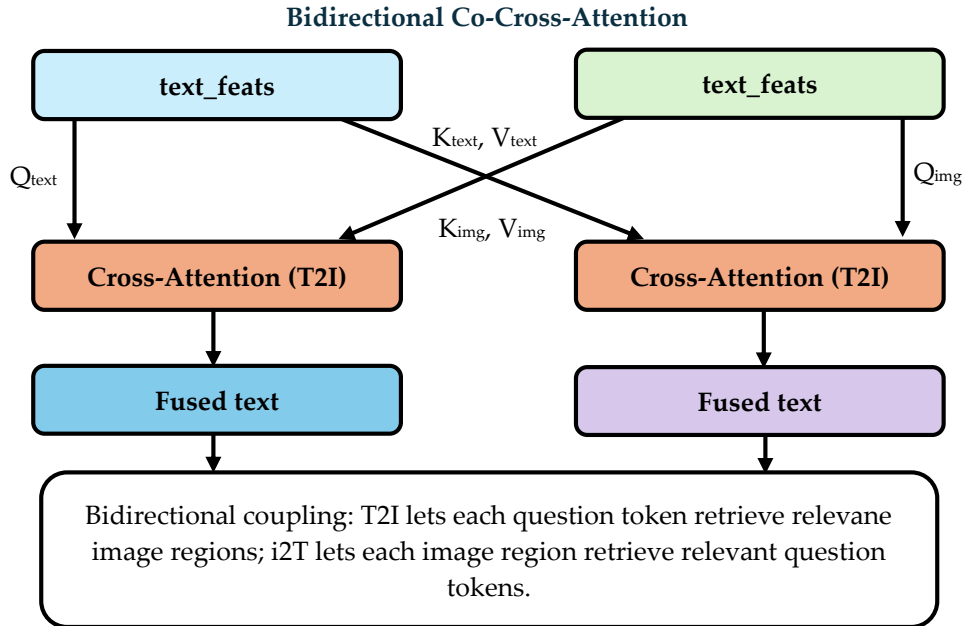


Fig. 6 Bidirectional Co-Cross-Attention mechanism between the text and image branches.

4.3. Feature Fusion and Classification

The two attention outputs still carry their sequence dimensions $O_{(T \rightarrow I)} \in \mathbb{R}^{(B \times L \times 768)}$ for text tokens enriched with visual information, and $O_{(I \rightarrow T)} \in \mathbb{R}^{(B \times 256 \times 768)}$ for visual tokens conditioned on the question. Mean pooling collapses each to a single vector, $v_{(T \rightarrow I)}$ and $v_{(I \rightarrow T)} \in \mathbb{R}^{(B \times 768)}$, which are then summed and normalized:

$$dv_{fused} = LayerNorm(v_{T \rightarrow I} + v_{I \rightarrow T}) \quad (6)$$

The fused vector then passes through a two-layer MLP classifier, with ReLU, LayerNorm, and a dropout rate of 0.1.

4.4. Training Configuration

Training uses the cross-entropy loss, with (N) the batch size and (C) the number of answer classes. Optimization is AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, paired with a cosine-annealing learning-rate schedule and a 10% linear warm-up. Table 4 lists the main hyperparameters. Early stopping with a patience of 10 epochs guards against overfitting.

Everything ran on Google Colab with a single NVIDIA Tesla T4 GPU (16 GB VRAM), PyTorch 2.1.0, and HuggingFace Transformers 4.35.2. To see how learnability holds up across model capacity, four otherwise identical configurations were trained, differing only in EVA02 backbone size (Tiny, Small, Base, and Large), with the text branch, fusion mechanism, and hyperparameters held fixed throughout.

Table 4. Training hyperparameter configuration

Hyperparameter	Value
Learning rate	5×10^{-5}
Batch size	32
Maximum epochs	100
Maximum sequence length	64
Dropout rate	0.1
Warm-up ratio	10%
Weight decay	0.01

5. Experiments and Results

5.1. Evaluation Metrics

Four standard metrics were used – Accuracy, Precision, Recall, and F1-score – with the latter three computed as weighted averages across answer classes to account for the class imbalance present in the dataset.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (7)$$

$$Precision = TP / (TP + FP) \quad (8)$$

$$Recall = TP / (TP + FN) \quad (9)$$

$$F1 - score = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (10)$$

5.2. Training Behavior

All four backbone variants converged quickly, within roughly the first five epochs, which is not surprising given that both branches start from pretrained weights. Past about epoch 12, training accuracy kept climbing toward 95–99%, while validation accuracy flattened out around 68–71% – a fairly typical sign of mild overfitting.

Validation loss did not spike back up sharply, though, which suggests weight decay and dropout were doing their job. Epochs 12–17 looked like the sweet spot for saving a checkpoint. Figure 7 shows the training curves for the EVA02-Base variant; the other three look much the same, which is itself a useful signal that training dynamics are being driven mainly by the dataset rather than by quirks of any particular backbone.

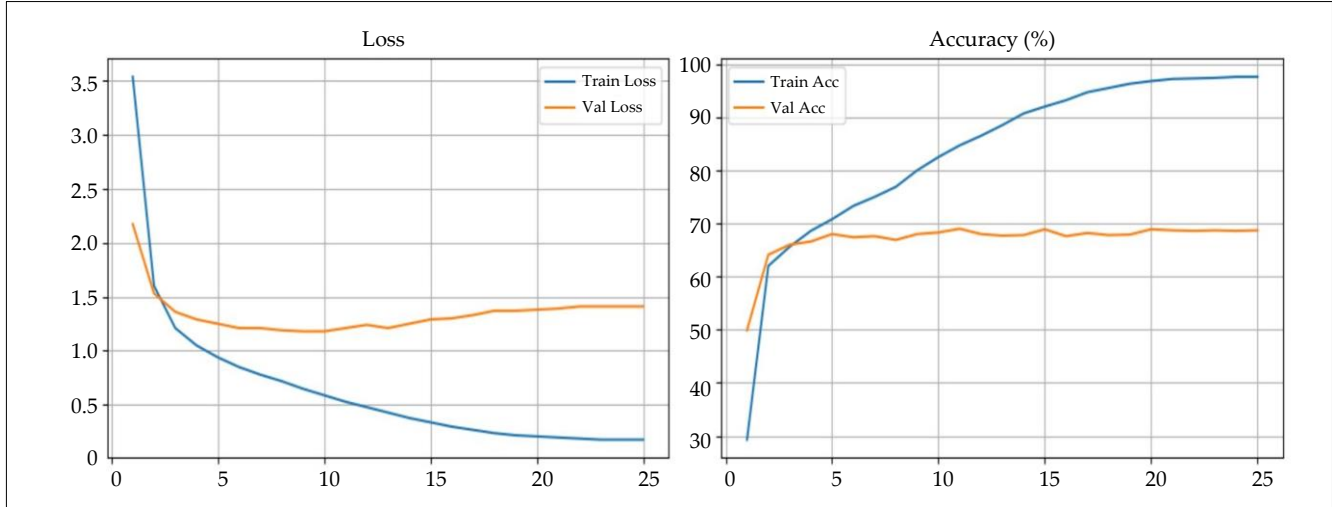


Fig. 7 Training/validation loss and accuracy curves for the EVA02-Base variant

5.3. Quantitative Results

Table 5. Evaluation results of the four model variants on the test set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EVA02-Tiny	70.78	68.45	70.78	69.06
EVA02-Small	71.44	68.79	71.44	69.42
EVA02-Base	71.33	69.80	71.33	69.80
EVA02-Large	70.89	69.05	70.89	69.21

Table 5 lays out how the four variants performed on the held-out test set. None of them stands out dramatically – accuracy ranges narrowly from 70.78% to 71.44%. EVA02-Small comes out on top for accuracy (71.44%), while EVA02-Base edges ahead on Precision and F1-score (69.80% each).

That narrow spread is itself informative: moving from a small backbone to a much larger one does not buy much here. EVA02-Large, despite considerably more capacity, reaches only 70.89% – essentially tied with the smaller models. Backbone size, in other words, does not appear to be the limiting factor; the dataset seems to be setting the ceiling on what this baseline can achieve.

EVA02-Tiny holds its own too, at 70.78%, hinting that a much smaller backbone could be a reasonable trade-off wherever compute is limited. One caveat is worth flagging here: each configuration was trained once, not averaged over multiple random seeds, so the sub-one-percentage-point gaps between backbones in Table 5 should be read as suggestive rather than statistically confirmed – proper significance testing across repeated runs is left for future work (Section 6.3).

5.4. Qualitative Analysis

Figure 8 gives a concrete example of the model in action. Shown a classroom image and the question “Is this classroom crowded with students?”, EVA02-Base answers “Yes” with 92.55% confidence – a reasonable call, given how many students are visibly seated in the frame. High-confidence, correct answers like this were typical for

confirmation questions; the model was noticeably less sure of itself on counting and action/state questions, consistent with the category-level pattern discussed next.



Q: “Is this classroom crowded with students?”

Fig. 8 Qualitative example: the model correctly answers a confirmation question with high confidence (92.55%)

5.5. Category-Level Error Analysis

Breaking accuracy down by category confirms what might be expected: confirmation (yes/no) questions score highest, which makes sense given how much of the training data falls into that category and how little ambiguity a yes/no answer space carries. Counting and action/state questions lag behind, since they demand more spatial and semantic reasoning – a pattern that shows up in other VQA work too [4, 19]. Looking through the errors themselves, counting mistakes tend to occur when the target is partly hidden behind something else or when several similar-looking entities compete for attention, leading the model to over- or under-count. Action/state errors, meanwhile, usually trace back to a subtle cue – a small hand gesture, a particular posture – that occupies only a small part of the image and gets outweighed by more visually dominant background content. Both patterns point toward the same fix: finer spatial supervision, such as region proposals or pose keypoints, would likely help exactly the categories that currently underperform. None of this counts against the dataset, though – if anything, having a genuine spread of difficulty across categories is what makes Classroom VQA useful as a diagnostic benchmark, rather than one where every model simply saturates near 100%.

6. Discussion

6.1. Significance and Applications

Taken together, the results suggest Classroom VQA has real potential as a benchmark for Vietnamese classroom VQA. The PhoBERT-EVA02 baseline holds steady across four backbone sizes, with under one percentage point separating them – evidence that the benchmark supports stable learning without leaning too heavily on any single architecture.

There are a few directions this could go from here. The fixed data split makes it possible to compare different architectures, language encoders, or vision backbones on equal footing. The annotation pipeline itself is reusable – it could extend to more classroom scenarios or other Vietnamese domain-specific VQA tasks without much rework. And the four question categories already give researchers a way to isolate harder reasoning problems, particularly counting and action/state, for further study.

6.2. Ethical Considerations

Given that this dataset is about monitoring classroom behavior, its ethical footprint deserves direct attention rather than a passing mention. Every image comes from the publicly released Places365 benchmark [18], not from footage recorded for this study, so no new personal data, student records, or identifiable information were collected, and no institutional ethical approval or consent process was needed for the dataset itself, since the underlying images were already public under Places365's terms of use. The baseline also does not do anything resembling facial recognition or individual identification; it answers questions about aggregate scene content – counts, actions, attributes – rather than about who specific people are. That said, a real deployment on actual, unpublished classroom footage would be a different situation entirely, one requiring informed consent from teachers, students, and guardians, institutional review, and compliance with applicable data-protection regulations; nothing in this study has been evaluated for that scenario, and it is explicitly out of scope. Regarding bias: Places365 was not curated with Vietnamese classrooms in mind, so whatever visual or cultural conventions dominate that source imagery could carry over into anything trained on Classroom VQA. The authors would caution against using this dataset or baseline for individual student evaluation, disciplinary decisions, or any high-stakes decision-making without considerably more validation; if a classroom-monitoring tool eventually grows out of this line of work, it should remain aggregate, formative, and teacher-facing rather than a tool for watching individual students.

6.3. Limitations

A few limitations are worth stating plainly. At 3,000 images and 9,000 question–answer pairs, the dataset is modest next to large-scale English VQA benchmarks. The images come from Places365, which offers realistic classroom scenes but was never curated to maximize behavioral variety, and this likely caps how well fine-grained actions and postures – especially in the action/state category – are represented. The annotation process also relied on manual verification rather than a formally computed agreement statistic such as Cohen's or Fleiss' kappa; that gap is acknowledged here, and computing such a statistic on a re-annotated sample is planned for a future dataset release. As noted in Section 5.3, the reported results are based on single-run training rather than multiple seeds or confidence intervals, so the small differences between backbones should not be read as more statistically solid than they are – multi-seed replication would be needed before claiming genuine backbone-invariance. And as discussed in Section 6.2, the dataset likely carries some visual and cultural bias from its source imagery, and its fairness across classroom types or demographics has not been externally audited. None of this rules out useful work with the dataset as it stands, but it does set a fairly concrete agenda: more diverse imagery, a proper annotation review, and richer spatial supervision would all directly address these gaps.

7. Conclusion and Future Work

This paper set out to build – and did build – Classroom VQA, a Vietnamese Visual Question Answering dataset for classroom learning-behavior monitoring, along with a PhoBERT–EVA02 baseline to validate it. The dataset covers 3,000 images and 9,000 question–answer pairs across four categories, built through a six-stage pipeline with verification built in. The baseline, using bidirectional Co-Cross-Attention to connect Text-to-Image and Image-to-Text reasoning, reaches 70.78 – 71.44% accuracy across four backbone sizes with very little variation between them. Together, the dataset, the annotation protocol, and the baseline give Vietnamese classroom VQA – and education-oriented multimodal research more broadly – a starting point to build on.

Future work has three fairly concrete directions. First, expanding the dataset with more, and more varied, classroom scenarios, particularly to shore up the counting and action/state categories, and where possible using imagery collected specifically for behavior understanding rather than repurposed scene-recognition photos, which would also help with the source-imagery bias discussed in Section 6.2. Second, strengthening annotation with independent secondary review and a formal inter-annotator agreement measurement. Third, exploring fusion approaches that bring in explicit spatial cues – bounding boxes, pose information – to help with the harder question categories, alongside proper multi-seed statistical evaluation and a look at real-time deployment for classroom

monitoring. More broadly, the intent behind this work is to support formative, teacher-facing tools – something that helps educators notice engagement patterns in aggregate, not something meant to replace their judgment. Turning that into a responsible, real-world tool will mean validating the ethical safeguards discussed in Section 6.2 together with actual educators and, if real classroom deployment is ever pursued, going through formal institutional review first.

Data Availability

The Classroom VQA dataset, including images, the six-stage annotation documentation, and the fixed train/validation/test split described in Section 3, is intended to be made available to the research community upon reasonable request to the corresponding author, subject to the licensing terms of the underlying Places365 imagery.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research received no external funding.

Authors' Contributions

Conceptualization, H.T. and L.L.; Methodology, H.T. and L.L.; Software, L.L.; Validation, H.T. and L.L.; Formal Analysis, H.T. and L.L.; Investigation, L.L.; Resources, L.L.; Data Curation, L.L.; Writing – Original Draft Preparation, H.T. and L.L.; Writing – Review & Editing, H.T.; Visualization, L.L.; Supervision, H.T.; Project Administration, H.T.

Acknowledgments

This study is based on an undergraduate thesis conducted at the Faculty of Information Technology, Ho Chi Minh City University of Education. The authors sincerely thank the faculty members who supported this research, as well as the annotators who contributed to the construction and verification of the Classroom VQA dataset.

References

- [1] William M. Baum, "What Counts as Behavior? The Molar Multiscale View," *The Behavior Analyst*, vol. 36, no. 2, pp. 283-293, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Qingtang Liu, Xinyu Jiang, and Ruyi Jiang, "Classroom Behavior Recognition using Computer Vision: A Systematic Review," *Sensors*, vol. 25, no. 2, pp. 1-22, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Stanislaw Antol et al., "VQA: Visual Question Answering," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425-2433, 2015. [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Yash Goyal et al., "Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904-6913, 2013. [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Mengye Ren, Ryan Kiros, and Richard Zemel, "Exploring Models and Data for Image Question Answering," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1-9, 2015. [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Justin Johnson et al., "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901-2910, 2017. [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Jason J. Lau et al., "A Dataset of Clinically Generated Visual Questions and Answers about Radiology Images," *Scientific Data*, vol. 5, no. 1, pp. 1-10, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [8] Khanh Quoc Tran et al., "ViVQA: Vietnamese Visual Question Answering," *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pp. 1-9, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Dat Quoc Nguyen, and Anh Tuan Nguyen, "PhoBERT: Pre-Trained Language Models for Vietnamese," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1037-1042, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Yuxin Fang et al., "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 19358-19369, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Byeong Su Kim et al., "Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges," *ACM Computing Surveys*, vol. 57, no. 10, pp. 1-35, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, "Ask your Neurons: A Neural-based Approach to Answering Questions about Images," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1-9, 2015. [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Haoyuan Gao et al., "Are you Talking to a Machine? Dataset and Methods for Multilingual Image Question," *Advances in Neural Information Processing Systems*, vol. 28, 2015. [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Jiayuan Mao et al., "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision," *arXiv preprint*, pp. 1-28, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Remi Cadene et al., "RUBi: Reducing Unimodal Biases for Visual Question Answering," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1-12, 2019. [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer, "Don't Take the Easy Way Out: Ensemble based Methods for Avoiding Known Dataset Biases," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4069-4082, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Mateusz Malinowski, and Mario Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1-9, 2014. [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Bolei Zhou et al., "Places: A 10 Million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452-1464, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Kushal Kafle, and Christopher Kanan, "Visual Question Answering: Datasets, Algorithms, and Future Challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3-20, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Kiet Van Nguyen et al., "UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis," *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, Ho Chi Minh City, Vietnam, pp. 19-24, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]