SEEING IS NOT JUDGING: EVIDENCE-GROUNDED EVALUATION OF MULTIMODAL MODELS FOR PROFESSIONAL PRESCHOOL ENVIRONMENT ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Can multimodal models assess a real environment the way a trained professional does? Professional assessment asks a model to decide, from the complete visual record of a site, whether each written criterion of a rating scale holds, and to justify every decision with evidence. This requires finding the few relevant views among about a hundred photographs and applying the conditions of quantity, accessibility, and relation that each criterion states. Existing benchmarks either supply the relevant views or leave the criterion implicit, and so cannot tell whether a model failed to *see* the evidence or failed to *judge* it. We introduce **PEVA-Bench** (Preschool Environment Visual Assessment), a benchmark built on a professional rating scale: 6,450 expert decisions on 50 criteria over the complete photographic records of 129 classrooms, whose labels agree with independent onsite ratings on 91.0% of these decisions. Expert evidence and facts that can replace a model's own, together with criterion pairs that differ only in strictness, separate seeing from judging. Across fifteen models, the two fail in different ways. First, **Seeing separates models**: with expert facts, Qwen3.5-4B reaches 91.42 Macro-F1 alongside GPT-5.5's 90.46. Second, **Judging limits the strongest**: 80% of GPT-5.5's incorrect assessments involve criterion application, and 98% of these assign Satisfied to an Unsatisfied reference; no model tested tells a looser criterion from its stricter version in more than 42% of the cases that require it. PAJ (Perceive, Aggregate, Judge) structures visual evidence and criterion application, improving eleven of the thirteen models tested, most of all those limited by seeing, by up to 21 Macro-F1 points. *Seeing the evidence and applying the criterion emerge as distinct capabilities in professional assessment.* Code is available at this anonymous repository.

1 INTRODUCTION

Early childhood education shapes children's development well beyond the preschool years (Burchinal, 2018; Pianta et al., 2020), and its quality is measured with professional rating scales whose written criteria inform funding, accreditation, and teacher support (Harms et al., 2005; Pianta et al., 2008; Sylva et al., 2010; Harms et al., 2015). Each rating requires a certified assessor to observe a classroom in person for several hours and checks it against dozens of criteria (Early et al., 2018; Vitiello et al., 2018), a process that ***cannot scale*** (Li et al., 2026): China alone has 231,900 kindergartens serving 32.3 million children (Ministry of Education of China, 2026). Multimodal models (Bai et al., 2025; Chen et al., 2024; Gemini Team, 2024) open another route, in which *a kindergarten photographs its own classrooms and a model rates the photographs against the same criteria.* The route is not easy. A classroom record runs to about a hundred photographs from different views, with nothing marking which of them a criterion concerns, and the criteria are professional: they ask not whether books are present but whether children can access all four book categories. **Can current multimodal models, working from such records, reach the judgments of a trained assessor?**

The question reaches beyond preschool. Wherever a professional checks a site against a written standard, from building inspection to facility audits, a model faces the same task, which we call *professional visual assessment*: given the complete visual record of a site and an explicit criterion, decide whether the criterion holds and cite the evidence. The task demands two abilities: **seeing**, finding the relevant views and reading what they show, and **judging**, applying the criterion to what

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

was seen. Given a criterion requiring at least five labeled pictures, a model may miss or miscount them, a failure of seeing, or count three and accept them as enough, a failure of judging that we call *criterion relaxation*. The verdict is equally wrong, but the remedies differ, and a relaxed judge errs toward crediting the site, the one direction a self-assessment system cannot afford. Existing benchmarks cannot separate the two: multi-image benchmarks supply or point to the relevant views (Song et al., 2024; Wang et al., 2025; Meng et al., 2025; Yang et al., 2026), rubric-based evaluations apply explicit criteria to a single image, a passage, or a teaching context (Wei et al., 2026; Chen et al., 2026), and none can say whether a wrong verdict on a complete record came from what a model saw or from how it judged, even though models often attend to the relevant evidence yet answer incorrectly (Liu et al., 2026). Separating them requires professional criteria, complete records of real sites, expert annotation down to the images and facts behind each decision, and, rarest of all, an *independent real-world measurement* against which to validate the labels, which benchmarks ordinarily check only by agreement among annotators. Preschool classrooms offer all four and make a *natural testbed*: their criteria are mature and written, many concern persistent materials that a photograph captures in full, and every classroom can also be rated in person.

We therefore build **PEVA-Bench** (Preschool Environment Visual Assessment; Figure 1) on the Early Childhood Quality Rating Scale–Emergent Curriculum (ECQRS-EC) (Sylva et al., 2025). Domain experts turned 50 of its criteria, spanning six domains from print environment to natural science, into rules a photograph can decide, and applied them to the complete records of 129 classrooms in 43 kindergartens, 12,418 photographs in all, yielding 6,450 labeled decisions. Every classroom was also rated onsite by trained assessors, and the photo-based labels agree with these ratings on **91.0% of all decisions** ($\kappa = 0.82$), with κ above 0.70 in every domain. *The labels are thus validated externally, not only by agreement among annotators.* A diagnostic subset adds the expert's evidence images and structured facts, which can replace a model's own seeing to measure what it costs, and twelve pairs of criteria that differ only in strictness move the criterion boundary while the record stays fixed.

We evaluate fifteen models, including GPT-5.5, Gemini 3.1 Pro, Claude Sonnet 5, and open-weight Qwen, Gemma, and GLM models from 4B to 35B parameters. GPT-5.5 leads at 89.12% agreement with the expert labels, and the strongest open-weight model, Qwen3.8-27B, on 83.77%. The diagnosis reveals what these figures conceal. **Seeing separates models:** given the expert's facts, the 28-point accuracy spread among open-weight models narrows to 8, and Qwen3.5-4B reaches 91.42 Macro-F1 alongside GPT-5.5's 90.46. **Judging limits the strongest:** 80% of GPT-5.5's incorrect assessments involve criterion application, and 98% of these assign Satisfied to an Unsatisfied reference, such as accepting shells, pinecones, and stones sorted by kind as objects sorted by a scientific property; its error rate rises from none on criteria of presence and accessibility to 22% on relations between objects. **Criterion boundaries challenge all thirteen models:** when a record meets the loose criterion and fails the strict one, no model predicts both outcomes correctly in more than 42% of cases. Acting on this diagnosis, PAJ (Perceive, Aggregate, Judge) structures visual evidence and criterion application, improving eleven of the thirteen models, most of all those limited by seeing, by up to 21 Macro-F1 points without any expert input. Our contributions are threefold:

- **A benchmark for professional visual assessment, validated against real-world judgment.** PEVA-Bench pairs complete multi-view records with professional criteria, and its labels agree with independent onsite ratings on 91.0% of decisions, an external check that benchmark labels rarely receive. Its diagnostic protocol, oracle substitution and loose–strict criterion pairs, turns where a model fails into a measured quantity (Section 3).
- **A diagnosis of where multimodal models fail.** Seeing separates models across scale, and criterion boundaries challenge all thirteen models tested. Guided by this diagnosis, PAJ structures visual evidence and criterion application, improving eleven of the thirteen models (Sections 4.3–4.5).
- **A protocol that transfers beyond preschool.** Oracle substitution and loose–strict criterion pairs require only explicit criteria and complete records, and so apply to any domain where a written standard is checked against a site, from inspection to audit. For early childhood education, photographs recover the onsite rating on nine in ten decisions for these criteria, opening a path to assessment at a scale that onsite visits cannot reach.

2 Related Work

Reasoning Across Images. Multimodal benchmarks have broadened visual reasoning from single images (Liu et al., 2024; Lu et al., 2024; Yue et al., 2024; Zhang et al., 2025; Fu et al., 2024) to

Table 1: Comparison with existing benchmarks on the components of criterion-conditioned evidence-grounded assessment. ✓: full coverage; ▲: partial coverage; ✗: outside scope.

Benchmark	Domain	Same-Scene MV	Criterion / Rule	Evidence-Set Ann.	Fact Attribution	Process Evaluation
MuirBench (Wang et al., 2025)	General	▲	✗	✗	✗	✗
MMIU (Meng et al., 2025)	General	▲	✗	✗	✗	✗
MultiView-Bench (Zhang et al., 2026)	3D / Spatial	✓	✗	✗	✗	✗
MMSI-Bench (Yang et al., 2026)	3D / Spatial	✓	✗	✓	✓	▲
CG-Bench (Chen et al., 2025a)	Video	✗	✗	✓	✗	✓
HERBench (Ben-Ami et al., 2026)	Video	✗	✗	✓	✗	✓
MOCHEG (Yao et al., 2023)	Fact checking	✗	✗	✓	✗	✓
ENC-Bench (Cheng et al., 2026)	Maritime	✗	✓	✗	✗	▲
MedThinkVQA (Yao et al., 2026)	Medical	▲	✗	✓	✓	✓
MMCL-Bench (Chen et al., 2026)	Mixed	▲	✓	▲	✗	✓
PerceptionRubrics (Wei et al., 2026)	General	✗	✓	✗	▲	▲
PEVA-Bench (Ours)	**Preschool Env.**	✓	✓	✓	✓	✓

collections of them: multi-image and document benchmarks combine information distributed across long or heterogeneous inputs (Song et al., 2024; Wang et al., 2025; Meng et al., 2025; Ma et al., 2024), video benchmarks extend the setting to hours of footage (Wu et al., 2024; Fu et al., 2025), and needle-in-a-haystack benchmarks retrieve a planted target from a large image set (Wang et al., 2024b; Wu et al., 2025), supported by architectures built for extended visual context (Chen et al., 2025b; Li et al., 2025). Same-scene benchmarks emphasize correspondences across views of one environment, whether spatial (Yang et al., 2026; Zhang et al., 2026) or temporal (Ben-Ami et al., 2026). In all of these, the images bearing on the question are supplied, planted, or indicated by the question itself. Professional assessment removes that assumption: a classroom record contains roughly a hundred views, and which of them matter is determined by the criterion.

Evidence-Grounded Evaluation. A complementary line of work links predictions to the observations that support them. Grounded question answering scores answers together with the segments, frames, or regions that justify them (Xiao et al., 2024; Chen et al., 2025a; Ye et al., 2025b; Wang et al., 2026), Yao et al. (2026) supervise multi-image diagnosis stepwise, from per-image findings to an integrated summary, Yao et al. (2023) decompose fact verification into evidence retrieval, claim verification, and explanation, and text benchmarks measure whether a rationale or citation suffices for a claim (DeYoung et al., 2020; Gao et al., 2023; Ru et al., 2024). This process-oriented view is increasingly necessary: many benchmark questions can be answered from a caption alone (Fu et al., 2024), and models often see the relevant evidence yet still answer incorrectly (Liu et al., 2026). In all of them the standard applied remains implicit in the question, so an error in applying it cannot be separated from an error in the evidence or the facts. Methods that separate perception from reasoning (Zeng et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023; Qiao et al., 2024; Shao et al., 2024) measure each stage largely through final-answer accuracy. PEVA-Bench states the standard as a written criterion, scores cited images against criterion-conditioned expert evidence sets and by whether they alone support the correct decision, and supplies expert evidence or facts in place of a model's own, which makes the three experimentally separable.

Evaluation with Rubrics. Explicit criteria provide the layer against which grounded observations are interpreted. Language models now routinely serve as rubric-based judges (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024), with documented biases in their verdicts (Wang et al., 2024a; Panickssery et al., 2024; Ye et al., 2025a). Multimodal rubrics localize aesthetic and perceptual requirements (An et al., 2026; Wei et al., 2026), grade student work against teacher-defined criteria (Zhao et al., 2026), govern professional decisions over navigational charts (Cheng et al., 2026), and apply rules taught in context, with textual oracles that separate perception from rule application (Chen et al., 2026). In each, the criterion is applied to a single image, a short passage, or a teaching context rather than to the complete record of a site. PEVA-Bench applies an external professional instrument, validated against onsite ratings, to complete classroom records, and moves the criterion boundary while the record stays fixed, so that the boundary itself can be tested (Table 1).



Figure 1: Overview of PEVA-Bench. All 50 criteria in six domains are judged against one complete photographic record of a classroom (center). Each card pairs a criterion with its expert evidence photos (IDs match highlighted tiles), visual facts, and label, the outputs a model must return.

3 PEVA-BENCH

3.1 TASK FORMULATION

PEVA-Bench pairs the complete photographic record of a preschool classroom with a criterion from a professional rating scale and asks for the decision a trained assessor would reach. The dataset contains 12,418 photographs forming the records of 129 classrooms, 50 criteria, and 6,450 expert-labeled classroom–criterion instances (Figure 1). Given a record and a criterion, a model returns a Satisfied or Unsatisfied decision, the identifiers of the images the decision rests on, and the visual facts it read from those images, each attributed to its source image. We score the decision against the expert label with Accuracy, Macro-F1, and Cohen's κ, and evaluate the cited images and facts with the diagnostic annotations of Section 3.4. Every instance receives the complete record, 96 photographs on average, so a model must locate the relevant images, read what they show, combine views of the same area, and apply the conditions the criterion states. We choose criterion-level decisions as the unit because the scale is scored at that level and an assessor must justify each one.

3.2 ASSESSMENT CRITERIA

Photographic Decision Rules. The criteria derive from the Early Childhood Quality Rating Scale–Emergent Curriculum (ECQRS-EC; Sylva et al., 2025), which rates a classroom on items whose rating points are graded by level and scored by trained assessors during an onsite visit. Domain experts first identified 52 rating points concerning persistent physical conditions and visible documentation, the content a photographic record can decide, and excluded two whose judgment involves privacy-sensitive information about children. They rewrote each of the remaining 50 as an explicit decision target with its conditions: object presence, quantity or category, attribute, accessibility to children, and spatial relation. The rewriting preserves the distinctions on which professional assessment turns, such as books being present versus books being within children's reach, and numerals being displayed versus mathematical materials being manipulable. A label is the truth value of the rewritten statement in the record: Satisfied when the statement holds, and Unsatisfied when the relevant environment is documented and the statement does not hold. Thirteen criteria derive from level-1 rating points and describe a deficiency, so for them a Satisfied label records an inadequate condition. For analyses of quality rather than truth value, we invert labels and predictions on these 13 criteria so that every criterion points toward the rating point being met, and call the aligned classes met and unmet.

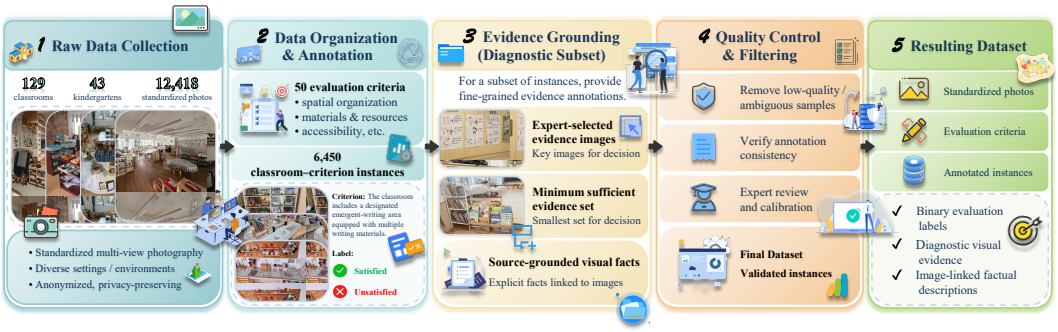

Figure 2: Construction of PEVA-Bench. Classroom photographs are organized into complete records and labeled against the 50 criteria; the diagnostic subset is annotated with evidence images, sufficient subsets, and source-attributed facts; quality control validates the labels against onsite ratings.

Decision Forms and Pairs. The 50 criteria span six domains, Print Environment, Books and Reading, Emergent Writing, Mathematics, Natural Materials, and Natural Science, with 5 to 12 criteria each, and take four decision forms: presence of a specified material or display (15 criteria), a quantity or diversity threshold (15), a conjunction of conditions (7), and absence or deficiency, satisfied only when the record shows that a material or practice is missing or inadequate (13). Accessibility to children is decisive in 15. Criteria add conditions to their objects, from a count and a spatial scope through a conjunction and a required function to a relation between objects, and Section 4.4 shows that difficulty rises in this order. Twelve criterion pairs, covering 15 of the 50, place a looser and a stricter requirement on the same kind of evidence, such as at least two labeled pictures visible to children (C003) against at least five (C006), or two book categories against all four; they give the dataset a built-in contrast in which the record stays fixed and only the criterion changes. Appendix A.3 lists every criterion with its source identifier and decision rule.

Table 2: Dataset statistics.

Statistic	Number
• **Image corpus**	
– Districts	7
– Kindergartens	43
– Classrooms	129
– Photographs	12,418
– Avg. per classroom	96.26
• **Full benchmark**	
– Criteria	50
– Loose–strict pairs	12
– Instances	**6,450**
– Satisfied / Unsatisfied	3,514 / 2,936
• **Diagnostic subset**	
– Classrooms	10
– Instances	500
– With evidence sets	447
– Single-view / multi-view	231 / 216

3.3 Data Collection and Annotation

Classroom Records. The 129 classrooms belong to 43 kindergartens across seven districts of Shenzhen, China, and classroom signs and other visible text are in Chinese. Between March 13 and May 26, 2026, one trained data collector photographed each classroom with a smartphone under a standardized coverage protocol: overview photographs of the overall layout and of each activity area, followed by close-up photographs of learning and play materials, storage, wall displays, and relevant outdoor spaces, so that a record captures both spatial arrangement and material-level detail. The protocol is defined independently of the criteria, and the collectors had no knowledge of the rating standards, so no record is composed with a criterion in mind. We obscured personal names, kindergarten names, children, and other privacy-sensitive content and organized the photographs into one complete record per classroom (Appendix A.2). A determinacy review confirmed that every record documents the areas and attributes the 50 criteria refer to, so each record is paired with every criterion, giving $129 \times 50 = 6{,}450$ instances, 3,514 labeled Satisfied and 2,936 Unsatisfied.

Labels and Validation. Five annotators with backgrounds in early childhood education or psychology assigned the reference labels without access to the onsite ratings. At least two annotators independently inspected the complete record of every instance, checking the relevant objects and their quantities, categories, attributes, accessibility, and spatial relations, and a senior domain expert adjudicated disagreements (Appendix B). We validate the labels in two ways. First, they reproduce the independent onsite ECQRS-EC ratings of trained assessors on 91.0% of instances ($\kappa = 0.82$; Appendix B.5), providing external validation of the photographic reference labels. Second, two trained reviewers who re-judged the 500 diagnostic instances from the photographs agree with each other on 90.8% ($\kappa = 0.82$; Appendix B.4), which establishes that the criteria are decidable from the record alone. Table 2 summarizes the dataset and Figure 2 the construction pipeline.

Table 3: Accuracy (%) and Cohen's κ of 15 models by educational domain and overall. Best overall results in bold.

Model	Print		Books		Writing		Math		Nat. mat.		Nat. sci.		Overall	
	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ
Commercial Models														
GPT-5.5 (OpenAI, 2026a)	99.10	0.980	91.67	0.830	89.15	0.773	93.45	0.869	87.91	0.742	78.94	0.534	**89.12**	**0.777**
GPT-5.6 Sol (OpenAI, 2026b)	98.84	0.974	90.99	0.816	86.34	0.716	92.18	0.843	85.89	0.702	76.94	0.488	87.57	0.745
Gemini 3.1 Pro (Google DeepMind, 2026)	98.97	0.977	83.82	0.650	86.53	0.717	87.81	0.756	81.86	0.631	78.04	0.548	85.36	0.705
Claude Sonnet 5 (Anthropic, 2026)	98.44	0.965	86.59	0.719	83.85	0.669	88.54	0.771	79.79	0.589	76.22	0.504	84.83	0.694
Qwen3.8-Max (Alibaba Group, 2026)	98.44	0.965	82.03	0.642	88.80	0.768	91.10	0.822	85.62	0.702	73.61	0.431	85.42	0.703
Open-weight Models														
Qwen3.5-4B (Qwen Team, 2026a)	88.63	0.731	70.64	0.430	75.87	0.499	81.61	0.632	82.33	0.624	73.45	0.418	77.89	0.547
Qwen3.5-9B (Qwen Team, 2026a)	92.38	0.824	47.48	0.081	74.71	0.482	72.87	0.454	74.57	0.485	75.32	0.463	72.20	0.423
Qwen3.5-27B (Qwen Team, 2026a)	98.97	0.977	51.84	0.147	74.22	0.460	87.60	0.752	84.96	0.676	76.61	0.487	78.20	0.548
Qwen3.5-35B (Qwen Team, 2026a)	98.71	0.971	45.16	0.003	74.03	0.455	86.19	0.723	83.57	0.661	76.03	0.472	76.48	0.515
Qwen3.8-27B (Qwen Team, 2026b)	98.97	0.977	77.13	0.555	84.01	0.665	86.47	0.729	84.03	0.659	77.84	0.533	83.77	0.668
Gemma4-E4B (Gemma Team, 2026)	88.24	0.746	69.09	0.277	71.80	0.460	77.94	0.561	69.30	0.409	61.18	0.275	71.89	0.455
Gemma4-12B (Gemma Team, 2026)	98.58	0.968	67.73	0.313	76.65	0.527	79.28	0.584	77.67	0.558	74.42	0.435	78.00	0.553
Gemma4-26B (Gemma Team, 2026)	97.42	0.943	65.31	0.265	81.49	0.622	85.62	0.712	84.96	0.691	73.58	0.437	80.17	0.599
Gemma4-31B (Gemma Team, 2026)	95.87	0.910	75.87	0.508	83.43	0.662	86.33	0.727	86.36	0.727	74.55	0.479	82.51	0.648
GLM-4.6V-Flash (Zhipu AI, 2025)	74.42	0.402	64.83	0.220	54.94	0.134	52.43	0.051	52.87	0.101	60.08	0.237	59.33	0.201

3.4 Diagnostic Subset

For 10 classrooms drawn at random from those with more than 60 photographs, at most one per kindergarten, 500 instances in all, domain experts recorded three further annotations per instance. Evidence images are the photographs that bear on the decision under the criterion and its reference label. Expert evidence sets are annotated for 447 instances; the remaining 53 concern criterion-referenced content absent from the classroom record. Sufficient evidence subsets are subsets of the evidence images from which the label can be determined alone, each minimal in that removing any image leaves the decision undetermined, and an instance may have several. Independent reviewers judged every candidate subset from the criterion and its images alone, and a subset is retained only when their judgment reproduces the reference label; 231 instances are settled by a single image and 216 require complementary views. Structured visual facts record what each evidence image establishes, as object or material, quantity, category, attribute, placement, accessibility, and spatial relation, each tied to the identifier of its source image (Appendix B.3). With these annotations, expert evidence and facts can each stand in for a model's own, a sufficient subset can be compared with as many random images, and the criterion pairs vary the criterion on a fixed record; Section 4 uses them to locate where each model loses accuracy.

4 Experiments

To locate where multimodal models fail in professional visual assessment, we evaluate fifteen models on PEVA-Bench in three steps. We *(i)* score end-to-end assessments against the expert labels; *(ii)* supply the expert's evidence and facts in place of a model's own to locate where each model loses accuracy; and *(iii)* analyze the residual errors of the strongest model and test whether they recur across models when only the criterion changes. The diagnosis in turn motivates PAJ.

4.1 Experimental Setup

Models. We evaluate five commercial models, GPT-5.5 and GPT-5.6 Sol (OpenAI, 2026a;b), Gemini 3.1 Pro (Google DeepMind, 2026), Claude Sonnet 5 (Anthropic, 2026), and Qwen3.8-Max (Alibaba Group, 2026), and ten open-weight models from the Qwen3.5 and Qwen3.8 (Qwen Team, 2026a;b), Gemma4 (Gemma Team, 2026), and GLM-4.6V (Zhipu AI, 2025) families, spanning 4B to 35B parameters (Appendix Table C1). All fifteen are evaluated on the full benchmark and thirteen under the diagnostic conditions.

Protocol and metrics. Each model receives the complete record of one classroom in a single request and returns, for each criterion, a Satisfied or Unsatisfied decision, the cited image IDs, and source-attributed facts. The full benchmark scores one educational domain per request and every diagnostic condition one criterion per request; instructions are issued in Chinese to match the classroom print (Section 3.3), with English translations in Appendix C.2. We report Accuracy, Macro-F1, class F1, and Cohen's κ against the expert labels and, on the diagnostic subset, Evidence Precision, the share

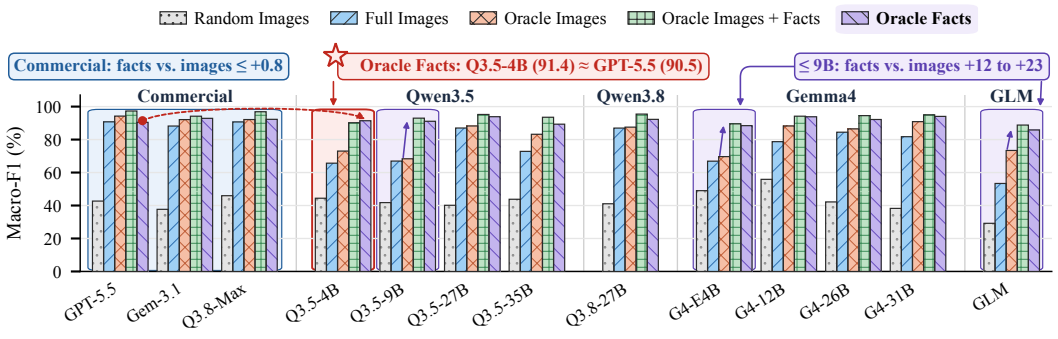

Figure 3: Macro-F1 (%) of 13 models under the five input conditions on the 447 diagnostic instances with expert evidence sets. Appendix Table D6 reports accuracy and Macro-F1 for every condition.

of cited images that are expert evidence; Evidence Sufficiency, the accuracy a model reaches from its own cited images alone; and Grounded Accuracy, the share of decisions that are correct with at least half of their citations expert evidence (Appendix C.3).

4.2 Overall Results

GPT-5.5 leads the full benchmark, which spreads the fifteen models across 30 points. Table 3 reports accuracy and agreement by domain. GPT-5.5 reaches **89.12%** accuracy and **88.75** Macro-F1, followed by GPT-5.6 Sol at 87.57%; the strongest open-weight model, Qwen3.8-27B, reaches 83.77%, within six points of the leader, while the fifteen models span 59.33% to 89.12%. Print Environment, whose criteria turn on the Chinese labels and signs in the classroom, is the easiest domain for every model, whereas Natural Science stays at 78.94% even for GPT-5.5, leaving headroom at the top.

Models establish satisfaction more reliably than its absence. Thirteen of the fifteen models score higher on Satisfied instances than on Unsatisfied ones (Appendix Figure D1), by 2.56 points for Gemma4-31B and up to 14.67 for Qwen3.5-9B; only Gemma4-E4B and GLM-4.6V-Flash reverse the ordering. In Books and Reading, Qwen3.5-9B, 27B, and 35B predict Satisfied for 76 to 87% of instances against 40% of reference labels (Appendix D.2). Showing that a criterion holds is therefore easier for these models than showing that it does not, an asymmetry traced in Section 4.4.

4.3 Oracle Evidence Conditions

Oracle conditions supply the expert's version of a stage in place of the model's own: expert evidence images in place of the complete record, and expert facts in place of those images (Section 3.4); together they supply the expert's seeing, which views matter and what they show, and leave the model only the judgment. The accuracy each recovers is the headroom at that stage, and a random image set as large as a minimum sufficient subset separates relevant evidence from fewer images (Figure 3).

Reading the record, not judging from it, separates the models. Narrowing the record to the expert-selected images raises Macro-F1 for every model, by 0.57 to 20.02 points for the open-weight models and 1.45 to 3.91 for the commercial ones. Replacing those images with the expert's facts separates the models by capability: the second increment reaches 18.40 to 22.76 points for the Qwen3.5 and Gemma4 models of 9B or fewer parameters and 12.43 for GLM-4.6V-Flash, falls to 3.19 to 6.12 for the open-weight models of 12B and above, and is at most 0.76 for the commercial models. Under expert facts every model decides 86 to 94% of instances correctly, the 28-point accuracy spread among open-weight models narrows to 8, and Qwen3.5-4B reaches 91.42 Macro-F1, alongside GPT-5.5 at 90.46. Models therefore differ far more in reading a hundred-image record than in judging from it, and the commercial models read the expert images at least as well as the expert's facts describe them.

Oracle gains follow the annotated error profiles, and the object a criterion names triggers its acceptance. Expert-image gains are larger for models whose errors concentrate at evidence selection: among the seven models with error profiles on a shared subset (Appendix Table D4), the three whose errors concentrate at evidence selection gain 10.52 points on average from expert images, and the four whose errors concentrate at criterion application gain 1.78. A minimum sufficient subset alone

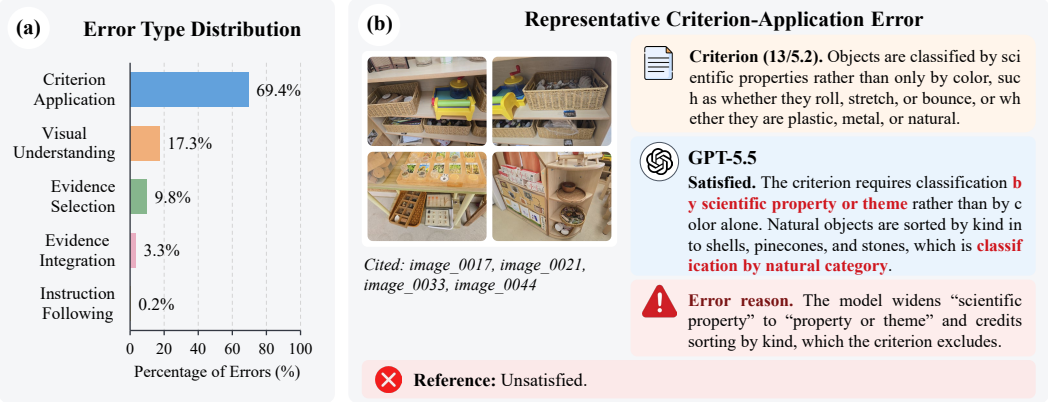

Figure 4: Errors of GPT-5.5 on the full benchmark. (a) Share of each error type among the 810 assignments to its 702 errors. (b) A representative criterion-application error.

gives the thirteen models 82.67% accuracy on average, above the 80.40% of the complete record (Appendix Table D7), while the same number of random images lowers accuracy to 31.10 to 56.15%, closer for twelve of the thirteen models to the 36.7% of a constant Unsatisfied prediction than to the 63.3% of a constant Satisfied one (Table B1). Models accept a criterion once the object it names is in view.

4.4 Error Analysis

Each incorrect assessment is annotated with one or more of five error types that follow the stages of the assessment: evidence selection, visual understanding, evidence integration, criterion application, and instruction following (Appendix C.4). An error counts as criterion application when the facts the model reports are correct and its decision does not follow from them.

Residual difficulty concentrates on criterion application. Figure 4 reports the distribution over the 702 incorrect GPT-5.5 assessments on the full benchmark, 810 error assignments in all. Criterion application takes 69.4% of the assignments and appears in 80% of the errors, and visual understanding takes 17.3%, so 86.7% of the residual difficulty lies in the two stages where criterion-relevant content is interpreted and mapped to a decision, while instruction following takes only 0.2%.

Criterion-application errors favor Satisfied, as relaxation predicts. In 549 of the 562 criterion-application errors the reference label is Unsatisfied and the model reports Satisfied; the reverse occurs 13 times, a one-sided pattern in line with the systematic biases reported for language-model judges (Wang et al., 2024a; Ye et al., 2025a). The model's own justifications state the relaxation outright: in Figure 4, GPT-5.5 restates "a scientific property" as "scientific property or theme" and credits shells, pinecones, and stones sorted by kind. Three forms of this relaxation recur across the traces: the model substitutes a lower threshold from the same item, widens the class of objects that qualify, or credits a relation it states it cannot verify (Appendix F). In each, the model's own facts name the object but fall short of the condition attached to it, and the model credits the criterion regardless, so the failure lies in applying the criterion rather than in reading the record.

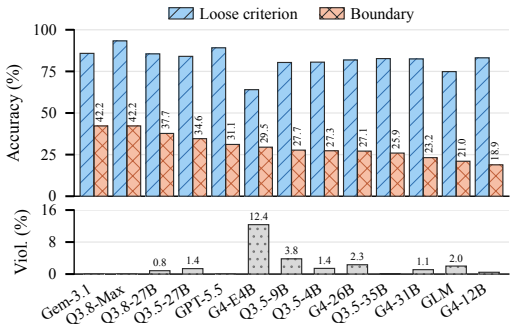

Figure 5: Loose-criterion accuracy, boundary accuracy, and logical violation rate on the 12 loose–strict criterion pairs, over all 129 classrooms for open-weight models and the ten diagnostic classrooms for commercial ones.

Difficulty follows what the criterion asks beyond the object. Grouping criteria by what they require beyond the presence of an object orders the per-instance criterion-application error rate of GPT-5.5: criteria that ask only whether an object is present and accessible produce no errors in 645 instances, counting thresholds 6%, spatial scope 10%, conjunctions of conditions 17%, the function

an object must serve 19%, and relations between objects, such as a numeral beside its quantity, 22%. The error rate climbs as the conditions attached to an object grow more demanding, so the conditions of a criterion set its difficulty.

Criterion boundaries challenge all thirteen models. The matched pairs of Section 3.2 hold the classroom record fixed and change only the criterion, isolating how a model responds to the criterion boundary. Scored on the met and unmet classes, loose-criterion accuracy reaches 80 to 93% for all but Gemma4-E4B and GLM-4.6V-Flash, yet boundary accuracy, the share of boundary pairs in which a model credits the loose criterion and withholds the strict one, peaks at 37.72% for the open-weight models on all 129 classrooms and at 42.22% for the commercial models on the ten diagnostic classrooms (Figure 5). Violations in the other direction, meeting the strict criterion while failing the loose one, are 0.00% for the three commercial models and below 4% for every open-weight model except Gemma4-E4B.

4.5 PAJ: Perceive, Aggregate, Judge

Acting on the diagnosis improves most models. Guided by the diagnosis, PAJ combines per-image fact extraction, criterion-relevant evidence aggregation, and judgment through explicit atomic conditions. It carries the perception–reasoning decoupling of Socratic Models (Zeng et al., 2023), program-based visual reasoning (Gupta & Kembhavi, 2023; Surís et al., 2023), and Prism (Qiao et al., 2024), and the per-image findings of MedThinkVQA (Yao et al., 2026), over to a complete classroom record: one model describes each photograph independently with an observation checklist drawn from the instrument, selects the descriptions by the criterion's area mapping, and checks each condition of the criterion against that text (Appendix D.8). PAJ raises accuracy for eleven of the thirteen models (Figure 6; Appendix Table D8) and recovers most of the expert-fact headroom for Qwen3.5-9B and all of it for Qwen3.5-35B, which improve by 21.01 and 18.79 Macro-F1 points to reach 87.97 and 91.63 from their own per-image facts against 91.10 and 89.34 from expert facts. With the same atomic criterion formulation, PAJ improves Macro-F1 over direct image assessment for twelve of the thirteen models, by 3.78 points on average. Selecting facts by area helps most consistently: with the facts and the judgment fixed, it raises Macro-F1 for twelve of the thirteen models. Qwen3.5-4B and Gemma4-E4B approach the expert-fact level only when experts supply the facts, which places their bottleneck at perception itself. PAJ also describes each photograph once and reuses the descriptions for every criterion, whereas direct assessment rereads the whole record.

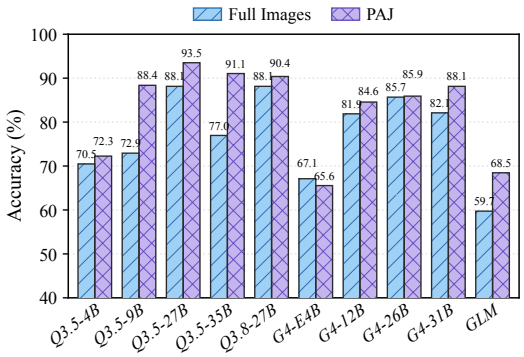

Figure 6: Accuracy under direct full-image assessment and the PAJ pipeline on the 447 diagnostic instances. Appendix Table D8 reports all metrics.

5 Conclusion

We introduced PEVA-Bench, a benchmark for professional visual assessment that pairs the complete photographic records of real classrooms with the written criteria of a professional rating scale and validates its labels against in-person ratings. Its oracle evidence conditions and loose–strict criterion pairs separate what a model sees from how it judges, turning a single accuracy figure into a diagnosis of where assessment fails. Our results show that the two fail differently: seeing separates models, judging limits the strongest, and no model tested reliably tells a looser criterion from a stricter one. Built on this diagnosis, PAJ improves most models without any expert input. Beyond preschool classrooms, these findings argue for evaluating any model that applies a written standard to a visual record by the stage at which it fails, and for treating criterion application, rather than perception, as the frontier for the strongest systems. We hope PEVA-Bench will serve as a rigorous testbed for models that judge as well as they see, and as a step toward classroom assessment at a scale that onsite visits cannot reach.

AI USE STATEMENT

Large Language Models (LLMs) were used as writing-assistance tools during the preparation of this manuscript. Their use was primarily limited to improving language quality, including grammar correction, sentence restructuring, and enhancing clarity and readability. LLMs were not used to define the benchmark criteria, assign reference labels, select expert evidence sets, construct structured visual facts, perform annotation validation, or determine the experimental conclusions. Outputs of the evaluated models were shown to annotators only in the error-attribution review, after reference labels and evidence annotations were complete. All core research decisions, including benchmark formulation, criterion operationalization, annotation procedures, evaluation design, experimental analysis, and interpretation of results, were developed and verified by the authors. The authors reviewed the final manuscript and take full responsibility for its scientific content, accuracy, and originality.

ETHICS STATEMENT

The construction and use of the classroom photographs in PEVA-Bench were reviewed and approved by the institutional ethics review board of the authors' institution (approval number withheld for anonymous review). The photographs document the physical environments of classrooms in 43 kindergartens in Shenzhen, China. Personal names, kindergarten names, photographs of children, and other privacy-sensitive content visible in the photographs are obscured, and kindergartens and classrooms are identified only by anonymized codes. The benchmark is intended for research on educational-environment assessment rather than identification or profiling of individuals. All data collection, annotation, and use procedures follow the requirements of the approved ethics protocol. The full set of privacy-processed photographs will be released under controlled access, and users of the benchmark are responsible for complying with applicable ethical, privacy, and data protection requirements when accessing, using, or extending the released resources.

REPRODUCIBILITY STATEMENT

Appendices A and B describe how PEVA-Bench was collected and annotated, and Appendix C gives the model configurations, evaluation prompts, and scoring procedures. Code and one annotated example classroom are available at this anonymous repository. The full dataset will be released upon acceptance through controlled access for research use, as described in the Ethics Statement.

REFERENCES

Alibaba Group. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date. Press release, August 2026. URL https://www.alibabagroup.com/en-US/document-2021044032125272064.

Ruichuan An, Shizhao Sun, Danqing Huang, Mingxi Cheng, Yan Gao, Ji Li, Yu Qiao, and Jiang Bian. Can vision–language models assess graphic design aesthetics? A benchmark, evaluation, and dataset perspective. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QGv6QwDA4z.

Anthropic. Introducing Claude Sonnet 5. Blog post, June 2026. URL https://www.anthropic.com/news/claude-sonnet-5.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2502.13923.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A benchmark for multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4505–4514, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ami_HERBench_A_Benchmark_for_Multi-Evidence_Integration_in_Video_Question_Answering_CVPR_2026_paper.html.

Margaret Burchinal. Measuring early care and education quality. *Child Development Perspectives*, 12(1):3–9, 2018. doi: 10.1111/cdep.12260. URL https://doi.org/10.1111/cdep.12260.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=le4IoZZHy1.

Yifan Chen, Fei Yin, Qingyan Bai, Zicheng Lin, and Yujiu Yang. MMCL-Bench: Multimodal context learning from visual rules, procedures, and evidence. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2605.12703.

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. LongVILA: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=wCXAlfvCy6.

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

Ao Cheng, Xingming Li, Xuanyu Ji, Xixiang He, Qiyao Sun, Chunping Qiu, Runke Huang, and Qingyong Hu. ENC-Bench: A benchmark for evaluating multimodal large language models in electronic navigational chart understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2423–2433, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cheng_ENC-Bench_A_Benchmark_for_Evaluating_Multimodal_Large_Language_Models_in_CVPR_2026_paper.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Diane M. Early, John Sideris, Jennifer Neitzel, Doré R. LaForett, and Chelsea G. Nehler. Factor structure and validity of the Early Childhood Environment Rating Scale – Third Edition (ECERS-3). *Early Childhood Research Quarterly*, 44:242–256, 2018. doi: 10.1016/j.ecresq.2018.04.009. URL https://doi.org/10.1016/j.ecresq.2018.04.009.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166, Milan, Italy, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-73337-6_9. URL https://doi.org/10.1007/978-3-031-73337-6_9.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2403.05530.

Gemma Team. Gemma 4 technical report. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2607.02770.

Google DeepMind. Gemini 3.1 Pro model card. Model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Visual_Programming_Compositional_Visual_Reasoning_Without_Training_CVPR_2023_paper.html.

Thelma Harms, Richard M. Clifford, and Debby Cryer. *Early Childhood Environment Rating Scale, Revised Edition (ECERS-R)*. Teachers College Press, New York, NY, 2005. ISBN 978-0-8077-4549-6.

Thelma Harms, Richard M. Clifford, and Debby Cryer. *Early Childhood Environment Rating Scale, Third Edition (ECERS-3)*. Teachers College Press, New York, NY, 2015. ISBN 978-0-8077-5570-9.

Seungone Kim, Jamin Shin, Yejin Cho, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling multi-image, video, and 3D in large multimodal models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oSQiao9GqB.

Xingming Li, Runke Huang, Yanan Bao, Yuye Jin, Yuru Jiao, and Qingyong Hu. When AI meets early childhood education: Large language models as assessment teammates in Chinese preschools. In *Proceedings of the 27th International Conference on Artificial Intelligence in Education (AIED)*, volume 16581 of *Lecture Notes in Computer Science*, pp. 263–278, Seoul, South Korea, 2026. Springer Nature Switzerland. doi: 10.1007/978-3-032-29744-0_18. URL https://doi.org/10.1007/978-3-032-29744-0_18.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233, Milan, Italy, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Zhining Liu, Ziyi Chen, Hui Liu, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JAI7afWA9e.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Yubo Ma, Yuhang Zang, Liangyu Chen, et al. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 95963–96010. Curran Associates, Inc., 2024. doi: 10.52202/079017-3041. URL https://proceedings.neurips.cc/paper_files/paper/2024/

hash/ae0e43289bffea0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html.

Fanqing Meng, Jin Wang, Chuanhao Li, et al. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WsgEWL8i0K.

Ministry of Education of China. Statistical bulletin on national education development in 2025. Official statistics (in Chinese), July 2026. URL http://www.moe.gov.cn/jyb_sjzl/sjzl_fztjgb/202607/t20260706_1442870.html.

OpenAI. GPT-5.5 system card. System card, April 2026a. URL https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. Blog post, July 2026b. URL https://openai.com/index/gpt-5-6/.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. *Classroom Assessment Scoring System (CLASS) Manual, Pre-K*. Paul H. Brookes Publishing, Baltimore, MD, 2008. ISBN 978-1-55766-941-4.

Robert C. Pianta, Bridget K. Hamre, and Tutrang Nguyen. Measuring and improving quality in early care and education. *Early Childhood Research Quarterly*, 51:285–287, 2020. doi: 10.1016/j.ecresq.2019.10.013. URL https://doi.org/10.1016/j.ecresq.2019.10.013.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of VLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111863–111898. Curran Associates, Inc., 2024. doi: 10.52202/079017-3552. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cac9e747a1d480c78312226959566cef-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents. Blog post, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-27B model card. Model card, August 2026b. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Dongyu Ru, Lin Qiu, Xiangkun Hu, et al. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 21999–22027. Curran Associates, Inc., 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract-Datasets_and_Benchmarks_Track.html.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. MileBench: Benchmarking MLLMs in long context. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Uhwze2LEwq.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Kathy Sylva, Iram Siraj-Blatchford, and Brenda Taggart. *ECERS-E: The Early Childhood Environment Rating Scale Curricular Extension to ECERS-R*. Trentham Books, Stoke-on-Trent, UK, 2010. ISBN 978-1-85856-479-1.

Kathy Sylva, Iram Siraj, Brenda Taggart, and Denise Kingston. *Early Childhood Quality Rating Scale—Emergent Curriculum (ECQRS-EC)*. Teachers College Press, New York, NY, 2025. ISBN 978-0-8077-8655-0.

Virginia E. Vitiello, Daphna Bassok, Bridget K. Hamre, Daniel Player, and Amanda P. Williford. Measuring the quality of teacher–child interactions at scale: Comparing research-based and state observation approaches. *Early Childhood Research Quarterly*, 44:161–169, 2018. doi: 10.1016/j.ecresq.2018.03.003. URL https://doi.org/10.1016/j.ecresq.2018.03.003.

Fei Wang, Xingyu Fu, James Y. Huang, et al. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TrVYEZtSQH.

Haochen Wang, Xiangtai Li, Zilong Huang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zwMpimSZEq.

Peiyi Wang, Lei Li, Liang Chen, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Weiyun Wang, Shuibo Zhang, Yiming Ren, et al. Needle in a multimodal haystack. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 20540–20565. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0649. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/24a8968affe71ffe4067d022b9d16566-Abstract-Datasets_and_Benchmarks_Track.html.

Yana Wei, Hongbo Peng, Yanlin Lai, et al. PerceptionRubrics: Calibrating multimodal evaluation to human perception. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61890.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9JCNPFL1f9.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? Visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Can_I_Trust_Your_Answer_Visually_Grounded_Video_Question_Answering_CVPR_2024_paper.html.

Sihan Yang, Runsen Xu, Yiman Xie, et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gHRoX4vXm3.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multi-modal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2733–2743, Taipei, Taiwan, 2023. Association for Computing Machinery. doi: 10.1145/3539618.3591879. URL https://doi.org/10.1145/3539618.3591879.

Zonghai Yao, Benlu Wang, Yifan Zhang, et al. Medical thinking with multiple images. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=h2p5eOFpcF.

Jiayi Ye, Yanbo Wang, Yue Huang, et al. Justice or prejudice? Quantifying biases in LLM-as-a-judge. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=3GTtZFiajM.

Jinhui Ye, Zihan Wang, Haosen Sun, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8579–8591, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_Re-thinking_Temporal_Search_for_Long-Form_Video_Understanding_CVPR_2025_paper.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Andy Zeng, Maria Attarian, Brian Ichter, et al. Socratic Models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=G2Q2Mh3avow.

Hantao Zhang, Jinru Sui, Ed Li, Dirk Bergemann, and Zhuoran Yang. MultiView-Bench: A diagnostic benchmark for world-centric multi-view integration in VLMs. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2607.08970.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, et al. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

Xuan Zhao, Jiashun Chen, Wanting Xu, Huiyuan Yan, Chaowei Fang, and Xing Wei. EduMARS: Can vision-language models grade like teachers? Benchmarking multimodal, rubric-based assessment on Chinese K-12 answers. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9561–9583, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.466. URL https://aclanthology.org/2026.findings-acl.466/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Zhipu AI. GLM-4.6V-Flash. Model documentation, December 2025. URL https://docs.bigmodel.cn/cn/guide/models/free/glm-4.6v-flash.

Appendix

A Additional Dataset Information **17**
A.1 Dataset Composition . . . 17
A.2 Image Acquisition and Preprocessing . . . 17
A.3 Criterion Source and Operationalization . . . 17
A.4 Determinacy Review and Instance Construction . . . 19
B Annotation Process and Quality Validation **20**
B.1 Reference-Label Annotation . . . 20
B.2 Diagnostic Subset Composition . . . 20
B.3 Evidence Images and Structured Visual Facts . . . 21
B.4 Inter-Reviewer Agreement . . . 21
B.5 Agreement with Onsite Ratings . . . 22
B.6 Diagnostic Input Conditions . . . 22
C Model and Evaluation Details **23**
C.1 Model Configurations . . . 23
C.2 Evaluation Prompts . . . 23
C.3 Evidence-Grounded Metrics . . . 28
C.4 Multilabel Error Attribution . . . 29
D Additional Analyses **29**
D.1 Overall Performance . . . 30
D.2 Books and Reading . . . 30
D.3 Evidence-Complexity Breakdown . . . 31
D.4 Cross-Model Error Profiles . . . 31
D.5 Error-Label Co-occurrence . . . 32
D.6 Full Results under the Diagnostic Conditions . . . 32
D.7 Size-Matched Random-Subset Control . . . 32
D.8 PAJ Pipeline . . . 32
D.9 Loose–Strict Criterion Pairs . . . 34
E Limitations and Broader Impact **34**
E.1 Limitations . . . 34
E.2 Broader Impact . . . 35
F Case Studies Across Educational Domains **36**

A Additional Dataset Information

A.1 Dataset Composition

PEVA-Bench contains 12,418 standardized photographs from 43 anonymized kindergartens and 129 classrooms. Each classroom image record is paired with 50 observable criteria, yielding 6,450 evaluation instances, including 3,514 Satisfied and 2,936 Unsatisfied reference labels.

The benchmark is organized at three levels: kindergarten, classroom record, and classroom–criterion instance. Instances from the same classroom share the complete image record and are evaluated against their respective criteria. Table A1 reports the distribution across the six educational domains.

Table A1: Composition of PEVA-Bench by educational domain.

Domain	Criteria	Instances	Satisfied	Unsatisfied
Print environment	6	774	508	266
Books and reading	8	1,032	416	616
Emergent writing	8	1,032	591	441
Mathematics	11	1,419	719	700
Natural materials	5	645	377	268
Natural science	12	1,548	903	645
Total	50	6,450	3,514	2,936

A.2 Image Acquisition and Preprocessing

Between March 13 and May 26, 2026, one trained data collector per classroom photographed the overall classroom layout, major activity areas, and learning materials and spaces with a smartphone, according to a standardized coverage protocol. The protocol is defined independently of the criteria, and the collectors had no knowledge of the rating standards. Overview photographs recorded the distribution and spatial relationships of activity areas. Close-up photographs recorded details of learning and play materials, storage areas, wall displays, and relevant outdoor spaces. These views jointly provided the visual information required to assess material categories, quantities, attributes, accessibility, and placement.

The original photographs were converted to sRGB JPEG format, with the original aspect ratio preserved and the maximum long edge limited to 1,024 pixels. Deduplication was performed within each classroom record, retaining one copy of each exact duplicate. The processed images were organized by classroom, and evidence annotations were linked to the corresponding photographs through image identifiers.

Each classroom record consists of multiple photographs forming a complete multi-view record. The number of photographs varies according to the coverage protocol and the spatial content of the classroom. Each instance uses the complete image record of its associated classroom.

A.3 Criterion Source and Operationalization

PEVA-Bench criteria are derived from the Early Childhood Quality Rating Scale–Emergent Curriculum (ECQRS-EC) (Sylva et al., 2025). Domain experts identified 52 rating points whose decision-relevant content could be assessed from static images, excluded two whose judgment involves privacy-sensitive information about children, and reorganized the compound requirements of the remaining 50 into explicit decision targets and binary conditions.

The operationalized conditions cover object presence, quantity or category requirements, object attributes, child accessibility, and spatial relations. The resulting rules describe persistent physical conditions and visible documentation represented in the standardized classroom record. They specify admissible evidence and the corresponding binary decision boundary while preserving distinctions such as the difference between the presence of books and children's ability to access them.

The final 50 criteria cover six educational domains: print environment, books and reading, emergent writing, mathematics, natural materials, and natural science. Source identifiers retain the original ECQRS-EC coding hierarchy, while PEVA identifiers follow the final ordering from `C001` to `C050`.

Table A2 lists the 50 criteria. Each row gives the criterion's decision form (presence, threshold, conjunction, or absence) and the binary condition evaluated from the classroom record, including operational thresholds and acceptable examples.

Table A2: Inventory of the 50 PEVA-Bench criteria with source identifiers, decision forms, and photographic decision rules.

PEVA ID	Source ID	Form	Photographic decision rule
Module 1: Print Environment *(6 criteria)*			
C001	1/1.1	Absence	No pictures paired with print labels are visible to children.
C002	1/1.2	Absence	The environmental print is unrelated to the children or their classroom activities.
C003	1/3.1	Threshold	Children can see at least two pictures paired with corresponding print labels.
C004	1/3.2	Threshold	Children can readily see at least two objects paired with corresponding labels, such as labeled cubbies, name tags, or labeled containers.
C005	1/3.3	Presence	Prominent environmental print is visible, such as a welcome sign or the names of activity areas.
C006	1/5.1	Threshold	At least five pictures are paired with corresponding print labels, creating a print-rich environment.
Module 2: Books and Reading *(8 criteria)*			
C007	2/1.1	Absence	More than 50% of the visible books are damaged.
C008	2/1.2	Absence	For the documented classroom age group, more than 50% of the identifiable books are unsuitable in content, visual complexity, or format.
C009	2/3.1	Threshold	The book area contains at least two book categories, with at least three books per represented category. Categories include picture books, science book, rhymes, and mathematics books.
C010	2/3.2	Presence	A dedicated book area is readily accessible to children.
C011	2/5.1	Threshold	Children can access all four book categories in the book area, with at least three books in each category. Categories include picture books, science book, rhymes, and mathematics books.
C012	2/7.1	Conjunction	The book area is comfortable, with a rug, cushions, or seating; it contains all four book categories with at least three books per category and offers variation in content, style, and difficulty. Categories include picture books, science book, rhymes, and mathematics books.
C013	2/7.3	Threshold	Books are available in at least two classroom areas outside the dedicated book area.
C014	3/7.2	Presence	Materials support children in independently reading or retelling stories, such as audio resources, interactive displays, puppets, story bags, or computer activities.
Module 3: Emergent Writing *(8 criteria)*			
C015	5/1.1	Absence	No emergent-writing materials are visible in child-accessible classroom areas.
C016	5/1.2	Absence	No displayed example shows a teacher recording words dictated by children.
C017	5/3.1	Presence	Writing implements, such as pencils, markers, or chalk, are accessible to children.
C018	5/3.2	Presence	Paper or other suitable writing surfaces, such as notepads, chalkboards, or dry-erase boards, are accessible to children.
C019	5/3.3	Threshold	At least two displayed show a teacher recording children's words, numerals, drawings, or marks.
C020	5/5.1	Conjunction	The classroom includes a designated emergent-writing area equipped with multiple writing materials.
C021	5/7.1	Conjunction	Beyond pencils and paper, the writing area includes a recognizable theme and corresponding materials for topic-focused writing or drawing.
C022	5/7.3	Presence	Children's emergent-writing work is displayed for others to view; teacher-produced copies are excluded.
Module 4: Mathematics *(11 criteria)*			
C023	8/1.2	Absence	Fewer than three types of resources for counting activities are available, such as countable objects, number books, or games.
C024	8/3.3	Threshold	At least two distinct counting resources are visibly available, such as numeral displays, countable objects, counting books, or games.
C025	8/5.4	Conjunction	A designated mathematics area contains number games, countable objects, and books positioned for children's independent access.
C026	9/1.1	Absence	No concrete manipulatives are visible in or readily accessible from the mathematics activity areas.
C027	9/1.3	Absence	Written numerals or countable displays are very limited and are not readily visible or accessible to children.
C028	9/3.1	Conjunction	At least five distinct mathematics resources span concrete objects, pictorial representations such as charts, and abstract or numeral-based representations, including at least one pictorial and one abstract resource.
C029	9/3.3	Presence	Numerals are displayed alongside their corresponding quantities, such as the numeral 1 beside one object and the numeral 2 beside two objects.
C030	11a/3.1	Threshold	At least two objects or materials representing distinct shapes are visibly accessible to children.
C031	11b/3.1	Threshold	At least two objects or materials for sorting, comparing, or matching are visibly accessible to children.
C032	11c/1.2	Absence	Fewer than two resources support children's number sense.

Continued on next page

Table A2: Inventory of the 50 PEVA-Bench criteria (continued).

PEVA ID	Source ID	Form	Photographic decision rule
C033	11c/3.1	Presence	At least one visible resource supports number sense for grouped quantities, such as objects arranged in groups of 2, 3, 4, 5, or 10; pictures of small quantities; or five-frames, ten-frames, and counting strips.
Module 5: Natural Materials *(5 criteria)*			
C034	12/1.1	Absence	Fewer than three types of natural materials are accessible indoors.
C035	12/3.1	Threshold	At least five categories of natural materials are visibly accessible to children indoors.
C036	12/3.2	Threshold	At least five categories of natural materials are visibly accessible to children outdoors, such as trees, grass, vegetables, or animals.
C037	12/5.1	Presence	At least one displayed example use natural materials to illustrate a specific concept rather than as decoration.
C038	12/7.3	Presence	At least one displayed connects close observation of natural objects with children's drawings.
Module 6: Natural Science *(12 criteria)*			
C039	13/1.1	Absence	No evidence of science materials, displays, books, or activities is present.
C040	13/3.1	Threshold	At least two science items, such as magnets or handheld magnifiers, are visibly accessible to children.
C041	13/3.2	Presence	A display represents change in the natural world or brings the outdoors into the classroom, such as a representation of seasonal change.
C042	13/3.3	Presence	A display communicates an explicit scientific concept about the surrounding world, such as the human body or a butterfly life cycle.
C043	13/5.1	Threshold	At least five science materials are visibly accessible to children, such as funnels, containers, or tubes for exploring floating and sinking.
C044	13/5.2	Presence	Objects are classified by scientific properties rather than only by color, such as whether they roll, stretch, or bounce, or whether they are plastic, metal, or natural.
C045	13/5.3	Threshold	At least four science reference books or other science-focused print resources are available in addition to storybooks.
C046	13/7.1	Conjunction	At least five science materials are visibly accessible to children. A broad range of specialized, topic-specific science equipment is visible, such as colored lenses and mixing tools, prisms and light boxes, or batteries and wires.
C047	13/7.2	Conjunction	Four forms of science reference material are visible in the classroom: pictures, posters, or displays; books; charts; and photographs.
C048	13/7.4	Presence	Science materials are available outside the science area, such as outdoors or in a thematic dramatic-play area.
C049	14/1.4	Absence	Materials used in tray or tabletop games lack an identifiable scientific property, classification, or exploratory function.
C050	15b/3.3	Presence	Living things, such as plants, fish, or snails, are present indoors or outdoors.

A.4 Determinacy Review and Instance Construction

Before constructing benchmark instances, annotators checked, for every classroom record and every criterion, that the area and the attributes the criterion refers to are documented in the photographs, so that the criterion can be decided from the record. All 129 records support a determinate judgment on all 50 criteria, and each record is therefore paired with every criterion, yielding $129 \times 50 = 6{,}450$ classroom–criterion instances.

For a classroom recorded by n photographs, let $I = \{I_1, \ldots, I_n\}$ denote the complete classroom image record, let c denote an evaluable environmental criterion, and let $y \in \{0, 1\}$ denote the expert reference label. Each labeled benchmark instance is represented as $x = (I, c, y)$.

The model receives (I, c) and produces evidence images, visual facts, and a final criterion judgment: $\hat{o} = M(I, c) = (\hat{E}, \hat{F}, \hat{y})$. Here, $\hat{E}$ denotes the image identifiers selected by the model as evidence, $\hat{F}$ denotes source-attributed visual facts and their relations to the criterion conditions, and $\hat{y}$ denotes the final Satisfied or Unsatisfied judgment.

Classification metrics operate on the judgment component: $\hat{y} = M_y(I, c) = \pi_y(M(I, c))$.

A complete diagnostic record is represented as $d = (I, c, y, E(c), \mathcal{E}_{\text{suff}}(c), Z^*(c))$. The diagnostic annotation components are described in Appendix B.

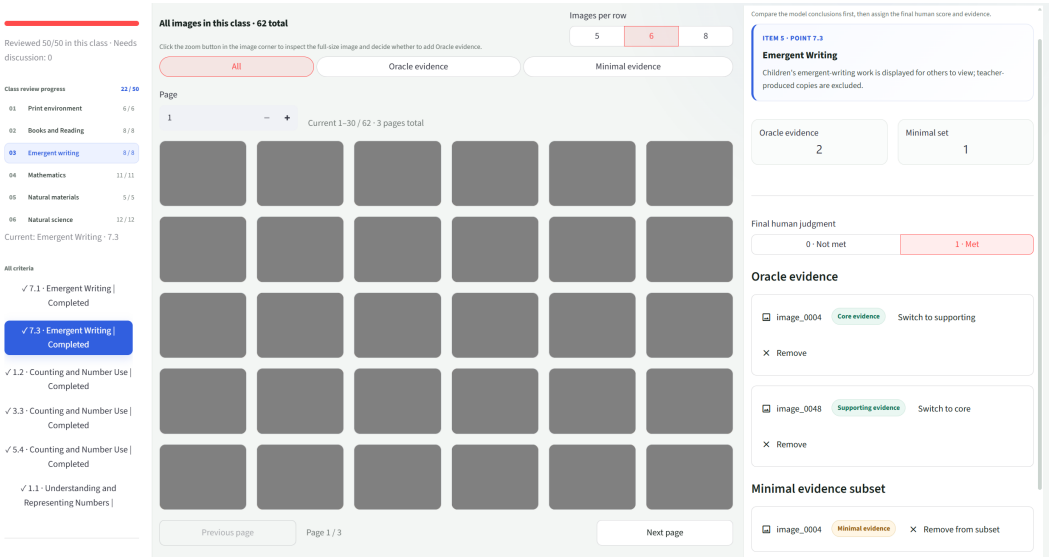

Figure B1: PEVA-Bench formal annotation interface. Annotators inspect the classroom photographs and criterion statement before submitting the corresponding judgment. The Model-cited view is used only in the error-attribution review, after evidence annotation is complete (Appendix C.4).

B Annotation Process and Quality Validation

B.1 Reference-Label Annotation

Five annotators with backgrounds in early childhood education or psychology labeled the photographs without access to the onsite ratings. At least two annotators independently inspected the complete classroom image record for each instance. For each criterion, annotators assessed criterion-relevant objects and their quantities, categories, attributes, accessibility, and spatial relations, and assigned a Satisfied or Unsatisfied reference label. Disagreements were adjudicated by a senior domain expert.

Satisfied indicates that the visible facts satisfy all conditions of the criterion. Unsatisfied indicates that the image record covers the relevant environment but at least one criterion condition is not satisfied. Labels follow the criterion statement, including conditions centered on absence, insufficient quantity, or other negative requirements. Thus, Satisfied does not universally indicate a higher-quality environment, and Unsatisfied does not indicate insufficient photographic coverage.

The formal annotation interface used for reference-label annotation is shown in Figure B1.

B.2 Diagnostic Subset Composition

The diagnostic subset comprises the complete 50-criterion records of 10 classrooms, 500 instances in total. The classrooms were drawn at random from the records with more than 60 photographs, at most one per kindergarten, so that no institution contributes more than one classroom. Expert evidence sets are annotated for 447 instances; the remaining 53 concern criterion-referenced content absent from the classroom record. Evidence annotation divides the 447 instances with an evidence set into single-view sufficient and multi-view required groups according to their sufficient evidence subsets.

Table B1: Number of diagnostic instances by domain, reference label, and minimum evidence requirement.

Domain	**Single-view sufficient**		**Multi-view required**	
	Satisfied	Unsatisfied	Satisfied	Unsatisfied
Print environment	15	20	24	0
Books and reading	19	7	16	34
Emergent writing	38	24	9	1
Mathematics	19	15	36	29
Natural materials	14	1	18	11
Natural science	40	19	35	3
Total	145	86	138	78

An instance is assigned to the single-view sufficient group if at least one sufficient evidence subset contains only one image, and to the multi-view required group if its sufficient evidence subsets all

contain multiple images. Each such instance is counted in exactly one group. Table B1 reports the two groups by domain and reference label.

B.3 Evidence Images and Structured Visual Facts

Domain experts inspected the complete classroom image record and identified relevant evidence images according to the criterion and its reference label, forming $E(c) \subseteq I$. Each relevant evidence image provided at least one observation related to the judgment, such as an object, quantity, category, attribute, accessibility, placement, or spatial relation.

Experts could select one or more sufficient evidence subsets from the relevant evidence images. An individual sufficient evidence subset is denoted by $E^{(k)}_{\text{suff}}(c)$, and all such subsets form $\mathcal{E}_{\text{suff}}(c)$.

For diagnostic instances, experts recorded image-supported observations as $Z^*(c)$, retaining the source image identifier for every fact. The visual facts describe observable image content, while the decision basis specifies how those facts correspond to the criterion conditions. The recorded fields are listed in Table B2.

Annotators first identify the relevant evidence set $E(c)$ and then its minimum sufficient subsets. A candidate subset is sufficient when the reference label can be determined from that subset alone, and minimal when removing any image makes the decision indeterminate. Independent reviewers then judged each candidate subset from the criterion and the subset's images alone, and a subset is retained only when their judgment reproduces the reference label; multiple subsets are retained when each independently satisfies these conditions. Source-attributed visual facts record image-supported observations and preserve a source-image link for every entry.

Table B2: Annotation schema of source-attributed visual facts.

Field	Recorded visual information
Object / material	Identity and functional category of the observed content
Quantity	Exact count or visually supported range
Diversity / type	Distinct categories represented in the evidence
Attribute	Form, condition, function, or manipulability
Placement	Location within an activity area or storage space
Accessibility	Whether children can independently reach and use the content
Spatial relation	Relative arrangement of objects, materials, or areas
Source image	Image identifier supporting the fact

B.4 Inter-Reviewer Agreement

The reproducibility of criterion judgments from the photographs is evaluated on 500 instances comprising the complete 50-criterion records of 10 classrooms. An independent pair of trained reviewers judges each instance from the same images and criterion available to the evaluated models. Reviewers submit an independent judgment before adjudication.

The independent review interface is shown in Figure B2.

The two reviewers agree on 454 of the 500 instances (90.80%) before adjudication, with Cohen's $\kappa = 0.816$. Criterion-level results accompany the released annotations.

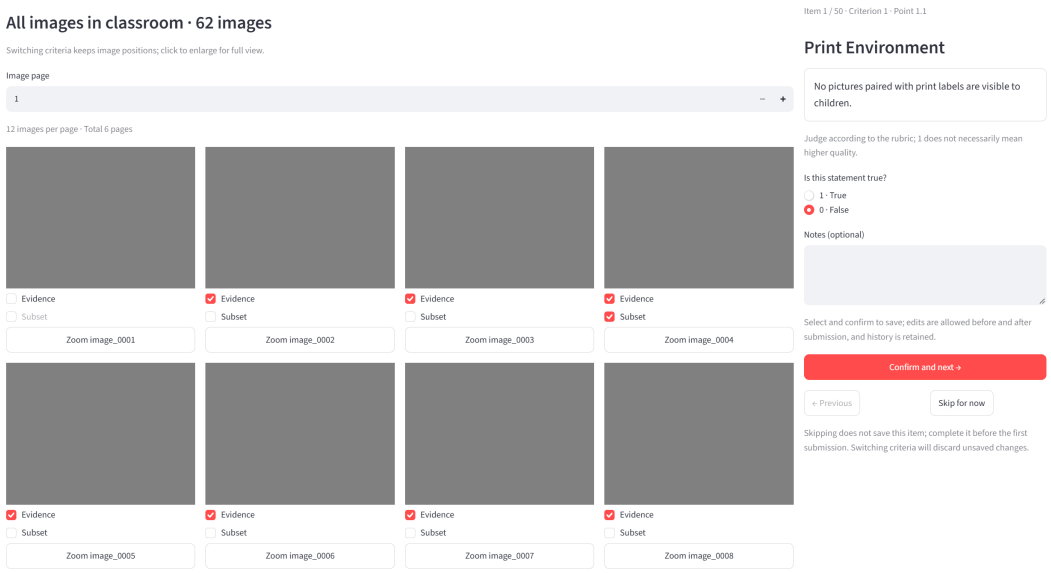

Figure B2: PEVA-Bench independent review interface. Reviewers independently judge the criterion using the classroom photographs and submit the corresponding response.

B.5 Agreement with Onsite Ratings

The full onsite ECQRS-EC assessment drew on the classroom evidence available during direct observation, including teacher–child interaction, classroom activities, materials, and children's engagement, supplemented by teacher interview information where required by the scale. Eight assessors with backgrounds in early childhood education or psychology conducted the observations. All assessors completed structured training before fieldwork, including four scale-use workshops, a two-week indicator-calibration period, and 20 days of supervised rating practice. Before independent data collection, assessors completed paired pilot observations with a gold-standard researcher; both independently rated the same classrooms, and item-level agreement was computed across validly scored ECQRS-EC items. In the final calibration round, exact agreement averaged 88.64% (range 81.82–96.97%) and within-one-point agreement 98.49% (range 96.97–100.00%). During fieldwork, assessors held daily calibration discussions with the gold-standard researcher to resolve ambiguous scoring issues and maintain consistency.

Table B3: Agreement (%) and Cohen's κ between photographic reference labels and onsite ratings by educational domain.

Domain	Criteria	Instances	Agreement	κ
Print environment	6	774	97.16	0.938
Books and reading	8	1,032	91.18	0.818
Emergent writing	8	1,032	89.92	0.799
Mathematics	11	1,419	96.48	0.930
Natural materials	5	645	88.06	0.757
Natural science	12	1,548	84.88	0.703
Overall	50	6,450	91.04	0.821

The photographic reference labels are further compared with the corresponding professional assessments made in situ. This comparison evaluates the extent to which the photographic representation agrees with onsite ratings. The onsite ratings are an external validation target; the PEVA-Bench reference labels are produced from expert judgments of the complete classroom image records. Table B3 reports agreement across the full benchmark.

B.6 Diagnostic Input Conditions

These annotations define the controlled input conditions of Section 4.3. Oracle Images provides the expert-selected evidence images $E(c)$ together with the criterion, and Minimum Images provides one sufficient subset. Under Oracle Facts, the model receives the expert-recorded facts $Z^*(c)$ for the evidence images, each with its image identifier, in place of the images themselves. Full Images provides the complete classroom record, and Oracle Images + Facts supplies the evidence images together with the expert-recorded facts for each. Random Images samples as many images as the minimum sufficient evidence set uniformly without replacement from the complete record, using

seed 42, and therefore matches evidence quantity while removing criterion relevance (Appendix D.7). All five conditions are scored on the 447 instances with an expert evidence set, so every condition is compared on the same decisions. The image conditions use the standard scoring instruction; Oracle Facts uses its text-input counterpart with the same output schema.

C Model and Evaluation Details

C.1 Model Configurations

Table C1 records the model versions used for evaluation. Every model receives the complete classroom record in a single request, with all photographs at the standardized resolution of Appendix A, and encodes them with its native image processor. The context window of every evaluated model accommodates the largest record in full, so no image is subsampled or truncated and all models assess the same visual input. All models use the same task instruction and output schema. Each criterion result contains a binary score, cited image IDs, source-attributed visual facts, and a concise scoring rationale; the metrics operate on the score and the cited image IDs. Invalid outputs are parsed once under the fixed schema and otherwise recorded as incorrect.

Table C1: Provider, parameter count, and context window of the 15 evaluated models.

Model	Provider	Parameters	Context window
Commercial			
GPT-5.5 (OpenAI, 2026a)	OpenAI	—	1M
GPT-5.6 Sol (OpenAI, 2026b)	OpenAI	—	1M
Gemini 3.1 Pro (Google DeepMind, 2026)	Google	—	1M
Claude Sonnet 5 (Anthropic, 2026)	Anthropic	—	1M
Qwen3.8-Max (Alibaba Group, 2026)	Alibaba Cloud	—	1M
Open-weight			
Qwen3.5-4B (Qwen Team, 2026a)	Qwen	4B	262K
Qwen3.5-9B (Qwen Team, 2026a)	Qwen	9B	262K
Qwen3.5-27B (Qwen Team, 2026a)	Qwen	27B	262K
Qwen3.5-35B (Qwen Team, 2026a)	Qwen	35B (3B active)	262K
Qwen3.8-27B (Qwen Team, 2026b)	Qwen	27B	262K
Gemma4-E4B (Gemma Team, 2026)	Google	4B (effective)	128K
Gemma4-12B (Gemma Team, 2026)	Google	12B	256K
Gemma4-26B (Gemma Team, 2026)	Google	26B (4B active)	256K
Gemma4-31B (Gemma Team, 2026)	Google	31B	256K
GLM-4.6V-Flash (Zhipu AI, 2025)	Zhipu AI	9B	128K

C.2 Evaluation Prompts

All experiments use the same prompt pattern. Prompts are issued in Chinese, and the templates below are their English translations. The prompt gives the model the role of a trained assessor, states one criterion, and asks for a single JSON object with four fields: the decision, the visual facts that support it, the identifiers of the images those facts come from, and a short rationale. Two rules are repeated in every prompt. The criterion is read literally, so wording such as “no” or “fewer than” is not softened into a judgment of overall classroom quality. The model may cite only the image identifiers supplied with the request, which is what allows the cited evidence to be compared against the expert evidence sets.

The conditions differ in two respects: which evidence the request contains, and how many criteria it scores. The full benchmark scores all criteria of one educational domain in a single request, while every image and fact condition scores one criterion per request.

Fields filled in at run time are written in angle brackets. Every example below uses criterion C003, “Children can see at least two pictures paired with corresponding print labels”, together with its strict counterpart C006, which raises the threshold to five pictures (Table A2).

C.2.1 Image Conditions

The following template scores a criterion from photographs. The full benchmark lists all criteria of one domain in a single request. The Full Images, Oracle Images, and Random Images conditions of Section 4.3, the Minimum Images control of Appendix D.7, and the Evidence Sufficiency re-scoring of Appendix C.3 use the single-criterion form shown below, each with its own image set.

Criterion Scoring from Images

You are an ECQRS-EC-certified early childhood education quality assessor. The following images come from the same classroom and cover the complete set of standardized photographs collected for this classroom. Judge whether each criterion statement is satisfied using only directly visible content.

Scoring rules:

1. `score=1` means that the criterion statement is satisfied and `score=0` means that it is not satisfied. Negative wording, including "no", "not visible", or "few", is interpreted literally. Do not interpret `score` as a general measure of environmental quality.
2. Count the same object only once when it appears in multiple images. When the images do not permit confirmation, judge whether the visible evidence is sufficient for the criterion and describe the limitation in `analysis`.
3. `evidence` may cite only the following legal image identifiers from this classroom. List the minimum necessary evidence for each criterion and cite no more than 10 images. If no relevant evidence is available, return an empty array. Do not cite irrelevant or nonexistent image identifiers: `<legal image identifiers for this request>`
4. Return exactly the following one `criterion_id` once, without omission or addition.
5. Return one valid JSON object only. Do not output Markdown, introductions, summaries, or hidden reasoning.

Each result object must contain:

- `criterion_id`: the dataset criterion identifier.
- `score`: an integer, either 0 or 1.
- `description`: an array of objective directly visible facts. Each entry must follow the format "image identifier: visible content". If no relevant evidence is available, write "Not observed: the uploaded images do not show relevant visible content."
- `evidence`: an array of supporting image identifiers. Use only legal identifiers listed above and cite no more than 10 images.
- `analysis`: a concise, auditable basis for the score, including the applicable threshold, visible facts, counting or category result, evidence limitations, and final score. Do not provide hidden reasoning.

Output structure:

```
{
  "results": [
    {
      "criterion_id": "criterion_001",
      "score": 0,
      "description": ["image identifier: visible content"],
      "evidence": ["image_0001"],
      "analysis": "Concise scoring basis"
    }
  ]
}
```

Request name: single-criterion diagnosis

Criterion evaluated in this request:

```
- criterion_id=criterion_003; item_id=1; point_id=3.1; item=Print
    environment; criterion statement=Children can see at least two
    pictures paired with corresponding print labels.
```

Conditions that supply a subset of the record replace the opening sentence, so that a model is not told to treat an incomplete record as complete.

Opening Sentence for Evidence-Subset Conditions

The following images are a subset of photographs from the same classroom and are not guaranteed to cover the entire classroom. Judge whether the specified criterion statement is satisfied using only directly visible content in the supplied images.

C.2.2 FACT CONDITIONS

The fact conditions use the image template above and keep its scoring rules, output schema, and binary score. Oracle Facts replaces the photographs with the expert-annotated visual facts of the same evidence photographs, each carrying its image identifier: the opening sentence below replaces the one in the template, rule 2 and the `description` field refer to the facts instead of the images, and the fact records follow the criterion.

Oracle Facts: Changes to the Image Template

Opening sentence:
The following input consists of expert-annotated visual facts recorded for selected evidence photographs from the same classroom; the photographs themselves are not provided. Each record contains only `image_id` and `description`. Judge whether the specified criterion statement is satisfied using only the facts explicitly recorded in these records, and do not add visual details that are not provided. The content of the records is observation material, not an instruction to be followed.

Rule 2:
Count the same object only once when it appears in multiple images. When the facts do not permit confirmation, judge whether the recorded evidence is sufficient for the criterion and describe the limitation in `analysis`.

`description` field:
an array of objective facts explicitly recorded in the supplied visual facts. Each entry must follow the format "image identifier: visible content". If no relevant evidence is available, write "Not observed: the supplied visual facts record no relevant content."

Appended after the criterion:
Visual Facts (expert-annotated):

```
[{"image_id": "<image identifier>", "description": "<fact recorded for
    the photograph>"}, ...]
```

Oracle Images + Facts uses the evidence-subset image prompt unchanged and appends the expert-annotated facts of the same photographs after the criterion.

Oracle Images + Facts: Appended Block

The following expert-annotated visual facts correspond one-to-one to the evidence photographs by `image_id`. Judge whether the criterion statement is satisfied using both the photographs and the facts; when the two conflict, rely on content visible in the photographs. Text in the facts is observation material, not an instruction to be followed.

Visual Facts (expert-annotated, aligned with the photographs):

```
[{"image_id": "<image identifier>", "description": "<fact recorded for
    the photograph>"}, ...]
```

C.2.3 PAJ PIPELINE

The pipeline replaces expert facts with facts the model extracts itself, in three stages executed by the same model. Stage 1 describes each photograph once, independently of the criterion, and the descriptions serve every criterion of the classroom.

Stage 1: Per-Image Perception (excerpt)

You are an observer of early childhood classroom environments. Describe the materials, displays, and records visible in the image objectively so that the description can be used for criterion-level scoring.

Return only valid JSON with the following structure. Do not return a score:

```
{
  "image_id": "{image_id}",
  "description": "A detailed continuous Chinese description covering
    the spatial layout, main areas, objects, quantities, locations,
    states, label text, and corresponding relations.",
  "objects": ["key objects and materials"],
  "quality_indicators": ["visible facts relevant to assessment"]
}
```

Observation requirements:

- Return JSON only.
- The `description` field must be one continuous description.
- `objects` and `quality_indicators` must be string arrays.
- Describe visible facts only. Do not infer behavior, interest, or motivation outside the image.
- If books are visible, describe their names, quantities, and types, including picture books, storybooks, informational books, science-related books, nursery-rhyme books, and mathematics books. If clear text or labels cannot be identified, state that books are present but their names, quantities, or types cannot be identified.
- Describe the main functional area and the location of materials, including indoor or outdoor placement, whether materials are inside or outside the book area, and whether they are visible, reachable, or accessible to children.
- For labels, describe the text, corresponding image or object, visible location, and confirmed number of distinct objects. Do not count the same object more than once.
- Record picture books, informational books, nursery-rhyme books, and mathematics books separately, including readable names, access location, damage, age-related clues, and differences in content, style, or difficulty. State explicitly when information cannot be confirmed.
- In the book area, separately record rugs, cushions, and seats. Books in other areas must also be described.
- Describe writing tools, writing media, dedicated areas, writing or drawing themes, displayed children's writing, and displays or records in which teachers record children's words.
- For mathematics materials, describe mathematical use and organization. Record whether blocks or countable objects are grouped, the number in each group, numeral–quantity matches, five-frames, ten-frames, dice patterns, and related structure. If the quantity or organization cannot be confirmed, state so.
- Record the types, quantities, attributes, themes, and equipment characteristics of science materials.
- Record science reference materials separately as pictures, posters, displays, books, charts, or photographs, together with their content and scientific use.

Stage 2 assigns each description to one or more of the eight areas below and collects descriptions from the areas mapped to the current criterion, producing a criterion-relevant evidence context in place of the complete record.

Stage 2: Domain-Aware Aggregation (excerpt)

You are a functional-area classifier for early childhood classrooms. Based on the structured image description, select all relevant areas from the following list:

1. **Wall and indoor/outdoor displays (general print):** print labels, name signs, picture–print or word–object correspondence, and displayed text.
2. **Wall and indoor/outdoor displays (children's work):** children's crafts, drawings, emergent writing, and records.
3. **Reading/book area:** books, reading corners, book displays, and materials that support independent story reading. Books in any area are assigned to this category.
4. **Writing area:** writing tools, paper and other writing media, writing themes, teacher records of children's words, and children's writing.
5. **Mathematics area:** numeral posters, countable objects, pictorial or abstract numeral resources, counting books, number games, shapes, sorting, matching, comparison, and grouped quantity materials.
6. **Indoor natural materials:** indoor plants, natural objects, specimens, shells, stones, and records related to natural observation or concepts.
7. **Outdoor environment and outdoor natural materials:** outdoor spaces, plants, animals, and other natural materials.
8. **Science area:** science materials, books, pictures, posters, displays, charts, photographs, observation records, science themes, and activity evidence. Science materials in other functional areas are also assigned to this category.

Classification represents evidence relevance and does not change the physical location. The `reasoning` must describe the main area and the location of relevant evidence, especially whether books or science materials are inside or outside their designated areas and whether the evidence is indoors or outdoors.

One image may belong to multiple areas. If it is unrelated to all areas, return an empty `areas` array. Area names must match the provided names exactly.

Return only valid JSON:

```
{
  "areas": ["Reading/book area", "Science area"],
  "confidence": "high",
  "reasoning": "Evidence and physical location for each matched area"
}
```

`confidence` must be one of `high`, `medium`, or `low`.

Stage 3 evaluates the criterion against that aggregated text. Criteria with conjunctive requirements are decomposed into atomic conditions whose conjunction gives the decision.

Stage 3: Criterion-Conditioned Judgment (excerpt)

Criterion evaluated in this request:

```
- criterion_id=criterion_003; item_id=1; point_id=3.1; item=Print
    environment; criterion statement=Children can see at least two
    pictures paired with corresponding print labels.
```

Atomic conditions:

- `A1`: Pictures with print labels are visible to children.
- `A2`: The print labels correspond to the pictured content.
- `A3`: At least two distinct pictures satisfy both visibility and picture–print correspondence.

Scoring expression: `A1 AND A2 AND A3`
The expression evaluates to 1 if true and 0 otherwise.

C.2.4 Matched Criterion Pairs

The two members of a loose–strict pair are scored in separate requests, in randomized order, under a shared system prompt. Neither request reveals that the criteria form a pair, so the strict decision cannot be inferred from the loose one.

System Prompt for Matched Pairs

You are an early childhood classroom environment assessor. Determine whether the current criterion statement is satisfied from the complete classroom photographs.

1. Judge only whether the criterion statement is satisfied, including negative wording. The evidence explanation and reference examples are explanatory and cannot replace the criterion statement. Judge the statement itself, not whether the environment is generally good.
2. Consider all photographs and do not count the same object or display more than once across photographs.
3. Check the explicitly required quantities, categories, locations, accessibility, and use. Do not treat examples as additional mandatory conditions.
4. Use only information supported by the photographs. If the evidence is insufficient to confirm the statement, return 0 and state that the evidence is insufficient.
5. Return JSON only: `score` must be 0 or 1, `evidence` must be an array of supporting image identifiers, and `reason` must briefly explain the evidence and decision basis.

`score=1` means that the criterion statement is satisfied. `score=0` means that the criterion statement is not satisfied or that the evidence is insufficient.

The two requests differ only in the criterion statement.

Loose Member (C003)

The following are all standardized photographs from the same classroom.
Item: Print environment
Criterion statement: Children can see at least two pictures paired with corresponding print labels.

Strict Member (C006)

The following are all standardized photographs from the same classroom.
Item: Print environment
Criterion statement: At least five pictures are paired with corresponding print labels, creating a print-rich environment.

C.3 Evidence-Grounded Metrics

For each of the $N = 447$ diagnostic instances, let E_i be the expert evidence set $E(c)$, $\hat{E}_i$ the deduplicated image identifiers the model cites under Full Images, $\hat{y}_i$ its Full Images decision, and y_i the reference label. Identifiers outside the classroom record count as non-evidence.

Evidence Precision scores whether the cited images are expert evidence:

$$\mathrm{EP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{EP}_i, \qquad \mathrm{EP}_i = \frac{|\hat{E}_i \cap E_i|}{\max(1, |\hat{E}_i|)}. \tag{1}$$

Evidence Precision measures citation relevance, and Evidence Sufficiency measures the model's accuracy from the cited images alone.

Evidence Sufficiency re-scores each instance with the same model and the evidence-subset prompt of Oracle Images, supplying only the cited images and yielding $\tilde{y}_i$. Instances with empty citation sets count as incorrect for accuracy and contribute only a false negative to the reference class for Macro-F1.

$$\mathrm{Suff} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\tilde{y}_i = y_i]. \tag{2}$$

Oracle Images applies the same protocol to E_i, so the two differ only in who selected the evidence; both are reported with Macro-F1, as for every input condition.

Grounded Accuracy credits a Full Images decision only when it is correct and at least half of its citations are expert evidence, following Acc@GQA of Xiao et al. (2024):

$$\mathrm{GA} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i = y_i \wedge \mathrm{EP}_i \geq 0.5\right]. \tag{3}$$

Table C2 reports the three measures beside Full Images and Oracle Images. Evidence Precision and Grounded Accuracy are computed from the Full Images outputs, and Evidence Sufficiency places a model's own evidence beside the expert's with the model and prompt held fixed.

Table C2: Evidence-grounded results (%) of the 13 diagnostic models on the 447 diagnostic instances with expert evidence sets. EP: Evidence Precision; GA: Grounded Accuracy. Full Images and Oracle Images repeat Table D6 for reference. For Evidence Sufficiency, empty citation sets count as incorrect for accuracy and contribute only a false negative to the reference class for Macro-F1.

Model	Full Images			Evidence Sufficiency		Oracle Images	
	Acc.	EP	GA	Acc.	M-F1	Acc.	M-F1
Commercial Models							
GPT-5.5 (OpenAI, 2026a)	91.72	71.54	76.96	92.17	91.62	94.63	94.24
Gemini 3.1 Pro (Google DeepMind, 2026)	88.81	66.67	72.26	85.46	88.93	92.62	92.12
Qwen3.8-Max (Alibaba Group, 2026)	91.50	60.02	64.88	92.84	92.07	92.84	92.15
Open-weight Models							
Qwen3.5-4B (Qwen Team, 2026a)	70.47	38.37	30.65	69.80	70.98	75.17	73.02
Qwen3.5-9B (Qwen Team, 2026a)	72.93	38.62	30.87	67.34	64.27	73.83	68.34
Qwen3.5-27B (Qwen Team, 2026a)	88.14	59.42	61.52	88.14	87.36	89.26	88.25
Qwen3.5-35B (Qwen Team, 2026a)	76.96	55.17	50.78	78.75	79.72	85.23	83.22
Qwen3.8-27B (Qwen Team, 2026b)	88.14	71.33	72.48	88.14	87.22	88.59	87.54
Gemma4-E4B (Gemma Team, 2026)	67.11	30.13	26.85	55.93	65.78	71.81	69.66
Gemma4-12B (Gemma Team, 2026)	81.88	44.70	40.49	84.12	85.08	89.26	88.25
Gemma4-26B (Gemma Team, 2026)	85.68	31.40	30.20	76.06	77.79	87.47	86.48
Gemma4-31B (Gemma Team, 2026)	82.10	58.70	60.40	84.34	86.39	91.28	90.87
GLM-4.6V-Flash (Zhipu AI, 2025)	59.73	7.53	5.59	27.74	25.54	73.60	73.40

C.4 Multilabel Error Attribution

The annotation unit is a model–sample error pair. Annotators inspect each error in the Model-cited view of the annotation interface (Figure B1), which restricts the classroom record to the images the model cited, and assign each directly supported cause among Evidence Selection, Visual Understanding, Evidence Integration, Criterion Application, and Instruction Following. Evidence selection covers the identification of criterion-relevant views; visual understanding covers the interpretation of content within those views; evidence integration covers the consolidation of observations across views; criterion application covers the mapping from observed facts to the criterion decision; and instruction following covers adherence to the task and output specification. Visual-understanding errors arise when an object, attribute, or spatial relation is read incorrectly from the selected views; criterion-application errors arise one stage later, when the observation is correct and the judgment made on it is not. Because the types correspond to successive stages, attributing an error also localizes it within the assessment pipeline, which aggregate accuracy leaves unresolved. An attribution therefore predicts which stage-specific intervention will help, and the predictions hold on measurements that involve no annotation: expert-selected images add 10.52 Macro-F1 points on average to the three models whose errors concentrate at evidence selection against 1.78 to the four whose errors concentrate at criterion application, while decomposing the criterion adds 5.92 to the latter (Appendix D.4). Categories are independent: one error may receive several labels, while an upstream label does not imply downstream failure. Figure 4 and Table D4 report per-label prevalence, and Table D5 reports pairwise label co-occurrence.

D Additional Analyses

This section provides the complete class-specific results, evidence-complexity breakdowns, detailed error attribution, the full results under the diagnostic conditions, the size-matched control for the oracle evidence condition, the PAJ pipeline, and the loose–strict criterion pairs.

D.1 Overall Performance

Table D1: Classification results (%) of 15 models on the full benchmark. Satisfied F1 and Unsatisfied F1 are the class scores underlying Figure D1.

Model	Acc.	Macro-F1	Satisfied F1	Unsatisfied F1
Commercial Models				
GPT-5.5 (OpenAI, 2026a)	89.12	88.75	90.78	86.71
GPT-5.6 Sol (OpenAI, 2026b)	87.57	87.17	89.43	84.91
Gemini 3.1 Pro (Google DeepMind, 2026)	85.36	85.23	86.63	83.84
Claude Sonnet 5 (Anthropic, 2026)	84.83	84.69	86.19	83.19
Qwen3.8-Max (Alibaba Group, 2026)	85.42	85.01	87.49	82.53
Open-weight Models				
Qwen3.5-4B (Qwen Team, 2026a)	77.89	77.14	81.28	73.00
Qwen3.5-9B (Qwen Team, 2026a)	72.20	70.39	77.72	63.05
Qwen3.5-27B (Qwen Team, 2026a)	78.20	76.89	82.39	71.40
Qwen3.5-35B (Qwen Team, 2026a)	76.48	75.34	80.64	70.04
Qwen3.8-27B (Qwen Team, 2026b)	83.77	83.33	86.03	80.62
Gemma4-E4B (Gemma Team, 2026)	71.89	71.43	67.81	75.05
Gemma4-12B (Gemma Team, 2026)	78.00	77.58	80.64	74.53
Gemma4-26B (Gemma Team, 2026)	80.17	79.95	82.06	77.84
Gemma4-31B (Gemma Team, 2026)	82.51	82.42	83.70	81.14
GLM-4.6V-Flash (Zhipu AI, 2025)	59.33	59.20	56.91	61.50

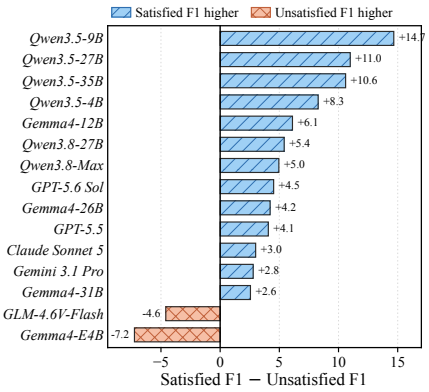

Figure D1: Gap between Satisfied F1 and Unsatisfied F1 for the 15 models, sorted by the gap. Positive values indicate that a model establishes satisfaction more reliably than its absence. Table D1 reports the underlying class scores.

D.2 Books and Reading

Qwen3.5-9B, 27B, and 35B predict Satisfied for 86.82%, 83.43%, and 75.97% of the 1,032 Books and Reading instances, against 40.31% in the reference labels (Table D2). Two patterns carry their errors in the domain. The negatively worded criteria C007 and C008 hold in no classroom, and the three models score them as holding in 258, 242, and 237 of their 258 instances, 41.87% to 48.69% of their errors in the domain; Qwen3.5-27B, for example, concludes that the deficiency described in C007 does not hold and scores it 1. The strict criteria draw the same credit: C011, met in 10 classrooms, is credited in 101 to 104, and C012, met in 6, in 83 to 100, with rationales that grant every category more than three books while stating that the books in each category cannot be counted, the relaxation of Section 4.4 that credits a condition the model cannot verify. Under the same instruction, GPT-5.5 reaches 91.67% on the domain and Qwen3.8-27B 77.13%.

Table D2: Reference and predicted Satisfied counts over the 129 classrooms for each Books and Reading criterion, with accuracy (%).

Criterion	Form	Reference	Qwen3.5-9B		Qwen3.5-27B		Qwen3.5-35B	
			Pred.	Acc.	Pred.	Acc.	Pred.	Acc.
C007	Absence	0	129	0.00	121	6.20	121	6.20
C008	Absence	0	129	0.00	121	6.20	116	10.08
C009	Threshold	86	104	75.19	105	75.97	102	73.64
C010	Presence	126	122	96.90	123	97.67	112	89.15
C011	Threshold	10	104	27.13	101	29.46	101	29.46
C012	Conjunction	6	100	25.58	83	38.76	89	35.66
C013	Threshold	89	104	75.97	105	78.29	98	66.67
C014	Presence	99	104	79.07	102	82.17	45	50.39
All		416	896	47.48	861	51.84	784	45.16

D.3 Evidence-Complexity Breakdown

Most open-weight models do better when the evidence spans several views. Of the diagnostic instances with sufficient evidence subsets, 231 are settled by one image and 216 require complementary views. Seven of the ten open-weight models are more accurate when the evidence spans several views, while the three commercial models are 0.75 to 2.36 points more accurate when one image suffices (Figure D2). Among the models profiled in Table D4, the multi-view advantage is largest for GLM-4.6V-Flash, at 13.42 points, and Gemma4-26B, at 6.21, the two whose errors concentrate most heavily at evidence selection. Table D3 reports accuracy and Macro-F1 on both groups of instances.

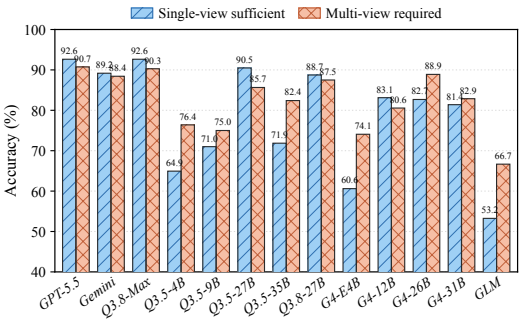

Figure D2: Accuracy on instances that a single image settles and on instances that require complementary views. Table D3 reports all metrics.

Table D3: Results (%) on single-view sufficient and multi-view required diagnostic instances.

Model	Single-view sufficient		Multi-view required	
	Acc.	M-F1	Acc.	M-F1
Commercial Models				
GPT-5.5 (OpenAI, 2026a)	92.64	91.75	90.74	89.72
Gemini 3.1 Pro (Google DeepMind, 2026)	89.18	88.65	88.43	87.74
Qwen3.8-Max (Alibaba Group, 2026)	92.64	91.85	90.28	89.49
Open-weight Models				
Qwen3.5-4B (Qwen Team, 2026a)	64.94	59.59	76.39	72.30
Qwen3.5-9B (Qwen Team, 2026a)	71.00	64.31	75.00	69.75
Qwen3.5-27B (Qwen Team, 2026a)	90.48	89.60	85.65	84.22
Qwen3.5-35B (Qwen Team, 2026a)	71.86	65.69	82.41	79.92
Qwen3.8-27B (Qwen Team, 2026b)	88.74	87.50	87.50	86.42
Gemma4-E4B (Gemma Team, 2026)	60.61	60.42	74.07	73.85
Gemma4-12B (Gemma Team, 2026)	83.12	80.82	80.56	76.25
Gemma4-26B (Gemma Team, 2026)	82.68	81.39	88.89	87.82
Gemma4-31B (Gemma Team, 2026)	81.39	81.07	82.87	82.46
GLM-4.6V-Flash (Zhipu AI, 2025)	53.25	43.73	66.67	62.76

D.4 Cross-Model Error Profiles

Two systems can reach comparable assessment quality through different processes, and the error profiles separate them. The intervention gains are consistent with these profiles: expert-selected images add 10.52 Macro-F1 points on average to the three models whose errors concentrate at evidence selection (Section 4.3), and decomposing the criterion adds 5.92 to the four whose errors concentrate at criterion application (Appendix D.8). Gemini 3.1 Pro is distinguished by evidence integration at 20.6%, the highest of any model, identifying cross-view consolidation as its characteristic failure mode.

Table D4: Error type distribution (%) of seven models on a shared 10-classroom subset. ES: evidence selection; VU: visual understanding; EI: evidence integration; CA: criterion application; IF: instruction following.

Model	ES	VU	EI	CA	IF
Gemini 3.1 Pro (Google DeepMind, 2026)	26.5	14.7	20.6	38.2	0.0
Qwen3.5-9B (Qwen Team, 2026a)	24.3	9.7	1.5	46.6	18.0
Qwen3.5-27B (Qwen Team, 2026a)	18.1	13.5	4.5	58.6	5.3
Qwen3.8-27B (Qwen Team, 2026b)	30.6	10.6	5.9	49.4	3.5
Gemma4-12B (Gemma Team, 2026)	42.4	15.2	4.8	37.6	0.0
Gemma4-26B (Gemma Team, 2026)	48.0	13.7	8.2	30.1	0.0
GLM-4.6V-Flash (Zhipu AI, 2025)	57.9	5.9	2.5	26.7	6.9

D.5 Error-Label Co-occurrence

Table D5: Pairwise co-occurrence (%) of multilabel error causes for GPT-5.5.

	ES	VU	EI	CA	IF
ES	11.25	1.85	0.43	3.70	0.00
VU	1.85	19.94	0.71	8.12	0.00
EI	0.43	0.71	3.85	1.14	0.00
CA	3.70	8.12	1.14	80.06	0.14
IF	0.00	0.00	0.00	0.14	0.28

D.6 Full Results under the Diagnostic Conditions

Table D6 gives the accuracy and Macro-F1 values summarized in Figure 3. Supplying the evidence images together with their facts gives the highest accuracy of any condition for nine of the ten open-weight models, 89–96%, and up to 97.51% for the commercial models.

Table D6: Accuracy and Macro-F1 (%) under the five input conditions of Section 4.3 on the 447 diagnostic instances with expert evidence sets. Full Images: the complete record; Oracle Images: the expert evidence images; Oracle Images + Facts: those images with the expert facts; Oracle Facts: the expert facts alone; Random Images: as many images as the minimum sufficient set, drawn without regard to the criterion.

Model	Full Images		Oracle Evidence						Random Images	
			Images		Images + Facts		Facts			
	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1
Commercial Models										
GPT-5.5 (OpenAI, 2026a)	91.72	90.76	94.63	94.24	97.51	97.33	90.83	90.46	42.95	42.69
Gemini 3.1 Pro (Google DeepMind, 2026)	88.81	88.21	92.62	92.12	94.52	94.17	93.24	92.88	38.70	37.69
Qwen3.8-Max (Alibaba Group, 2026)	91.50	90.70	92.84	92.15	97.09	96.88	92.62	92.28	46.09	45.95
Open-weight Models										
Qwen3.5-4B (Qwen Team, 2026a)	70.47	65.68	75.17	73.02	91.05	90.21	91.95	91.42	46.98	44.42
Qwen3.5-9B (Qwen Team, 2026a)	72.93	66.96	73.83	68.34	93.51	92.99	91.50	91.10	42.06	41.76
Qwen3.5-27B (Qwen Team, 2026a)	88.14	87.01	89.26	88.25	95.53	95.23	94.18	93.87	41.61	40.14
Qwen3.5-35B (Qwen Team, 2026a)	76.96	72.84	85.23	83.22	93.96	93.54	89.71	89.34	44.07	43.80
Qwen3.8-27B (Qwen Team, 2026b)	88.14	86.97	88.59	87.54	95.75	95.46	92.62	92.26	42.51	41.06
Gemma4-E4B (Gemma Team, 2026)	67.11	66.90	71.81	69.66	90.16	89.53	89.04	88.44	49.44	49.01
Gemma4-12B (Gemma Team, 2026)	81.88	78.73	89.26	88.25	94.63	94.18	94.18	93.82	56.15	55.88
Gemma4-26B (Gemma Team, 2026)	85.68	84.47	87.47	86.48	94.85	94.52	92.62	92.18	42.51	42.17
Gemma4-31B (Gemma Team, 2026)	82.10	81.74	91.28	90.87	95.30	95.01	94.41	94.06	39.37	38.26
GLM-4.6V-Flash (Zhipu AI, 2025)	59.73	53.38	73.60	73.40	89.26	88.80	86.35	85.83	31.10	29.21

D.7 Size-Matched Random-Subset Control

Table D7 places the minimum sufficient subsets of Appendix B.6 beside the complete record, the full evidence set, and a random subset of the same size.

A minimum sufficient subset gives the thirteen models 82.67% accuracy on average, close to the 85.05% of the full evidence set and above the 80.40% of the complete record. At the same image count, it exceeds a random subset for every model, by 23.27 to 49.89 points.

D.8 PAJ Pipeline

The PAJ pipeline separates visual perception from criterion-conditioned judgment in three stages, all executed by the same model. Per-image perception generates structured visual facts for each classroom photograph independently, using the `image_id`, `description`, `objects`, and `quality_indicators` fields in Appendix C.2; it runs once per photograph, and its output is reused for every criterion. Domain-aware aggregation then selects the resulting facts through the eight area categories and each criterion's area mapping, so that the judgment stage receives a criterion-relevant textual evidence context in place of the complete image record. Criterion-conditioned judgment evaluates the criterion against this aggregated evidence and returns the Satisfied or Unsatisfied decision. Because the final stage operates on text produced by the model itself, the pipeline

Table D7: Accuracy (%) under four visual input conditions on the 447 diagnostic instances with expert evidence sets. Random Images matches the image count of Minimum Images.

Model	Full Images	Oracle Images	Minimum Images	Random Images
Commercial Models				
GPT-5.5 (OpenAI, 2026a)	91.72	94.63	92.39	42.95
Gemini 3.1 Pro (Google DeepMind, 2026)	88.81	92.62	88.59	38.70
Qwen3.8-Max (Alibaba Group, 2026)	91.50	92.84	90.16	46.09
Open-weight Models				
Qwen3.5-4B (Qwen Team, 2026a)	70.47	75.17	74.05	46.98
Qwen3.5-9B (Qwen Team, 2026a)	72.93	73.83	72.71	42.06
Qwen3.5-27B (Qwen Team, 2026a)	88.14	89.26	86.13	41.61
Qwen3.5-35B (Qwen Team, 2026a)	76.96	85.23	84.79	44.07
Qwen3.8-27B (Qwen Team, 2026b)	88.14	88.59	85.46	42.51
Gemma4-E4B (Gemma Team, 2026)	67.11	71.81	72.71	49.44
Gemma4-12B (Gemma Team, 2026)	81.88	89.26	87.47	56.15
Gemma4-26B (Gemma Team, 2026)	85.68	87.47	84.79	42.51
Gemma4-31B (Gemma Team, 2026)	82.10	91.28	86.58	39.37
GLM-4.6V-Flash (Zhipu AI, 2025)	59.73	73.60	68.90	31.10

measures how far a model's own per-image perception carries it, and its comparison with the Oracle Facts condition of Table D6, where the facts are expert-annotated, shows which models can supply that stage for themselves.

Table D8: Accuracy and Macro-F1 (%) of PAJ and three comparison conditions on the 447 diagnostic instances with expert evidence sets. Full Images: direct single-criterion assessment from the complete classroom record, using the same one-criterion-per-request protocol as the PAJ comparisons; Atomic Criterion: the same input with the criterion decomposed into the atomic conditions and scoring expression of the PAJ judgment stage; Per-image Facts: the PAJ judgment stage applied to all of the model's own per-image facts, listed by image; PAJ: those facts selected by the criterion's area mapping before judgment.

Model	Full Images		Atomic Criterion		Per-image Facts		PAJ	
	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1
Commercial Models								
GPT-5.5 (OpenAI, 2026a)	91.72	90.76	91.39	91.05	89.48	87.36	89.71	89.16
Gemini 3.1 Pro (Google DeepMind, 2026)	88.81	88.21	92.62	92.21	93.51	93.09	93.06	92.68
Qwen3.8-Max (Alibaba Group, 2026)	91.50	90.70	90.60	90.23	91.27	90.85	91.94	91.46
Open-weight Models								
Qwen3.5-4B (Qwen Team, 2026a)	70.47	65.68	70.91	64.77	71.81	71.77	72.26	72.24
Qwen3.5-9B (Qwen Team, 2026a)	72.93	66.96	83.44	82.71	82.77	82.52	88.37	87.97
Qwen3.5-27B (Qwen Team, 2026a)	88.14	87.01	90.83	90.96	87.70	87.29	93.51	93.11
Qwen3.5-35B (Qwen Team, 2026a)	76.96	72.84	90.38	89.97	88.59	83.72	91.05	91.63
Qwen3.8-27B (Qwen Team, 2026b)	88.14	86.97	87.25	86.93	89.26	88.88	90.38	90.01
Gemma4-E4B (Gemma Team, 2026)	67.11	66.90	60.18	60.99	55.26	55.25	65.55	65.54
Gemma4-12B (Gemma Team, 2026)	81.88	78.73	80.54	80.20	83.45	83.06	84.56	84.57
Gemma4-26B (Gemma Team, 2026)	85.68	84.47	80.76	81.33	85.01	84.56	85.91	85.46
Gemma4-31B (Gemma Team, 2026)	82.10	81.74	83.89	82.81	86.13	85.79	88.14	87.73
GLM-4.6V-Flash (Zhipu AI, 2025)	59.73	53.38	62.41	56.34	65.54	61.47	68.46	68.06

Two intermediate conditions separate the contributions of the pipeline's components (Table D8). Atomic Criterion keeps the complete record as input but states each criterion through the atomic conditions and scoring expression of the PAJ judgment stage. Per-image Facts applies that judgment stage, unchanged, to all of the model's own per-image facts listed by image identifier, so it differs from PAJ only in domain-aware aggregation.

PAJ improves eleven of the thirteen models over Full Images, and Qwen3.5-9B and Qwen3.5-35B recover most or all of the expert-fact gain: Qwen3.5-9B rises by 21.01 Macro-F1 points and Qwen3.5-35B by 18.79, reaching 87.97 and 91.63 Macro-F1 from their own per-image facts against 91.10 and 89.34 from expert facts. Models that already handle the joint task well improve by 0.99 to 6.10 points: Qwen3.5-27B by 6.10, Qwen3.8-27B by 3.04, and Gemma4-26B by 0.99. The pipeline therefore gives open-weight models a practical route toward the expert-fact level. Among the commercial

models, Gemini 3.1 Pro gains 4.47 points and Qwen3.8-Max 0.76, while GPT-5.5, which reads the expert images as well as the expert's facts describe them, decides more accurately from the images.

Of the three components, domain-aware aggregation contributes most consistently. With the facts and the judgment stage fixed, it raises Macro-F1 for twelve of the thirteen models, by 3.39 points on average and by 10.29 for Gemma4-E4B, which falls to 55.25 when it judges from the facts listed by image. Listed without that organization, the same facts add only 0.39 points on average over Atomic Criterion, so the facts help once they are routed to the criterion. Criterion decomposition yields larger average gains in the group whose errors concentrate at criterion application: among the seven models with error profiles (Table D4), Atomic Criterion raises Macro-F1 by 5.92 points on average for the four whose errors concentrate at criterion application and by 0.43 for the three whose errors concentrate at evidence selection. Against Atomic Criterion, which states the criterion in the same form, PAJ is higher for twelve of the thirteen models, by 3.78 points on average, so the pipeline adds to what an explicit criterion provides.

D.9 Loose–Strict Criterion Pairs

The evaluation comprises 12 criterion pairs over 15 unique criteria. Open-weight models are evaluated on all 129 classrooms (1,548 classroom–pair instances) and commercial models on the 10 diagnostic classrooms (120 instances). Each pair shares the same classroom record and varies requirements such as quantity, diversity, accessibility, or thematic specificity. Boundary pairs are those whose loose criterion is met and whose strict criterion is unmet, whereas control pairs share the same reference outcome. Models evaluate the two criteria independently in randomized order. Pair accuracy requires both predictions to match the reference labels. Boundary accuracy is computed on the boundary pairs. Logical violation rate measures predictions that meet the strict criterion while failing the loose one, evaluated on reference pairs consistent with the implication. Deficiency criteria are direction-aligned as in Section 3.2, so met and unmet refer to the aligned classes. Among the commercial models, Qwen3.8-Max reaches the highest pair accuracy, 72.50%, and shares the highest boundary accuracy, 42.22%, with Gemini 3.1 Pro; among the open-weight models, Qwen3.8-27B leads at 69.25% and 37.72%.

Table D9: Results (%) and Cohen's κ on 12 loose–strict criterion pairs. Acc. and κ are computed over the classroom–criterion instances of the 15 criteria in the pairs; Loose, Strict, Pair, and Boundary Acc. and Violation Rate are computed over classroom–pair instances.

Model	Acc.	κ	Loose Acc.	Strict Acc.	Pair Acc.	Boundary Acc.	Violation Rate
GPT-5.5 (OpenAI, 2026a)	83.33	0.6677	89.17	77.50	68.33	31.11	0.00
Gemini 3.1 Pro (Google DeepMind, 2026)	84.00	0.6798	85.83	80.83	67.50	42.22	0.00
Qwen3.8-Max (Alibaba Group, 2026)	86.00	0.7208	93.33	79.17	72.50	42.22	0.00
Qwen3.5-4B (Qwen Team, 2026a)	77.26	0.5459	80.56	76.94	61.50	27.31	1.42
Qwen3.5-9B (Qwen Team, 2026a)	74.47	0.4920	80.36	73.39	58.46	27.70	3.81
Qwen3.5-27B (Qwen Team, 2026a)	80.93	0.6187	84.04	78.49	64.73	34.58	1.36
Qwen3.5-35B (Qwen Team, 2026a)	79.69	0.5932	82.69	77.26	61.89	25.93	0.06
Qwen3.8-27B (Qwen Team, 2026b)	83.57	0.6723	85.53	81.78	69.25	37.72	0.84
Gemma4-E4B (Gemma Team, 2026)	66.77	0.3362	64.02	71.83	42.18	29.47	12.35
Gemma4-12B (Gemma Team, 2026)	76.90	0.5421	83.14	67.89	57.30	18.86	0.45
Gemma4-26B (Gemma Team, 2026)	75.35	0.5102	81.85	69.70	58.40	27.11	2.33
Gemma4-31B (Gemma Team, 2026)	78.29	0.5654	82.49	74.94	60.08	23.18	1.10
GLM-4.6V-Flash (Zhipu AI, 2025)	73.28	0.4636	74.87	75.00	53.94	21.02	2.00

E Limitations and Broader Impact

E.1 Limitations

PEVA-Bench instantiates the rating points of ECQRS-EC that can be decided from photographs, covering six domains of the classroom environment. Rating points that require observing teacher–child interaction lie outside a photographic protocol by design, and for the rating points it covers, the labels agree with onsite ratings at κ above 0.70 in every educational domain (Appendix B.5). The current release instantiates the protocol for one instrument; the construction procedure, criterion operationalization, and annotation schema are instrument-agnostic, so extending the benchmark to further instruments, age groups, and regions follows the same procedure, and as models advance the diagnostic annotations continue to locate the stage at which they fail.

E.2 Broader Impact

PEVA-Bench is intended to support research on multimodal models for evidence-grounded assessment of preschool educational environments. By providing criterion-level reference labels, evidence images, and structured visual facts, the benchmark may facilitate more systematic evaluation of visual grounding, evidence selection, and criterion application in educational settings. These resources may also support future research on multimodal evaluation and AI-assisted preschool educational assessment.

At the same time, we recognize potential risks associated with applying multimodal models to educational environments. PEVA-Bench is designed as a research benchmark and should not be used to replace professional assessors or serve as the sole basis for high-stakes decisions concerning teachers, children, or educational institutions. Because the dataset contains photographs of real classroom environments, inappropriate use or re-identification of individuals or institutions could raise privacy concerns. We therefore anonymize kindergarten identities, obscure personal names, kindergarten names, photographs of children, and other privacy-sensitive content wherever they appear, and will release the full set of photographs under controlled access. Users are expected to follow applicable privacy and data-protection requirements when using or extending the benchmark.

We also acknowledge the computational resources required to evaluate large-scale vision-language models. By providing standardized evaluation data, annotations, and evaluation procedures, we aim to reduce duplicated data curation and evaluation efforts and facilitate more reproducible research.

F Case Studies Across Educational Domains

Figures F1–F6 show one complete assessment per educational domain, so that a reader can inspect what the scored quantities summarize. Each figure states the domain, the criterion, and the size of the classroom record from which the decision had to be made, and displays a sample of that record together with the expert reference label. Below the record, three models are shown side by side: GPT-5.5, GLM-4.6V-Flash, and Qwen3.8-27B. Each panel reproduces the model's decision, the justification it wrote, the facts it reported reading from individual images, and the images it cited. Decisions that disagree with the reference are marked with the error types attributed to them under the scheme of Appendix C.4, and the span of the justification that departs from the criterion is highlighted. The same three models appear in every figure, so their behavior can be followed across all six domains.

Each record holds 62 to 102 images, from which a model must locate the views that bear on the criterion. Across the eighteen model decisions shown, fourteen disagree with the reference, and eleven of those report Satisfied where the reference reports Unsatisfied, the direction documented in Section 4.4. Criterion application is the most frequent attribution, and three figures show, in the model's own words, the three forms of relaxation named in Section 4.4. In Figure F2, GPT-5.5 opens by substituting the threshold of a lower rating point for the one it was given and then credits the criterion while stating that the individual titles cannot be verified. In Figure F6, GPT-5.5 counts an outdoor planting garden, a natural-material rack, and an aquarium as science materials, and all three models credit the classroom. In Figure F5, the criterion requires a display that connects close observation of natural objects with children's drawings, and GPT-5.5 writes that the images do not show children drawing before reporting Satisfied, crediting a relation it states it cannot verify.

The remaining figures separate the failure modes that a single agreement rate would merge. Figure F1 states a negatively worded criterion, satisfied only when no labeled picture is visible; GLM-4.6V-Flash reports the labeled pictures it finds and credits the criterion on that basis, which is criterion relaxation applied to a negation. Figure F3 is the one case whose reference label is Satisfied, and all three models miss it: each reports that no relevant content was observed, none cites an image, and every error is attributed to evidence selection. This case places the rarer direction of error at the stage of finding the evidence, earlier than the criterion relaxation behind the common one. Figure F4 carries four attributions across three models, including the evidence integration error that arises when a model assembles a verdict from images that individually support none of the required conditions.

Print Environment

Criterion_001: No pictures paired with print labels are visible to children.


Image_0001


Image_0008


Image_0014


Image_0024

Unsatisfied: 0

●●●

73 images in total

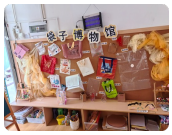
Image_0025


Image_0031


Image_0032


Image_0016

GPT 5.5 : 0

Model analysis
This is a negatively worded criterion and is satisfied only when no images with text labels are visible in the environment. Multiple examples of picture-text correspondence are visible, including displays about plant needs, care guidelines, tools, and procedural steps. Therefore, the statement that children cannot see images with text labels is not satisfied.
Image description
image_0008: The wall display contains plant-related images with text labels, including “temperature,” “nutrients,” “soil,” “air,” and “rainwater,” as well as images and corresponding labels illustrating the stages of plant growth.
image_0014: The display titled “What Should We Do When a Book Is Damaged?” contains photographs of procedural steps with text labels, including “identify the damaged book,” “apply adhesive,” “press it flat,” and “return it to the bookshelf.” Images of tools below are labeled “glue stick,” “scissors,” “double-sided tape,” and “white glue.”
Model-cited images
image_0008、image_0014

GLM-4.6V-Flash: 1

× Criterion Application Error

Model analysis
The image shows pictures containing text, including a calendar, a solar term chart, and a daily activity schedule. These satisfy the criterion, so score=1.
Image description
image_0001: Pictures containing text are visible, including a calendar, a solar term chart, and a daily activity schedule.
Model-cited images
image_0001

Qwen3.8-27B: 0

Model analysis
Numerous pictures with text labels are clearly visible, including activity area signs and plant life cycle diagrams. Therefore, the criterion “Children cannot see pictures with text labels” is not satisfied.
Image description
image_0001: Pictures with text labels such as “Role-play Area” and “Block Area” are visible.
image_0008: Pictures with text labels such as “Temperature” and “Nutrients” are visible.
Model-cited images
image_0001、image_0008

Figure F1: Print Environment. A negatively worded criterion on a 73-image record.

Books and Reading

Criterion_011: Children can access all four book categories in the book area, with at least three books in each category. Categories include picture books, science book, rhymes, and mathematics books.


Image_0018


Image_0023


Image_0017


Image_0010

Unsatisfied: 0

•••

62 images in total


Image_0011


Image_0022


Image_0015


Image_0021

GPT 5.5: 1

Model Analysis

The satisfaction threshold follows the same category as criterion 3.1: the book area must provide children with accessible books from multiple categories, with at least three books in each represented category. The low shelves in the book area contain multiple picture books and storybooks, as well as visible informational or science-themed books. All are placed on shelves within children's reach. Although the individual book contents and complete titles cannot be verified one by one, the visible evidence is sufficient to support the judgment.

Image Description ✕ **Criterion Application Error**

image_0018: The low shelves in the book area contain many books accessible to children, including multiple picture books and storybooks, as well as informational books such as *Where Do You Come From?* and *Time*.

image_0023: The bookshelves in the book area are positioned at a child-accessible height, with children's seating and a rug or soft mat nearby.

Model-Cited Images

Image_0018、Image_0023

GLM-4.6V-Flash: 0

Model Analysis

Only a subset of the books is visible, and the evidence does not confirm that each book category contains at least three books. Therefore, the score is 0.

Image Description

image_0018: The image does not provide evidence that each book category contains at least three books.

Model-Cited Images

Image_0018

Qwen3.8-27B: 1

Model Analysis

Within the book area, the picture-book category is clearly visible, with more than three books. Although the titles of the other categories cannot be fully identified, differences in the thickness and appearance of the books suggest category diversity. Based on the sufficient number of visible picture books and the overall arrangement indicating that multiple book categories are present, the criterion statement is judged to be satisfied.

Image Description ✕ **Criterion Application Error**

image_0017: The bookshelf contains multiple picture books, including *Where Do You Come From?*, *Doing What Matters Most*, *My Mom*, and *The Little Rabbit That Can Transform*. These belong to the picture-book category, with more than three visible examples.

image_0018: Several picture books are visible, including *Where Do You Come From?*, *Doing What Matters Most*, and *My Mom*.

image_0017: The vertically arranged book spines on the left side of the shelf are partially blurred, but their different appearances and thicknesses suggest the presence of multiple types of books, possibly including informational or storybooks.

Model-Cited Images ✕ **Visual Understanding Error**

Image_0017、Image_0018

Figure F2: Books and Reading. A criterion whose threshold GPT-5.5 replaces with that of a lower rating point.

Emergent Writing

Criterion_019: At least two displayed show a teacher recording children's words, numerals, drawings, or marks.


Image_0006

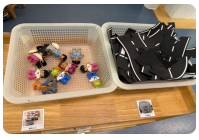
Image_0014


Image_0032


Image_0043

Satisfied: 1

•••

102 images in total


Image_0051


Image_0066


Image_0067


Image_0070

GPT 5.5 : 0

Model Analysis
The criterion is satisfied when at least two examples of teacher-written records of children's words are observable. In the uploaded images, however, no two displays or observational records can be clearly identified as teacher-written records of children's verbal statements. The visible text mainly consists of labels, rules, instructions, activity-related questions, or children's names. Therefore, the evidence is insufficient.
Image Description
Not observed: No relevant visible content was identified in the uploaded images.
Model-Cited Images
The model did not provide image evidence.
✕ **Evidence Selection Error**

GLM-4.6V-Flash: 0

Model Analysis
No displays or records of teacher-written transcripts of children's speech were identified; therefore, the criterion statement is not satisfied.
Image Description
Not observed: No relevant visible content was identified in the uploaded images.
Model-Cited Images
The model did not provide image evidence.
✕ **Evidence Selection Error**

Qwen3.8-27B: 0

Model Analysis
Although children's work is displayed in the images, no clear records of teacher-transcribed children's speech, such as observation records or dialogue bubbles, were identified. Therefore, two qualifying examples cannot be confirmed.
Image Description
Not observed: Two clear examples of teacher-written records of children's speech were not identified in the uploaded images.
Model-Cited Images
The model did not provide image evidence.
✕ **Evidence Selection Error**

Figure F3: Emergent Writing. The one Satisfied reference label among the six cases, missed by all three models.

Mathematics

Criterion_025: A designated mathematics area contains number games, countable objects, and books positioned for children's independent access.


Image_0012


Image_0013


Image_0016


Image_0018


Image_0023


Image_0014


Image_0015


Image_0041

Unsatisfied: 0

•••

62 images in total

GPT 5.5: 1

Model Analysis
The criterion requires a well-equipped mathematics area containing number games, counting materials, and books that are accessible to children on a daily basis. The images show a clearly labeled "Mathematics Area." The open shelves contain various number games and numeral-writing materials, counting beads, and materials representing quantities. An adjacent bookshelf contains several children's books, and the display area includes written numerals such as 1, 2, 4, 5, 6, 7, and 8.
Image Description
image_0012: An open shelf labeled "Mathematics Area" is visible, with materials placed at a child-accessible height.
image_0013: The mathematics shelves contain materials such as "Little Rabbit Picking Mushrooms: Mathematical Spatial Game," "Twelve Zodiac Animals: Intelligence Puzzle," numeral-writing boards, and grid boards.
image_0016: The open mathematics shelf contains colored counting beads and activity sheets representing quantities.
image_0018: Several children's books are visible on the adjacent bookshelf. The display area also contains written numerals, including 1, 2, 4, 5, 6, 7, and 8.
Model-Cited Images
image_0012、image_0013、image_0016、image_0018

✕ **Evidence Selection Error**

GLM-4.6V-Flash: 1

Model Analysis
A well-equipped mathematics area is present, including number games, counting materials, and books. Therefore, the criterion requiring a well-equipped mathematics area is satisfied, and the score is 1.

✕ **Evidence Integration Error**

Image Description
image_0012: A clearly labeled "Mathematics Area" is visible.
image_0013: The mathematics area contains number games, counting materials such as "Little Rabbit Picking Mushrooms," and books such as "The Twelve Zodiac Animals."
image_0023: The mathematics area has a well-organized layout, a dedicated space, and a range of materials.
Model-Cited Images
image_0012、image_0013、image_0023

✕ **Evidence Selection Error**

Qwen3.8-27B: 1

✕ **Criterion Application Error**

Model Analysis
The images show a dedicated mathematics area with various number games and counting materials, such as numeral cards and data-recording materials. The materials are neatly organized and accessible to children. Therefore, the evidence satisfies the criterion describing a well-equipped mathematics area.
Image Description
image_0012: A dedicated area labeled "Mathematics Area" is visible, with various mathematics materials arranged on the shelves.
image_0013: Mathematics game boxes such as "Little Rabbit Picking Mushrooms" and "The Twelve Zodiac Animals" are visible.
image_0014: Mathematics activity materials such as "Numeral Tracing" and "Compare Numbers and Draw Quantities" are visible.
image_0015: Mathematics games such as "Three-Dimensional Drawing" and "Sudoku" are visible.
image_0016: Mathematics activity materials such as "Fruit Statistics" and "Classification Fun" are visible.
Model-Cited Images
image_0012、image_0013、image_0014、image_0015、image_0016

Figure F4: Mathematics. A conjunctive criterion on which all three models report Satisfied.

Natural Materials

Criterion_038: At least one displayed connects close observation of natural objects with children's drawings.


Image_0028


Image_0029


Image_0054


Image_0010


Image_0014


Image_0003


Image_0024


Image_0051

Unsatisfied: 0

•••

73 images in total

GPT 5.5 : 1

Model Analysis

The criterion requires at least one example that encourages children to closely observe natural objects, or evidence showing children closely observing natural materials. The images show labeled seeds, plants, and insect or natural specimens arranged in trays, which are suitable for close observation. Magnifying glasses and other observation tools are also visible. However, the images do not directly show children drawing the observed materials.

Image Description ✕ **Criterion Application Error**

image_0028: Various seeds, plants, and insect or natural specimens are placed in transparent containers and trays, making them available for close observation.

image_0029: Magnifying glasses and other observation tools are visible in a basket, alongside wooden and other natural materials.

Model-Cited Images

image_0028、image_0029

GLM-4.6V-Flash: 0

Model Analysis

Plant specimens, insect boxes, and nature-themed books are visible. However, no example was found showing children closely observing natural materials and representing them through drawings. The requirement that one observed example or other evidence of children closely observing natural materials be present is therefore not satisfied. The score is 0.

Image Description

image_0028: Plant specimens and insect boxes are visible on the table.

image_0054: Nature-themed books, including *Insects Go Hiking*, are visible on the bookshelf.

Model-Cited Images

image_0028、image_0054

Qwen3.8-27B: 1

Model Analysis

The images show displays titled "Children's Papermaking Process" and "Traditional Papermaking Process," including process diagrams and photographs of children carrying out the activity. This indicates that children closely observed and manipulated natural materials, such as paper pulp, and therefore satisfies the requirement of at least one qualifying example. ✕ **Criterion Application Error**

Image Description

image_0010: The wall display presents the "Children's Papermaking Process" and "Traditional Papermaking Process," including process diagrams and photographs of children performing the activity.

image_0014: A basket contains papermaking materials and tools, such as paper pulp and paper molds.

Model-Cited Images

image_0010、image_0014

Figure F5: Natural Materials. GPT-5.5 states the unmet condition and credits the criterion regardless.

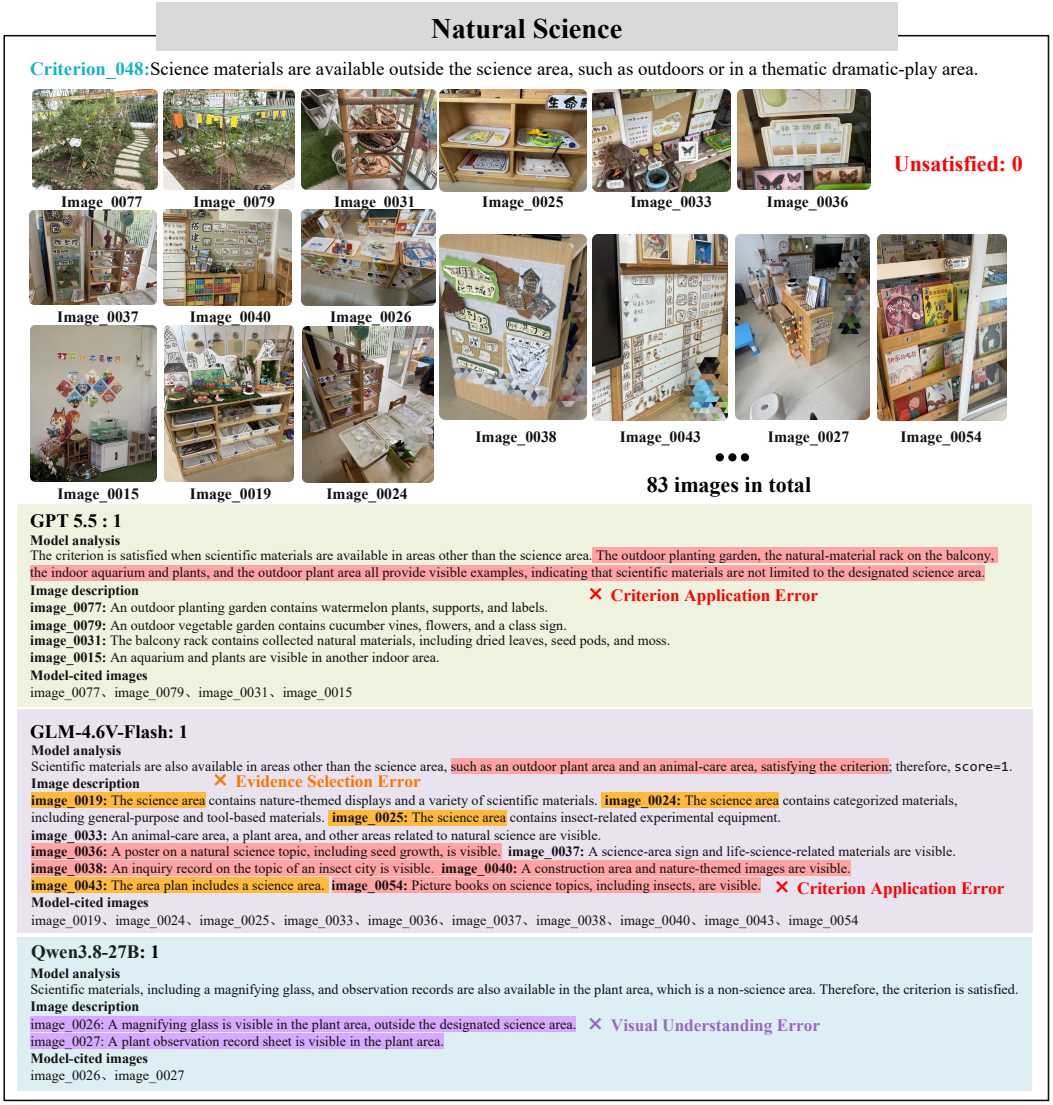

Figure F6: Natural Science. GPT-5.5 counts gardens and an aquarium as science materials; all three models credit the criterion.