Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Captioning Daily Activity Images in Early Childhood Education: Benchmark and Algorithm

Sixing Li
Beijing Union University
Zhibin Gu
Hebei Normal University
Ziqi Zhang
Institute of Automation, Chinese Academy of Sciences
Weiguo Pan
Beijing Union University
Bing Li
Institute of Automation, Chinese Academy of Sciences
Ying Wang
People Youhe Education Technology Co., Ltd.
Hongzhe Liu
liuhongzhe@buu.edu.cn

Beijing Union University

Research Article

Keywords: Multimedia systems, image captioning, early childhood education, multimodal large language models, controlled-access benchmark, reinforcement learning

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10234796/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on October 23rd, 2026. See the published version at https://doi.org/10.1007/s00530-026-02704-0.

Captioning Daily Activity Images in Early Childhood Education: Benchmark and Algorithm

Sixing Li[1†], Zhibin Gu[2†], Ziqi Zhang[3], Weiguo Pan[1], Bing Li[4], Ying Wang[5], Hongzhe Liu[1*]

[1]School of Robotics, Beijing Union University, Beijing, 100101, China.
[2]College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang, 050010, China.
[3]Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China.
[4]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences; PeopleAI, Inc., Beijing, 100190, China.
[5]People Youhe Education Technology Co., Ltd., Shandong, 250033, China.

*Corresponding author(s). E-mail(s): liuhongzhe@buu.edu.cn;
Contributing authors: lisixing@buu.edu.cn; guzhibin@hebtu.edu.cn; zhangziqi2017@ia.ac.cn; ldtweiguo@buu.edu.cn; bli@nlpr.ia.ac.cn; 15210019181@163.com;
†These authors contributed equally to this work.

Abstract

Purpose: Early childhood education (ECE) produces large volumes of daily activity images that document classroom routines, play, peer interaction, and teacher-guided activities. Converting such multimedia records into educationally meaningful captions remains difficult because ECE scenes contain fine-grained activity areas, visually ambiguous teaching toys, and professional terminology that are underrepresented in generic captioning datasets. **Methods:** We present KinderMM-Cap, a multimedia captioning system built around ECAC, a controlled-access benchmark containing 256,121 real-world kindergarten images with expert-level captions, activity-area labels, and structured teaching-toy annotations. We introduce Teaching Toy Recognition Score (TTS) for evaluating professional toy-name accuracy and propose RSRS, a reward-conditional hybrid

training framework that switches between reinforcement learning and supervised fine-tuning when sparse rewards cause zero-advantage groups. **Results:** Based on ECAC and RSRS, KinderMM-Cap-3B achieves a TTS of 51.06 and an overall caption-quality score of 86.20, outperforming strong multimodal baselines. Precision analysis further shows that RSRS improves measured toy-recognition precision among the compared training strategies. **Conclusion:** The results suggest that child-sensitive multimedia captioning benefits from domain data, specialized evaluation, controlled data access, and reward-aware training.

Keywords: Multimedia systems, image captioning, early childhood education, multimodal large language models, controlled-access benchmark, reinforcement learning

1 Introduction

Kindergarten classrooms are increasingly documented through multimedia records such as photographs, videos, speech, and teacher notes. Among these modalities, daily activity images are a compact and widely collected source of information: they capture children working with teaching toys, participating in group routines, moving across functional activity areas, and interacting with peers and teachers. If these images can be automatically converted into accurate captions, they can support classroom documentation, family-school communication, activity retrieval, and downstream educational assessment.

General-purpose multimodal large language models (MLLMs) have made substantial progress in image understanding and natural-language generation [1–4]. They have also shown strong potential in vertical domains such as medical imaging [5, 6], finance [7–9], fashion product description [10], and agriculture [11]. Educational multimedia analysis has similarly benefited from recent vision-language systems, including classroom observation, multimodal teaching analytics, and child-centered interaction modeling [12–19]. However, image captioning for early childhood education remains less explored than recognition, tracking, or conversational support.

ECE daily activity captioning is not simply a generic image description problem. A usable caption should identify the activity area, describe the main action, and name teaching toys using professional and context-appropriate terminology. For example, a generic model may describe an object as "craft material" or "item", while teachers expect more precise terms such as "modeling clay", "cotton swab", or "fruit pattern card". These differences matter because the caption is intended to support educational records rather than merely provide a natural sentence. ECE images also contain small objects, partial occlusions, repeated background props, and children-centered interactions, making fine-grained object naming difficult.

Two obstacles limit existing captioning systems in this setting. First, there is a shortage of large-scale ECE-specific benchmarks with activity-area labels, expert captions, and structured annotations for teaching toys. Without such data, models cannot learn the visual-semantic associations that distinguish ECE scenes from generic indoor images. Second, conventional training paradigms are not well suited to

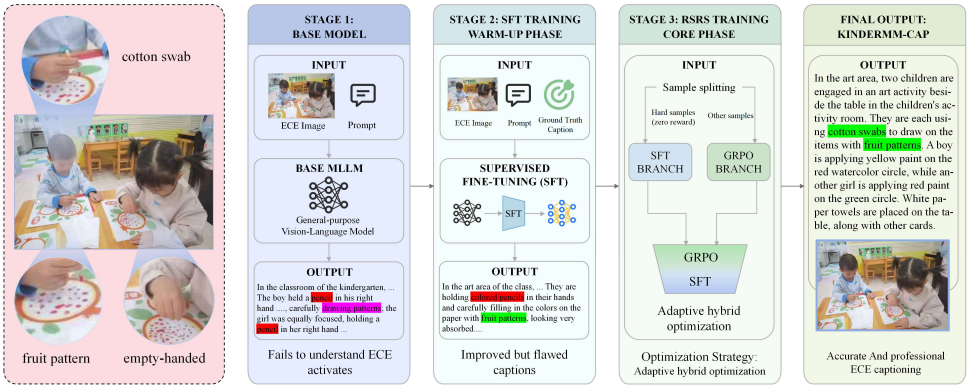


Fig. 1 Overview of the KinderMM-Cap multimedia captioning system. Daily activity images in early childhood education contain fine-grained teaching objects and functional activity areas. A general-purpose MLLM often produces generic descriptions. The proposed system uses ECAC for domain adaptation, TTS for professional toy-name evaluation, and RSRS for hybrid SFT/RL optimization.

sparse, domain-specific reward signals. Supervised fine-tuning (SFT) can teach common domain expressions but often favors frequent terms, while reinforcement learning can stagnate when all sampled captions receive zero reward on hard samples.

This paper reformulates the task as a multimedia system problem: how to build a captioning pipeline that collects authentic ECE images, governs child-sensitive data access, constructs domain annotations, evaluates pedagogically relevant details, and trains a compact MLLM that can produce professional captions. Figure 1 summarizes the overall system. The proposed KinderMM-Cap system combines (1) a controlled-access benchmark, ECAC; (2) a domain-specific evaluation protocol, Teaching Toy Recognition Score (TTS); (3) a reward-conditional hybrid training algorithm, RSRS; and (4) a domain-adapted model, KinderMM-Cap-3B.

The main contributions are as follows.

- We introduce ECAC, a large-scale benchmark for ECE daily activity image captioning. The benchmark contains 256,121 real-world kindergarten images, expert-level captions, activity-area labels, and structured teaching-toy annotations. Because the images involve children, the full image set is governed by a controlled-access procedure, while code, annotation schemas, split files, evaluation scripts, and de-identified examples are intended for public release.
- We propose TTS, a domain-oriented evaluation protocol that measures whether generated captions correctly name teaching toys at multiple precision levels and foreground/background roles. TTS complements general caption-quality assessment by focusing on educationally meaningful details.
- We design RSRS, a reward-conditional switch of reinforcement learning and supervised fine-tuning. RSRS routes zero-reward hard samples to supervised optimization while applying reward-driven exploration to informative samples, mitigating the zero-advantage problem in sparse-reward captioning.

- We build KinderMM-Cap-3B on Qwen2.5-VL-Instruction-3B and evaluate it against strong MLLM baselines. The model achieves a TTS of 51.06 and improves fine-grained ECE object naming, while precision analysis shows that the gains are not merely caused by indiscriminate toy-name insertion.

2 Related Work

2.1 Multimodal large language models

MLLMs connect language models with visual, audio, and other sensory inputs, enabling open-ended perception and generation. Representative systems such as Flamingo [1], MiniGPT-4 [2], LLaVA [22], and Qwen2.5-VL [23] have pushed forward instruction-following multimodal understanding across visual question answering, document understanding, OCR, and video-language tasks [24–28]. Recent work has also adapted MLLMs to vertical domains where domain terminology and visual distributions differ from generic web data, including medical [30, 32], remote sensing [29], financial [31], and agricultural [11] applications. This work follows the same vertical adaptation motivation, but focuses on child-centered educational multimedia and the fine-grained captioning needs of ECE scenarios.

2.2 Computer vision in early childhood education

Computer vision has been used in ECE for facial expression analysis, engagement estimation, face tracking, classroom monitoring, and child-robot interaction [33–37]. These studies demonstrate the value of visual perception for understanding child behavior and classroom activities. Nevertheless, most existing work emphasizes recognition, classification, or interaction control rather than generating natural-language descriptions that can be used by teachers and families. ECE captioning requires a model to identify the activity setting, the main participants' actions, and pedagogically relevant objects, making it both a multimedia understanding task and a domain communication task.

2.3 Image captioning in vertical domains

Image captioning bridges computer vision and natural language processing by transforming visual content into semantically consistent text [20, 21, 38]. Vertical-domain captioning has been studied in news reporting [39], biomedical images [40], underwater imagery [41], fashion products [10], and other application areas. These domains show that generic captioning metrics and datasets are often insufficient when specialized terminology and domain-specific correctness matter. In ECE, captions must not only be fluent but also distinguish similar teaching toys, identify functional activity areas, and ignore irrelevant background objects. ECAC and TTS are designed for this gap.

Table 1 Distribution of ECAC images across top-level activity categories.

Category	Number of images
Outdoor	56,665
Classroom	195,397
Specialty	4,059
Total	256,121

3 ECAC Benchmark and System Pipeline

3.1 Controlled benchmark construction

ECAC was constructed from authentic kindergarten daily activity images recorded during routine educational observation. After removing damaged and blurred images, 256,121 images were retained. The images cover 35 functional areas and are grouped into three scenario categories: outdoor, classroom, and specialty activities. Table 1 gives the top-level distribution.

Because ECAC involves real images of children, the benchmark is designed as a controlled-access resource rather than an unrestricted public image repository. The released resources should include source code, annotation schemas, split files, evaluation scripts, and de-identified examples. Full images should be made available only through a controlled request process subject to institutional approval, privacy review, and data-use agreements. This access design is part of the multimedia system rather than an afterthought: it supports reproducible research while respecting the sensitivity of child-centered classroom data.

3.2 Structured caption construction

The annotation pipeline combines MLLM-assisted captioning with expert review. First, ECE experts designed prompts with high-quality exemplar captions. GPT-4o was then used to generate initial free-form captions and structured annotations. The structured prompt identifies teaching toy names, assigns naming precision levels (low, medium, high), and classifies each toy as foreground or background depending on whether it is involved in the child's main interaction. Eight graduate students specializing in early childhood education reviewed and refined the generated structures, especially the teaching-toy names.

This pipeline reflects the desired output semantics of the system. Captions should center on the child-relevant activity, avoid irrelevant background details, and use professional object terminology when such objects support the activity description. Figure 3 shows representative daily activity images in the dataset.

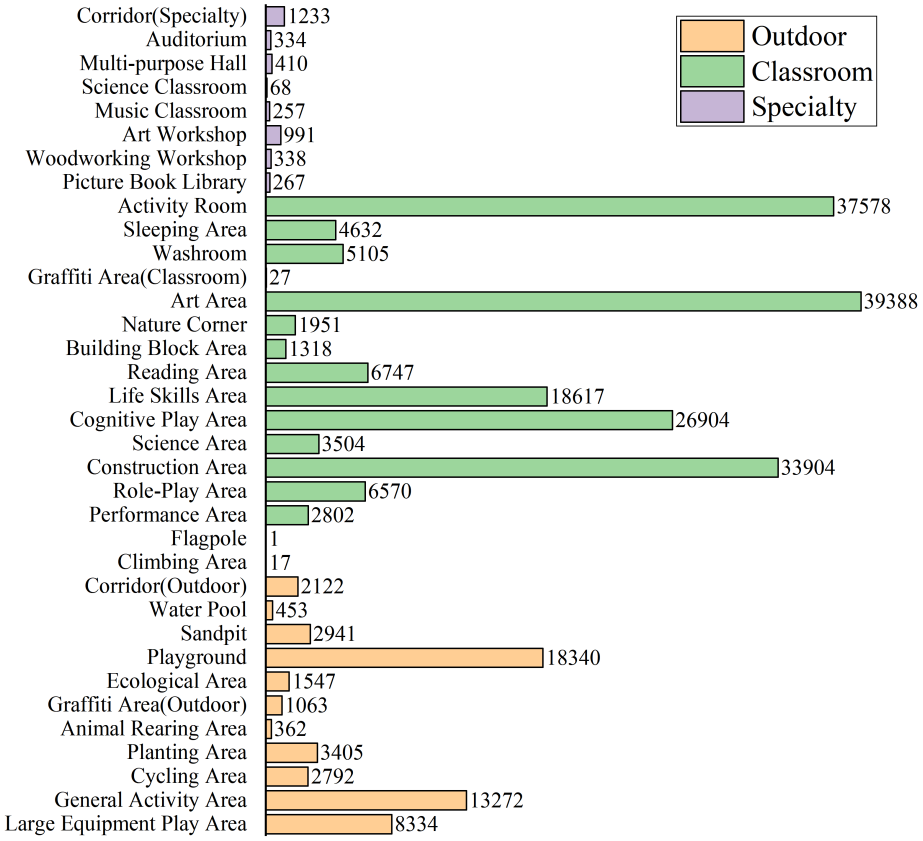


Fig. 2 Distribution of images across the 35 functional regions in ECAC. The long-tailed distribution reflects real kindergarten activity records and motivates domain-specific evaluation.

3.3 Teaching Toy Recognition Score

General caption metrics are useful for evaluating fluency and overlap with references, but they do not directly measure whether a caption uses professional names for teaching toys. We therefore define Teaching Toy Recognition Score (TTS). For an image, let $T = \{t_1, t_2, \ldots, t_{|T|}\}$ denote annotated toy names. For each toy t_i, let $A_i = \{a_{i1}, a_{i2}, \ldots, a_{iN}\}$ be its candidate descriptions at different precision levels. A generated caption is considered correct for t_i if it matches at least one candidate description:

$$\delta(a) = \begin{cases} 1, & \text{if the candidate description matches the caption,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$$\gamma(t_i) = \mathbf{1}\left(\sum_{j=1}^{N} \delta(a_{ij}) \geq 1\right), \tag{2}$$



Fig. 3 Examples of daily activity images across ECE scenario categories. ECE captioning requires identifying functional regions, child actions, and specialized teaching toys rather than producing generic indoor descriptions.

$$\mathrm{TTS} = \frac{1}{|T|} \sum_{i=1}^{|T|} \gamma(t_i). \tag{3}$$

Based on foreground/background status and low/medium/high naming precision, TTS is further decomposed into TTS-FL, TTS-FM, TTS-FH, TTS-BL, TTS-BM, and TTS-BH. These fine-grained metrics reveal whether a model only recognizes easy foreground objects or also learns rare, precise, and background teaching objects.

4 RSRS: Reward-Conditional Hybrid Training

4.1 Motivation

Domain-adaptive SFT helps a general MLLM learn ECE terminology, but it tends to reproduce high-frequency descriptions and may still miss rare or visually subtle teaching toys. Reinforcement learning with a toy-name reward can encourage more precise descriptions, but sparse rewards create a training problem: when all sampled captions in a group fail to mention any annotated toy, all rewards become zero and the normalized advantage collapses. In this case, reward learning provides no useful gradient for hard samples.

RSRS addresses this problem by combining supervised correction and reward-driven exploration. Informative samples with non-zero reward distributions continue through GRPO-style optimization, while zero-reward hard samples are placed in a buffer and later optimized with supervised fine-tuning. Figure 4 illustrates the mechanism.

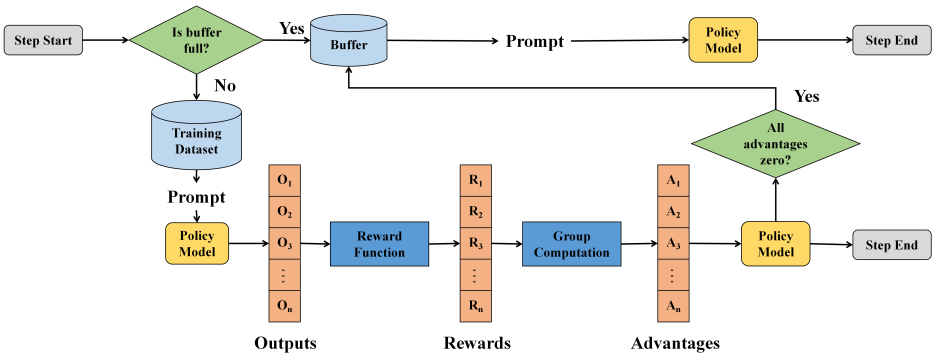


Fig. 4 RSRS training mechanism. When generated groups produce informative rewards, the model is optimized through reinforcement learning. When all advantages collapse to zero, the sample is routed to a buffer and later trained by supervised fine-tuning.

4.2 Warm-up supervised fine-tuning

The warm-up stage adapts Qwen2.5-VL-Instruction-3B to ECE captioning using 247,579 image-text pairs from ECAC. Given image I, prompt p, and ground-truth caption sequence $g = [a_1, a_2, \ldots, a_L]$, the supervised objective minimizes negative log-likelihood:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(I,p,g)\sim D}\left[\sum_{t=1}^{L} \log \pi_\theta(a_t|s_t)\right], \tag{4}$$

where s_t denotes the decoding state at step t and π_θ is the model policy.

4.3 Reward design

For a caption C, the reward function uses structured teaching-toy annotations. Let Ω denote the set of teaching toys in an image, divided into foreground set $\mathcal{F} = \{f_1, \ldots, f_n\}$ and background set $\mathcal{B} = \{b_1, \ldots, b_m\}$. Each toy name has a precision score $s(\cdot) \in \{1, 2, 3\}$. Foreground toys are weighted by $w_f = 1$, while background toys are weighted by $w_b = 0.5$:

$$\mathcal{R}(C) = \sum_{i=1}^{n} w_f s(f_i) h_{f,i} + \sum_{j=1}^{m} w_b s(b_j) h_{b,j}, \tag{5}$$

where $h_{f,i}$ or $h_{b,j}$ equals 1 if the corresponding toy appears in the generated caption and 0 otherwise. The reward therefore encourages captions to mention more precise and pedagogically relevant teaching toys, with stronger emphasis on foreground objects.

Algorithm 1 RSRS: Reward-Conditional Switch of RL and SFT

Require: SFT data D_{SFT}, hybrid data D_{RSRS}, warm-up steps W, hybrid steps T, group size G

Ensure: Final policy π_θ

Initialize policy π_θ
for $i = 1$ to W **do**
  Sample (I, p, g) from D_{SFT}
  Update θ by minimizing $\mathcal{L}_{\text{SFT}}$
end for
Initialize empty hard-sample buffer
for $t = 1$ to T **do**
  if buffer size is smaller than batch size **then**
    Sample (I, p, g, Ω) from D_{RSRS}
    Generate $\{o_j\}_{j=1}^{G}$ using $\pi_\theta(I, p)$
    Compute rewards $\{r_j\}_{j=1}^{G}$ using $\mathcal{R}$
    if $\sum_{j=1}^{G} r_j = 0$ **then**
      Append (I, p, g, Ω) to the buffer
    end if
    Update θ with the GRPO branch
  else
    Sample a batch from the buffer
    Update θ with the SFT branch
  end if
end for
return π_θ

4.4 Reward-conditional switch

In the hybrid stage, each image and prompt generates G candidate captions. The reward values are normalized within the group to obtain advantages. If all generated captions receive zero reward, the sample is stored in a buffer. Once the buffer reaches the training batch size, samples are trained using the SFT objective. Otherwise, samples with informative reward distributions are optimized using the GRPO branch. Algorithm 1 summarizes this process.

5 Experiments

5.1 Experimental setup

Qwen2.5-VL-Instruction-3B was selected as the backbone because of its strong Chinese multimodal performance. In the warm-up stage, 247,579 ECAC image-text pairs were used for SFT with LLaVA-Factory. Training ran for 3 epochs on 8 NVIDIA V100 GPUs with a learning rate of 1.0×10^{-5}. In the hybrid stage, 12,000 ECAC images with structured toy annotations were used. Four GPUs were assigned to training and

Table 2 Performance comparison on ECE daily activity image captioning.

Model	TTS	Overall score
GLM-4v-9B	29.36	78.19
Gemma3-27B	10.74	12.00
Qwen2.5-VL-3B	38.07	84.63
Qwen2.5-VL-7B	45.25	85.70
KinderMM-Cap-3B	**51.06**	**86.20**

four GPUs to online inference through vLLM, generating 4 candidate responses per step.

We evaluate mainstream MLLMs and trained variants using TTS and an automated overall caption-quality score. The overall score is produced by a vision-language evaluator using three dimensions: activity-area recognition, teaching-toy recognition, and action/expression description. The evaluation focuses on TTS because professional object naming is the core domain requirement in ECE captioning.

5.2 Main results

Table 2 compares KinderMM-Cap-3B with representative MLLM baselines. KinderMM-Cap-3B achieves the highest TTS score of 51.06, outperforming Qwen2.5-VL-3B by 34.1% relative improvement and exceeding the larger Qwen2.5-VL-7B baseline. This result supports the central system claim: in ECE captioning, domain data and targeted optimization are more effective than relying only on larger general-purpose MLLMs.

5.3 Reward validation

To test whether the reward function aligns with ECE expert judgment, we randomly sampled 100 ECAC test images. For each image, five captions were generated by Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-7B-Instruct with SFT, Qwen2.5-VL-32B-Instruct, and Qwen2.5-VL-72B-Instruct. Four graduate students with ECE expertise independently ranked the captions based on the accuracy and professionalism of teaching-toy names. Table 3 shows that the proposed reward function has the strongest correlation with human rankings.

5.4 Ablation on training strategies

Figure 5 and Table 4 compare baseline, SFT, GRPO, and RSRS variants. SFT improves the baseline by adapting the model to ECE terminology. GRPO further improves TTS with far fewer training samples, indicating the data efficiency of reward-driven learning. RSRS obtains the best overall TTS and improves all fine-grained toy-recognition categories over pure GRPO, including foreground low-precision recognition, medium/high-precision recognition, and background toy recognition.

Table 3 Correlation between automatic ranking methods and human judgments.

Method	Kendall's τ	Spearman's ρ
CN-CLIP	0.11	0.13
GPT-4o	0.43	0.52
Gemma3-27B	0.26	0.32
Ours	**0.67**	**0.72**

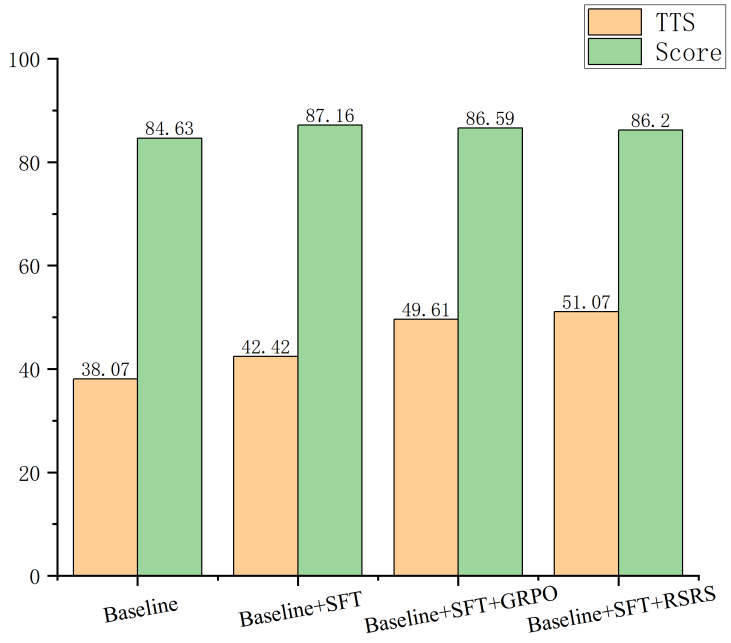


Fig. 5 Ablation study on training strategies for ECE daily activity captioning.

RSRS achieves the best value on all six fine-grained TTS metrics. The gains are especially meaningful in TTS-FM, TTS-FH, TTS-BL, and TTS-BM, where pure GRPO struggles with precise or background teaching-toy names. This supports the rationale of RSRS: when reinforcement learning cannot learn from zero-reward groups, supervised correction can still expose the model to expert annotations for hard samples.

Table 4 Fine-grained toy-recognition metrics across training strategies.

Model	TTS-FL	TTS-FM	TTS-FH	TTS-BL	TTS-BM	TTS-BH
Baseline	57.47	8.88	0.03	24.52	1.53	0.01
Baseline+SFT	66.59	15.71	0.04	24.15	2.01	0.00
Baseline+SFT+GRPO	65.66	14.39	0.02	37.97	2.86	0.02
Baseline+SFT+RSRS	**68.21**	**17.18**	**0.09**	**39.69**	**3.87**	**0.03**

Table 5 Toy recognition precision across training strategies.

Model	Avg. toy count	Precision (%)
Baseline	1.45	64.15
Baseline+SFT	1.66	65.12
Baseline+SFT+GRPO	1.84	64.43
Baseline+SFT+RSRS	2.08	**69.59**

5.5 Precision analysis

Because reward-driven training can in principle increase the number of mentioned teaching toys, we evaluate toy-recognition precision to verify whether the TTS gains come from genuine recognition rather than indiscriminate toy-name insertion:

$$\text{Precision} = \frac{|\mathcal{T}_{\text{caption}} \cap \mathcal{T}_{\text{groundtruth}}|}{|\mathcal{T}_{\text{caption}}|}, \tag{6}$$

where $\mathcal{T}_{\text{caption}}$ is the set of toy names extracted from a generated caption and $\mathcal{T}_{\text{groundtruth}}$ is the annotated set of toys in the image. Table 5 shows that RSRS increases the average number of predicted toys while also improving precision. Compared with SFT, RSRS raises precision from 65.12% to 69.59%, indicating that the additional toy mentions are more often correct rather than merely inflated.

This result is important for deployment. RSRS improves professional toy-name coverage under sparse rewards while maintaining the highest measured precision among the compared training strategies. The increased average toy count therefore reflects better recovery of relevant teaching toys rather than a simple tendency to over-describe the scene.

6 Discussion

The proposed system shows that ECE captioning requires a different design point from generic image captioning. A caption useful to teachers should be accurate, concise, child-centered, and professionally named. ECAC provides the domain data required for this goal; TTS measures the details that generic metrics may miss; and RSRS addresses the sparse-reward training issue that appears when the model faces rare or visually subtle teaching toys.

The system-level contribution lies in the integration of data governance, annotation, evaluation, and training into a single multimedia captioning pipeline. The controlled-access design is especially important. Fully open data release is not appropriate for child images, but reproducibility can still be supported through code, annotation schemas, split files, evaluation scripts, de-identified examples, and a reasonable data-access application process.

Several issues remain. First, the current evaluation focuses on toy-name correctness and automated caption-quality scoring, while reference-overlap captioning metrics can be further used to improve comparability with broader captioning literature. Second, although RSRS improves both TTS and measured toy-recognition precision in the current experiments, future work should further validate this precision gain under broader activity areas and unseen kindergartens. Third, ECAC currently focuses on image captioning; future work can extend the system to temporal video understanding, emotion-aware activity interpretation, and teacher feedback generation.

7 Conclusion

This paper presented KinderMM-Cap, a multimedia captioning system for daily activity images in early childhood education. The system includes ECAC, a large-scale controlled-access benchmark; TTS, a domain-specific evaluation protocol for teaching-toy recognition; RSRS, a reward-conditional hybrid training framework; and KinderMM-Cap-3B, a domain-adapted MLLM. Experiments show that KinderMM-Cap-3B improves professional teaching-toy recognition and outperforms strong baselines on TTS. Precision analysis further shows that RSRS improves toy-name coverage while achieving the highest measured toy-recognition precision among the compared training strategies. Overall, the work demonstrates how multimedia captioning systems can be adapted to child-sensitive educational scenarios through domain data, privacy-aware release, specialized evaluation, and targeted optimization.

Funding This work was supported in part by the National Natural Science Foundation of China under Grant 62506116 and Grant U24A20331.

Data availability The full ECAC image set contains real daily activity images of children and is therefore not released as an unrestricted public download. Code, annotation schemas, split files, evaluation scripts, and de-identified examples are planned for release at https://github.com/SixingLI030/KinderMM-Cap. Access to the full image set can be requested through a controlled-access procedure subject to institutional approval, privacy review, and data-use agreements.

Ethics approval and consent to participate The images used in this study were collected from routine educational activities in participating kindergartens with institutional authorization. Written informed consent for research use of the children's daily activity images was obtained from their parents or legal guardians. No interventional procedure was conducted in this study. Full image access requires confirmation that the requester will comply with child privacy, data security, and non-redistribution requirements.

Consent for publication Written consent for publication of de-identified example images was obtained from the children's parents or legal guardians and the participating institutions.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions Sixing Li and Zhibin Gu contributed equally to this work. The authors contributed to dataset construction, algorithm design, experiments, writing, and manuscript revision. Hongzhe Liu supervised the work and served as corresponding author.

References

[1] J. B. Alayrac, J. Donahue, L. Luc, et al., Flamingo: a visual language model for few-shot learning, Advances in Neural Information Processing Systems 35 (2022) 23716–23736.

[2] D. Zhu, J. Chen, X. Shen, et al., MiniGPT-4: Enhancing vision-language understanding with advanced large language models, arXiv preprint arXiv:2304.10592 (2023).

[3] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[4] L. Ouyang, J. Wu, X. Jiang, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[5] Q. Jin, F. Chen, Y. Zhou, et al., Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine, NPJ Digital Medicine 7(1) (2024) 190.

[6] J. Zhou, X. He, L. Sun, et al., Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4, Nature Communications 15(1) (2024) 5649.

[7] G. Bhatia, E. M. B. Nagoudi, H. Cavusoglu and M. Abdul-Mageed, FinTral: A family of GPT-4 level multimodal financial large language models, Findings of the Association for Computational Linguistics: ACL 2024, 13064–13087.

[8] Z. Gan, Y. Lu, D. Zhang, et al., MME-Finance: A multimodal finance benchmark for expert-level understanding and reasoning, arXiv preprint arXiv:2411.03314 (2024).

[9] J. Li, Y. Zhu, D. Cheng, et al., CFBenchmark-MM: Chinese financial assistant benchmark for multimodal large language model, arXiv preprint arXiv:2506.13055 (2025).

[10] Y. Tang, L. Zhang, Y. Yuan and Z. Chen, Describe fashion products via local sparse self-attention mechanism and attribute-based re-sampling strategy, IEEE Transactions on Circuits and Systems for Video Technology 33(7) (2023) 3409–3424.

[11] M. Awais, A. H. S. A. Alharthi, A. Kumar, et al., AgroGPT: Efficient agricultural vision-language model with expert tuning, IEEE/CVF Winter Conference on Applications of Computer Vision (2025) 5687–5696.

[12] U. Lee, Y. Jeong, J. Koh, et al., I see you: teacher analytics with GPT-4 vision-powered observational assessment, Smart Learning Environments 11(1) (2024) 48.

[13] S. Xu, X. Huang, C. K. Lo, et al., Evaluating the performance of ChatGPT and GPT-4o in coding classroom discourse data, Computers and Education: Artificial Intelligence 7 (2024) 100325.

[14] S. Yu, A. Androsov and H. Yan, Exploring the prospects of multimodal large language models for automated emotion recognition in education, Computers & Education 232 (2025) 105307.

[15] G. G. Lee and X. Zhai, NERIF: GPT-4V for automatic scoring of drawn models, arXiv preprint arXiv:2311.12990 (2023).

[16] G. G. Lee, E. Latif, L. Shi, et al., Gemini Pro defeated by GPT-4V: Evidence from education, arXiv preprint arXiv:2401.08660 (2023).

[17] U. Lee, M. Jeon, Y. Lee, et al., LLaVA-docent: Instruction tuning with multimodal large language model to support art appreciation education, Computers and Education: Artificial Intelligence 7 (2024) 100297.

[18] W. Shi, H. V. Le and K. T. W. Choo, Towards multimodal large-language models for parent-child interaction: A focus on joint attention, Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (2025) 1–6.

[19] B. Wang, Y. Song, J. Cao, et al., DinoCompanion: An attachment-theory informed multimodal robot for emotionally responsive child-AI interaction, arXiv preprint arXiv:2506.12486 (2025).

[20] K. Xu, J. Ba, R. Kiros, et al., Show, attend and tell: Neural image caption generation with visual attention, International Conference on Machine Learning (2015) 2048–2057.

[21] X. Chen and C. L. Zitnick, Mind's eye: A recurrent visual representation for image caption generation, IEEE Conference on Computer Vision and Pattern Recognition (2015) 2422–2431.

[22] H. Liu, C. Li, Y. Li, et al., Improved baselines with visual instruction tuning, IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 26296–26306.

[23] S. Bai, K. Chen, X. Liu, et al., Qwen2.5-VL technical report, arXiv preprint arXiv:2502.13923 (2025).

[24] Y. Goyal, T. Khot, D. Summers-Stay, et al., Making the V in VQA matter: Elevating the role of image understanding in visual question answering, IEEE Conference on Computer Vision and Pattern Recognition (2017) 6904–6913.

[25] K. Marino, M. Rastegari, A. Farhadi, et al., OK-VQA: A visual question answering benchmark requiring external knowledge, IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019) 3195–3204.

[26] M. Mathew, D. Karatzas and C. V. Jawahar, DocVQA: A dataset for VQA on document images, IEEE/CVF Winter Conference on Applications of Computer Vision (2021) 2200–2209.

[27] Y. Liu, Z. Li, M. Huang, et al., OCRBench: On the hidden mystery of OCR in large multimodal models, Science China Information Sciences 67(12) (2024) 220102.

[28] J. Xu, T. Mei, T. Yao, et al., MSR-VTT: A large video description dataset for bridging video and language, IEEE Conference on Computer Vision and Pattern Recognition (2016) 5288–5296.

[29] X. Li, J. Ding and M. Elhoseiny, VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding, Advances in Neural Information Processing Systems 37 (2024) 3229–3242.

[30] Y. Zong, Y. Yang and T. Hospedales, MEDFAIR: Benchmarking fairness for medical imaging, arXiv preprint arXiv:2210.01725 (2022).

[31] Z. Liu, X. Guo, H. Xia, et al., VisFinEval: A scenario-driven Chinese multimodal benchmark for holistic financial understanding, arXiv preprint arXiv:2508.09641 (2025).

[32] Y. Ji, H. Bai, C. Ge, et al., AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation, Advances in Neural Information Processing Systems 35 (2022) 36722–36732.

[33] H. Yi, T. Liu and G. Lan, The key artificial intelligence technologies in early childhood education: A review, Artificial Intelligence Review 57(1) (2024) 12.

[34] M. Del Coco, M. Leo, P. Carcagni, et al., A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders, IEEE International Conference on Computer Vision Workshops (2017) 1401–1407.

[35] O. Rudovic, Y. Utsumi, J. Lee, et al., CultureNet: A deep learning approach for engagement intensity estimation from face images of children with autism, IEEE/RSJ International Conference on Intelligent Robots and Systems (2018) 339–346.

[36] D. Li, W. Li, S. Chen, et al., Intelligent robot for early childhood education, International Conference on Information and Education Technology (2020) 142–146.

[37] Y. Xia, D. Huang and Y. Wang, Detecting smiles of young children via deep transfer learning, IEEE International Conference on Computer Vision Workshops (2017) 1673–1681.

[38] Y. Wang, N. Xu, A. A. Liu, W. Li and Y. Zhang, High-order interaction learning for image captioning, IEEE Transactions on Circuits and Systems for Video Technology 32(7) (2022) 4417–4430.

[39] T. Liu, Q. Cai, C. Xu, et al., Image captioning in news report scenario, arXiv preprint arXiv:2403.16209 (2024).

[40] J. Pavlopoulos, V. Kougia and I. Androutsopoulos, A survey on biomedical image captioning, Proceedings of the Second Workshop on Shortcomings in Vision and Language (2019) 26–36.

[41] H. Li, H. Wang, Y. Zhang, et al., Underwater image captioning: Challenges, models, and datasets, ISPRS Journal of Photogrammetry and Remote Sensing 220 (2025) 440–453.