Does Visual Privacy Leakage Compromise Fairness in Multimodal Automated Feedback? Extending Group Fairness to Privacy Contexts

Hai Li[1], Wanli Xing[2], Chenglu Li[3], and Neil Heffernan[4]

[1] University of Florida, United States
li.ha@ufl.edu
[2] University of Miami, United States
wanli.xing@miami.edu
[3] University of Utah, United States
chenglu.li@utah.edu
[4] Worcester Polytechnic Institute, United States
nth@wpi.edu

Abstract. Multimodal automated feedback systems are core components of next-generation K-12 mathematics learning interfaces, processing textual responses and handwritten images for large-scale personalized learning. When students upload assignment images in uncontrolled environments, visual privacy information such as faces and hands may be inadvertently captured, potentially correlating with contextual factors including photo-taking environment, image quality, student effort, assignment type, and device access, raising critical algorithmic fairness concerns in high-stakes educational assessment.

This study extends group fairness from traditional demographic attributes to privacy contexts by treating privacy leakage status and type as group-defining characteristics. We propose a component-driven multimodal feedback model simulating teacher cognitive processes through four components: Mathematical Element Masking (MEM), Cross-Modal Consistency Verification (CMCV), Question-Answer Interaction (QAI), and Scoring Prototype Contrastive Learning (SPCL). The optimal configuration, MatCha with complete components, achieves MSE = 0.0906 and $R^2 = 0.1686$, representing a 257.2% improvement in explanatory power over baseline, highly competitive given the inherent subjectivity in open-ended mathematical scoring.

Using a dual-perspective fairness framework distinguishing model fairness (opportunity allocation) from error fairness (service quality), we identify a privacy leakage paradox: samples containing privacy leakage exhibit superior predictive accuracy yet systematically lower high-score opportunity rates. The overall privacy leakage disparate impact ratio (0.7565) falls below the legal threshold of 0.80, while facial leakage (0.8274) approaches the fairness boundary. No substantial gender bias was detected under facial leakage conditions.

Keywords: Multimodal Automated Feedback · Visual Privacy Leakage · Algorithmic Fairness · K-12 Mathematics Education

1 Introduction

Multimodal learning interfaces that incorporate automated scoring have become an important part of the infrastructure for large-scale personalized learning in K-12 mathematics education [17]. Recent advances have demonstrated the efficacy of integrating multimodal architectures with fine-tuning approaches to provide automated feedback for student mathematics responses [10]. Designing effective multimodal learning interfaces requires integrating AI capabilities with human-centered interaction design principles and pedagogical insights from the learning sciences, treating automated feedback as one important component of human–AI collaborative educational systems. Open-ended questions are particularly valuable because they capture students' problem-solving strategies and reasoning processes, yet manual grading is time-intensive and challenging to standardize [1]. Contemporary K-12 mathematics naturally involves multimodal responses combining textual explanations with visual elements such as geometric constructions, annotated graphs, and handwritten calculations, making joint text-image interpretation essential for effective automated scoring.

While large multimodal models offer promising opportunities for educational scoring [17], several challenges persist. Mathematical work is visually dense with small symbols and complex structures, and these general models are rarely optimized for messy K-12 handwriting. More critically, deployment in authentic educational settings introduces significant privacy and fairness concerns. Students frequently capture work in uncontrolled environments where faces, hands, and backgrounds are inadvertently included, potentially revealing sensitive demographic information. Once embedded in multimodal representations, these privacy cues risk influencing predictions in ways unrelated to academic ability, creating systematic disadvantages for certain subgroups in high-stakes contexts.

This study addresses two interconnected research questions through a complete technical-to-ethical evaluation loop.

RQ1 (Technical) How do component-driven architectural enhancements specifically designed to simulate teacher cognitive processes improve multimodal mathematical scoring performance compared to standard backbone networks?

RQ2 (Ethical) When deployed as a component of a multimodal learning interface, how are fairness outcomes distributed across student groups with different visual privacy leakage states, examining both opportunity allocation and prediction accuracy? We investigate these associations through three progressive comparisons: (2.1) privacy presence versus absence, (2.2) facial versus hand leakage types, and (2.3) gender-related patterns under facial leakage scenarios, treating findings as observational associations that may reflect underlying contextual factors rather than direct causal effects.

2 Related Work

Early automated scoring relied on rule-based keyword matching, evolving to machine learning approaches with hand-crafted features and eventually BERT-

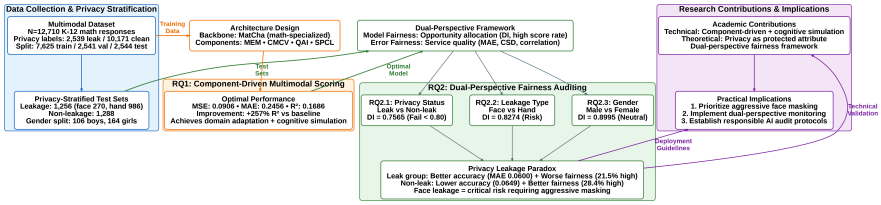

Fig. 1: Research methodology framework from privacy-stratified data preparation through component-driven scoring to dual-perspective fairness evaluation.

based architectures for text scoring [9, 5]. However, text-only approaches face fundamental limitations in mathematics where key solution steps appear in diagrams and handwritten equations. Early multimodal attempts used OCR-then-score pipelines [18], which broke spatial relationships inherent in mathematical work.

Multimodal pretrained models like CLIP [15], BLIP [13], and mathematics-specialized models like MatCha [14] enabled direct joint encoding. Despite these advances, naively applying general models to mathematical scoring remains challenging as they often attend to visually salient but educationally irrelevant regions.

Building on these foundations, recent work has successfully demonstrated how multi-modality and collaborative filtering can be leveraged to support automatic scoring in mathematics education [12], highlighting the necessity of domain-specific adaptations for complex mathematical problem-solving contexts.

Recent scholarship on adaptive learning technologies and human–AI collaboration in education emphasizes the importance of embedding AI capabilities within pedagogically grounded learning environments and human-centered classroom practices [2, 7]. Research on human–AI interaction in educational settings further demonstrates how interface design choices, including automated agent personality and interaction styles, significantly shape student engagement and learning outcomes in mathematics learning contexts [11]. Research on algorithmic bias indicates that fairness should be considered throughout the design, evaluation, and deployment of educational AI systems rather than treated solely as a post-deployment concern [3]. Our work contributes an empirical fairness audit of multimodal learning feedback interfaces, examining how visual privacy leakage status associates with differential outcomes and informing principles for responsible interface design at the intersection of AI, HCI, and the learning sciences.

3 Methodology

Figure 1 presents our research methodology framework, systematically integrating component-driven multimodal scoring (RQ1) with dual-perspective fairness auditing (RQ2) through a complete technical-to-ethical evaluation loop.

3.1 Dataset Construction and Ethical Framework

Ethics and Privacy Statement This study follows established protocols for educational data collection and privacy protection [4]. We use only existing, de-identified multimodal response records obtained under formal institutional agreements.

This study utilizes 12,710 multimodal mathematical response records from authentic K-12 classroom settings. Trained annotators systematically labeled student response images for visual privacy leakage (faces, hands, background elements) based on a standardized coding protocol. This identified 2,539 samples with privacy leakage and 10,171 without. Figure 2 illustrates the two primary privacy leakage types hand leakage (exposing skin tone) and facial leakage (revealing gender, age, race). Mixed cases are classified as facial leakage due to richer demographic information.

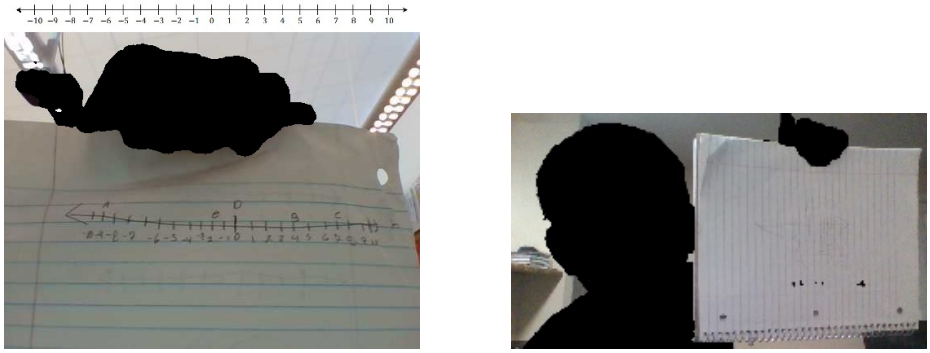

Fig. 2: Privacy leakage examples (left) hand leakage, (right) facial leakage.

To ensure robust statistical power for fairness analysis, we constructed a privacy-balanced test set (N=2,544) through stratified oversampling, yielding approximately equal leakage (1,256) and non-leakage (1,288) samples. The remaining samples were allocated to training (7,625) and validation (2,541) sets using strict group-based partitioning to prevent data leakage.

3.2 Component-Driven Multimodal Architecture

We enhance five multimodal backbones (CLIP, BLIP, DePlot, Pix2Struct, MatCha) with four components designed to simulate teacher cognitive processes.

Mathematical Element Masking (MEM) identifies and amplifies mathematically salient tokens through binary masks and type embeddings, forcing attention toward equations and symbols while suppressing generic visual textures.

Table 1: Ablation Results Across Backbone Networks (Test Set, N=2,544)

CLIP, BLIP, DePlot						Pix2Struct, MatCha					
Model	**Config**	**MSE**	**RMSE**	**MAE**	R^2	**Model**	**Config**	**MSE**	**RMSE**	**MAE**	R^2
CLIP	Base	0.1084	0.3293	0.2526	−0.006	Pix2Struct	Base	0.1024	0.3200	0.2726	0.050
	Complete	0.1269	0.3562	0.2757	−0.177		Complete	0.0959	0.3097	0.2608	0.110
BLIP	Base	0.1014	0.3184	0.2717	0.060	**MatCha**	Base	0.1027	0.3205	0.2746	0.047
	Complete	0.1015	0.3186	0.2695	0.058		**Complete**	**0.0906**	**0.3010**	**0.2456**	**0.1686**
DePlot	Base	0.1029	0.3207	0.2728	0.046						
	Complete	0.1002	0.3165	0.2645	0.071						

Cross-Modal Consistency Verification (CMCV) detects conflicts between textual explanations and visual work by computing consistency scores that gate global feature contributions, simulating teachers' cross-checking behavior.

Question-Answer Interaction (QAI) models conditional scoring through bidirectional attention between question and student response features, ensuring scoring depends on whether work addresses the specific question asked.

Scoring Prototype Contrastive Learning (SPCL) maintains learnable prototypes for each score level, using contrastive loss to create well-separated clusters in representation space, mimicking teachers' mental score categories.

We train using AdamW optimizer with learning rate 10^{-4}, cosine annealing schedule, and early stopping based on validation performance.

4 Experimental Results

4.1 Technical Performance Analysis (RQ1)

Table 1 presents comprehensive ablation results across backbone networks and component configurations.

Optimal Configuration and Performance Interpretation MatCha with complete components achieves optimal performance across all metrics MSE 0.0906 (11.8% reduction from baseline), MAE 0.2456 (10.6% reduction), and R^2 0.1686 (257.2% improvement over baseline). MatCha's excellence stems from its specialized pretraining on mathematical charts, functions, and plots, providing crucial inductive biases for interpreting the dense symbolic and geometric information characteristic of student mathematical work. This performance is highly competitive for open-ended mathematical scoring tasks.

Component Effectiveness The architecture enhancements demonstrate clear value for domain-adapted backbones. General-purpose models like CLIP actually degrade under complete configuration, highlighting that domain-adapted pretraining provides advantages that architectural modifications alone cannot overcome. The QAI module demonstrates stable improvement across most backbones. For MatCha, QAI alone reduces MSE from 0.1027 to 0.0926 and improves R^2 from 0.047 to 0.141, with the complete configuration achieving further gains through positive synergy among all four components.

Table 2: Comprehensive Fairness Analysis Results (RQ2.1 to RQ2.3)

		RQ2.1 Privacy Status			RQ2.2 Leakage Type			RQ2.3 Gender (Face)	
Dimension	**Metric**	**Leak**	**Non**	**Result**	**Face**	**Hand**	**Result**	**Male**	**Female**
Model Fairness	High Score Rate (%)	21.5	28.4	Gap	21.5	26.0	Gap	23.6	26.2
	DI Ratio	0.7565		Fail	0.8274		Risk	0.8995	
Error Fairness	MAE	0.0600	0.0649	Better	0.0616	0.0595	Similar	0.0609	0.0621
	CSD	0.0033**		Significant	0.0004		NS	0.0002	

4.2 Privacy-Based Fairness Analysis (RQ2)

Using the optimal MatCha configuration, we implement a dual-perspective fairness framework distinguishing **model fairness** (opportunity allocation via high-score access and disparate impact) from **error fairness** (prediction accuracy consistency across groups). We adopt the 4/5-rule threshold (DI < 0.80 as warning trigger) while acknowledging its contextual limitations.

Table 2 presents comprehensive fairness analysis results across all three research questions.

The Privacy Leakage Paradox (RQ2.1) Results reveal a counterintuitive pattern where privacy leakage samples demonstrate superior prediction accuracy (MAE 0.0600 vs. 0.0649, $p = 0.0078$) yet face systematic disadvantages in high-score opportunities (21.5% vs. 28.4% high-score rate, DI=0.7565 < 0.80 threshold), representing a 24.3% relative reduction in high-score access. The conditional scoring difference (CSD=0.0033, $p = 0.0035$) indicates a statistically detectable but small incremental effect of privacy status on errors after controlling for answer quality, with effect size below the predefined concern threshold (CSD < 0.01). This accuracy-fairness separation suggests that improved technical metrics do not guarantee equitable outcomes.

Facial Leakage as Critical Risk Factor (RQ2.2) Within privacy leakage samples, facial leakage shows significantly lower high-score rates than hand leakage (21.5% vs. 26.0%, $\chi^2 = 4.35$, $p = 0.0360$), with DI=0.8274 approaching the legal boundary with only 0.0274 safety margin. Both types show similar prediction accuracy (MAE difference only 3.55%), indicating the mechanism operates through decision threshold shifts.

Gender Neutrality Achievement (RQ2.3) Under facial leakage scenarios, the most privacy-sensitive condition, the system maintains gender neutrality with DI=0.8995, substantially exceeding the 0.80 threshold and showing nearly identical error rates across male and female students (MAE 0.0609 vs. 0.0621, CSD=0.0002).

5 Discussion

Technical Contributions and Educational Validity The component-driven architecture shifts from generic end-to-end learning toward explicit cognitive simulation of teacher grading processes. MEM, CMCV, QAI, and SPCL collectively align system behavior with documented pedagogical reasoning workflows [8, 19],

achieving substantial performance gains that validate enhanced processing of mathematical visual content. Future work should explicitly evaluate intermediate visual recognition accuracy to distinguish mathematical elements from privacy features.

Observed Associations and Potential Contributing Factors The observed accuracy-fairness separation likely reflects multiple interacting factors rather than a single causal pathway. Models pretrained on general image-text corpora may allocate disproportionate attention to visually salient human features relative to sparse mathematical content, and this attentional imbalance may be associated with the observed differences in high-score opportunity allocation. However, privacy leakage status may also correlate with contextual variables such as photo-taking environment, image quality, student effort levels, assignment type, and device access patterns. Students whose submissions contain privacy information may disproportionately represent particular home learning contexts, mobile device usage patterns, or submission behaviors that co-vary with scoring distributions in ways not fully captured by the current analysis. This complexity aligns with broader findings in educational AI research suggesting that model behavior in authentic deployment contexts emerges from the interaction between algorithmic design and real-world usage conditions. Disentangling these potential pathways requires controlled experimental designs and attention visualization techniques beyond the scope of this observational study.

Design Implications for Responsible Learning Interfaces These findings offer concrete guidance for designing next-generation learning interfaces that integrate AI capabilities with human-centered design and pedagogical principles. Prior research on human–AI interaction in mathematics learning has shown that interface-level design decisions, such as agent interaction styles and feedback mechanisms, meaningfully influence student experience and learning outcomes [11]. The present findings extend this perspective by demonstrating that the visual processing pipeline underlying automated feedback is equally consequential for equitable outcomes. Interface designers should prioritize proactive facial detection and masking over hand masking given the differential fairness risks observed. Evaluation frameworks should adopt dual-perspective monitoring that distinguishes opportunity allocation from error consistency, as accuracy optimization alone can obscure allocation disparities.

Findings support a systematic deployment framework (1) prioritize aggressive facial detection and masking over hand masking given differential fairness risks, (2) adopt dual-perspective monitoring separating opportunity allocation from error consistency, as accuracy optimization alone can obscure allocation biases [16], (3) incorporate privacy-aware training objectives penalizing attention on non-task-relevant demographic regions, and (4) establish human oversight protocols when fairness metrics approach warning thresholds in high-stakes contexts [6].

Limitations and Future Directions The modest R^2 reflects inherent assessment difficulty and human inter-rater variability ceilings. Stratified oversampling alters natural distributions and may affect absolute fairness metric magnitudes. Secondary data analysis limits causal inference about paradox mechanisms. Future

research should develop privacy-aware architectures that intrinsically decouple academic assessment from visual demographic cues, validate findings across diverse contexts, and employ attention visualization and counterfactual techniques to illuminate causal pathways.

6 Conclusion

This study demonstrates that multimodal educational AI systems require integrated technical-ethical design and proactive fairness auditing. Our component-driven architecture achieves strong technical performance by explicitly simulating teacher cognitive processes, yet fairness analysis reveals that privacy leakage creates systematic opportunity disadvantages despite improved prediction accuracy. The findings underscore that responsible deployment of multimodal educational AI demands comprehensive privacy protection protocols and continuous fairness monitoring as fundamental system requirements. Future research must prioritize privacy-aware architectures that decouple academic assessment from visual demographic cues to ensure equitable educational opportunities for all students.

References

1. Aldriye, H., Alkhalaf, A., Alkhalaf, M.: Automated grading systems for programming assignments: A literature review. International Journal of Advanced Computer Science and Applications **10**(3) (2019)
2. Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R.: Instruction based on adaptive learning technologies. In: Mayer, R.E., Alexander, P.A. (eds.) Handbook of Research on Learning and Instruction, pp. 522–559. Routledge, 2 edn. (2017). https://doi.org/10.4324/9781315736419
3. Baker, R.S., Hawn, A.: Algorithmic bias in education. International Journal of Artificial Intelligence in Education **32**(4), 1052–1092 (2022). https://doi.org/10.1007/s40593-021-00285-9
4. Baral, S., Botelho, A.F., Erickson, J.A., Benachamardi, P., Heffernan, N.T.: Improving automated scoring of student open responses in mathematics. International Educational Data Mining Society (2021)
5. Ghavidel, H.A., Zouaq, A., Desmarais, M.C.: Using bert and xlnet for the automatic short answer grading task. In: CSEDU (1). pp. 58–67 (2020)
6. Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., Rienties, B.: The promise and challenges of generative ai in education. Behaviour & Information Technology **44**(11), 2518–2544 (2025)
7. Holstein, K., McLaren, B.M., Aleven, V.: Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. Journal of Learning Analytics **6**(2), 27–52 (2019). https://doi.org/10.18608/jla.2019.62.3
8. Lachner, A., Nückles, M.: Tell me why! content knowledge predicts process-orientation of math researchers' and math teachers' explanations. Instructional Science **44**(3), 221–242 (2016)

9. Lan, A.S., Vats, D., Waters, A.E., Baraniuk, R.G.: Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In: Proceedings of the second (2015) ACM conference on learning@ scale. pp. 167–176 (2015)
10. Li, H., Li, C., Xing, W., Baral, S., Heffernan, N.: Automated feedback for student math responses based on multi-modality and fine-tuning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. pp. 763–770. ACM (2024)
11. Li, H., Xing, W., Li, C., Zhu, W., Lyu, B., Zhang, F., Liu, Z.: Who should be my tutor? analyzing the interactive effects of automated text personality styles between middle school students and a mathematics chatbot. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. pp. 910–917. ACM (2025)
12. Li, H., Xing, W., Zhu, W., Li, C., Lyu, B., Liu, Z., Heffernan, N.: Leveraging multimodality and collaborative filtering for supporting automatic scoring in mathematics education. In: International Conference on Artificial Intelligence in Education. pp. 313–320. Springer (2025)
13. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
14. Liu, F., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Altun, Y., Collier, N., Eisenschlos, J.: Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12756–12770 (2023)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
16. Starke, C., Baleis, J., Keller, B., Marcinkowski, F.: Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data & Society **9**(2), 20539517221115189 (2022)
17. Su, J., Yang, W.: Unlocking the power of chatgpt: A framework for applying generative ai in education. ECNU Review of Education **6**(3), 355–366 (2023)
18. Wong, K.Y., Oh, K.S., Ng, Q.T.Y., Cheong, J.S.K.: Linking it-based semi-automatic marking of student mathematics responses and meaningful feedback to pedagogical objectives. Teaching Mathematics and its Applications: An International Journal of the IMA **31**(1), 57–63 (2012)
19. Wong, T.K., Tao, X., Konishi, C.: Teacher support in learning: Instrumental and appraisal support in relation to math achievement. Issues in Educational Research **28**(1), 202–219 (2018)