An Automated Aesthetic Assessment Framework of Mathematical Story Images Validated by Click Counts

Hai Li
University of Florida
Gainesville, Florida, USA
li.ha@ufl.edu

Wanli Xing
University of Florida
Gainesville, Florida, USA
wanli.xing@coe.ufl.edu

Bailing Lyu
University of Utah
Salt Lake City, Utah, USA
u6057059@utah.edu

Wangda Zhu
University of Florida
Gainesville, Florida, USA
wangdazhu@ufl.edu

Zifeng Liu
University of Florida
Gainesville, Florida, USA
liuzifeng@ufl.edu

Hongming Li
University of Florida
Gainesville, Florida, USA
hli3@ufl.edu

Abstract

Some online learning platforms frequently recommend educational materials to attract student engagement, with visual elements playing a critical role in capturing attention. To optimize the visual design of mathematical stories, this study examines the relationship between visual features and click frequency, based on log data from a U.S. platform featuring AI-generated mathematical stories for elementary students. Our methodology involves a multi-level visual feature extraction framework, categorizing features into low-, mid-, and high-level. Low-level features capture fundamental visual elements like color, texture, shape, and composition, commonly used for their simplicity. Mid-level features, inspired by psychological and artistic theories, more directly link to emotional impact, including attributes like brightness and contrast. High-level features focus on semantic content, using AI models to extract aesthetic scores and identify entities. Based on the correlation analysis between visual features and clicks, our findings indicate that images featuring characters and natural landscapes positively correlate with student interest, aligning with theories of situational interest. In contrast, images with pronounced brightness contrasts negatively impact engagement, likely due to increased cognitive load. The study highlights the limited influence of mid-level aesthetic features on elementary students' engagement, emphasizing the importance of visual clarity and educational relevance over purely aesthetic considerations.

CCS Concepts

• **Applied computing → Interactive learning environments**; *Computer-assisted instruction.*

Keywords

visual aesthetic assessment, click counts, mathematical story

ACM ISBN 979-8-4007-1291-3/25/07
https://doi.org/10.1145/3698205.3733923

ACM Reference Format:
Hai Li, Wanli Xing, Bailing Lyu, Wangda Zhu, Zifeng Liu, and Hongming Li. 2025. An Automated Aesthetic Assessment Framework of Mathematical Story Images Validated by Click Counts. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25), July 21–23, 2025, Palermo, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3698205.3733923

1 Introduction

Mathematical stories support the development of mathematical skills, foster critical and creative thinking, and promote equity in learning mathematics [15, 37]. In particular, multimodal mathematical stories (combining text and images) leverage dual coding theory [27] to enhance information processing, with empirical studies reporting improved academic outcomes and motivation [10, 31].

On online learning platforms, story cards featuring images and titles function similarly to advertisements [9], where visual presentation can influence student engagement. Emotional design theory indicates that carefully crafted visual elements can evoke positive emotions and stimulate situational interest, which in turn boosts learning motivation [13, 18, 33]. Motivated by these insights, our study focuses on identifying which types of images in the story card format are most likely to attract student clicks and encourage further exploration.

To achieve this, we adopt a visual aesthetic quality assessment framework [1] that extracts a range of image features from basic attributes (e.g., color, texture, composition) to higher-level semantic content and aesthetic scores [14, 21, 25]. In the humanities, such visual appeal is often termed visual interest [3] and is understood to depend on factors like visual novelty, complexity, emotional arousal, and cognitive load [2, 3]. Previous work on online platforms confirms that aesthetic attributes significantly influence click behavior [21, 24, 35], yet evaluations of instructional materials have traditionally been text-centric. Given the expanding role of multimedia in education, there is a pressing need for systematic evaluation of visual design in educational content. Accordingly, this paper addresses the following research questions (RQs):

(RQ1) How can we extract image features from the perspective of automated aesthetic quality assessment?

(RQ2) Which features of images in elementary mathematical stories are significantly correlated with the number of clicks?

2 Related Works

2.1 Theoretical Foundation

Visual interest is defined as an image's ability to engage users [2] and is examined in education as situational interest—a temporary state triggered by engaging tasks [7]. Analogous to epistemic emotions that foster exploration and learning [12], multimedia learning research posits that image attributes evoke both cognitive and affective responses, a process termed visual emotional arousal [3, 5, 26]. While emotional design can enhance engagement [33], overly vivid images may increase cognitive load [22]. Moreover, integrating narrative elements (e.g., mathematical stories) with attention-capturing images further stimulates situational interest [4, 29]. This work focuses on how images alone impact learner engagement.

2.2 Automatic Aesthetic Quality Assessment of Image Attractiveness

Previous studies have correlated image features with attractiveness, using metrics such as views and comments as proxies for user interest [25]. High-quality, professionally captured images—especially those featuring faces—tend to garner greater engagement, though color effects can be context-dependent [21]. Computational methods that analyze gradients, textures, and colors, alongside AI-predicted image entities, have effectively predicted image popularity [14]. Similarly, research on image memorability has shown that images with human and object entities are more memorable than those of natural scenes [11]. Our study extends these approaches to investigate the link between visual features and their appeal within learning environments.

3 Dataset

Log data represents records of student activities within the learning system. The advantages of using log data to measure student engagement include its low intrusiveness (data is captured in the background without disturbing students) and scalability (it can be widely implemented to obtain fine-grained student operation data) [6, 20]. Our dataset originates from an online math story learning platform in the US. Based on click logs collected over approximately one year in 2024, we analyzed 177 math stories with some stories used across multiple grade levels: 166 from fifth grade, 40 from fourth grade, and 21 from third grade. In the United States, fifth grade is a pivotal year before transitioning to middle school in some regions, making it essential to cultivate an interest in mathematics. This is why the platform prioritizes fifth-grade students, resulting in a higher proportion of stories tailored to this grade level. Each math story is presented in a multimodal format, incorporating both text (stories) and visuals (images). The texts are generated by GPT-4, while the images are generated by DALL-E-3. This multimodal approach was designed by a team of professionals on the platform, and the math story texts and titles generated through this process have been reviewed by several educational experts to ensure they meet the required standards. The webpage randomly displays different math stories with equal probability, showing only one image and one title per story, and students can continue switching through the random displays. After students click on the image, they are directed to the math story learning page; such clicks can be interpreted as active engagement with the math story. Among these math stories, the highest click count is 263, while the lowest is 1, with a total of 946 clicks. The average number of clicks per story is 5.34 (SD=21.62). (After removing stories with only one click, the average number of clicks per story increases to 9.84 (SD = 30.35).) A total of 199 students participated in the study, primarily from fifth grade, with additional representation from third and fourth grades. On average, each student clicked on 4.75 stories (SD = 10.93). Student privacy was protected during data processing by removing all personally identifiable information.

There are two limitations in our click logs. First, student interactions with the stories occur in two different scenarios: one where students freely explore any story, and another where they are instructed to select specific stories assigned by their teachers. Since teacher-assigned stories are generally of higher quality, we did not differentiate between these two scenarios, aligning with the characteristics of real online learning platforms and our quality assessment objectives. Additionally, some stories may have been removed from the platform shortly after being published, resulting in lower click counts. However, since the primary reason for removal is 'low' quality (the number of displayed stories was limited due to the availability of better math stories), this also aligns with our post-hoc evaluation of quality in this study.

4 Visual Factors Correlated with Clicks

4.1 Methods

Image appeal is closely related to visual features, but the underlying mechanisms are complex. For instance, the same scene or objects with different colors can evoke distinct emotional experiences in viewers, while different scenes or objects may elicit similar emotional responses. Designing a discriminative representation that encompasses all emotional factors is challenging. Moreover, some images, such as abstract pictures and artistic paintings, are not easily constrained by precise rules [34]. Given that this work aims to develop reasoning on the attractiveness of images, it is crucial to extract various types of visual features for comprehensive image analysis. Building on previous research [8, 32, 35, 36], we designed a visual feature extraction framework that captures visual features of different levels and generalities for each image, ranging from low-level, mid-level, to high-level features. Based on these extracted features, we employed Spearman correlation analysis to assess the relationship between these features and click counts. Our research framework is illustrated in Figure 1.

Low-level features typically describe the elements of an image and are among the most commonly used visual features due to their computational simplicity. Previous studies have found that low-level visual features can stimulate associations and evoke emotions to some extent, making them useful for expressing the emotional effects of images. However, these features are often difficult to interpret and may have ambiguous correlations with the psychological and artistic perceptions of the image [36]. Low-level features can be broadly categorized into four types: color, texture, shape, and composition.

Mid-level features are inspired by psychology and art theory, and are generally more interpretable than low-level features, often having a stronger connection to image emotions [36]. In our

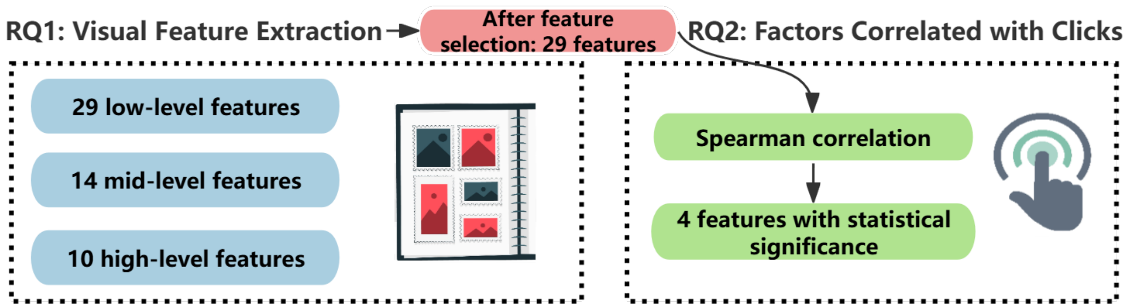

Figure 1: Research framework used in this study.

Table 1: Feature selection categorized by level, including selected and unselected features.

Level	Selected	Name	Description
Low-level (29 → 11 features)	Yes	Gray_features (2)	Combined metrics of grayscale values, including mean and standard deviation, representing the overall brightness and range of brightness variations in the image.
	Yes	Contrast_features (2)	Combined contrast metrics, including the number of peaks in the brightness histogram and the distance between the highest peaks, illustrating the span and distribution of brightness changes.
	Yes	Edge_features (3)	Combined edge characteristics, including density, average distance between edge points, and size of the bounding box containing edge points, providing insights into edge distribution and sparsity.
	Yes	N_segments_0.05 (1)	The number of segments with areas larger than 5% of the image area, helping to understand the presence of dominant regions in the image.
	Yes	Balance_features (3)	Combined balance features, including vertical, horizontal, and center of mass distances, indicating the overall visual balance of the image. Smaller values suggest better balance.
	No	19 others	Other basic features include metrics such as contrast range, lower and upper limits of contrast range, Simpson color diversity index, and saliency attributes (total, block, box size, consistency), as well as line features (total lines, horizontal, vertical, slant), providing comprehensive insights into image brightness contrasts, color distribution, and structural patterns.
Mid-level (14 → 8 features)	Yes	Art_features (5)	Combined artistic metrics, including symmetry, emphasis, harmony, variety, and gradation analysis, capturing the balance, stability, and visual complexity in the image.
	Yes	Color_features (3)	Combined colorfulness and diversity metrics, including colorfulness computed by Hasler and Suesstrunk's formula, earth mover's distance, and Shannon diversity index, representing the richness, uniformity, and diversity of colors.
	No	6 others	Includes metrics such as sharpness, contrast, brightness, saturation, Simpson color diversity index, and Art6_movement, which provide additional insights into visual properties affecting image quality and viewer engagement.
High-level (10 → 10 features)	Yes	Aesthetic scores (2)	Combined aesthetic scores from the AVA dataset and the NIMA model, incorporating multi-scale perception to assess image aesthetics comprehensively.
	Yes	Content_features (8)	Combined content detection metrics, including the number of people, and the presence of plants, animals, buildings, scenery, food, objects, and activities, providing comprehensive context about the image content and potential engagement factors.

Table 2: Four features with statistically significant correlations based on number of clicks.

Level	Feature	Description	Coeff	Mean (SD)
Low-level	Contrast_peak_distance	The maximum peak distance, indicating the span of brightness changes in the image.	-0.228**	-4911.06 (22071.07)
High-level	Scenery	Whether the image contains scenery.	0.221**	0.12 (0.38)
High-level	Plant	Whether the image contains plants.	-0.163**	0.10 (0.36)
High-level	People	Number of people in the image.	0.162*	1.05 (1.16)

*Note. *p<0.05, **p<0.01. If the Bonferroni correction is applied for multiple comparisons, the entity category features Plant and People will no longer be significant, thereby weakening their explanatory power.*

framework, we selected two categories of mid-level features: the first category is based on representations of image attributes, which are popular in visual recognition for their intuitive explanations and cross-category generalization. The second category includes mid-level image features inspired by artistic principles, such as balance, contrast, harmony, variety, gradient, and motion [35].

High-level features encompass the semantic content within images, where the conveyed meanings are easily understood by viewers. These are usually identified through semantic feature recognition, achieved by AI models pretrained on large-scale datasets to generate semantic embeddings. Examples include aesthetic scores [8, 32] and semantic entities. For aesthetic scores, open-source models can predict scores for images. For semantic entities, open-source models extract all objective objects, scenes, and other semantic information within an image based on its visual content, typically expressed as nouns. The BLIP generation model [19] has demonstrated strong feature extraction capabilities for images. BLIP multimodal generation model accepts an image as input and produces a textual description based on a given prompt. Therefore, we employ

BLIP to generate image descriptions and use a predefined dictionary to categorize the extracted entities into different classes. We follow the official prompt design outlined in the foundational paper for extracting image entities, using the prompt: "An image of". Based on the generated descriptions, we construct a dictionary to systematically classify different entity types.

A total of 53 features were extracted. Considering correlation methods used subsequently, there might be collinearity among the extracted features. Backward feature selection was employed to filter features based on the linear matrix until the maximum Variance Inflation Factor (VIF) was less than 5. The process stopped when the maximum VIF reached 3.44, which met the criteria, resulting in 29 selected features, including 11 basic features, 8 intermediate features and 10 advanced features. A summary of all features is presented in the Table 1. After addressing the feature extraction method for RQ1, we conduct a Spearman correlation analysis between the extracted features and click frequency for RQ2 to identify important features.

4.2 Results

Using the 29 filtered features, Spearman correlation analyses were conducted in relation to number of clicks. Among all the features, four were identified as having significant Spearman correlations with the number of clicks, as presented in Table 2. Although the correlation coefficients are not particularly high, they are statistically significant and provide insights into the design of instructional materials. The feature contrast peak distance refers to the distance between the two most prominent peaks in the image brightness histogram (i.e., the primary and secondary brightness peaks). A positive or negative mean value indicates the direction in which the distance between the primary and secondary peaks shifts. The mean values of the features scenery, plant, and people represent the average presence or frequency of these elements in the images. Based on the analysis of these four features from Table 2, we can draw the following opinions: (1) Preference for simple light and shadow levels: Designs featuring simple light and shadow variations, with less complex layering, are more appealing to these students. This suggests that elementary school students are more easily attracted to straightforward and direct visual information, which may be easier for them to process and understand compared to images with intricate light dynamics or strong contrasts. (2) Popularity of people and scenery elements: Images containing elements such as people and scenic landscapes align well with the visual preferences of elementary school students. This could be attributed to the inherent interest of children in this age group toward human figures and natural scenery. However, it is important to note that the term scenery here refers to broader perspectives, such as landscapes or wide views. In contrast, images that contain only plant elements may have an inverse effect.

5 Discussion and Limitations

We introduced an image-based analytics approach, grounded in visual-appeal theory, to link mathematical-story image features with student click rates. RQ1 applied automated aesthetic assessments to extract comprehensive image characteristics, and RQ2 correlated these features with learning logs to highlight those most predictive of engagement.

Our results mirror earlier work showing that images with characters or landscapes significantly boost student clicks—likely by providing relatable narrative anchors that align with learners' emotions—while plant motifs have little impact, perhaps because they don't contextualize math effectively. In short, storyline-driven visuals act as cognitive scaffolds, grounding abstract mathematics in more tangible, real-world scenarios and deepening engagement [13][4].

The finding that high brightness contrasts correlate with fewer clicks underscores the role of cognitive load: abrupt lighting shifts increase visual complexity and siphon attention away from the math, whereas more uniform brightness minimizes extraneous load and lets students concentrate on core concepts. Thus, design choices should reinforce—and not compete with—cognitive processing in educational materials [22].

Limitations include reliance on a single platform's images and modest correlations for many features, implying additional moderators—such as individual preferences [17] or prior knowledge—are unaccounted for. Future work should integrate richer learner profiles, physiological or self-reported engagement measures [28, 30], qualitative click-behavior analyses, and multimodal (text+image) contexts [16]. We also plan to triangulate further engagement indicators and apply causal analyses [23].

6 Conclusion

Our automated framework revealed that character and landscape imagery enhance elementary students' engagement by evoking interest, while excessive contrast hinders it by imposing cognitive load. Mid- and high-level aesthetic scores showed no significant impact, perhaps reflecting developmental limits. These findings underscore prioritizing visual clarity and educational relevance over pure aesthetics. The framework offers a basis for AI-driven, real-time personalization of educational visuals. Future research should verify whether engagement gains translate into measurable learning improvements and refine causal links between design choices and mathematical understanding.

Acknowledgement

This work is supported by the Learning Engineering Virtual Institute under the grant number G-23-2137070 and Jaffe Foundation under the grant number AGR00026932. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] Azimi, J., Zhang, R., Zhou, Y., Navalpakkam, V., Mao, J., and Fern, X. Visual appearance of display ads and its effect on click through rate. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (2012), pp. 495–504.

[2] Chamaret, C., Demarty, C.-H., Demoulin, V., and Marquant, G. Experiencing the interestingness concept within and between pictures. *Electronic Imaging 28* (2016), 1–12.

[3] Constantin, M. G., Redi, M., Zen, G., and Ionescu, B. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR) 52*, 2 (2019), 1–37.

[4] Fox, E. The role of reader characteristics in processing and learning from informational text. *Review of Educational Research 79*, 1 (2009), 197–261.

[5] Heidig, S., Müller, J., and Reichelt, M. Emotional design in multimedia learning: Differentiation on relevant design features and their effects on emotions and learning. *Computers in Human behavior 44* (2015), 81–95.

[6] Henrie, C. R., Bodily, R., Larsen, R., and Graham, C. R. Exploring the potential of lms log data as a proxy measure of student engagement. *Journal of Computing in Higher Education 30* (2018), 344–362.

[7] Hidi, S., and Anderson, V. Situational interest and its impact on reading and expository writing. *The role of interest in learning and development 11* (1992), 213–214.

[8] Hosu, V., Goldlucke, B., and Saupe, D. Effective aesthetics prediction with multi-level spatially pooled features. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 9375–9383.

[9] Hsieh, Y.-C., and Chen, K.-H. How different information types affect viewer's attention on internet advertising. *Computers in human Behavior 27*, 2 (2011), 935–945.

[10] Ida, S., Aziz, R., and Irawan, W. H. Critical and creative thinking skills to solving math story problems in elementary school students. *Jurnal Tatsqif 19*, 2 (2021), 98–113.

[11] Isola, P., Xiao, J., Torralba, A., and Oliva, A. What makes an image memorable? In *CVPR 2011* (2011), IEEE, pp. 145–152.

[12] Izard, C. E., and Ackerman, B. P. Motivational, organizational, and regulatory functions of discrete emotions. *Handbook of emotions 2* (2000), 253–264.

[13] Kaya, N., and Epps, H. H. Relationship between color and emotion: A study of college students. *College student journal 38*, 3 (2004), 396–405.

[14] Khosla, A., Das Sarma, A., and Hamid, R. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web* (2014), pp. 867–876.

[15] Li, H., Guo, R., Li, C., and Xing, W. Automated quality assessment of multimodal mathematical stories generated by generative artificial intelligence. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale* (2024), pp. 110–121.

[16] Li, H., Li, C., Xing, W., Baral, S., and 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* (2024), pp. 763–770.

[17] Li, H., Xing, W., Li, C., Zhu, W., Lyu, B., Zhang, F., and 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* (2025), pp. 910–917.

[18] Li, H., Xing, W., Li, C., Zhu, W., and Oh, H. Are simpler math stories better? automatic readability assessment of gai-generated multimodal mathematical stories validated by engagement. *British Journal of Educational Technology* (2024).

[19] Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning* (2022), PMLR, pp. 12888–12900.

[20] Li, Q., and Baker, R. The different relationships between engagement and outcomes across participant subgroups in massive open online courses. *Computers & Education 127* (2018), 41–65.

[21] Li, Y., and Xie, Y. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of marketing research 57*, 1 (2020), 1–19.

[22] Liew, T. W., Tan, S.-M., Gan, C. L., and Pang, W. M. Colors and learner's gender evoke different emotional and cognitive effects in multimedia learning. *Human Behavior and Emerging Technologies 2022*, 1 (2022), 1235732.

[23] Lyu, B., Li, C., Li, H., Zhu, W., and Xing, W. Explaining technical, social, and discursive participation in online mathematical discussions. *Distance Education* (2024), 1–24.

[24] Maros, A., Belém, F., Silva, R., Canuto, S., Almeida, J. M., and Gonçalves, M. A. Image aesthetics and its effects on product clicks in e-commerce search. In *Proceedings of the SIGIR 2019 eCom workshop. CEUR-WS. http://ceur-ws. org* (2019), vol. 2410.

[25] McParlane, P. J., Moshfeghi, Y., and Jose, J. M. " nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of international conference on multimedia retrieval* (2014), pp. 385–391.

[26] Münchow, H., and Bannert, M. Feeling good, learning better? effectivity of an emotional design procedure in multimedia learning. *Educational Psychology 39*, 4 (2019), 530–549.

[27] Paivio, A., Clark, J. M., et al. Dual coding theory and education. *Pathways to literacy achievement for high poverty children 1* (2006), 149–210.

[28] Papamitsiou, Z., Pappas, I. O., Sharma, K., and Giannakos, M. N. Utilizing multimodal data through fsqca to explain engagement in adaptive learning. *IEEE Transactions on Learning Technologies 13*, 4 (2020), 689–703.

[29] Park, S. The effects of social cue principles on cognitive load, situational interest, motivation, and achievement in pedagogical agent multimedia learning. *Journal of Educational Technology & Society 18*, 4 (2015), 211–229.

[30] Schmitt, H., Witmer, S., and Rowe, S. Text readability, comprehension instruction, and student engagement: Examining associated relationships during text-based social studies instruction. *Literacy Research and Instruction 61*, 1 (2022), 62–83.

[31] Sulistio, W., and Qohar, A. Development of instructional media "game math comic story" based android on number. *Journal of Education Research and Evaluation 4*, 2 (2020), 109–113.

[32] Talebi, H., and Milanfar, P. Nima: Neural image assessment. *IEEE transactions on image processing 27*, 8 (2018), 3998–4011.

[33] Um, E., Plass, J. L., Hayward, E. O., Homer, B. D., et al. Emotional design in multimedia learning. *Journal of educational psychology 104*, 2 (2012), 485.

[34] Zhang, H., Xu, D., Luo, G., and He, K. Learning multi-level representations for affective image recognition. *Neural Computing and Applications 34*, 16 (2022), 14107–14120.

[35] Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., and Sun, X. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia* (2014), pp. 47–56.

[36] Zhao, S., Yao, H., Gao, Y., Ding, G., and Chua, T.-S. Predicting personalized image emotion perceptions in social networks. *IEEE transactions on affective computing 9*, 4 (2016), 526–540.

[37] Zhu, W., Xing, W., Lyu, B., Li, C., Zhang, F., and Li, H. Bridging the gender gap: The role of ai-powered math story creation in learning outcomes. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (2025), pp. 918–923.