Enhancing Creative Expression Skills in Digital Media Education Using CLIP-Driven Multimodal Generation Systems

Yajing Jia
Academy of Fine Arts, Shanxi College of Applied Science and Technology, Taiyuan 030000, Shanxi, China
Corresponding author: Yajing Jia (e-mail: xbyj123456@126.com)

ABSTRACT Current digital media education lacks a quantitative assessment method for creativity that simultaneously considers semantic fidelity, aesthetic quality, and stylistic diversity. This paper proposes a CLIP-driven multimodal generation and evaluation framework for improving creative expression skills. In the generation phase, the CLIP text encoder extracts semantic vectors from student prompts and feeds them into a diffusion model as conditional inputs. During sampling, embeddings of intermediate images and corresponding prompts are compared to guide semantic correction. In the assessment phase, CLIPScore is used to quantify semantic fidelity, a neural image assessment model outputs aesthetic quality scores, and principal component analysis is performed on CLIP visual embeddings from multi-round generated images to compute style dispersion through the covariance matrix trace. These three indicators are weighted to form a comprehensive creative-expression score. Propensity score matching and difference-in-differences estimation are further applied to identify the net causal effect of the system intervention. Experiments involving digital media students show a median semantic fidelity of 0.631, mean visual expressiveness of 0.774, style exploration breadth of 0.629, and average comprehensive creative-expression score of 0.782. The framework enables three-dimensional quantitative assessment and causal evaluation of creative expression in digital media education.

INDEX TERMS clip, multimodal generation, digital media education, creative expression, visual signal analysis

I. INTRODUCTION

WITH the rapid development of generative artificial intelligence technology, multimodal generative models are gradually being applied in the field of digital media education. In particular, text-driven image generation technology is reshaping the teaching paradigm of creative expression. Traditional digital media education relies heavily on hand-drawing or software operation training, and the evaluation method focuses on the presentation of results but lacks systematic quantitative analysis of the creative generation process and expression quality [1-2]. At the same time, students are often limited by technical barriers and performance capabilities in the creative expression process, making it difficult to effectively transform abstract ideas into visual results. In recent years, cross-modal semantic alignment models represented by CLIP have provided the possibility of high-level semantic mapping between text and images. However, in educational scenarios, there is still a lack of unified standards on how to scientifically evaluate their role in promoting creative expression ability. Existing research mostly stays at the level of subjective evaluation or single indicator analysis, lacking comprehensive quantitative tools that take into account both semantic consistency and aesthetic quality, making it difficult to objectively verify the teaching intervention effect and internal mechanism [3-4]. This paper provides a quantitative analysis of how students express creativity in their learning with digital media. The primary areas of data collection are:

(1) The use of a CLIP enabled multimodal generation system for teaching;

(2) Visual output from students when making things with the help of AI;

(3) A system of indices that will reflect how students are assessed for the ability to express their creativity in multiple dimensions [5-6].

The research uses students enrolled in digital media courses, examines their performance differentials based upon their engagement with both traditional hand-drawn creation methods and AI-assisted creation modes, and reveals how technology intervenes within the creative ex-

pression process [7-8]. In addition to determining whether the creative product represents the semantic meaning of a series of words, the creation of visual and stylistic diversity in visualizing that meaning is an important part of being creative. Because of this, it is very difficult to determine a student's level of creativity with one dimensional measures; therefore, it is essential to create a comprehensive assessment framework that combines all three facets of creativity as enumerated above [9-10]. In addition to a research project's research object, a project also has an evaluation mechanism and process to give feedback to the project on itself. All three of these factors together build a three-in-one analytical system of: "technology, learner, and work", thus providing precise and focused ways for the researcher to address problems, and design new methods of conducting research based on their research problem.

Researchers have developed many different algorithms and models in past work on multimodal generation quality assessment. One common approach to evaluate the semantic consistency of generated images and text prompts is using CLIP-based similarity scoring to measure how well the generated images match the provided text prompts. This approach quantitatively expresses semantic alignment through calculation of the cosine similarity between the cross-modal embeddings in their respective embedding spaces [11-12]. In addition to this, model optimization across all generative models has seen new CLIP guidance to use text embeddings as conditional signals during sampling for diffusion networks or GANs which ultimately helps to provide more semantically related outputs [13-14]. NIMA (Neural Image Assessment) model is a tool used in the area of visual quality assessment that allows for an effective way to quantify visual expressiveness by predicting the aesthetic quality of an image based on data provided by human evaluators across a number of different sources [15-16]. These strategies remain inadequate in their application to digital media education: almost all studies examined discriminate only along one dimension (either semantic or visual) and fail to represent creative expression across multiple dimensions. The assessments used to determine whether educational interventions were effective do not consistently separate the contribution from technical tools compared to educational intervention, making it difficult to identify the impact of the technical tool(s). In addition, there is a continuing lack of empirical research on the contributions of image-style diversity and divergent thinking; therefore the reliability and applicability of the reported indicators to education is uncertain [17]. Therefore, it is necessary to construct a systematic evaluation model that integrates semantic, aesthetic and innovative dimensions, and to introduce rigorous experimental design to fill the research gap in the educational context. To address the aforementioned issues, this paper aims to construct a quantitative assessment method for creative expression ability in digital media education. This method centers on cross-modal semantic alignment and aesthetic quality prediction. Through a CLIP-guided diffusion generation mechanism, dynamic semantic matching between text and images is achieved during the generation phase. In the assessment phase, CLIP Score is calculated to quantify semantic fidelity, and NIMA scoring is introduced to measure visual aesthetic performance. Principal component dimensionality reduction analysis of image embeddings is performed to extract style dispersion to reflect creative divergence ability. Subsequently, a weighted joint adaptation function is constructed to integrate the three types of indicators into a unified quantitative comprehensive score for creative expression. At the method validation level, this paper further introduces propensity score matching and a difference-in-differences strategy, using traditional hand-drawn works as a benchmark to identify the net causal effect of the CLIP-driven generation system on students' creative expression ability.

II. QUANTITATIVE EVALUATION MODEL CONSTRUCTION

A. FORMAL DEFINITION OF THE PROBLEM

Traditional CLIP applications in multimodal generation systems are largely limited to post-generation quality assessment. This paper embeds CLIP into the closed-loop control loop of the generation process: the semantic vector is obtained from the student's input text prompts via the CLIP text encoder. This vector serves as both a conditional input to the diffusion model and, at each sampling step, calculates the gradient between the image embedding corresponding to the current latent variable and the semantic vector, guiding the sampling trajectory towards a region with high semantic alignment [18-19].

The student's creative behavior in a CLIP-driven multimodal generation system is abstracted as a mapping process of triplets (T, I, S). The student inputs creative text prompts, where T represents the finite text space of natural language description; the generation system outputs corresponding images, where I represents the RGB (Red, Green, Blue) image tensor space; and the student iterates through multiple rounds of image generation to form an image sequence, where S represents the trajectory space of the generated image sequences.

The quantitative assessment of creative expression ability is defined as a problem of measuring the quality of the triplet mapping function $Q(T, I, S)$. This function numerically characterizes the creative process from three orthogonal dimensions: semantic alignment, aesthetic expressiveness, and stylistic divergence. Formal representation:

$$W = \lambda_1 \cdot \Phi + \lambda_2 \cdot \Psi + \lambda_3 \cdot \Omega \tag{1}$$

In equation (1), Φ is the semantic fidelity function, which is calculated by extracting the cosine similarity between the text embedding and the image embedding using CLIP dual encoder; Ψ is the aesthetic quality function, which is scored by the NIMA convolutional network on the aesthetic output of the image; Ω is the style exploration breadth function [20-

21], which is calculated by reducing the covariance matrix trace of the CLIP visual embedding matrix of all images in the sequence after principal component dimensionality reduction; λ_1, λ_2, λ_3 are preset weight coefficients. Hand-drawn works are processed in an adapted framework. While their trajectory space is traditionally a single instance, for fair comparison, we define their 'style exploration breadth' as the diversity among the multiple draft versions produced during their manual creation

process, rather than treating them as a single element set. This ensures the covariance matrix calculation for the control group remains mathematically valid. Semantic fidelity corresponds to the ability of "creative translation", which is whether learners can accurately externalize the internal concepts into visual symbols, which is the basic skill of digital media creation; aesthetic quality corresponds to the ability of "visual judgment", which is whether learners can identify and produce works with artistic appeal, which is the embodiment of professional quality; style exploration breadth corresponds to the ability of "divergent thinking" and "iterative optimization", which is the degree to which learners try different visual solutions in multiple rounds of human-computer interaction, reflecting the exploratory behavior in the creative process.

B. CLIP SEMANTIC FIDELITY EXTRACTION MODULE

All parameters [22-23]. The text encoder uses a byte-pair-based vocabulary to segment the input prompt words into a word sequence {w1, ..., wL} of length L. A [SOS] marker is inserted at the beginning, and a [EOS] marker is appended at the end. After passing through a 12-layer semantic fidelity extraction module, the text prompt words input by the student in a single creation are compared with the corresponding images output by the generation system. The output scalar represents the alignment degree between the two in the CLIP joint embedding space. The specific implementation is as follows:

Load the pre-trained CLIP ViT (Vision Transformer)-B/32 model weights, freeze the Transformer encoder to extract the context representation, and take the hidden state corresponding to the [EOS] position, mapping it to a 512-dimensional joint embedding space through a linear projection layer to obtain the text embedding vector $\mathbf{t}$. The visual encoder scales the input image to 224×224 pixels and divides it into 7×7 fixed-size blocks. After 12 layers of multi-head self-attention encoding by ViT-B/32, the output corresponding to the category label is taken and mapped through a linear projection layer of the same dimension to obtain the image embedding vector $\mathbf{v}$. To quantify the semantic matching degree between the text description and the generated image, the cosine value Φ of the angle between the two embedding vectors is calculated: $\mathbf{v}^\top \mathbf{t}$

$$\Phi = \frac{\mathbf{v}^\top \mathbf{t}}{\|\mathbf{v}\|_2 \cdot \|\mathbf{t}\|_2} \tag{2}$$

In equation (2), $\mathbf{v}^\top \mathbf{t}$ is the vector inner product, and $\|\cdot\|_2$ represents the Euclidean norm. The cosine similarity value ranges from [-1, 1] , with negative values indicating semantic contradictions, zero values indicating

orthogonality and inconsistency, and positive values indicating semantic convergence. When Φ is close to 1, it indicates that the generated image is highly faithful to the visual semantics implied by the text description. For each pair of prompt word-image data in this method, the above embedding extraction and similarity calculation process is performed, and the resulting Φ is recorded as the original semantic fidelity score of the sample. To address the 'hubness' problem and the inherent misalignment between CLIP's visual and textual spaces, the module implements a temperature-scaled cosine similarity with a learnable projection layer fine-tuned on a digital media education dataset. In addition to mean centering, this specialized layer maps general embeddings into an education-specific vocabulary space, ensuring more accurate fidelity scores for pedagogical assessment[24]. Finally, the centered Φ is linearly scaled to the [0,1] interval as the standardized input for subsequent weighted fusion.

The CLIP model is not only used for semantic fidelity evaluation after generation, but also participates in the image generation process. After the student inputs the creative text prompt, the system first uses the CLIP text encoder to extract the semantic embedding vector, which is also used as the conditional input of the diffusion model. In the iterative denoising process of the diffusion model, each sampling step maps the current intermediate latent variable to an intermediate image through a differentiable decoder, and calculates the semantic alignment gradient between the image and the text prompt in the CLIP joint embedding space. Then, the sampling trajectory is corrected along the gradient descent direction [25-26].

C. NIMA AESTHETIC QUALITY ASSESSMENT MODULE

The aesthetic quality assessment module takes the generated image as input and outputs a scalar $\Psi \in [0, 1]$ representing a comprehensive score of the image's technical quality and artistic beauty. The core of the module employs a NIMA convolutional neural network pre-trained on the AVA (Aesthetic Visual Analysis) dataset. By training on the entire score distribution instead of just one peak, the network becomes more robust to the natural uncertainty and vagueness that occurs with subjective scoring. The module does not perform gradient or weight updates during forward propagation or freeze the pre-trained weights; it serves as an aesthetic feature extractor. The backbone architecture is MobileNet, but the last layer of MobileNet was replaced with a fully connected layer to produce a vector representation of the 10-dimensional score distribution[27-28].

The input image first undergoes a preprocessing pipeline: the RGB three-channel pixel values are normalized to the [-1,1] interval and scaled to a 224×224 pixel resolution to match the dimensionality requirements of the network input layer. The preprocessed tensor is processed layer by layer

by the MobileNet depthwise separable convolutional module to extract multi-scale visual features. After global average pooling in the final layer, a 1280-dimensional feature vector is obtained. This vector is then mapped to 10 neurons through a fully connected layer with a Dropout rate of 0.5. A Softmax activation function is applied to output a rating probability distribution vector, where p_i represents the probability that the image is rated at the i-th rating level. The rating levels are linearly divided from 1 point (lowest quality) to 10 points (highest quality) [29-30]. To aggregate the discrete probability distribution into a scalar aesthetic score, the expectation value calculation method is used:

$$\Psi = \frac{1}{9}\left(\sum_{i=1}^{10} p_i \cdot s_i - 1\right) \tag{3}$$

In equation (3), s_i is the original score of the i-th rating level, $\sum_{i=1}^{10} p_i \cdot s_i$ is the expected score value, and its value range is [1, 10]. After linear transformation, subtracting 1 and dividing by 9, Ψ is mapped to the normalized interval [0, 1], consistent with the dimension of the semantic fidelity index. This transformation maintains the relative order of the score distribution, only adjusting the numerical scale to adapt to the subsequent weighted fusion calculation.

D. FRAMEWORK FOR WEIGHTED FUSION AND CAUSAL INFERENCE OF CREATIVE EXPRESSION QUANTITATIVE COMPREHENSIVE SCORE

The creative expression quantitative comprehensive score weighted fusion module aggregates three heterogeneous indicators—semantic fidelity Φ, aesthetic quality Ψ, and style exploration breadth Ω—into a single comprehensive score. The calculation of style exploration breadth Ω is performed on the image sequence generated in K iterations of a student's single creative task: each image is processed by the CLIP visual encoder to extract a 512-dimensional embedding vector, forming an embedding matrix $\mathbf{E}$[31-32]. Principal component analysis is performed on the transpose of $\mathbf{E}$, and principal components with a cumulative variance explanation rate of 95% are selected to obtain the dimensionality-reduced embedding matrix $\mathbf{E}_{\text{red}}$. Its covariance matrix $\Sigma = \frac{1}{K-1}\mathbf{E}_{\text{red}}^{\top}\mathbf{E}_{\text{red}}$ is calculated, where K represents the number of generated images. Its trace is taken as the style dispersion, which measures the mathematical dispersion of images. To bridge this to 'creative divergence' in education, we introduce a 'Quality-Weighted Diversity' filter: variations are only

included in the covariance matrix if their NIMA aesthetic score exceeds a student's baseline. This ensures that the breadth reflects intentional exploratory growth rather than random, low-quality variations [33-34]. After normalization, the three indicators are linearly combined according to preset weights to form a quantitative comprehensive score for creative expression:

$$CEQ = \lambda_\Phi \cdot \Phi + \lambda_\Psi \cdot \Psi + \lambda_\Omega \cdot \bar{\Omega} \tag{4}$$

In equation (4), $\lambda_\Phi = 0.4$ is the semantic fidelity weight, $\lambda_\Psi = 0.4$ is the aesthetic quality weight, $\lambda_\Omega = 0.2$ is the style exploration breadth weight, and $\bar{\Omega}$ is the style dispersion value after maximum-minimum normalization to the [0,1] interval. The weight allocation ($\lambda_\Phi = 0.4$, $\lambda_\Psi = 0.4$, $\lambda_\Omega = 0.2$) is determined based on the 'Semantic-Aesthetic Primacy' theory in digital media education, which posits that fundamental communication (semantics) and visual quality (aesthetics) are the primary prerequisites for creative expression. However, these weights are task-adaptive; for abstract narrative tasks where divergent thinking is paramount, λ_Ω can be increased to 0.4 to better reflect creative exploration. The causal inference framework adopts a strategy combining propensity score matching and difference-in-differences (DID). To satisfy the parallel trends assumption, we use Propensity Score Matching (PSM) to select control participants with similar baseline technical skills and prior artistic performance, thereby ensuring that the two groups possess comparable potential for improvement regardless of the technical medium used. The net causal effect of the system intervention is given by the average treatment effect on the treated group (ATT) estimate [35], that is, the difference between the mean ΔCEQ of the method in this paper and the mean ΔCEQ of the matched control group, thereby removing the confounding effect of natural maturation, external factors, and the 'novelty effect.' To isolate pure skill enhancement from temporary engagement with a new tool, we conducted a 'Staged Adaptation' pre-test where students in all groups were familiarized with their respective tools (AI or traditional) for a week prior to formal data collection, ensuring that the DID estimator captures sustained creative improvement. The algorithm architecture diagram is shown in Figure 1.

Figure 1 illustrates the complete algorithm flow from student creative data to the evaluation of the system intervention effect. The input on the left includes the image sequence generated by students in multiple iterations, the image generated in a single iteration, and the corresponding text prompts. Three processing modules are arranged side-by-side in the middle: the semantic fidelity module calculates the semantic alignment by matching text and images; the aesthetic quality module independently evaluates the visual expressiveness of the images; and the style exploration breadth module analyzes the distribution dispersion of the multi-round image sequence in the feature space, reflecting the style variation and exploratory behavior of students' creations. The outputs of these three modules are then fed into a fusion module, where they are weighted and summed to obtain a quantitative comprehensive score for each student's creative expression. The causal inference module on the right takes the quantitative comprehensive scores of creative expression from the proposed method and the control group, as well as pre-test covariates, as input. It first balances the background differences between the two groups of students through propensity score matching, then uses the difference-in-differences method to calculate the

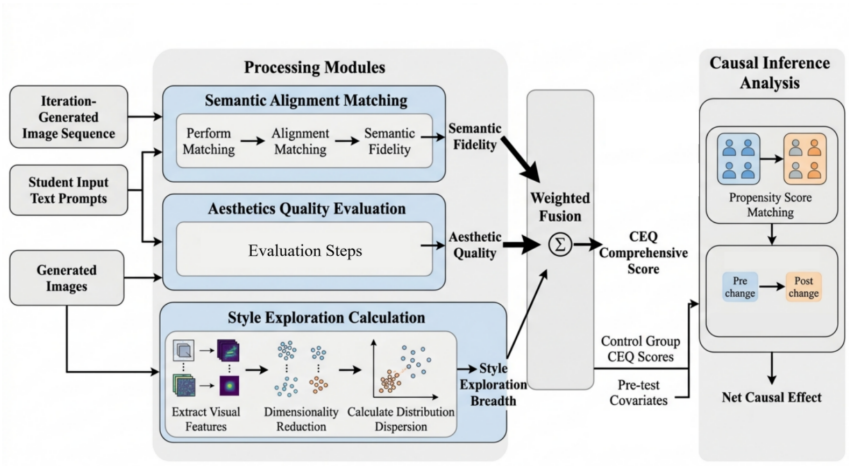

FIGURE 1. Algorithm architecture diagram of weighted fusion and causal inference of creative expression quantitative comprehensive score

changes in pre- and post-test scores, and finally outputs the net causal effect brought about by the system intervention.

III. TEACHING EXPERIMENTS AND EFFECTIVENESS EVALUATION

A. EXPERIMENTAL DESIGN

The experiment recruit 64 students from a university majoring in digital media art, including 28 males and 36 females. All participants have completed two prerequisite courses, basic sketching and color composition, and possessed basic experience in digital image processing. They had never used a CLIP-driven multimodal generation system. A stratified randomization method is used to assign the 64 students to the proposed method and three control groups: students are divided into high, medium, and low levels based on their baseline scores in the pre-test for creative expression ability, and within each level, they are randomly assigned in a 1:1:1:1 ratio. Ultimately, there are 16 students in the proposed method, 16 in Control A, 16 in Control B, and 16 in Control C. This study was approved by the Ethics Committee of Shanxi College of Applied Science and Technology, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

The experimental tasks consist of three progressively challenging creative tasks: Task 1, "Concrete Scene Description," requires students to transform a given textual passage into a visual image; Task 2, "Abstract Concept Expression," requires students to interpret abstract vocabulary through visual metaphors; and Task 3, "Narrative Sequence Construction," requires students to create a narrative sequence containing three images based on a short plot.

This paper employs a self-developed multimodal generation system to complete the three creative tasks. The backend of the system is based on a diffusion model. It uses the CLIP ViT-B/32 as a module to guide text encoding and generation while the frontend provides a web-based interactive interface. Students access the system via a browser. The interface has four overall areas: prompt word input box, generated image display area, semantic alignment feedback panel, and historical iteration thumbnail bar. The workflow of the system is as follows: the CLIP text encoder extracts a semantic vector from the student-provided prompt word which serves as the conditional input for the diffusion model and then at each step of sampling, the generation direction is corrected using CLIP-guided gradient. After generating the image, the system provides the user with a CLIPScore and semantic heatmap that can be used to iteratively optimize their prompt word. The complete creative trajectory is automatically saved by the system.

Control A: A reverse evaluation scheme driven by the BLIP-2 model is adopted. To mitigate systemic bias favoring the experimental group, we cross-validated all assessments by applying the CLIPScore metric to scanned hand-drawn works and the BLIP-2/BLEU-4 metric to AI-generated works. This dual-tool verification ensures that the observed performance gap is not an artifact of the specific evaluation model (CLIP vs. BLIP-2) but a reflection of actual creative output quality.

Control B: A single NIMA aesthetic quality scoring scheme is adopted. The creative expression ability of students' hand-drawn works is represented solely by the aesthetic score output by the NIMA model, without considering semantic fidelity and style dispersion dimensions.

Control C: A generated image diversity evaluation scheme based on Inception Score is adopted. The scanned images of students' hand-drawn works are processed by the Inception v3 network to extract the category conditional probability distribution. The expected value of the marginal distribution of the KL (Kullback- Leibler) divergence is calculated as the sole quantitative basis for creative expression ability.

B. SEMANTIC FIDELITY EVALUATION

This module quantifies the semantic alignment between text and images. The proposed method involves calculating the cosine similarity of each student's cue word-image pair for

each task using the CLIP encoder and scaling it to [0,1]. The highest score across multiple iterations is taken, and the average of the three tasks is used as the individual semantic fidelity index, reflecting the optimal semantic alignment achieved through cue word optimization. Control A: BLIP-2 is used to generate descriptions for hand-drawn artworks, with BLEU-4 matching of the original cue words as a substitute metric. Control B: This index is not calculated. Control C: Projected similarity after dimensionality reduction using CLIP text embedding and Canonical Correlation Analysis (CCA) of Inception v3 image features is used as the estimate. One-way ANOVA and Dunnett's post-hoc test are used to compare inter-group differences. The correlation coefficient between the proposed method score and the pre-test baseline score is calculated to test sensitivity, and the consistency between CLIPScore and human-computer scoring is verified based on teacher evaluation. The semantic fidelity index is shown in

Figure 2 shows the evaluation results of the semantic fidelity index across the three groups. Figure 2(a) shows that the median for Proposed Method is 0.631, for Control A it is 0.459, and for Control C it is 0.545; the interquartile range of Proposed Method is 0.149, which is greater than that of Control A (0.114) and Control C (0.101). Figure 2(b) shows that the peak value for Proposed Method is 0.638, for Control A it is 0.469, and for Control C it is 0.542. The median and peak values of the three groups show the following order: Proposed Method has the highest value, followed by Control C, and then Control A has the lowest value.

C. VISUAL EXPRESSIVENESS INDICATORS AND ASSESSMENT TOOL VALIDITY

The visual expressiveness assessment module uses the NIMA model to quantitatively measure the technical quality and artistic aesthetics of the images generated by the proposed method and three sets of control works. The proposed method uses the highest NIMA score generated through multiple iterations for each student in each task, and then uses the average of the highest scores from the three tasks as the individual indicator to reflect the student's judgment and selection ability regarding the aesthetic quality of the gener ated images. Control A and Control B both use NIMA scoring for the hand-drawn scanned images; the former uses the arithmetic mean of the three tasks, while the latter uses the NIMA score as the sole assessment dimension. Control C supplements the NIMA score as a reference indicator but does not participate in the overall score calculation.

To verify the applicability of the NIMA model in digital media education scenarios, twenty generated images and twenty hand-drawn scanned images are selected. Three professional teachers independently assessed the aesthetic quality using a ten-point scale. The Spearman rank correlation coefficient between the model score and the average teacher score is calculated to assess the consistency between human and machine ranking. Simultaneously, the Pearson correlation coefficient between the model score and the students' prior course grades is calculated to examine the discriminant validity of the indicator for basic art skills. Visual expressiveness, human-computer consistency, and criterion-related validity are shown in Figure 3.

Regarding the three assessment results in Figure 3, Figure 3(a) shows the inter-group comparison of visual performance among the four groups of students: the original mean of Proposed Method is 0.774, Control A is 0.648, Control B is 0.635, and Control C is 0.671; the adjusted means after adjusting for covariates are 0.771, 0.634, 0.627, and 0.668, respectively. The mean of Proposed Method is the highest among the four groups, while the mean of Control B is the lowest. In the human-machine consistency testing found in Figure 3(b) there was a Spearman correlation coefficient of 0.899 between NIMA scores and the average teacher's scores showing a very high level of consistency in the model's aesthetic scores compared to the human rating of the aesthetic. The criterion validity analysis found in Figure 3(c) shows that there is a moderate level of positive correlation between the NIMA scores and prerequisite course grades with a Pearson correlation coefficient of 0.563.

D. EVALUATION OF THE BREADTH INDEX OF STYLE EXPLORATION

The style exploration breadth index quantifies the spatial distribution of images from multiple creation iterations, that display variation in styles and allow for individual artist's exploration of visual styles with cue word modifications. Samples from the control (no multiple iteration samples) do not contribute to this index. The images generated for each student and for each task will be collected in multiple rounds and are processed in a sequential manner before measuring their geometric characteristics using CLIP visual embeddings, then standardized and then reduced using Principal Component Analysis (PCA) dimensionality reduction techniques. The style dispersion is then calculated based on the trace of the covariance matrix of the dimensionality-reduced embedding matrix. Logarithmic transformation and normalization are then applied to obtain the style exploration breadth score. Finally, the average of the three tasks is taken as the individual index. To verify the index's effectiveness, the cumulative effect of iteration rounds on dispersion is analyzed, and the correlation between style exploration breadth and semantic fidelity, as well as its correlation with visual expressiveness, are examined. The style exploration breadth assessment is shown in Figure 4.

Figure 4 illustrates the evaluation results of the style exploration breadth index. Figure 4(a) shows that as the number of iterations expanded from two rounds to all rounds, the mean style dispersion gradually increased from 0.38 to 0.629, indicating that the diffusion of image distribution within the CLIP embedding space by students continuously expanded during multi-round interactions. In Figure 4(b), the Pearson correlation coefficient between style exploration breadth and semantic fidelity is -0.524

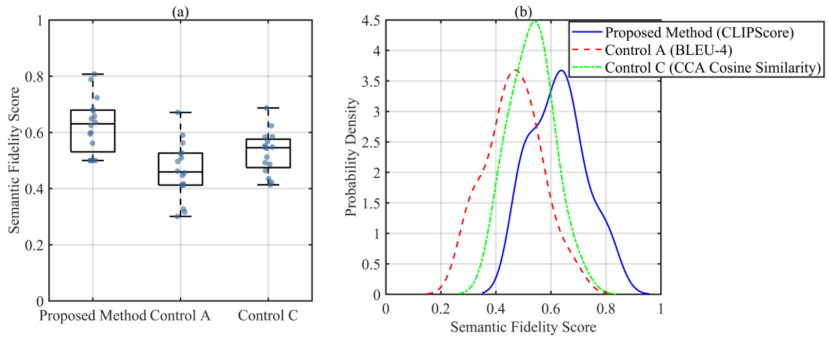

FIGURE 2. Evaluation results of semantic fidelity index: (a) Distribution of semantic fidelity in each group; (b) Kernel density curve of scores in each group

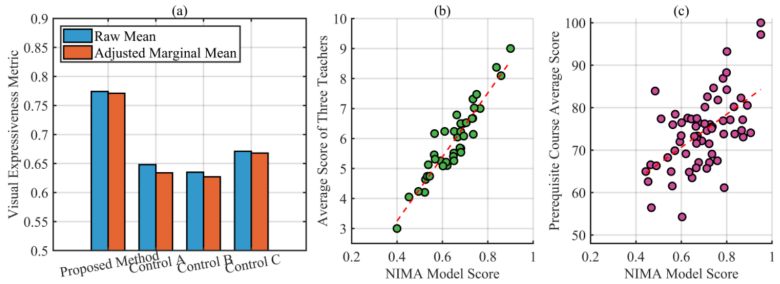

FIGURE 3. Visual expressiveness, human-computer consistency, and criterion-related validity assessment: (a) Comparison of visual expressiveness between groups; (b) Human-computer consistency; (c) Criterion-related validity

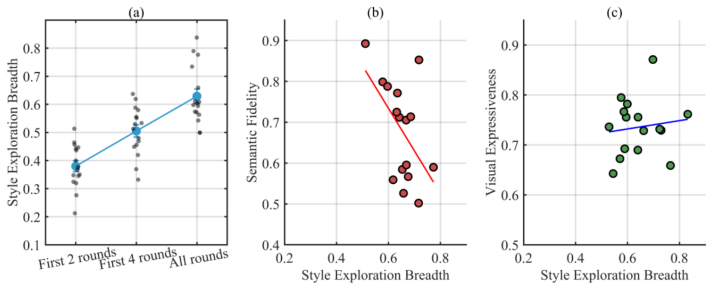

FIGURE 4. Evaluation of the breadth of style exploration: (a) Cumulative effect of iterative rounds of style exploration breadth; (b) Style exploration breadth vs. semantic fidelity; (c) Style exploration breadth vs. visual expressiveness

(p=0.037), with a negative correlation strength higher than the preset target value of -0.32, indicating that students with a strong tendency for divergent exploration had relatively lower semantic alignment levels, suggesting a certain trade-off between the two. In Figure 4(c), the Spearman correlation coefficient between style exploration breadth and visual expressiveness is 0.085 (p=0.755), lower than the target value of 0.45 and statistically insignificant, indicating that the visual expressiveness generated in multiple rounds does not depend on style breadth.

E. COMPREHENSIVE ASSESSMENT OF CREATIVE EXPRESSION ABILITY

The comprehensive assessment module for creative expression ability integrates three heterogeneous indicators—semantic fidelity, visual expressiveness, and breadth of style exploration—according to preset weights into a single quantitative comprehensive score for creative expression, achieving a multi-dimensional and unified quantification of students' creative expression ability. The quantitative score for creative expression is shown in Figure 5.

According to Figure 5, the original mean of the Proposed Method is 0.782 (adjusted to 0.779), while the original means of the control groups A, B, and C are 0.653, 0.641, and 0.660, respectively (adjusted to 0.656, 0.644, and 0.662). The score difference between the proposed method and each control group is greater than 0.1, while the maximum difference among the three control groups is small, indicating that the initial covariate differences between groups are small, and the original results already showed good robustness. This set of data shows that the ternary weighted fusion model proposed in this paper can effectively distinguish the creative expression level of students under the CLIP-driven generation system from that under traditional or single-dimensional evaluation schemes based on the comprehensive score.

NIMA×9 Rating refers to the score after the normalized aesthetic score (range [0,1]) output by the NIMA model is

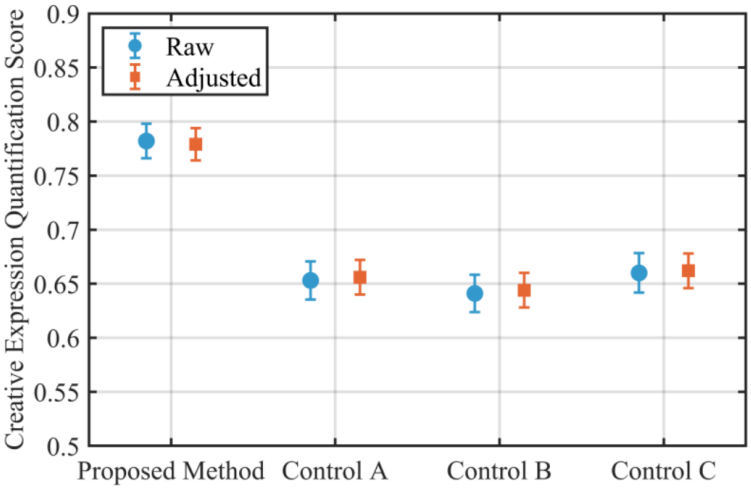

FIGURE 5. Comparison of quantitative scores for creative expression

restored to a 10-point scale (range [1,10]) through linear transformation. The analysis results are shown in Table 1.

TABLE 1. Bland-Altman Analysis Results

Statistic	Value	95% Confidence Interval / Limits
Mean Teacher Rating (10-point scale)	6.48	[6.00, 6.96]
Mean NIMA×9 Rating (10-point scale)	6.35	[5.88, 6.82]
Mean Difference	0.13	[-0.08, 0.34]
Standard Deviation of Differences	0.66	—
Upper Limit of Agreement	1.42	[1.12, 1.72]
Lower Limit of Agreement	-1.16	[-1.46, -0.86]
Percentage of Points within Limits	95.0%	—

The Bland-Altman analysis in Table 1 shows that the mean teacher rating is 6.48, the mean NIMA×9 rating is 6.35, and the average difference between the two is 0.13. The 95% confidence interval is [-0.08, 0.34] (crossing zero), and 95.0% of the data points fell within the consistency bounds. This indicates that the aesthetic ratings output by the NIMA model and human judgments have no systematic deviation at the overall level and exhibit good consistency. Mechanistically, NIMA learns the probability distribution of ratings rather than a single peak, thus adapting to the ambiguity in subjective aesthetics. The average scores and number of participants in each group's pre-course are shown in Tables 2 and 3.

TABLE 2. Average scores of prerequisite courses for each group after grouping by NIMA score quartiles

Quartile Group	NIMA Score Range	Mean Prior Course Grade	p Value
Q1	[0.40, 0.59]	68.2	<0.001
Q2	[0.60, 0.70]	72.5	0.012
Q3	[0.71, 0.80]	76.8	0.187
Q4	[0.81, 0.95]	80.1	—

TABLE 3. Distribution of the number of participants in the method and control groups within each NIMA score quartile

Quartile Group	Proposed Method	Control A	Control B	Control C
Q1 (lowest)	0	6	5	5
Q2	2	5	5	4
Q3	6	3	4	3
Q4 (highest)	8	2	2	4

According to Table 2, students with higher NIMA aesthetic scores in the top two quartiles (Q3 & Q4) achieved higher average scores in the pre-course compared to students with lower scores; the Q4 group had an average score of 80.1 and the Q1 group had an average score of 68.2, demonstrating that there was a positive correlation between NIMA aesthetic scores and the level of basic hand-drawing skill levels. Also, as shown

on Table 3, while there were 14 total students in the top-scoring groups (Q3 and Q4) using the technique (CLIP-based generation system), there were only 8 students in the top-scoring groups in the control groups. In contrast, the control group had 30 students in the low-scoring groups (Q1 and Q2) and there were only 2 in the low-scoring groups using the proposed method. This disparity in the composition of the sample groups shows that while basic skills correspond with aesthetic quality, the proposed method (through CLIP-related semantic guidance and iterative feedback processes) reduces the technical level, allowing a greater number of students to create work with high aesthetic scores, thus highlighting the system's contribution to improving creativity-capable skills.

F. ESTIMATION OF THE MEAN TREATMENT EFFECT IN THE TREATMENT GROUP

The propensity score matching (ATT) and difference-in-differences (DID) method are used to isolate the causal effects of the systematic intervention. Pre-test covariates are selected, and for each student in this study, one student from

each of the three control groups is matched with the student whose propensity score is closest to the propensity score. After matching, the mean treatment effect in the treatment group is calculated, which is the difference between the pre- and post-test quantitative composite scores of creative expression in this study and the matched control groups. The ATT estimation results are shown in Table 4.

TABLE 4. Estimated results of the average treatment effect in the treatment groups

Group	Proposed Method △Score (Mean)	Control group △Score (Mean)	ATT	p-value	Cohen's d
Control A	0.312	0.084	0.228	<0.001	2.68
Control B	0.312	0.112	0.2	<0.001	2.35
Control C	0.312	0.098	0.214	<0.001	2.52

Table 4 shows that the pre- and post-test difference in the quantitative comprehensive score of creative expression for students using the proposed method is 0.312, while the differences for the three control groups are 0.084, 0.112, and 0.098, respectively, with the mean treatment effect (ATT) ranging from 0.2 to 0.228 ($p < 0.001$, Cohen's $d > 2.3$). This difference is attributed to the model's shift of CLIP semantic alignment from passive evaluation to active guidance in the generation process: in each sampling step, CLIP text embedding serves both as a conditional input to the diffusion model and as a gradient-corrected generation trajectory, while simultaneously integrating NIMA aesthetic quality and style dispersion indices to form a multi-dimensional closed-loop feedback. By using this method, students can advance not only in semantic accuracy but

also in visual expressibility and breadth of style discovery during their experience with human-computer creative collaborations, therefore achieving a far greater overall improvement in their capacity to express themselves creatively than the group that did not use this system.

IV. CONCLUSION

The paper develops a framework for an educational intervention that integrates CLIP-guided diffusion generation and quantitative assessment methods in three dimensions. The generation phase uses CLIP semantic vectors as conditional inputs to the diffusion model, and dynamically corrects the sampling trajectory through the use of gradients. In the assessment phase, they create a weighted quantitative score for creative expression by combining the CLIPScore, NIMA aesthetic scores, and style dispersion derived from the trace of the covariance matrix of the principal component analysis. They also use a bias score matching approach, and apply a difference-in-differences methodology to estimate the net causal effect of the intervention. The authors conduct an experiment with 64 students in digital media, finding semantic fidelity of 0.631, visual expressiveness of 0.774, style exploration breadth of 0.629, and a total score for comprehensive creative expression of 0.782, and average treatment effects ranging from 0.2 to 0.228. This framework provides a reproducible technical approach to objectively quantify and causally assess the ability to express creativity in digital media education.

AUTHOR CONTRIBUTIONS

Yajing Jia designed, collected and analyzed the data, and drafted the manuscript. Yajing Jia conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yajing Jia participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Shanxi College of Applied Science and Technology, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. Liu, W. Pang, J. Guo, and Y. Zhang, "A meta-analysis of the effect of multimedia technology on creative performance," Education and Information Technologies, vol. 27, no. 6, pp. 8603-8630, 2022, DOI: 10.1007/s10639-022-10981-1.

[2] H. Long, A. Kerr B, E. Emler T, and M. Birdnow, "A critical review of assessments of creativity in education," Review of Research in Education, vol. 46, no. 1, pp. 288-323, 2022, DOI: 10.3102/0091732X221084326.

[3] Y. Zhang and Z. Li, "Assessing GenAI-assisted digital multimodal composing: Reconceptualizing a genre-based framework through self-assessment and peer assessment," Assessing Writing, vol. 68, no. 1, pp. 101017-101031, 2026, DOI: 10.1016/j.asw.2026.101017.

[4] C. Laupichler M, A. Aster, O. Perschewski J, and J. Schleiss, "Evaluating AI courses: A valid and reliable instrument for assessing artificial-intelligence learning through comparative self-assessment," Education Sciences, vol. 13, no. 10, pp. 978-995, 2023, DOI: 10.3390/educsci13100978.

[5] Y. Zhuang, R. Zhao, Z. Xie, and L. Yu P, "Enhancing language learning through generative AI feedback on picture-cued writing tasks," Computers and Education: Artificial Intelligence, vol. 1, no. 1, pp. 100450-100462, 2025, DOI: 10.1016/j.caeai.2025.100450.

[6] J. Jiang, "When generative artificial intelligence meets multimodal composition: Rethinking the composition process through an AI-assisted design project," Computers and Composition, vol. 74, no. 1, pp. 102883-102896, 2024, DOI: 10.1016/j.compcom.2024.102883.

[7] A. Ringvold T, I. Strand, P. Haakonsen, and S. Strand K, "The AI generative text-to-image creative learning process: An art and design educational perspective," Design and Technology Education: An International Journal, vol. 29, no. 2, pp. 359-379, 2024, DOI: 10.24377/DTEIJ.article2433.

[8] G. Donnici, G. Galiè, and L. Frizziero, "Rethinking Sketching: Integrating Hand Drawings, Digital Tools, and AI in Modern Design," Designs, vol. 9, no. 5, pp. 119-131, 2025, DOI: 10.3390/designs9050119.

[9] L. Li, T. Zhu, P. Chen, Y. Yang, Y. Li, and W. Lin, "Image aesthetics assessment with attribute-assisted multi-modal memory network," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, pp. 7413-7424, 2023, DOI: 10.1109/TCSVT.2023.3272984.
[10] J. Liu, "Modeling analysis of perceived differences in product appearance design based on visual communication," International Journal for Simulation and Multidisciplinary Design Optimization, vol. 16, no. 1, pp. 12-26, 2025, DOI: 10.1051/smdo/2025016.
[11] Z. Tan, X. Yang, Z. Ye, Q. Wang, Y. Yan, A. Nguyen, and K. Huang, "Semantic similarity distance: Towards better text-image consistency metric in text-to-image generation," Pattern Recognition, vol. 144, no. 1, pp. 109883-109896, 2023, DOI: 10.1016/j.patcog.2023.109883.
[12] F. Jahara, M. Khoshnoodi, Y. Lu, M. Saxon, A. Sharma, and Y. Wang W, "Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (ts2)," Advances in Neural Information Processing Systems, vol. 37, no. 1, pp. 85630-85657, 2024, DOI: 10.52202/079017-2719.
[13] R. Gal, O. Patashnik, H. Maron, H. Bermano A, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," ACM Transactions on Graphics (TOG), vol. 41, no. 4, pp. 1-13, 2022, DOI: 10.1145/3528223.3530164.
[14] H. Ma, M. Li, J. Yang, O. Patashnik, D. Lischinski, D. Cohen-Or, and H. Huang, "CLIP-Flow: Decoding images encoded in CLIP space," Computational Visual Media, vol. 10, no. 6, pp. 1157-1168, 2024, DOI: 10.1007/s41095-023-0375-z.
[15] I. Santos, A. Casal M, J. Correia, Á. Torrente-Patiño, P. Machado, and J. Romero, "Towards robust evaluation of aesthetic and photographic quality metrics: Insights from a comprehensive dataset," Complexity, vol. 2024, no. 1, Art. no. 8223586, 2024, DOI: 10.1155/2024/8223586.
[16] M. Daryanavard Chounchenani, A. Shahbahrami, R. Hassanpour, and G. Gaydadjiev, "Deep learning based image aesthetic quality assessment-a review," ACM Computing Surveys, vol. 57, no. 7, pp. 1-36, 2025, DOI: 10.1145/3716820.
[17] J. Thomson T, D. Pfurtscheller, K. Lobinger, N. Laba, and K. Christ, "Visual literacies in transition: multimodal co-production with generative AI," Journal of Visual Literacy, vol. 44, no. 4, pp. 384-405, 2025, DOI: 10.1080/1051144x.2025.2566592.
[18] Y. Wang, "Self-Supervised CLIP-Based Image Recognition and Analysis for Electronic Data Forensics," IEEE Access, vol. 14, no. 1, pp. 3062-3077, 2025, DOI: 10.1109/ACCESS.2025.3648849.
[19] P. Shabari Nath and R. Chouhan, "Multimodal Quality Assessment of AI-Generated Images using BERT-CLIP Feature Fusion: S," Nath, R. Chouhan. Signal, Image and Video Processing, vol. 19, no. 14, pp. 1254-1271, 2025, DOI: 10.1007/s11760-025-04830-0.
[20] H. Su, Z. Li, and J. Hu, "AI-driven Art Learning System: Automated Color Assessment Through Image Recognition," Computer-Aided Design and Applications, vol. 1, no. 1, pp. 147-161, 2025, DOI: 10.14733/cadaps.2026.147-161.
[21] R. Tang, Q. Li, and S. Tang, "Comparison of visual features for image-based visibility detection," Journal of Atmospheric and Oceanic Technology, vol. 39, no. 6, pp. 789-801, 2022, DOI: 10.1175/jtech-d-21-0170.1.
[22] H. Ye, X. Ge, J. Wang, J. Fu, X. Xin, J. Xue, Y. Chen, et al., "Beyond efficient fine-tuning: Efficient hybrid fine-tuning of CLIP models guided by explainable ViT attention," Information Processing & Management, vol. 63, no. 4, pp. 104628-104653, 2026, DOI: 10.1016/j.ipm.2026.104628.
[23] L. Liu, L. Yang, F. Yang, F. Chen, and F. Xu, "CLIP-Driven Few-Shot Species-Recognition Method for Integrating Geographic Information," Remote Sensing, vol. 16, no. 12, pp. 2238-2257, 2024, DOI: 10.3390/rs16122238.
[24] K. Tyshchuk, P. Karpikova, A. Spiridonov, A. Prutianova, A. Razzhigaev, and A. Panchenko, "On isotropy of multimodal embeddings," Information, vol. 14, no. 7, pp. 392-404, 2023, DOI: 10.3390/info14070392.
[25] H. Zhao, G. Jin, X. Jiang, and M. Li, "SDE-RAE: CLIP-based realistic image reconstruction and editing network using stochastic differential diffusion," Image and Vision Computing, vol. 139, no. 1, pp. 104836-104851, 2023, DOI: 10.1016/j.imavis.2023.104836.
[26] J. Kim, E. Kim, and K. Lee, "GODiff: region-specific semantic editing with CLIP-guided diffusion models," IEEE Access, vol. 13, no. 1, pp. 112818-112834, 2025, DOI: 10.1109/access.2025.3580263.
[27] Y. Hu, W. Dong, Y. Zhang, and L. Lu, "Image aesthetic quality assessment: A method based on deep convolutional capsule network," PLoS One, vol. 20, no. 9, Art. no. e0331897, 2025, DOI: 10.1371/journal.pone.0331897.
[28] S. Yang, Z. Wang, G. Wang, Y. Ke, F. Qin, J. Guo, and L. Chen, "A self-supervised image aesthetic assessment combining masked image modeling and contrastive learning," Journal of Visual Communication and Image Representation, vol. 101, no. 1, pp. 104184-104197, 2024, DOI: 10.1016/j.jvcir.2024.104184.
[29] M. Alkanan and Y. Gulzar, "Enhanced corn seed disease classification: leveraging MobileNetV2 with feature augmentation and transfer learning," Frontiers in Applied Mathematics and Statistics, vol. 9, no. 1, pp. 1320177-1320189, 2024, DOI: 10.3389/fams.2023.1320177.
[30] K.S, "A, Singh R P, Panda M K, Palaniappan K," An Ensemble Approach using Self-attention based MobileNetV2 for SAR classification. Procedia Computer Science, vol. 235, no. 1, pp. 3207-3216, 2024, DOI: 10.1016/j.procs.2024.04.303.
[31] T. Fakoya J and O. Ajinaja M, "A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing," Cureus Journals, vol. 3, no. 1, pp. 1-14, 2026, DOI: 10.7759/s44389-026-00049-3.
[32] J. Olaniyan, F. Verkijika S, and C. Obagbuwa I, "Cross-Modal Few-Shot Learning via Siamese Similarity Networks on CLIP Embeddings for Fine-Grained Image Classification," Applied Sciences, vol. 16, no. 7, pp. 3181-3192, 2026, DOI: 10.3390/app16073181.
[33] G. Kwon and C. Ye J, "One-shot adaptation of gan in just one clip," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 12179-12191, 2023, DOI: 10.1109/tpami.2023.3283551.
[34] Z. Zhou, S. Dong, C. Ding, X. Gao, Y. He, and Y. Gong, "Diversity covariance-aware prompt learning for vision-language models," Pattern Recognition, vol. 1, no. 1, pp. 112806-112815, 2025, DOI: 10.1016/j.patcog.2025.112806.
[35] I. Naimi A and W. Whitcomb B, "Defining and identifying average treatment effects," American Journal of Epidemiology, vol. 192, no. 5, pp. 685-687, 2023, DOI: 10.1093/aje/kwad012.