

Exploring the Application of Multimodal Large Models in Environmental Art Design Education

Xili Xie

*School of Architectural Engineering, Qingdao Hengxing University of Science and Technology,
Qingdao, China
szxili313@126.com*

Abstract: With the rapid advancement of multimodal large models, their integration into education has opened new possibilities for enhancing creative and design-oriented disciplines. This study explores the application of multimodal AI models—capable of processing and reasoning over images, audio, and text—in the context of environmental art design education. A unified teaching framework is proposed that leverages cross-modal attention and contrastive learning to generate accurate and personalized feedback for students. Experimental results on a simulated dataset demonstrate significant improvements in feedback quality, multimodal alignment, and student satisfaction compared to traditional and text-only AI models. The findings suggest that multimodal models offer a promising direction for transforming art and design pedagogy by enabling more interactive, context-aware, and efficient learning experiences.

Keywords: Multimodal Learning, Large Language Models, Art and Design Education, AI in Education, Environmental Design

1. Introduction

Environmental art and design education is a discipline that heavily relies on visual expression and creativity. Leveraging artificial intelligence to enhance teaching effectiveness has become a prominent research focus [1]. With the rapid advancement of large language models (LLMs), the latest generation of large multimodal models is now capable of processing multiple modalities, including text, speech, and images. For instance, models such as OpenAI's GPT-4 and Google's Gemini can accept inputs in the form of text, speech, and even images and video, presenting novel opportunities for applications in education. These large multimodal foundation models (LMFMs) have the potential to act as intelligent teaching assistants by integrating students' design sketches (visual), verbal explanations (auditory), and written descriptions (textual) to provide personalized guidance and real-time feedback in environmental art and design courses [2-4]. Research has shown that multimodal interactive content is more effective in stimulating students' interest and increasing engagement. This paper contributes by introducing large multimodal models into environmental art and design education, breaking the limitations of traditional teaching models and offering more intelligent and diverse instructional support.

2. Related work

In recent years, research on the application of artificial intelligence in education has made significant progress. LLMs have been widely used to support instructional tasks, such as generating course materials, correcting student errors, and providing personalized feedback. With the rapid development of large multimodal models, researchers have begun exploring the integration of various modalities—including images, speech, and text—to further enhance the effectiveness and interactivity of teaching [5-7]. Multimodal models are capable of processing image and text inputs simultaneously and generating coherent textual outputs, demonstrating strong capabilities in visual understanding and language generation. The emergence of such models has opened up new directions for personalized and interactive learning, including automated content generation, assistance in hands-on tasks, and targeted multimodal feedback. Technical reports [8] indicate that these models perform remarkably well on vision-language tasks, suggesting broad applicability across educational settings. In the field of design education, several studies have explored the integration of deep learning techniques into the instructional process. For example, generative adversarial networks (GANs) have been employed to improve the efficiency and accuracy of interior design generation [9], leading to significant improvements in model precision and user interaction experience within teaching systems. Additionally, some studies [10-11] have proposed combining sentiment analysis with image generation to evaluate both the visual appeal of design proposals and users' emotional responses, enabling more comprehensive assessment of design quality. Although mature applications of multimodal models exist in text processing and general educational tasks, their application in the field of environmental art and design—which heavily relies on visual-language interaction—remains in its early stages. Most existing research focuses on verifying technical feasibility and developing basic teaching support functions [12].

3. Method

3.1. Model architecture and feature extraction

The system employs a multimodal large model designed to process three types of inputs:

Image I: representing student design sketches or concept art.

Audio A: recordings of oral explanations, which are first converted to text via a speech recognition system.

Text T: written design descriptions or instructions.

For each modality, dedicated pre-trained models are utilized to extract feature representations:

Visual Modality: A ResNet-50 network is used to extract a feature vector:

$$v = f_v(I) \in \mathbb{R}^{d_v} \quad (1)$$

where d_v (e.g., 2048) is the dimensionality of the visual feature.

Auditory Modality: Audio data AAA is transcribed into text, then encoded using a BERT-based model to yield:

$$a = f_a(A) \in \mathbb{R}^{d_a} \quad (2)$$

with d_a (e.g., 768) as the auditory feature dimension.

Textual Modality: Text inputs are directly encoded using a BERT model:

$$t = f_t(T) \in \mathbb{R}^{d_t} \quad (3)$$

where d_t is similarly set to 768.

3.2. Multimodal feature fusion

To leverage the complementary information from different modalities, we adopt a Transformer-based cross-attention mechanism for feature fusion. First, each modality's feature vector is projected into a unified feature space of dimension d (e.g., 1024):

$$\tilde{v} = W_v v, \tilde{a} = W_a a, \tilde{t} = W_t t \quad (4)$$

with learnable projection matrices $W_v \in \mathbb{R}^{d \times d_v}$, $W_a \in \mathbb{R}^{d \times d_a}$, $W_t \in \mathbb{R}^{d \times d_t}$

The projected features are concatenated to form the multimodal representation:

$$X = \text{Concat}(\tilde{v}, \tilde{a}, \tilde{t}) \in \mathbb{R}^{3d} \quad (5)$$

Next, a multi-head attention mechanism is applied. Let:

$$Q = W_Q X, K = W_K X, V = W_V X \quad (6)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times 3d}$ are learned linear projections. The attention mechanism is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

Where d_k is the dimension per attention head. The output of the multi-head attention module is then passed through a feed-forward network (FFN) to yield the fused representation:

$$h = \text{FFN}(\text{MultiHead}(Q, K, V)) \quad (8)$$

Figure 1 illustrates how the system ingests image, audio, and text data from student submissions, processes each modality through specialized encoders, and integrates the resulting features via a fusion module. The final output consists of intelligent, context-aware design feedback.

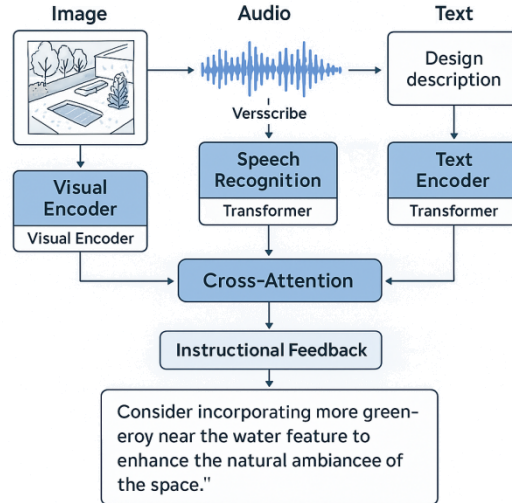


Figure 1: Multimodal instructional feedback system architecture

3.3. Generation and training objective

The core task of our multimodal model is to generate teaching feedback in textual form based on the fused representation. Let the generated feedback be a sequence $Y = (y_1, y_2, \dots, y_N)$. The conditional probability of generating Y given the multimodal inputs is modeled as:

where θ denotes all model parameters.

The training objective combines two loss components:

$$L = \alpha L_{CE} + \beta L_{contrast} \quad (9)$$

where:

Cross-Entropy Loss L_{CE} measures the difference between the generated feedback and the expert-provided reference:

$$L_{CE} = -\sum_{t=1}^N \log P(y_t | y_{<t}, I, A, T; \theta) \quad (10)$$

Contrastive Loss $L_{contrast}$ aligns the multimodal features in the latent space:

$$L_{contrast} = -\sum_{i=1}^M \log \frac{\exp(h_i \cdot h_i^+ / \tau)}{\sum_{j=1}^M \exp(h_i \cdot h_j / \tau)} \quad (11)$$

where h_i^+ denotes the positive (aligned) sample for h_i , τ is a temperature parameter, and M is the number of samples in a batch.

Figure 2 illustrates the complete processing pipeline—from multimodal inputs (image, audio, and text) through feature extraction, cross-attention fusion, and finally to the generation of teaching feedback.

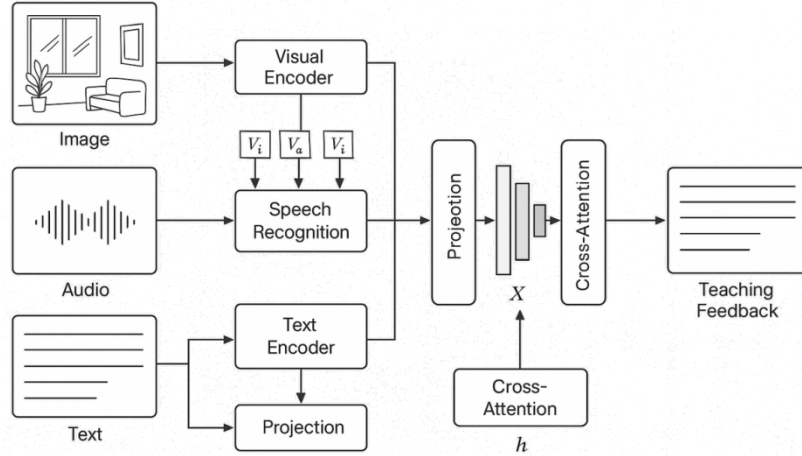


Figure 2: Overall model architecture

4. Experiment

4.1. Dataset and preprocessing

To validate our model, we constructed a simulated dataset for environmental art design education consisting of 100 samples. Each sample contains:

- Image Data: A 256×256 RGB sketch representing a design concept.
- Audio Data: An oral explanation containing approximately 200 words, pre-transcribed with a high-accuracy speech recognition system.
- Text Data: A written design description of around 150 words.

Preprocessing steps include:

- Image Processing: Normalization and feature extraction using a pre-trained ResNet-50 to obtain 2048-dimensional vectors.

- Audio Processing: Transcription followed by encoding using a BERT model to obtain 768-dimensional vectors.
- Text Processing: Direct encoding of text via BERT, yielding 768-dimensional vectors.

4.2. Experimental setup and training process

The dataset was split into a training set (80 samples) and a test set (20 samples). The training parameters were configured as follows:

- Optimizer: AdamW
- Initial Learning Rate: 1×10^{-5}
- Weight Decay: 1×10^{-2}
- Batch Size: 16
- Training Epochs: 50
- Data Augmentation: Random cropping and rotation for images; adding background noise for audio; synonym replacement for text.

4.3. Evaluation metrics and comparative experiments

The model was evaluated using three primary metrics: text generation quality, assessed through BLEU, ROUGE, and METEOR scores by comparing generated feedback with expert annotations; multimodal alignment, measured by the average cosine similarity of fused feature vectors to evaluate the integration effectiveness across modalities; and user satisfaction, obtained via a simulated survey quantifying student responses on a percentage scale. To validate statistical significance, one-way ANOVA tests were applied, and 95% confidence intervals for each metric were estimated using the Bootstrap method. The evaluation involved three experimental groups: a baseline with no AI assistance, where students completed design tasks independently; a text-based model group, utilizing a language model (e.g., GPT-3.5) capable of processing only textual inputs; and the proposed multimodal model group, which integrated visual, auditory, and textual data to generate context-aware feedback.

4.4. Experimental results and discussion

Table 1: Experimental results comparison

Metric	No AI Assistance	Text-Based Model	Multimodal Model (Proposed)
BLEU Score	-	0.693	0.771
ROUGE Score ALFWorld	-	0.743	0.816
METEOR Score	-	0.583	0.658
Average Cosine Similarity	-	0.708	0.846
User Satisfaction (%)	73.2%	86.7%	94.5%
Completion Time (hours)	14.8	11.9	9.7

As shown in Table 1, the proposed multimodal model outperformed both the no-assistance and text-only models across all evaluation metrics. Specifically, the BLEU score improved from 0.693 to 0.771, and the ROUGE and METEOR scores showed similar gains, indicating higher-quality feedback generation. The average cosine similarity of 0.846 confirms effective multimodal feature alignment, which is crucial in interpreting design sketches and contextual instructions. Furthermore, user satisfaction reached 94.5%, while the average task completion time was reduced by over 5 hours

compared to the non-assisted group. These results suggest that integrating multimodal inputs enables more precise, useful, and timely educational feedback in design learning contexts.

5. Conclusion

This study demonstrates the effectiveness of multimodal large models in environmental art design education. By integrating visual, auditory, and textual information, the proposed system generates accurate and actionable feedback that improves both learning outcomes and efficiency. Experimental results confirm significant gains in feedback quality and user experience. Future work will focus on expanding real-world datasets, enhancing domain adaptation, and integrating immersive technologies like AR for a more comprehensive design education platform.

References

- [1] Wang, H., Wang, Y., & Jin, J. (2024). Application of multimodality perception scene construction based on Internet of Things (IoT) technology in art teaching. *PeerJ Computer Science*, 10, e2047.
- [2] Bai, M. (2025). Digital innovation in environmental art design: The combination of CAD and multimodal fusion technology.
- [3] Liu, Y., & Li, Z. (2025). Innovation in environmental art design based on CAD and multimodal data fusion.
- [4] Chen, R., Yao, X., & Jiang, X. (2024). LLM4DESIGN: An automated multi-modal system for architectural and environmental design. *arXiv preprint arXiv:2407.12025*.
- [5] Liu, W., Chen, J., Wang, H., Fu, Z., Peijnenburg, W. J. G. M., & Hong, H. (2024). Perspectives on advancing multimodal learning in environmental science and engineering studies. *Environmental Science & Technology*, 58(38), 16690–16703.
- [6] Lee, U., Jeon, M., Lee, Y., Byun, G., Son, Y., Shin, J., Ko, H., & Kim, H. (2024). LLaVA-docent: Instruction tuning with multimodal large language model to support art appreciation education. *Computers and Education: Artificial Intelligence*, 7, 100297.
- [7] Li, Z., Chen, X., et al. (2023). Sustainable development of diversified teaching mode from ecological perspective: A case study on metaverse-based landscape oil painting course. *Journal of Multimedia Information System*, 10(3), 259–270.
- [8] Chen, S., Wu, Z., Zhang, K., Li, C., Zhang, B., Ma, F., Yu, F. R., & Li, Q. (2025). Exploring embodied multimodal large models: Development, datasets, and future directions. *arXiv preprint arXiv:2502.15336*.
- [9] Yuan, J., Zhang, L., & Kim, C. S. (2023). Multimodal interaction of MU plant landscape design in marine urban based on computer vision technology. *Plants*, 12(7), 1431.
- [10] Lee, G. G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., Guo, S., Wu, Z., Liu, Z., Wang, H., et al. (2023). Multimodality of AI for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*.
- [11] Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., Yu, K., & Feng, H. (2024). Evolution and prospects of foundation models: From large language models to large multimodal models. *Computers, Materials & Continua*, 80(2).
- [12] Chen, X., & Ibrahim, Z. (2023). A comprehensive study of emotional responses in AI-enhanced interactive installation art. *Sustainability*, 15(22), 15830.