





## Computer-Aided Multimodal Fusion in Product Creative Design

Yan Yang<sup>1</sup>  and Huan He<sup>2</sup> 

<sup>1,2</sup> School of Art And Science, Chengdu College of University of Electronic Science And Technology of China, Chengdu 611731, China, <sup>1</sup>[yangyan@cduetec.edu.cn](mailto:yangyan@cduetec.edu.cn), <sup>2</sup>[hehuan@cduetec.edu.cn](mailto:hehuan@cduetec.edu.cn)

Corresponding author: Yan Yang, [yangyan@cduetec.edu.cn](mailto:yangyan@cduetec.edu.cn)

**Abstract.** Product innovation design can not only optimize the function of the product itself but also bring a better user experience to the user. Traditional product innovation design has strong subjectivity and great dependence on the designer's experience, which limits the design. Therefore, this paper builds a multi-modal fusion model based on the CLIP model and obtains design image and text data features through GCN and Transformer to provide more design element data for product innovation design. The experimental results show that the model has better multi-modal fusion performance than other models, which can effectively improve all aspects of the model performance and provide more accurate and multi-dimensional basic data for the model application. Traditional product innovation design is subjective and dependent on the designer's experience, which limits the design. Therefore, this paper builds a multi-modal fusion model based on the CLIP model and obtains design image and text data features through GCN and Transformer to provide more design element data for product innovation design. The experimental results show that the model has better multimodal fusion performance than other models, which can effectively improve the performance of all aspects of the model and provide more accurate and multi-dimensional basic data for the model application.

**Keywords:** Computer Aided; Multimodal Data; Product Ideas; Clip Model; GCN; Transformer

**DOI:** <https://doi.org/10.14733/cadaps.2025.S3.66-77>

### 1 INTRODUCTION

With the acceleration of the pace of economic transformation and the improvement of residents' demand for quality of life, the market's demand and requirements for products are no longer confined to the basic product function level but pay more attention to the personalized, emotional and cultural connotation of products, and have increasingly high requirements for product quality, design, and brand. This change in demand has prompted the rapid development of the product creative design industry and opened an economic era to enhance the added value of products. Product creative design is an important means to change the added value of products, which can organically combine different cultural elements, design concepts, materials, etc., to present different product functions for consumers, improve consumers' sense of experience in the use process, and meet consumers'

personalized, diversified and quality needs. In the past, the product creative design mode was based on the market feedback information and the designer was core to realizing the product design, but many of the results of designing products for the market could not reach the expected effect [1]. The market feedback information in the previous design model is not dynamic and has a certain lag, and it cannot extract the elements and characteristics of the demand and preference of the market and consumers, resulting in inaccurate product positioning or failure to meet the market demand, and unable to provide reliable data basis for design decisions. At the same time, in many product designs, designers modify traditional cultural elements at will in order to meet market demand, resulting in damage and ambiguity to traditional culture [2]. Or simply implant traditional cultural elements into products, ignoring the inherent essence and spirit of traditional culture, lacking unique innovation and uniqueness, resulting in serious product homogeneity and failing to attract consumers' attention.

The rapid development of artificial intelligence (AI) and computer-aided design (CAD) technology. The introduction of AI enables the design process to be based on advanced technologies such as big data analysis and machine learning, predicting market trends, user needs, and behavioural patterns, providing a scientific basis for product innovation design [3]. Inspire students' innovative thinking and enable them to design products that are more in line with market demand and forward-looking. CAD technology, with its powerful modelling, rendering, and simulation capabilities, greatly improves the efficiency and accuracy of product design. At the same time, guide students to explore the deep integration of CAD and AI applications, such as intelligent parametric design, automated optimization design, etc., to achieve rapid verification and optimization of design iterations [4]. Not only has it profoundly changed the way visual communication design is created and expressed, but it has also greatly promoted innovation in the field of product innovation design, bringing unprecedented opportunities for changes in educational concepts, methods, and means. AI and CAD technology provide a powerful platform for interdisciplinary collaboration, enabling seamless integration of knowledge from different fields and jointly driving product innovation. In product innovation design education, students' proficiency in operating CAD software should be strengthened. Product innovation design often involves knowledge from multiple disciplines such as mechanics, electronics, materials, and art [5].

With the comprehensive penetration of digitalization, the pace of innovation in the field of product design has significantly accelerated. However, in the conceptual stage of product innovation design, visual representation is still relatively conservative. Although immersive 3D technology, especially virtual reality (VR), has achieved significant success in entertainment, education, and other fields with its powerful immersion and intuitive operability, it has not been widely applied in the conceptual representation of product innovation design. The lack of unified quantitative indicators to objectively evaluate the application effects of different VR systems in product innovation design makes it difficult for designers to make the best choices. Some scholars have classified VR as new methods applicable to product innovation design, with a particular focus on how to enhance the flexibility of geometric representation, improve the naturalness of interaction, and explore how to optimize the VR experience by combining ergonomic principles [6]. And attempt to propose a new method classification and framework for representing product innovation design concepts through the analysis of the latest technology. The lack of tight integration between geometric representations (such as parametric models or polygonal meshes) and interaction methods in current VR systems has led to limitations for designers when exploring design spaces. Highly dependent on traditional tools such as pencils and paper, or their digital alternatives for operation on a two-dimensional interface. The experimental results show that although VR systems bring users a novel and attractive design experience, they are not significantly better than traditional 2D sketching tools in actual product innovation design concept representation [7]. Subsequently, a series of experimental sessions were designed, inviting users from different design backgrounds to participate. By comparing traditional 2D sketching with two VR-based 3D sketching and carving tools, their effectiveness in representing product innovation design concepts was evaluated. Although VR environments can provide a high degree of immersion, ergonomic issues such as device weight and comfort, as well as spatial fidelity (i.e. accurate reflection of the virtual environment on the real world), become obstacles during long-term use. In addition, long-term use of VR systems has also caused physical fatigue, which is a

negative effect that traditional 2D sketching tools do not possess. Although the sample size limits statistical significance, these findings still provide valuable insights into the application prospects of VR in product innovation design [8].

In the process of product creative design, traditional design modes mainly focus on a single or limited number of design means and methods, which are often based on the designer's personal experience, professional skills and intuitive judgment, and cannot meet the diversified design needs. With the development of computer-aided technology and multi-modal fusion technology, multi-modal fusion of product creative design has become an important development trend. Multi-modal fusion technology refers to a technical method that combines data from different modes to improve information processing and understanding [9]. In the field of product creative design, this means that users' design intentions and preferences can be more fully understood by integrating information from multiple sensors and modes, thus assisting or enhancing the process of product creative design. Therefore, computer-aided multimodal fusion technology can effectively help designers improve design efficiency, optimize user experience, and iterate products on the basis of user feedback and real-time product use information. Therefore, this paper combines the CLIP model and multi-modal data to build a product creative design model, uses deep learning models such as graph Convolutional neural network (GCN) and Transformer to encode images and texts into vectors of the same dimension, and realizes vector fusion through multi-modal fusion module. To provide more accurate, diversified and comprehensive design decision data support for product creative design. The innovations of this paper are as follows:

First of all, this paper realizes the fusion of picture data information and text data information through the CLIP model, enriching the designer's innovative design elements and inspiration sources, and comprehensively improving the intelligent level of product innovative design.

Secondly, GCN in the model can not only analyze and extract user preferences but also provide designers with decision-making direction for design. It can also help designers analyze the development trend of the product market and obtain the elements that affect the development of design.

Finally, the Transformer model enriches the information features of innovative product design, such as text and audio, improves the efficiency of data processing, enables rapid response to market changes and user needs during product design, and accelerates product iteration and upgrade.

## 2 RELATED WORK

In the development process of product innovation design, computer algorithm technology plays a crucial role. These technologies not only improve design efficiency but also break through the boundaries of design innovation. In the early stages of product innovation design development, Pelliccia et al. [10] established corresponding product models through CAD software modelling algorithms, ensuring the accuracy of the models and improving the editability of the design models. In order to better present product design effects, Rapp et al. [11] combined 3D printing technology with CAD technology, creating 3D models through CAD software and then converting these models into instructions that can be recognized by 3D printers, thereby achieving precise manufacturing of products. The advantage of this method is that it reduces the requirements for product form and can maximize the designer's design effect. However, at the same time, this product design method requires high accuracy for 3D models and incurs significant time and labour costs. In order to better design products, Sarkon et al. [12] used algorithm-driven design to analyze user input information (such as industry, style preferences, etc.) and automatically recommend suitable templates. After selecting the template, the algorithm will further optimize content layout, colour matching, font selection, etc. to ensure design consistency and aesthetics. Some tools also use A/B testing algorithms to evaluate the effectiveness of different design schemes and select the optimal solution. Algorithm-driven design not only improves design efficiency but also makes the design more personalized and intelligent. At the same time, it also lowers the design threshold, allowing more people to participate in design innovation. With the development of artificial intelligence technology,

designers have introduced artificial intelligence algorithms in the design process to analyze user data (such as browsing history, purchase history, etc.), predict user interests, preferences, and needs, and provide data support for product design. In the product design process, artificial intelligence algorithms can also be used to optimize design solutions, such as evaluating the effectiveness of different design solutions through simulation experiments, in order to select the optimal solution.

The wave of Industry 4.0 is profoundly reshaping product design and the thinking process behind it, especially in the field of product innovation design. However, while pursuing high intelligence, Williams et al. [13] realized that the current technological system is still insufficient to support participatory design, an innovative model that emphasizes user-centricity and multi-party collaboration. The Human Machine Collaboration (HRC) environment greatly accelerates the transition from concept to product, significantly improving time efficiency and cost control capabilities. The rise of digital technology, especially advanced simulation techniques, enables us to simulate complex systems with unprecedented accuracy and efficiency. It focuses on exploring how to combine the most advanced 3D factory simulation software with participatory methods for product innovation design, aiming to develop new software tools that can support computer-aided participatory design conferences. Although existing computer-aided design (CAD) and simulation software are powerful, they often fail to fully integrate the core concepts of participatory design, limiting their effective application in product innovation design seminars. The participatory design emphasizes the direct incorporation of opinions and feedback from end-users, stakeholders, and multidisciplinary teams in the design process to promote the comprehensiveness and practicality of product innovation. Wang et al. [14] not only focus on technical compatibility and scalability, but also aim to understand and meet the unique requirements for intuitiveness, interactivity, flexibility, and data integration in participatory design processes.

With the widespread application of multimodal data, Yoo et al. [15] applied multimodal data to product innovation design, providing more comprehensive and in-depth insights and innovation space for product design. Some designers have introduced multimodal data in smart home design to improve user experience and intelligence level. For example, by integrating voice, text, and visual data, smart speakers can achieve functions such as voice interaction, smart home control, and personalized recommendations. Designers are also applying multimodal data fusion technology in the design of virtual reality (VR) and augmented reality (AR) products, providing users with a more immersive and interactive experience. For example, VR games integrate visual, auditory, and tactile data, making players feel as if they are in the game world. At present, the application of multimodal data in product innovation design has achieved significant results. By integrating data from different modes, product design can be more closely aligned with user needs and market trends, improving product intelligence and user experience levels. Meanwhile, with the continuous advancement of technology and the expansion of application scenarios, the application prospects of multimodal data in product innovation design will be even broader.

### **3 PRODUCT CREATIVE DESIGN MODEL**

#### **3.1 The Technique and Function of CLIP Model Multimodal Fusion**

CLIP model is a multimodal fusion technology based on contrast learning, whose core components include an image encoder, text encoder, and multimodal fusion module. The image encoder is used to encode the image as a vector, the text encoder is used to encode the text as a vector, and the multimodal fusion module is responsible for fusing the two vectors to calculate the similarity between them. CLIP model is pre-trained by contrast learning. During training, the model is asked to map image and text embeddings from the same sample to similar locations, as well as embeddings from different samples to distant locations. This learning mode enables the model to learn the common features between the image and the text so as to understand and describe the image content. A key innovation of the CLIP model is that both images and text are mapped into the same vector space, which allows the model to calculate the similarity between images and text directly in the vector

space. This ability to align across modes provides a rich source of information and flexible data processing for innovative product design. Figure 1 shows the basic framework of the CLIP model.

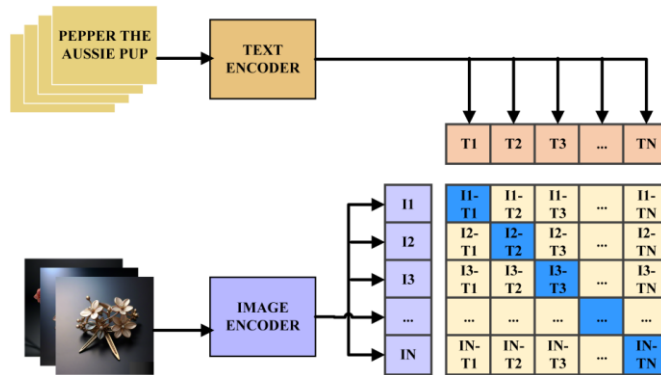


Figure 1: Basic framework of CLIP model.

In the process of product innovation design, CLIP model fusion enables the model to understand and process the image and text information at the same time and integrate the image and text information more efficiently. Designers can enter relevant text descriptions, and CLIP models are able to quickly find images that match them, saving a lot of searching and filtering time. At the same time, the model can also automatically generate relevant text descriptions based on the image content, providing designers with inspiration and reference. The CLIP model is able to process image and text data from different fields and styles, which helps to expand the idea of innovative product design. In addition, the model can also make intelligent recommendations according to user preferences and market trends, providing designers with more accurate design suggestions. The contrast learning mechanism of the CLIP model enables it to discover potential associations and novel combinations between images and text. This ability helps to stimulate the creativity of designers and promote the continuous development of innovative product design. Designers can use the CLIP model for creative collision and inspiration, resulting in more unique and competitive design solutions.

### 3.2 Image Processing Module Based on Graph Convolutional Network

Compared with traditional discrete convolution, GCN has translation invariance, which can realize convolution operation on topological data and obtain important spatial features. Product innovation design includes a lot of design nodes, different nodes will have a certain interaction, and GCN can capture the complex relationship between nodes and extract useful feature information. It can accelerate the design iteration process by analyzing a large amount of historical design data and user feedback to provide design recommendations and optimization directions. In addition, GCN can improve the overall quality of the product by avoiding potential problems during the design phase based on the predictive performance of the product design solution.

Let the topology data adjacency matrix be described as  $G$  node features are expressed as  $H$  The adjacency matrix after adding the connected edges is described as  $\hat{G}$  The node characteristics are shown in (1):

$$H^{(l+1)} = \sigma(J^{-\frac{1}{2}}G\hat{J}^{-\frac{1}{2}}H^lW^l) \tag{1}$$

Where the balance ratio after matrix normalization is described as  $J^{-\frac{1}{2}}G\hat{J}^{-\frac{1}{2}}$ ,  $\hat{J}$  Be matrix  $\hat{G}$  The weight matrix is expressed as  $W$ .

$$Y = f(H,G) = \text{softmax}(G \cdot \text{ReLU}(\hat{G}HW^0)W^1) \tag{2}$$

Where, the normalized self-joining adjacency matrix is described as  $\hat{G}$ , GCN The learnable weight parameters with sequence numbers one and two are represented as  $W^0, W^1$ , and the activation function is expressed as  $ReLU(\cdot)$  The final output is the probability that each node belongs to a class.

When the graph tasks processed in the training stage and the testing stage of the neural network model are different, GCN cannot process the direct inference task. Therefore, this paper introduces the attention mechanism in GCN and assumes that the input node features are represented as  $H = [\vec{h}_1, \vec{h}_2, \dots, \vec{h}_i]$ . The calculation of the attention coefficient of each node and adjacent node self-test is shown in (3):

$$c_{nm} = \vec{a}([W\vec{h}_n][W\vec{h}_m]) \tag{3}$$

Where the learnable weight matrix is expressed as  $W, \vec{a}$ .

Pass  $softmax$  The attention coefficient after normalization of the activation function is shown in (4):

$$\alpha_{nm} = \frac{\exp(LeakyReLU(c_{nm}))}{\sum_{g \in N_n} \exp(LeakyReLU(c_{ng}))} \tag{4}$$

Where the activation function is expressed as  $LeakyReLU(\cdot)$ , the sequence number is  $n$  The set of neighboring nodes is described as  $N_n$ .

The output characteristics of each node are shown in (5) after summing the feature vector with formaldehyde:

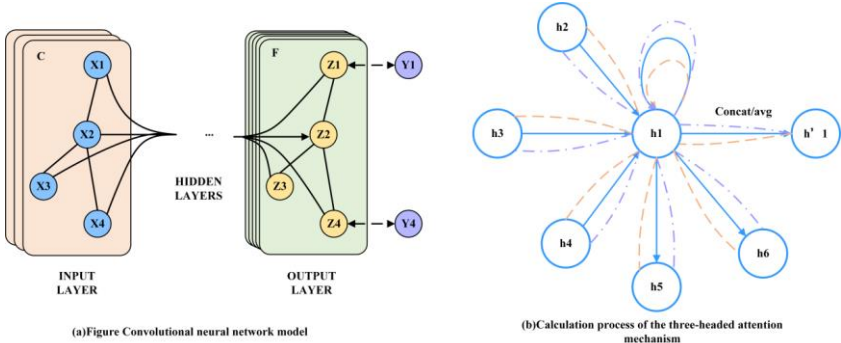
$$\vec{h}_n = \sigma(\sum_{m \in N_n} \alpha_{nm} W\vec{h}_m) \tag{5}$$

In the formula, the new node features obtained after the fusion of domain features are described as  $\vec{h}_n$  Can be learned as  $W$ .

The multi-head attention mechanism is shown in formula (6):

$$\vec{h}'_n(I) = \parallel \sigma(\sum_{m \in N_n} \alpha_{nm}^i W^i \vec{h}_m) \tag{6}$$

Where the number of heads of the multi-head attention mechanism is expressed as  $I$ . Figure 2 shows the schematic diagram of the GCN model and the calculation process of the three-head attention mechanism.



**Figure 2:** Schematic diagram of GCN model and calculation process of three-head attention mechanism.



In this paper, the performance of the GCN model was adjusted through training according to actual requirements. In order to verify the feature extraction performance of the GCN model after adjustment, the accuracy curves of two graph feature extraction models were compared with that of this model in the verification set, and the results are shown in Figure 3. In addition to the three graph feature extraction models, the benchmark model SNE is also included in the comparison experiment, and the performance of the other three models is analyzed based on its accuracy curve. The results show that the accuracy of the three graph feature extraction models is rapidly improved in the early stage and gradually tends to be stable in the later stage, and there is a certain degree of fluctuation in the whole process. Compared with the other two models, the accuracy of the proposed model is significantly higher than that of the benchmark model, and the accuracy of the other two models is higher than that of the benchmark model only in some states. This shows that the GCN model after adjustment has a good feature extraction model, which can provide reliable basic data for multi-modal fusion.

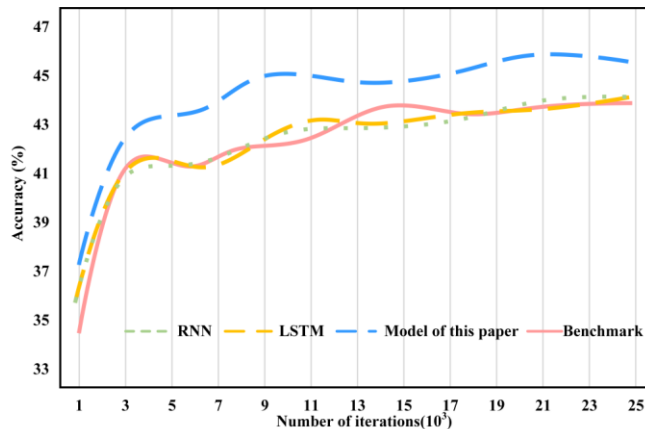


Figure 3: Comparison results of feature extraction accuracy of the three models.

### 3.3 Text Processing Module and Multimodal Fusion Based on Transformer

The Transformer model is an attention-mechanism-based sequential model with a unique and efficient structure that can perform well in a variety of natural language processing tasks. In innovative product design, the Transformer model captures complex dependencies between elements in the input sequence, regardless of their distance in the sequence. This capability enables the model to deeply understand the contextual information of the sequence data, which is particularly important for product innovation design where the internal logic and relationships of the sequence need to be understood. The Transformer model adopts an encoder-decoder structure, where the encoder is stacked with multiple identical layers, each layer contains a multi-head self-attention mechanism and fully connected feedforward network, and residuals and layer normalization connections are used between the sub-layers to map the input sequence into continuous representation. Each layer of the decoder contains three sub-layers, namely a multi-head self-attention mechanism, encoder-decoder attention mechanism, and a fully connected feedforward network, which gradually generates an output sequence based on the output of the encoder.

A core part of the Transformer model is the self-attention mechanism, which allows the model to process each element in the input sequence by comparing it to other elements in order to process each element correctly in different contexts, as shown in (7):

$$attention(Q, K, V) = \text{soft max}(QK^T / \sqrt{d_k})V \quad (7)$$

The word vector input in the self-attention mechanism will have a query vector, a key vector and a value vector, and their corresponding matrices are respectively represented in the formula  $Q, K, V$ . Representing the key vector dimension.

The key link of multimodal fusion is to align the data of different modes. In this paper, the attention mechanism will be used to group the features of different modes in pairs and take turns as query vector, key vector and value vector. In the graph and text processing module, this paper adopts a multi-head attention mechanism to increase the feature representation of different modes. In the process of fusion, this paper adopts a key-value pair attention mechanism. Set  $(K, U) = [(k_1, u_1), (k_2, u_2), \dots, (k_i, u_i)]$ . It represents the format of the key-value input vector, and the query vector is described as  $q$  when The attention function is shown in formula (8):

$$atta((K, U), q) = \sum_{i=1}^N \frac{\exp(s(k_i, q))}{\sum_j \exp(s(k_j, q))} u_i \quad (8)$$

Formula,  $s(k_i, q)$  Represents the attention-scoring function.

Let the input sample description be  $X = [x_1, x_2, \dots, x_i]$  There are four common attention scoring functions, namely additive model, dot product model, scaled dot product model and bilinear model, whose formulas are shown in (9) - (12):

$$s(x, q) = b^T \tanh(Wx + Dq) \quad (9)$$

$$s(x, q) = x^T q \quad (10)$$

$$s(x, q) = \frac{x^T q}{\sqrt{L}} \quad (11)$$

$$s(x, q) = x^T Wq \quad (12)$$

Where the dimension of the input vector is expressed as  $L$  can learn the parameter expressed as  $W, D, b$ .

## 4 EXPERIMENT RESULTS

### 4.1 Experimental Results of the Multimodal Fusion Model

In order to test the performance of the multi-modal fusion model based on the CLIP model, this paper selected three commonly used multi-modal fusion models for comparative experiments and evaluated the model performance indicators. The results are shown in Figure 4. The results show that the accuracy and accuracy of the MVAE model are the lowest, reaching more than 70%, while other models are reaching more than 80%, among which the accuracy of this model is the highest, reaching more than 90%. In terms of recall rate, CAFE and this model reach more than 80%, and the recall rate of this model is significantly higher than that of the CAFE model. In terms of F1 value, the model in this paper has the best performance compared with the other three models. Comprehensive analysis shows that the evaluation indexes of the model in this paper are higher than those of the other three models, which indicates that the model has better multimodal fusion performance and can integrate graph features and text features to provide better data information for product innovative design and realize design quality optimization.

In order to test the effectiveness of different modules of the multi-modal fusion model based on the CLIP model further, this paper introduces variants to different modules of the model for comparison. In the feature extraction part of the figure, this paper introduces HOG, LSTM, and CNN as variants, as well as TF-IDF and Topic Model as variants in text feature extraction. The results are shown in Figure 5. The results in the figure show that HOG, as a traditional graph feature extraction



model, has a high sensitivity to noise and a high dimension, so the two evaluation indexes of the model are low.

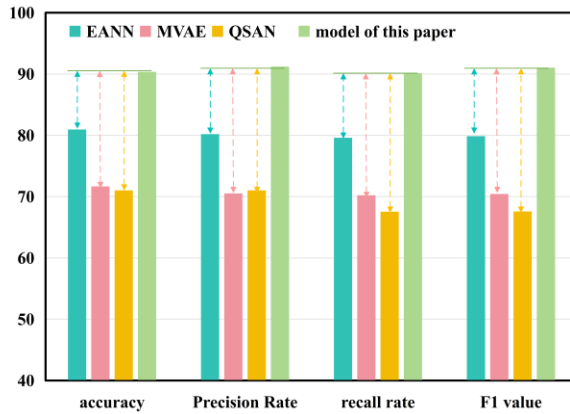


Figure 4: Comparison of evaluation index results of four multimodal fusion models.

The LSTM model and CNN model have similar values, and both of them can effectively extract graph features. However, in product innovation design, graph data has high complexity, and their ability to process the correlation between the feature data is relatively weak. The numerical performance of the evaluation index of the model in this paper is the best, which indicates that it can significantly improve the feature extraction performance of the model graph and enhance the fusion of multi-modal data. In terms of text feature extraction, In order to further test the validity of different modules in the multi-modal fusion model based on the CLIP model, variables are introduced in different modules of the model for comparison. In the feature extraction part of the figure, this paper introduces HOG, LSTM, and CNN as variants and TF-IDF and Topic Model as variants of text feature extraction. The result is shown in Figure 5. As can be seen from the figure, HOG, as a traditional graph feature extraction model, has a high sensitivity to noise and high dimension, so the two evaluation indexes of the model are low. The LSTM model and CNN model have similar values, and both can effectively extract the features of the graph. However, in product innovation design, graph data has high complexity, and its ability to deal with correlation between feature data is relatively weak. The numerical performance of the evaluation index of the model in this paper is the best, indicating that it can significantly improve the feature extraction performance of the model graph and enhance the multi-modal data fusion.

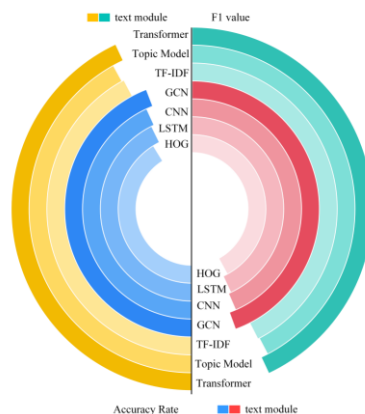


Figure 5: Ablation experiment results of multi-modal fusion model based on CLIP model.

To sum up, multi-modal fusion based on the CLIP model can effectively acquire multi-modal data features and align them through attention mechanism, as well as obtain the correlation between different data, providing effective and multi-dimensional data information for innovative product design.

#### 4.2 Experimental Results of Multimodal Fusion Model Applied in Product Innovation Design

In order to test the application effect of the multi-modal fusion model in product innovation design, this paper chooses to design related derivative products with flowers as the theme. Firstly, the elements need to be classified and extracted through the multi-modal fusion model. As shown in Table 1, the statistics of visual form elements are extracted from the flower theme resource section.

<i>ASSETS</i>	<i>Concrete resource</i>	<i>Extractable element</i>
Tangible resources	Floral appearance	Form, pattern, colour
	Flower material	Materials, graphics, colors
	Flower functionality	Graphics, text
Intangible resources	Flower culture connotation	Graphics, text
	Flower festival connotation	Graphics, text
	Traditional flower customs	Graphics, text, color

**Table 1:** Statistics of morphological elements extracted by subject resource section.

According to the needs of product designers, consumers, and the market, designers obtain new prompters through the multimodal fusion model based on relevant text data or image data, as shown in Figure 6. It can be seen from Figure (a) that before designing floricular-related designs, designers build corresponding databases through multi-modal fusion models and conduct data processing, and the corresponding data is mapped in space. Designers who input the required image or text data through the model can obtain the corresponding results, as shown in Figure (b), so as to obtain more design elements and inspiration. This indicates that the model in this paper can provide designers with more innovative design elements and inspiration, provide corresponding data for design decisions, and optimize design schemes.



**Figure 6:** Spatial distribution of multi-modal data for design with flowers as the theme.

According to the above data, the related products designed by designers with flowers as the theme are shown in Figure 7. As can be seen from the design results in the figure, the multi-modal fusion model based on the CLIP model can provide designers with more theme elements, help designers open design ideas, optimize design schemes, and form products that coexist with aesthetics and functionality to meet the individual needs of different users.



**Figure 7:** Product innovation design with a flower theme.

## 5 CONCLUSIONS

Product innovation design is an important way to promote product iteration optimization, which can improve user experience. The traditional product innovation design is based on the experience and level of the designer to achieve the design effect, but such a design model has some problems such as information lag and product homogeneity. Therefore, this paper combines computer-aided multi-modal fusion technology to build a product innovation design model, obtains product image and text data features through GCN and Transformer models, and realizes multi-modal data alignment and fusion through the CLIP model, providing a richer data basis for design. Model performance experiment results show that GCN and Transformer models can effectively improve the performance of multimodal fusion models and enhance the alignment of image data and text data. Compared with other multimodal fusion models, this model shows better accuracy and precision and has better fusion performance. Through the application experiment, the model can extract the features of image and text data by combining the existing design-related data, providing designers with more abundant design elements and broadening the design ideas according to the needs of users and the market. At the same time, the model can also obtain new prompts in the mapping space according to the data similarity, providing designers with more inspiration and optimizing the design scheme. At present, the model in this paper is mainly applied to some product innovation designs and lacks certain generalizations. Therefore, multi-modal data modules need to be optimized in future research to enhance the generalization of the model.

*Yan Yang*, <https://orcid.org/0000-0003-0362-0232>

*Huan He*, <https://orcid.org/0009-0007-8396-5543>

## REFERENCES

- [1] Alabdali, M.-A.; Salam, M.-A.: The impact of digital transformation on supply chain procurement for creating competitive advantage: An empirical study, *Sustainability*, 14(19), 2022, 12269. <https://doi.org/10.3390/su141912269>
- [2] Ben, Y.; Cengiz, K.: Research on visual orientation guidance of industrial robot based on CAD model under binocular vision, *Computer-Aided Design and Applications*, 19(S2), 2021, 52-63. <https://doi.org/10.14733/cadaps.2022.S2.52-63>

- [3] Frutiger, J.; Cignitti, S.; Abildskov, J.: Computer-aided molecular product-process design under property uncertainties - A Monte Carlo based optimization strategy, *Computers & Chemical Engineering*, 122(3), 2019, 247-257. <https://doi.org/10.1016/j.compchemeng.2018.08.021>
- [4] Gilal, F.-G.; Zhang, J.; Gilal, R.-G.: Integrating intrinsic motivation into the relationship between product design and brand attachment: a cross-cultural investigation based on self-determination theory, *European Journal of International Management*, 14(1), 2020, 1-27. <https://doi.org/10.1504/EJIM.2020.103800>
- [5] Liu, F.: Fast industrial product design method and its application based on 3D CAD system, *Computer-Aided Design and Applications*, 18(S3), 2020, 118-128. <https://doi.org/10.14733/cadaps.2021.S3.118-128>
- [6] Liu, Q.; Zhang, L.; Liu, L.: OptCAMD: An optimization-based framework and tool for molecular and mixture product design, *Computers & Chemical Engineering*, 124(28), 2019, 285-301. <https://doi.org/10.1016/j.compchemeng.2019.01.006>
- [7] Liu, X.; Yao, R.: Design of visual communication teaching system based on artificial intelligence and CAD technology, *Computer-Aided Design and Applications*, 20(S10), 2023, 90-101. <https://doi.org/10.14733/cadaps.2023.S10.90-101>
- [8] Lorusso, M.; Rossoni, M.; Colombo, G.: Conceptual modeling in product design within virtual reality environments, *Computer-Aided Design and Applications*, 18(2), 2020, 383-398. <https://doi.org/10.14733/cadaps.2021.383-398>
- [9] Lu, W.; Ni, Y.; Cai, Z.: User review data-driven product optimization design method, *Journal of Computer-Aided Design & Computer Graphics*, 34(03), 2022, 482-490. <https://doi.org/10.3724/SP.J.1089.2022.19097>
- [10] Pelliccia, L.; Bojko, M.; Prielipp, R.: Applicability of 3D-factory simulation software for computer-aided participatory design for industrial workplaces and processes, *Procedia CIRP*, 99(1), 2021, 122-126. <https://doi.org/10.1016/j.procir.2021.03.019>
- [11] Rapp, M.; Amrouch, H.; Lin, Y.; Yu, B.; Pan, D.-Z.; Wolf, M.; Henkel, J.: MLCAD: A survey of research in machine learning for CAD keynote paper, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(10), 2021, 3162-3181. <https://doi.org/10.1109/TCAD.2021.3124762>
- [12] Sarkon, G.-K.; Safaei, B.; Kenevisi, M.-S.; Arman, S.; Zeeshan, Q.: State-of-the-art review of machine learning applications in additive manufacturing; from design to manufacturing and property control, *Archives of Computational Methods in Engineering*, 29(7), 2022, 5663-5721. <https://doi.org/10.1007/s11831-022-09786-9>
- [13] Williams, G.; Meisel, N.-A.; Simpson, T.-W.; McComb, C.: Design repository effectiveness for 3D convolutional neural networks: Application to additive manufacturing, *Journal of Mechanical Design*, 141(11), 2019, 111701. <https://doi.org/10.1115/1.4044199>
- [14] Wang, W.; Su, J.; Zhang, X.: Research on product primitives recognition in a computer-aided brand product development system, *Computer-Aided Design and Applications*, 18(6), 2021, 1146-1166. <https://doi.org/10.14733/cadaps.2021.1146-1166>
- [15] Yoo, S.; Lee, S.; Kim, S.; Hwang, K.-H.; Park, J.-H.; Kang, N.: Integrating deep learning into CAD/CAE system: generative design and evaluation of 3D conceptual wheel, *Structural and Multidisciplinary Optimization*, 64(4), 2021, 2725-2747. <https://doi.org/10.1007/s00158-021-02953-9>