Journal of King Saud University Computer and Information Sciences (2025) 37:183
https://doi.org/10.1007/s44443-025-00185-1

ORIGINAL PAPER

Enhancing the efficiency of patent classification: a multimodal classification approach for design patents

Xiaodong Xie[1] · Jie Wu[1] · Mengjia Xiang[2] · Jianting Tang[1] · Yongxiang Sheng[1]

Received: 7 April 2025 / Accepted: 15 July 2025 / Published online: 19 August 2025

Abstract

With the rapid increase in the number of design patent applications, traditional patent classification systems encounter significant challenges in terms of both efficiency and scalability. This paper introduces a multimodal feature fusion approach that aims to improve the classification of design patents and address the growing need for faster and more accurate patent examination processes. By extracting modality-specific features from design patent texts, images, and metadata, a multimodal representation is constructed to optimize the feature representations of each modality. This approach effectively captures the interactions among modalities, thereby increasing the expressive power of the features. Furthermore, an attention mechanism is employed to integrate these multimodal features into a unified representation, facilitating the automatic classification of design patents. The empirical results demonstrate that the proposed method significantly outperforms baseline models, achieving substantial improvements in accuracy, precision, recall, and the F1 score. This study provides an innovative solution for automating patent classification, increasing both the accuracy and efficiency of patent examination in practical applications.

Keywords Multimodal fusion · Design patents · Patent classification · Feature optimization · Attention mechanism

1 Introduction

The rapid advancement of technology in the modern world has been accompanied by a remarkable surge in the number of patents (Fink et al. 2017). This increase in patent filings, which has reached millions, poses significant challenges to traditional patent classification systems (Cohen et al. 2016; Liu et al. 2020). At present, patent examination and classification largely rely on the efforts of examiners, applicants, and inventors, who assign relevant patent classification codes to new patents. However, this manual process is not only time-consuming but also inefficient (Haghighian Roudsari et al. 2022; Luo et al. 2021; Miric et al. 2023). As the number of patent applications continues to grow, improving patent examination and achieving rapid and effective patent classification have become key areas of research in the field of patent analysis and applications (Lee and Hsiang 2020; Trappey et al. 2019).

Patent classification is a critical component of the patent examination process as it is an effective means for organizing, searching, analyzing, and managing vast amounts of patent literature (Tseng et al. 2007). Its efficiency directly impacts the overall performance of intellectual property management systems (Wu et al. 2025). While recent advances have yielded substantial improvements in the automated classification of invention patents, the domain of design patents presents a markedly different set of challenges that remain insufficiently addressed by existing intelligent classification systems. Unlike invention patents, which are centered on detailed technical narratives and functional disclosures, design patents aim to safeguard the visual and ornamental aspects of products, including their appearance, shape, surface patterns, and color. This shift in protective focus necessitates a different representational structure: design patents typically consist of sparse, templated textual content and a set of schematic illustrations that convey the core aesthetic elements. Consequently, traditional text-centric classification approaches fall short in capturing the multimodal semantics inherent in design patents. Accurate

✉ Jie Wu
211110405106@stu.just.edu.cn

[1] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, China

[2] Department of Industrial & Systems Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

classification requires the effective integration of schematic visual representations, limited textual cues, and domain-specific structured knowledge to construct meaningful and context-aware representations tailored to the unique characteristics of design patent data (Jiang et al. 2022).

Previous studies have explored the application of machine learning methods in patent classification, where patent features are manually constructed and then fed into machine learning models to determine patent labels (Caldas and Soibelman 2003). While these methods are relatively simple and practical, they are limited in that they cannot capture the deep semantic information between patent texts or effectively handle the current large-scale automatic patent classification tasks. Furthermore, considering the unique nature of design patent classification, machine learning methods fail to fully leverage the deep interactive information provided by different modalities, resulting in suboptimal classification performance.

Deep learning-based methods can adaptively extract patent features, enabling better capture of the deep semantic relationships between patent texts. Additionally, these methods significantly outperform traditional machine learning approaches in terms of accuracy and generalizability (Aristodemou and Tietze 2018). While existing research on the application of deep learning methods to the International Patent Classification (IPC) classification of invention patents is relatively mature, current models primarily focus primarily on the textual modality of patents, neglecting other multimodal information, such as image modality and metadata modality (Haghighian Roudsari et al. 2022; Li et al. 2018). Furthermore, although current multimodal deep learning models have demonstrated strong feature learning capabilities and model performance, they are rarely applied in the field of patent analysis. The challenge remains how to transfer advanced multimodal deep learning models to the domain of patent analysis, particularly for the practical application of design patent classification, a problem that requires urgent attention in the field of patent analysis today.

To address these challenges, this paper introduces a novel classification method based on multimodal feature fusion, which integrates textual, image, and metadata features to achieve a more comprehensive classification of design patents. By optimizing the fusion of multimodal data, the model can capture the complex interactions between various data types, thereby improving classification accuracy. This method aims not only to improve the classification performance of design patents but also to increase patent examination efficiency through automation, reducing reliance on manual efforts and increasing processing capacity. The proposed method has significant practical implications for automating patent examination and improving the overall efficiency of intellectual property management systems.

The main contributions of this paper are as follows:

- Development of an intelligent classification method for design patents

This paper introduces a classification method for design patents that integrates textual, image, and metadata features. By combining multimodal features, the method captures complementary information provided by each data type, thereby increasing classification accuracy.

- Domain-specialized multimodal feature extraction

To address the limited expressiveness of design patent documents, we design tailored extraction strategies for each modality, guided by their structural and semantic characteristics. For textual data, domain-relevant keywords are first distilled from Locarno corpora and then embedded in context-aware representations that preserve essential semantic cues. For visual data, we jointly encode local geometric details and global shape structures to retain both fine-grained and holistic design features. For metadata, the applicant's historical distribution across Locarno subclasses is transformed into a normalized vector that functions as a semantic prior. These three specialized pipelines collectively mitigate the semantic sparsity of design patents and significantly enhance classification performance.

- Optimizing design patent classification performance

The proposed approach refines the fusion mechanism of multimodal features, allowing the model to focus more effectively on the most informative parts of each modality, thus improving overall feature representation and classification performance. Given the current lack of publicly available multimodal datasets specifically for design patent classification, particularly Chinese datasets, a design patent classification dataset is constructed. Additionally, comparative experiments demonstrate that the proposed method significantly outperforms baseline models.

- Improving patent examination efficiency

The proposed method offers a promising solution for automated patent classification, reducing reliance on manual examination and increasing the efficiency of patent review systems. This approach holds significant practical value for intellectual property management, particularly in the context of large-scale patent applications.

The remainder of this paper is structured as follows. Section 2 reviews prior research on patent classification and recent advancements in multimodal deep learning. Section 3 details the proposed methodology, including multimodal feature extraction, modality improvement and fusion, and classification design. Section 4 presents empirical evaluation

and discussion, covering dataset description and training configuration, model performance evaluation, and example analysis and discussion. Section 5 concludes the paper with a summary of findings and future directions.

2 Related work

2.1 Research on patent classification using intelligent methods

Patent classification is one of the key steps in patent management, serving as a method to categorize patents based on similar subjects or technical fields (Haghighian Roudsari et al. 2022). Widely used classification systems for invention patents and utility model patents include the IPC and the Cooperative Patent Classification (CPC). In contrast, design patents are classified according to the Locarno Classification (LOC), the international standard for design patents established under the Locarno Agreement (Shalaby and Zadrozny 2019).

Current research on automatic patent classification can be divided into traditional machine learning-based methods and deep learning-based methods. Traditional machine learning approaches involve manually constructing patent features and feeding them into machine learning models to determine patent labels (Cui 2024; Dib et al. 2024). Methods such as k-nearest-neighbor (Fall et al. 2003), support vector machines (D'hondt, et al., 2013); Fall et al. 2003; Wu et al. 2010), naive bayes (D'hondt, et al., 2013); Fall et al. 2003), K-means (Kim et al. 2008), and artificial neural networks (Trappey et al. 2006) have all been applied to patent classification tasks. Some scholars have combined expert knowledge with machine learning methods for patent classification. For example, Wu et al. (Wu et al. 2010) integrated expert selection methods with a genetic-based hybrid support vector machine model to build a patent classification model. Although traditional machine learning methods are simple and practical, they are unable to capture the deep semantic information between patent texts and are not well suited for handling today's large-scale automatic patent classification tasks.

Deep learning-based methods can adaptively extract patent features and better capture the deep semantic relationships between patent texts. Additionally, they significantly outperform traditional machine learning methods in terms of accuracy and generalizability. In existing research, the use of deep learning methods for IPC classification of invention patents is relatively well established. For example, Li et al. (Li et al. 2018) proposed a deep learning patent classification method, DeepPatent, which combines convolutional neural networks and word embedding techniques. This method outperforms all existing algorithms trained with the same information. Haghighian et al. (Haghighian Roudsari et al. 2022) introduced a deep learning patent classification algorithm based on a convolutional neural network (CNN) and word embedding vectors. Bekamiri et al. (Bekamiri et al. 2024) proposed a hybrid patent classification method that combines a sentence transformer model with traditional classification methods to analyze query patents and perform classification.

2.2 Multimodal deep learning models

The main objective of multimodal deep learning is to enable computers to extract information from various domains, such as text, images, video, and speech, and to effectively fuse multimodal information to improve model performance (Xu et al. 2023a). With the maturation of deep learning technologies, multimodal deep learning has been widely applied in tasks such as rumor detection (Li et al. 2024; Wu et al. 2021; Yin et al. 2025), sentiment analysis (Das and Singh 2023; Liu et al. 2022), and named entity recognition (Li et al. 2025; Tian et al. 2021), demonstrating excellent performance in these applications.

A persistent challenge is multimodal fusion: aligning, interacting with, and jointly encoding signals that differ in structure, scale, and information density. Early-/intermediate-/late-fusion taxonomies capture the processing stage but obscure the algorithmic diversity of modern deep fusion. Recent surveys therefore advocate functional, mechanism-centric classifications (Zhao et al. 2024). Below, we summarize three representative families that constitute today's technical backbone.

2.2.1 Attention-based fusion

Attention mechanisms, rooted in the transformer framework (Vaswani et al. 2017), have become prominent in current multimodal systems owing to their ability to learn context-adaptive weights across and within modalities (Madaan et al. 2024). Intra-modality self-attention is designed to explicitly model the relationships within a single modality. In this approach, attention operations, such as dot-product attention or additive gate-based attention, consider data from the same modality (Qin et al. 2021; Shul and Choi 2024). Intermodal (cross/co)-attention explicitly aligns modalities, e.g., pairing text tokens with visual regions in vision-language tasks (Li et al. 2019; Tan and Bansal 2019). Bidirectional variants leverage mutual dependencies (Fang et al. 2022; Wu et al. 2023). Stacked self-/cross-attention transformers jointly capture the global context and fine-grained interactions, achieving state-of-the-art results in video captioning, visual question answering (VQA), and image–text retrieval (Xu et al. 2023a). While these models excel at flexible feature

selection, they incur substantial computational overhead and can be sensitive to misaligned or noisy modalities.

2.2.2 Graph neural network-based fusion

Graph neural networks (GNNs) introduce relational inductive biases that are well suited to multimodal data with explicit or latent structures (Ektefaie et al. 2023; Li et al. 2023a; Zhang et al. 2024). Broadly, graph-based fusion methods can be categorized into two principal strategies. (1) Modality-specific graph encoding: Intra-modal relations are first modeled using appropriate graph neural network architectures, and the resulting embeddings are subsequently integrated with other modalities through co-attention or concatenation. This design supports misinformation detection by unifying rumor-propagation graphs with visual cues (Qi et al. 2025; Xu et al. 2023b). (2) Joint multimodal graph construction: Nodes from all modalities populate a heterogeneous graph whose intra-/inter-edges encode semantic or temporal links; fusion proceeds through message passing (Cai et al. 2022). GNNs naturally capture higher-order dependencies and remain robust to sparse or irregular data. Their main drawbacks—graph-building overhead and reliance on domain knowledge—necessitate careful scalability considerations for large-scale deployments.

2.2.3 Other constraint-based fusion

Constraint-oriented and hybrid schemes inject explicit priors or structural objectives to improve robustness. Coordinated representation learning aligns modality-specific embeddings via similarity constraints such as canonical-correlation analysis (CCA) or cosine distance, balancing modularity with cross-modal coherence (Khattar et al. 2019). Tensor-based fusion models high-order interactions through outer products; recent low-rank decompositions curb parameter growth without sacrificing discriminative capacity (Wu et al., 2024). Channel-exchange networks (e.g., CEN++) implicitly integrate features by dynamically swapping channels across unimodal subnetworks, achieving strong multitask performance with negligible parameter overhead (Wang et al. 2022).

Attention-centric models offer fine-grained, data-driven alignment; GNNs embed relational structures; and constraint methods integrate explicit priors and lightweight couplings. The choice between these models depends on task-specific trade-offs in data quality, computational budget, and scalability.

2.3 Research gaps

Despite the growing interest in patent classification, most existing efforts remain centered on invention and utility patents. These models are often built on assumptions that align with invention-patent characteristics: rich textual descriptions and, in some cases, images that support technical claims. However, design patents operate under a fundamentally different classification regime that emphasizes the protection of visual design elements, such as a product's appearance, shape, patterns, and colors, rather than technical functionality.

This divergence introduces several challenges. Design patents typically contain highly templated and semantically sparse text, and their images consist of schematic or ornamental drawings that lack the naturalistic features common in vision models that have been pretrained based on general-purpose datasets. Furthermore, structured metadata such as Locarno codes or applicant-specific filing patterns, which are often available and informative in design patent cases, are frequently neglected by current classification pipelines.

While multimodal deep learning offers a promising avenue for bridging heterogeneous information sources, general-purpose models fail to generalize well to the unique data characteristics and classification demands of design patents. Their reliance on verbose textual input, natural-image assumptions, and omission of domain-specific metadata signals creates a substantial gap when they are applied to appearance-centric patent classification. In this context, developing a domain-adapted multimodal framework tailored specifically to design patent data has become an urgent and underexplored problem in the field of patent analysis.

3 Methodology

In this study, an automatic classification method is constructed for design patents based on textual, image, and metadata modality features. The model architecture is shown in Fig. 1. The framework consists of three steps: (1) multimodal feature extraction, (2) multimodal feature enhancement and fusion, and (3) design patent classification, which are described in Sects. 3.1.1, 3.2, and 3.3, respectively.

3.1 Multimodal feature extraction

3.1.1 Textual Modal Features

Compared with invention patents and utility model patents, which include detailed textual descriptions such as titles, abstracts, claims, and specifications, design patent documents contain relatively limited textual information. The textual features available for patent classification are restricted to the title and abstract. Furthermore, the abstracts of design patents often includes redundant or repetitive content. For example, the abstract of the design patent"Ping Pong Paddle (CN308829903S)"reads:"1. The name of this design

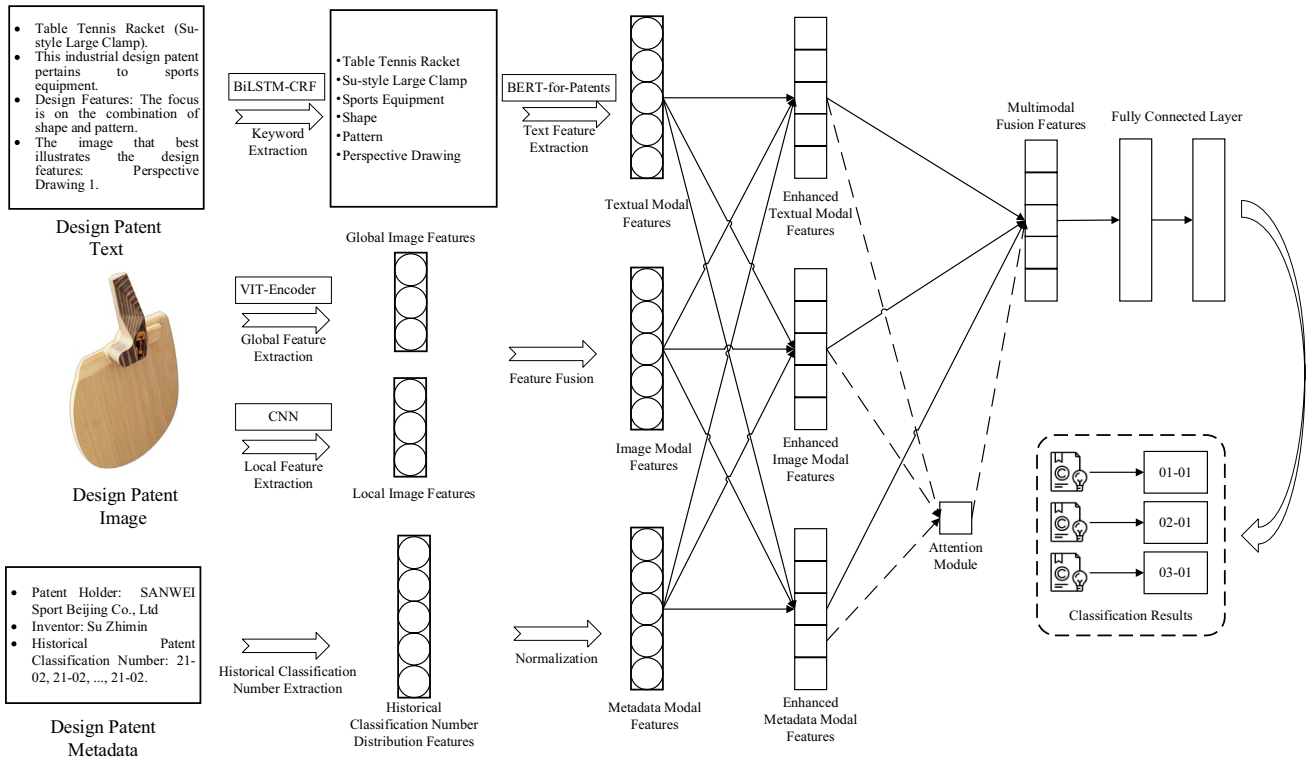

Fig. 1 Model Framework Diagram

product: Ping Pong Paddle. 2. The use of this design product: For striking a ping pong ball. 3. The key design feature of this product: The shape. 4. The image or photo that best represents the key design feature: A 3D view."

Given the limited and often indistinct textual features of design patents, this work addresses this issue by extracting domain-specific keywords from the titles and abstracts of design patents through patent keyword extraction. To construct a classification lexicon for design patents, we use the bidirectional long short-term memory with conditional random fields (BiLSTM-CRF) model in conjunction with relevant corpora such as the"14th Edition of the International Design Classification—Major and Minor Classes Table"and the"14th Edition of the International Design Classification—Product Item List."The BiLSTM-CRF model combines bidirectional long short-term memory (BiLSTM) and conditional random fields (CRF), making it suitable for sequence labeling tasks such as named entity recognition and keyword extraction. Specifically, the BiLSTM model captures context-dependent feature representations from both forward and backward LSTM processes to model long-range dependencies within patent texts, while the CRF considers label dependencies, ensuring that the generated label sequence is globally optimal.

Building on this approach, the pretrained BERT-for-Patents model developed by Google is employed to convert the unstructured textual features of patents into structured features. BERT-for-Patents is a language model specifically optimized for the patent domain that is built on the bidirectional encoder representations from transformers (BERT) architecture. It is particularly effective in processing the complex linguistic characteristics and domain-specific terminology commonly found in patent texts.

Specifically, the keyword sequence $X_{P_i} = (x_1, x_2, \cdots, x_n)$ extracted from the patent P_i is first annotated with the [CLS] and [SEP] tokens to denote the beginning and end of the text, respectively, as shown in Eq. (1):

$$X\prime_{P_i} = [CLS]x_1, x_2, \cdots, x_n[SEP] \tag{1}$$

The processed patent text $X\prime_{P_i}$ is fed into the BERT-for-Patents model, and the [CLS] token from the final hidden layer is extracted. This token is used as the corresponding vector, representing the textual state feature P_i^{text} of the patent P_i.

3.1.2 Image Modal Features

This paper adopts a hybrid method based on convolutional neural networks (CNN) and vision transformers (ViT) to extract the image modal features of design patents, extracting features from both local and global perspectives to obtain a

more comprehensive image representation. The model architecture is shown in Fig. 2.

In design patents, image features such as details, edges, and shapes play a significant role in recognition and classification. Therefore, this paper uses convolutional neural networks (CNNs) to extract local features from the images. The CNN architecture consists of several convolutional layers, pooling layers, and fully connected layers. The convolutional layers extract features by sliding convolutional filters across the image and the pooling layers reduce the data dimensionality and computational cost.

For the specialized image X, the operation through the convolution layer is expressed by formula (2) as follows:

$$f_{i,j} = \sum_{m=-k}^{k} \sum_{n=-k}^{k} W_{m,n} X_{i+m,j+n} + b \tag{2}$$

where $f_{i,j}$ represents the feature value after convolution, $W_{m,n}$ denotes the convolution kernel weights, and b is the bias term. The size of the convolution kernel is denoted by k.

This paper uses a vision transformer (ViT) to extract the global features of images. Compared with traditional convolutional networks, the ViT is better at capturing the relationships between image patches, which helps enhance the understanding of the overall design and shape of the appearance. The ViT model divides the input image into fixed-size patches, each of which is linearly projected into a feature vector. These vectors are then processed by the transformer encoder layers to learn the global dependencies between the patches.

First, the input image X is divided into N image patches, with each patch having a size of $P \times P$, as shown in formula (3).

$$x_i = \text{PatchEmbed}(X) \tag{3}$$

where x_i represents the input of the $i-th$ image patch.

Then, the position embedding (Position Embedding) is added to each image patch's feature representation, as shown in formula (4):

$$z_0 = [x_{\text{cls}}; x_1 + E_1; x_2 + E_2; \ldots; x_N + E_N] \tag{4}$$

where x_{cls} is the classification token, and E_i is the position embedding.

Finally, through multiple transformer encoder layers, the image patch features are encoded to capture global dependencies, as shown in formulas (5) and (6):

$$z_{l+1} = \text{MSA}(\text{LN}(z_l)) + z_l \tag{5}$$

$$z_{l+1}' = \text{MLP}(\text{LN}(z_{l+1})) + z_{l+1} \tag{6}$$

where MSA represents the multihead self-attention layer, MLP represents the multilayer perceptron, and LN is layer normalization.

To achieve a comprehensive understanding of the image, the global and local features extracted from the image of the patent P_i are combined. The fused feature representation is shown in Eq. (7) as follows:

$$P_i^{image} = \eta F_{P_i}^{CNN} + (1-\eta) F_{P_i}^{VIT} \tag{7}$$

where F_{CNN} represents the local features extracted by a convolutional neural network and where F_{ViT} represents the global features extracted by the vision transformer. The parameter η serves as the fusion coefficient, which controls the balance between the global and local feature weights.

3.1.3 Features of metadata modalities

Design patent holders often concentrate their innovations in specific fields (corresponding to the particular Locarno classification codes). Therefore, the historical distribution of Locarno classification codes associated with a patent

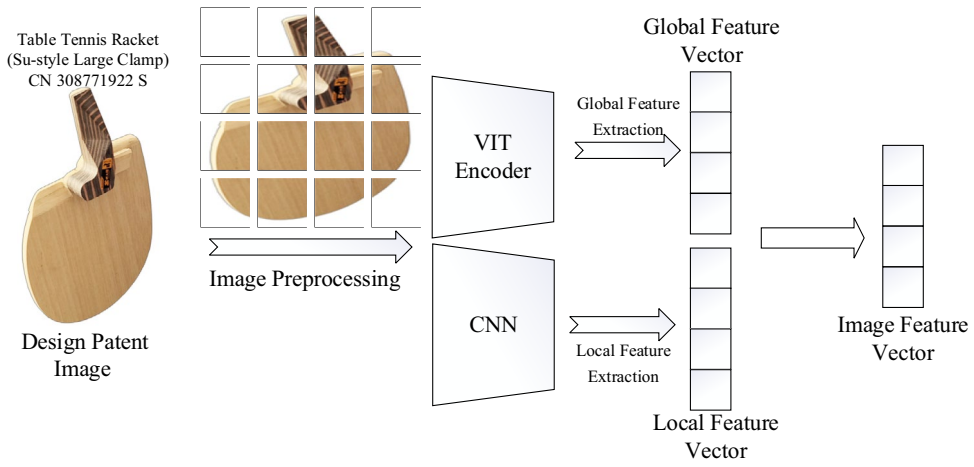

Fig. 2 Image Modality Feature Extraction Process

holder can serve as prior information for the classification of current patents. Incorporating historical distribution data can assist the classification model in making more accurate predictions.

In this work, the classification code distribution is constructed using metadata modality features by obtaining the historical patent distribution of holders of design patents to be recognized. By analyzing the Locarno classification code distribution F_A of patent holder A in previous patent applications, the metadata modality features can be derived, with the process for obtaining F_A shown in Eq. (8) as follows:

$$F_A = \{f_A(L_1), f_A(L_2), ..., f_A(L_n)\} \tag{8}$$

where $f_A(L_i)$ represents the number of patents held by patent holder A in the Locarno classification code L_i.

The number of classification codes is normalized to obtain the distribution frequency of each classification code. The normalization process is shown in Eq. (9) as follows:

$$w_A(L_i) = \frac{f_A(L_i)}{\sum_{j=1}^{n} f_A(L_j)} \tag{9}$$

The final obtained metadata modality features for patent P_i are shown in Eq. (10) as follows:

$$P_i^{meta} = \{\omega_A(L_1), \omega_A(L_2), ..., \omega_A(L_n)\} \tag{10}$$

3.2 Enhancement and fusion of multimodal representations

In the classification task of design patents, text modality features, image modality features, and metadata modality features reflect different aspects of the patent's classification characteristics. Multimodal representations are enhanced to optimize the feature representations of each modality. By capturing the interaction relationships between the modalities, the expression power of the features is strengthened, thereby increasing the accuracy of the classification task. The process of multimodal enhancement is shown in Fig. 3.

For each patent, three modality feature representations (text feature P_i^{text}, image feature P_i^{image}, and metadata feature P_i^{mata}) are extracted from different perspectives. First, the features of each modality undergo linear transformation and dimensionality reduction. Subsequently, the three modality features are fused to form the enhanced model input P_i, as shown in Eq. (11):

$$P_i = \left[P_i^{text}, P_i^{image}, P_i^{mata}\right] \tag{11}$$

Next, the reduced-dimensional representations are used to construct the covariance matrix Φ, which captures the correlations between modalities, as shown in Eq. (12):

$$\Phi = f(P_i) \cdot \bar{I} \cdot f(P_i)^T \tag{12}$$

where $f(P_i)$ is the modality feature matrix after linear dimensionality reduction, and $\bar{I}$ is the correction matrix used to eliminate scale differences and strengthen the off-diagonal elements.

The attention distribution of modality features is computed using the covariance matrix, which captures the importance of each modality to other modalities. The attention scores are normalized using the Softmax function, as expressed by Eq. (13):

$$U_i = softmax\left(\frac{\phi}{\sqrt{d/r}}\right) g(P_i) \tag{13}$$

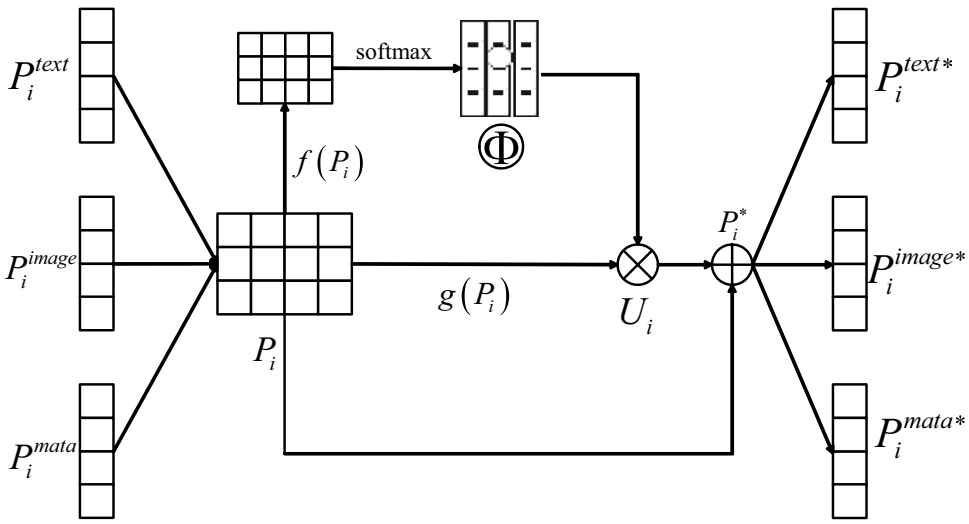

Fig. 3 Modality Enhancement Process

where U_i represents the enhanced modality representation, $g(P_i)$ is the feature obtained after applying a nonlinear mapping to P_i, d is the feature dimension, and r is the scaling factor.

Finally, the enhanced representation is concatenated with the original representation to obtain the enhanced representation for each modality P_i^*, as shown in Eq. (14):

$$P_i^* = U_i \oplus P_i \tag{14}$$

where $\oplus$ denotes the feature concatenation operation.

After enhancing the features of each modality, the features from the various modalities are fused into a unified feature representation. During the fusion of each modality's features, an attention mechanism is used to quantify the weight of each modality's contribution to the final classification. Specifically, the enhanced representation P_i^* for each modality is computed with the attention weight att_m, and the weighted sum is taken to obtain the fused multimodal representation, as shown in Eq. (15):

$$P_i^{fusion} = \sum_{i=1}^{m} att_m \cdot P_i^{m*} \tag{15}$$

where m represents the number of modalities, att_m is the attention weight for the $i-th$ modality, and P_i^{m*} is the enhanced modality representation.

The weight att_i for each modality is computed through the attention mechanism, and the Softmax function is used to normalize the weights, ensuring that the sum of all modality weights is 1. The process for calculating the modality weights is shown in Eq. (16):

$$att_m = W_2^m \cdot tanh\left(W_1^m \cdot P_i^m + b_1^m\right) + b_2^m \tag{16}$$

where W and b represent the weight matrix and bias parameters, respectively, and tanh is the activation function.

3.3 Design patent classification

After obtaining the multimodal feature fusion vector P_i^{fusion} for design patents, a fully connected layer is used for classification, as shown in Eq. (17):

$$\hat{y}_i = \text{softmax}(W_{fc} \cdot P_i^{fusion} + b_{fc}) \tag{17}$$

where $\hat{y}_i$ represents the predicted classification result for the design patent and where W_{fc} and b_{fc} are the weights and biases of the fully connected layer, respectively.

Additionally, binary cross-entropy loss is used to optimize the classification task, with the optimization function shown in Eq. (18):

$$L_{class} = -\sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{18}$$

where L_{class} is the classification loss function, N is the total number of samples, y_i is the true label for sample i, and $\hat{y}_i$ is the predicted value for sample i.

4 Empirical evaluation and discussion

In this section, we present the results of an empirical analysis that was conducted using the constructed design patent dataset. The performance of the proposed model is evaluated through comparative experiments and ablation studies. Additionally, specific case analyses are performed to discuss the results presented in this paper.

4.1 Dataset description and training configuration

4.1.1 Dataset Description

The experimental data in this paper are sourced from the PatSnap patent database. To ensure the balance of the samples, 1000 design patents from each subcategory were randomly selected as the experimental data source for this paper, resulting in a total of 241,000 design patents that were used for subsequent experiments. The data extraction time was set for October 2024. On this basis, the dataset was split into training, validation, and test sets at an 8:1:1 ratio.

The titles, abstracts, summary diagrams, patent owners, and original classification numbers of each design patent document were extracted for additional experiments. The basic information of the patent dataset constructed in this paper is shown in Table 1.

4.1.2 Training Configuration

We implemented all the experiments using PyTorch 2.1 with automatic mixed precision (AMP), running on a Windows 11 platform equipped with an NVIDIA RTX 4060 GPU (8 GB VRAM), Intel Core i7-13650HX CPU, and 32 GB RAM.

For textual encoding, we fine-tuned the BERT-for-Patents model comprising 12 transformer layers, 768 hidden units, and 12 attention heads. The maximum sequence length was set to 128 tokens, which was sufficient to represent the concatenated titles and abstracts of design patents.

The visual representation module integrates both local and global descriptors. Local features were extracted using a lightweight three-layer CNN with progressively increasing channel sizes ($64 \rightarrow 128 \rightarrow 256$), each followed by batch normalization and max pooling, effectively capturing

Table 1 Experimental Data Information

Class Number	Class Heading	Subclass Number	Subclass Heading	Quantity
01	Foodstuffs	01	BAKERS'PRODUCTS, BISCUITS, PASTRY, PASTA AND OTHER CEREAL PRODUCTS, CHOCOLATES, CONFECTIONERY, ICES	1000
		02	FRUIT, VEGETABLES AND PRODUCTS MADE FROM FRUITS AND VEGETABLES	1000
		03	CHEESES, BUTTER AND BUTTER SUBSTITUTES, OTHER DAIRY PRODUCE	1000
		…	…	…

edge- and region-level patterns. Global visual context was modeled using a ViT-Base/16 encoder (12 layers, 768 hidden units, patch size of 16), applied to images resized to 224×224 and normalized using standard ImageNet statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]).

For multimodal fusion, features from both modalities were projected into a 256-dimensional shared space, followed by a four-head cross-modal attention module with gaussian error linear unit (GELU) activation. The final visual embedding was computed as a weighted sum of local and global features, with the balance coefficient $\eta = 0.5$, determined via validation performance.

We employed the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay = 0.01. The learning rate was set to 2×10^{-5} for the text encoder and 1×10^{-4} for the visual and fusion branches. The learning schedule included a 10% linear warm-up followed by cosine decay. Dropout was applied to both the textual and fusion layers with a probability of 0.1.

Training was performed with a batch size of 8 using 16-bit floating-point precision. The model was trained for up to 8 epochs, with early stopping based on validation F1 score (patience = 2). Gradient clipping ($\|g\|_2 \leq 1.0$) was applied to stabilize training, and a fixed random seed (42) was used throughout to ensure reproducibility.

4.2 Model performance evaluation

The performance of the proposed model was evaluated using four metrics: accuracy, precision, recall, and F1 score.

4.2.1 Comparative Study

To validate the performance of the model proposed in this study, comparative study was conducted by selecting five types of baseline models: traditional machine learning models, text modality models, image modality models, patent-specific multimodal models, and general-domain document classification models. The selected models for comparison are listed below, and the results of the comparison are shown in Table 2.

Table 2 Comparison of Experimental Results

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.6745	0.6921	0.6453	0.6679
RF	0.7134	0.7328	0.6805	0.7057
BERT-for-Patents	0.8312	0.8413	0.8205	0.8308
DeepPatent	0.8147	0.8283	0.8001	0.8139
PatentSBERTa	0.8220	0.8302	0.7949	0.8122
VIT + CNN	0.6250	0.6354	0.6138	0.6244
VGG19	0.6045	0.6155	0.5895	0.6022
ResNet50	0.5919	0.6012	0.5815	0.5912
TechDoc	0.8478	0.8581	0.8368	0.8473
PatentLVLM	0.8426	0.8540	0.8204	0.8369
IMPACT	0.8341	0.8447	0.8005	0.8220
VLCDoC	0.8137	0.8248	0.7761	0.7997
PMF	0.8203	0.8320	0.7835	0.8070
GlobalDoc	0.8015	0.8131	0.7610	0.7860
Proposed model	0.8865	0.8982	0.8724	0.8851

(1) Machine learning models: Support vector machine (SVM) and random forest (RF)

In the machine learning model setup, features from patent titles and abstracts are extracted using the BERT-for-Patents model, whereas image modality features are extracted using a pretrained VIT + CNN model. Metadata modality features are obtained by analyzing the Locarno classification distribution of historical patents by patent holder. The feature vectors from the three modalities are then reduced in dimensionality and concatenated into a long vector as the model input. Finally, the performance of the model is evaluated using two machine learning models, SVM and RF, which are based on the concatenated features.

(2) Text modality models: BERT-for-Patents (Thakur et al. 2021), DeepPatent (Li et al. 2018), and PatentSBERTa (Bekamiri et al. 2024)

BERT-for-Patents, a patent text analysis model developed by Google that is based on the BERT Large architecture, is fine-tuned and applied in this study to the downstream task

of appearance-based design patent classification. DeepPatent employs deep learning algorithms by combining CNNs and word embeddings for patent classification. PatentSBERTa generates semantic embeddings of patent texts using the micro SBERT model and performs patent classification in conjunction with the K-nearest neighbors (KNN) algorithm.

(3) Image modality models: VIT + CNN, VGG19, and ResNet50

In this study, the image features of design patents obtained from the VIT + CNN, VGG19, and ResNet50 models are input into a fully connected layer for classification, with the Softmax activation function used to obtain the probability distribution.

(4) Patent-Specific Multimodal Models: TechDoc (Jiang et al. 2022), IMPACT (Shomee et al. 2024) and PatentLVLM (Awale et al. 2025)

TechDoc integrates VGG19-based image encoders and Bi-GRU text embeddings. These features are fused through a hierarchical attention mechanism and classified using a GraphSAGE-based network. IMPACT employs ResNet-50 for image encoding and RoBERTa for textual representation. Inputs comprise patent titles, abstracts and Large Language-and-Vision Assistant (LLaVA)- generated image captions. The modalities are fused through late concatenation, and the unified features are optimized in an end-to-end manner. PatentLVLM incorporates a pretrained ViT-g/14 backbone and RoBERTa textual features and introduces a Q-Former module for cross-modal alignment. Domain adaptation is achieved by further optimizing the Q-Former and task-specific classification parameters under a hybrid contrastive–supervised objective.

(5) General-Domain Document Classification Models: VLCDoC (Bakkali et al. 2023), PMF (Li et al. 2023b) and GlobalDoc (Bakkali et al. 2025)

VLCDoC uses ViT-Base and RoBERTa encoders pretrained with dual-contrastive objectives. Patent-specific adaptation is performed via a linear classification head that operates atop the preserved pretrained representations. PMF utilizes frozen ViT-Base and BERT-Base encoders, with three types of deep-layer prompts—query, query-context, and fusion-context—inserted to facilitate cross-modal alignment. In this work, domain-specific adaptation is performed by fine-tuning the prompts and classification head on multimodal patent data. GlobalDoc employs a cross-modal transformer pretrained via contrastive meta-learning, which encodes rich multimodal representations. In this work, for domain-specific adaptation, a task-oriented classification module is integrated atop the pretrained backbone to facilitate effective patent categorization.

As shown in the comparison results in Table 2, the proposed model significantly outperforms the baseline models across all the metrics. Compared with those of the best-performing baseline model, TechDoc, the proposed model achieves a 4.56% increase in accuracy, a 4.67% increase in precision, a 4.25% increase in recall, and a 4.46% increase in the F1 score, indicating that the proposed model has strong practical effectiveness in the automatic classification of design patents.

Traditional machine learning models have limited performance when handling multimodal patent data. These models primarily rely on feature engineering and traditional classification algorithms, which are suitable for simple, high-dimensional feature spaces but struggle to effectively model complex multimodal data (such as combinations of text, images, and metadata). As a result, their accuracy and precision tend to be lower.

Single-text and single-image modality models each have their own advantages in handling patent text and image features. However, since the illustrations in design patents are often abstract, text modality models outperform image modality models. The multimodal fusion model proposed in this paper effectively leverages the complementary information between text, image, and metadata modalities. Compared with single-modality models, the multimodal fusion model can better understand patent classification features from different perspectives. By optimizing feature fusion between modalities using an attention mechanism, the model improves overall performance.

Among patent-specific multimodal baselines, TechDoc, PatentLVLM, and IMPACT achieve the strongest performance, highlighting the advantages of architectures explicitly tailored to the characteristics of patent data. In contrast, general-purpose systems such as PMF, VLCDoC, and GlobalDoc consistently underperform. This performance gap can be attributed to three main factors: (1) Domain mismatch: Design patents typically combine sparse technical descriptions with highly abstract visual illustrations, which differ substantially from the semantically rich documents used to pretrain general-purpose models. (2) Text sparsity: Patent titles and short claims provide limited semantic context, diminishing the benefit of large language encoders that expect verbose input. (3) The distinctive style and high level of abstraction in design drawings differ markedly from the natural images or scanned documents typically used in general-purpose document classification, thereby constraining the ability of visual encoders to extract discriminative features.

The proposed model addresses these limitations by jointly encoding textual and visual cues through a unified attention-based framework, resulting in richer and more task-relevant

multimodal representations. Overall, the results suggest that while general-purpose models possess a degree of transferability, domain-specific adaptations remain critical for achieving state-of-the-art performance in patent classification.

4.2.2 Ablation Study

To validate the impact of each component of the proposed model on its overall performance and to identify the key driving factors behind its improved performance, ablation experiments were conducted by systematically removing certain components. The experimental setup and results of the ablation study are presented in Table 3.

The results of the ablation experiment presented in Table 3 demonstrate that the three modalities—text, image, and metadata—each contribute to varying degrees to the overall performance of the proposed automatic classification model for design patents, with all the modalities being indispensable.

Among the components, the removal of the text modality leads to the most significant decline in performance across all the metrics. Textual features are essential for understanding the subject and technical details of a patent, playing a pivotal role in identifying the category of design patents. The removal of the image modality results in the second-largest decrease in model performance, highlighting the unique contribution of the image modality. The image modality provides information regarding the product's appearance, shape, and structure, offering complementary classification details that enhance the text-based features.

When the modality enhancement module is removed, the model exhibits a moderate decline in performance. The modality enhancement module optimizes the feature representations of each modality, allowing for a clearer expression of the information from each modality and improving the representational power of the final fused features.

The removal of the metadata modality causes a slight reduction in performance, indicating that the metadata serve a more auxiliary role. While this module provides prior guidance for specific categories, its contribution to overall classification is relatively limited. Direct concatenation of the enhanced modality features has a minimal effect on performance, as the modality enhancement module already considers the interactions between modalities during the feature strengthening process. However, direct concatenation significantly increases the computational complexity of the model.

In summary, the text modality serves as the primary source of information and makes the greatest contribution to patent classification. The image modality provides valuable auxiliary information, whereas the metadata modality, although contributing less overall, offers prior guidance that can improve model performance in certain categories. The modality enhancement and attention mechanisms further optimize the synergy between modalities, resulting in a significant improvement in overall performance for design patent classification.

Table 3 Ablation Experiment Results

Model	Accuracy	Precision	Recall	F1 Score
Remove text modality	0.6410	0.6500	0.6282	0.6390
Remove image modality	0.8246	0.8321	0.8140	0.8230
Remove metadata modality	0.8571	0.8623	0.8507	0.8565
Remove modality enhancement	0.8392	0.8476	0.8315	0.8395
Modality enhancement with direct concatenation	0.8638	0.8697	0.8574	0.8634
Proposed model	0.8865	0.8982	0.8724	0.8851

4.2.3 Sensitivity analysis

To assess the robustness of our model to key hyperparameter settings, we conducted a comprehensive sensitivity analysis on six representative parameters. The results are presented in Fig. 4, which illustrates the F1 score variations under different configurations.

Figure 4 illustrates the F1 score response to individual variations in six key hyperparameters. The model achieves peak performance at a text encoder learning rate of 2×10^{-5}, an image/fusion learning rate of 1×10^{-4}, and a CNN–ViT fusion ratio of 0.5. Notable performance degradation is observed under excessive dropout, large weight decay, or overly small/large hidden dimensions.

4.2.4 Computational cost and scalability

To ensure fair comparison, only multimodal classification models were selected as baselines. All models were evaluated under consistent conditions on an RTX 4060 GPU. Table 4 summarizes the parameter counts and average inference latencies for seven representative models.

While TechDoc remains the most efficient in terms of latency (9.5 ms/sample) and performs competitively in classification accuracy, our model demonstrates a more favorable balance across key dimensions—achieving higher accuracy with only a moderate increase in parameter count and maintaining real-time inference capability (13.9 ms/sample). Compared to more complex systems like IMPACT and PatentLVLM, which require significantly larger computational resources (491 M and 3.4B parameters respectively) and suffer from high latency (52.1 ms and 160.8 ms), the proposed model is both lighter-weight and faster, making it better suited for

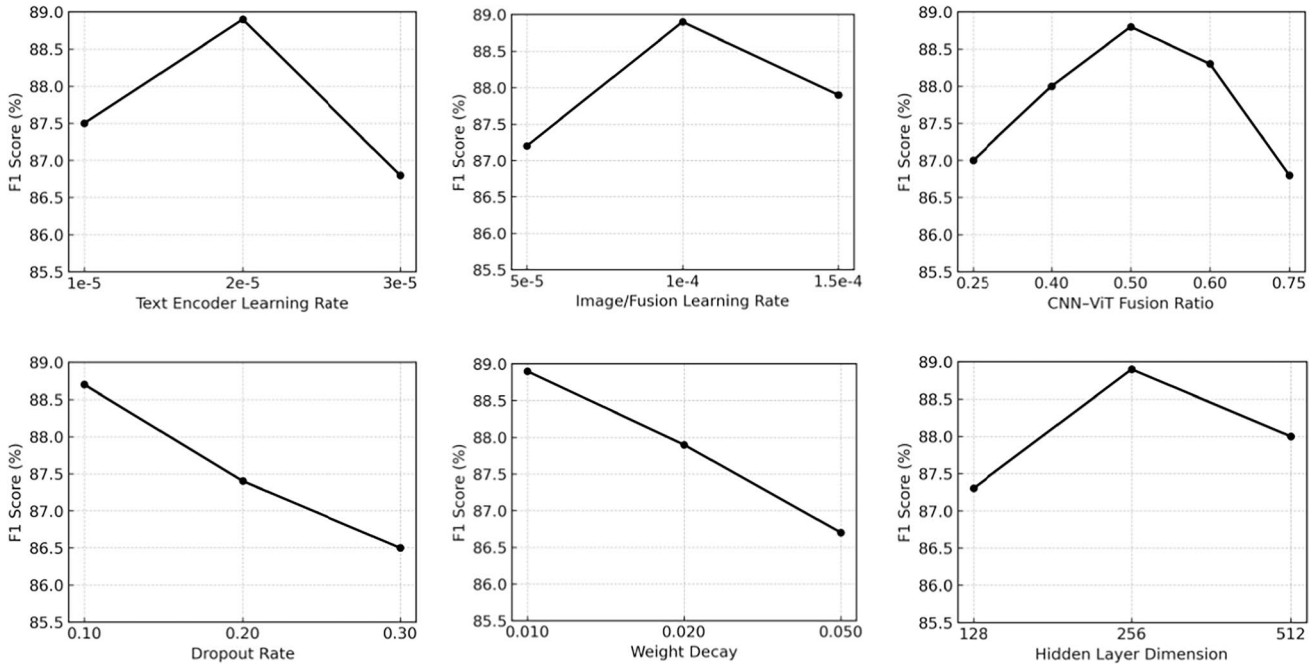

Fig. 4 Sensitivity Analysis Results

Table 4 Model Size and Inference Latency

Model	Parameter Count (M)	Inference Latency (ms/sample)
TechDoc	146.7	9.5
IMPACT	491.5	52.1
PatentLVLM	3401.5	160.8
VLCDoC	217.2	21.4
PMF	199.8	16.4
GlobalDoc	225.6	26.3
Proposed model	197.3	13.9

deployment in production environments. It also outperforms models such as PMF, VLCDoC, and GlobalDoc in inference speed, despite having comparable or smaller model sizes.

According to the China National Intellectual Property Administration (CNIPA), over 536,000 invention patent applications were filed in 2024 alone. These figures highlight the need for scalable, low-latency classification systems to support high-volume patent examination workflows. The compact design and efficient runtime of our model ensure practical deployability on standard hardware platforms, offering an effective solution for real-world, large-scale applications.

Table 5 Evaluation on Multilingual Design Patent Datasets

Dataset	Accuracy	Precision	Recall	F1 Score
CN-Design-D	0.8865	0.8982	0.8724	0.8851
US-2024-D	0.8563	0.8729	0.8381	0.8556
EU-2024-D	0.8254	0.8417	0.8123	0.8264

4.2.5 Cross-Domain Evaluation

To assess the robustness and cross-domain transferability of the proposed model, we extended our experiments to two additional design patent datasets from different legal and linguistic jurisdictions. Specifically, US-2024-D consists of 10,906 English-language design patents filed with the U.S. Patent and Trademark Office (USPTO) in 2024, where textual content includes only titles and structured claims. EU-2024-D contains 101,738 design patents submitted to the European Union Intellectual Property Office (EUIPO) in the same year. Notably, a substantial portion of entries in the EU dataset lack abstract information and are represented by titles alone, requiring the model to rely more heavily on non-textual modalities. The results are summarized in Table 5.

As summarized in Table 5, our model consistently maintains strong performance across all three datasets, achieving a F1 score of 0.8556 based on US-2024-D and 0.8264 based on EU-2024-D. Compared with that of CN-Design-241 K, which provides full-text Chinese descriptions and extensive assignee metadata, the slight drop in performance with the

US and EU corpora can be attributed to reduced text completeness, lower metadata coverage, and domain-specific differences in patent drafting styles. Nevertheless, the model demonstrates strong resilience in scenarios with limited textual input, highlighting its ability to adaptively leverage visual semantics and contextual metadata for reliable cross-lingual classification.

4.3 Example analysis and discussion

To investigate the classification performance of the proposed model in particular subdomains and assess its practical applicability, class 14 was selected as an example for analysis. This class includes"Recording, Telecommunications, or Data Processing Equipment,"which consists of 7 subclasses. The detailed information is provided in Table 6.

To ensure sample balance, 1,000 patents were selected randomly from each subcategory for the example analysis. To improve the interpretability of the proposed model, the baseline models with the best performance in single-modal features—BERT-for-Patents for the text modality and VIT + CNN for the image modality—are also included for comparative experiments. The final results of the example analysis are presented in Fig. 5.

The experimental results demonstrate that the proposed multimodal fusion model outperforms the single-modality baseline models across all the metrics, indicating that the model exhibits good robustness during the actual classification process. Furthermore, to investigate the typical errors made by both the proposed model and the single-modality models in the design patent classification task, as well as to examine the impact of each modality's features on the classification, ten patents that were misclassified by each model were selected for inspection. The recognition results for the ten patents that were obtained from the model proposed in this study and the two baseline models are shown in Table 7.

In common error cases, both the proposed model and the single-modality models (text and image) exhibit a certain number of misclassifications, as shown in Table 7:

For patents with multiple Locarno classification codes

Both the proposed model and the two baseline models can only recognize one classification code. For example, patent CN308898803S, which serves both as a clock and

Table 6 Classification Information for Class 14

Subclass Number	Subclass Heading
14–01	EQUIPMENT FOR THE RECORDING OR REPRODUCTION OF SOUNDS OR PICTURES
14–02	DATA PROCESSING EQUIPMENT AS WELL AS PERIPHERAL APPARATUS AND DEVICES
14–03	TELECOMMUNICATIONS EQUIPMENT, WIRELESS REMOTE CONTROLS AND RADIO AMPLIFIERS
14–04	GRAPHICAL USER INTERFACES AND ICONS
14–05	RECORDING AND DATA STORAGE MEDIA
14–06	HOLDERS, STANDS AND SUPPORTS FOR ELECTRONIC EQUIPMENT, NOT INCLUDED IN OTHER CLASSES
14–99	MISCELLANEOUS

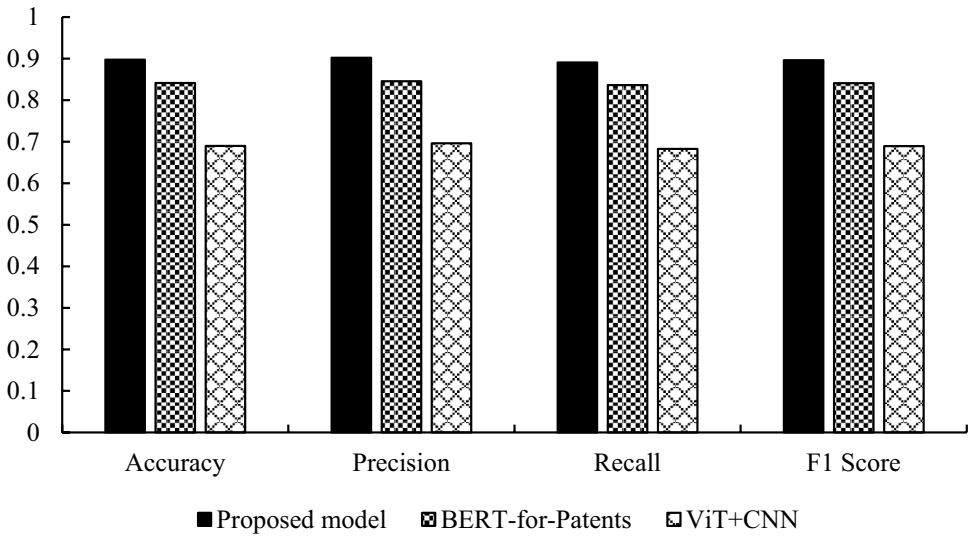

Fig. 5 Sample Analysis Experimental Results

Table 7 Recognition Results of Each Model

Serial number	Patent number	Locarno classification	Proposed model	BERT-for-Patents	ViT + CNN
1	CN308905045S	14–01、26–05	14–01	14–01	14–01
2	CN308898803S	14–01、10–01	10–01	14–01	10–01
3	CN308905059S	14–01	14–01	14–01	13–03
4	CN308862271S	14–01	16–01	14–01	16–01
5	CN308898700S	14–01	21–01	21–01	21–01
6	CN308879892S	14–03	14–03	19–07	14–03
7	CN308585220S	14–02	14–02	14–01	14–02
8	CN308879847S	14–03	14–03	10–06	10–04
9	CN308823181S	14–99	14–02	14–02	14–02
10	CN308856311S	14–99	23–01	13–03	23–01

a speaker, has features that are more aligned with a clock when observed from the appearance modality. Consequently, the single-image model classifies it as 10–01 (Clocks and Alarm Clocks), the single-text model classifies it as 14–01 (Recording or Reproducing Apparatus for Sound or Image), and the proposed model classifies it as 10–01 (Clocks and Alarm Clocks). For classifying design patents with multiple uses, the principle of"appearance first"should be followed, meaning that the product should be classified based on its visual similarity to a particular category. In the case of patent CN308898803S, its appearance more closely resembles that of a clock, so the classification code should primarily be 10–01 (Clocks and Alarm Clocks). For design patents with multiple Locarno classification codes, the proposed model, which incorporates image modality features, proves to be a valuable reference in practice.

For patents with discrepancies between textual descriptions or image content and their main innovative aspects

Patent CN308905059S, which is a recording box, may be classified as a button, control device, switch, or similar product based purely on the image features, resulting in the image modality model assigning it the classification code 13–03 (Distribution or Power Control Equipment, including wires, conductors, switches, circuit breakers, and distribution panels). For patents where the textual description or image content deviates from the main innovation (such as CN308862271S, CN308898700S, CN308879892S, and CN308585220S), single-modality models are likely to make errors. The proposed model, which combines both text and image modality features, can more accurately identify the complementary aspects and correctly classify patents such as CN308905059S and CN308879892S. However, for patents such as CN308862271S and CN308898700S, noise from modality features may interfere with the correct classification. Additionally, for patent CN308879847S, in which both the single-text and single-image models made errors, the proposed model correctly identified its classification by incorporating metadata features. Overall, the proposed model demonstrates better robustness and yields more accurate classification results.

For patents that do not fit well within the existing classification system

With the rapid pace of product iterations and technological innovations, the update frequency of the International Classification for Design Patents has not kept up with the speed of patent updates. As a result, the current classification system cannot accommodate all types of products. For example, patents CN308823181S (Identity Authentication Card) and CN308856311S (Waterproof and Anti-Silicone Oil Plug Device) do not have a suitable classification code, leading to their classification under 14–99 (Other Miscellaneous Categories of Recording, Telecommunications, or Data Processing Equipment). Both the proposed model and the single-modality models misclassified these patents. To address such errors, additional empirical data and training samples should be incorporated to improve the classification performance. However, the fundamental solution lies in the need for intellectual property management authorities to explore a more refined, multilevel design patent classification system that is better suited to the national context and more user friendly.

5 Conclusions

This study presents an automatic classification method for design patents that integrates text modality features, image modality features, and metadata modality features. It effectively mitigates the issue of sparse text modality features in design patents and enriches the current research on patent classification by expanding the scope of modality feature selection. This study makes four application-driven

contributions that address persistent challenges in design patent classification:

(1) Textual sparsity compensation: The textual content of design patents is extremely brief and highly redundant. We address the extreme brevity and redundancy of design patent texts through a two-stage representation strategy. First, domain-specific keywords are extracted using a BiLSTM-CRF model trained on authoritative Locarno classification corpora. This step filters out generic expressions and identifies salient semantic cues. Second, the keyword sequence is encoded via BERT-for-Patents, enhancing contextual and domain-specific representations. This hybrid approach produces compact yet informative embeddings tailored to the characteristics of design patents.

(2) Appearance-centric visual modelling: Design patents place primary emphasis on product appearance. To fully capture both geometric details and holistic structural patterns, we propose a hybrid image encoder that integrates Convolutional Neural Networks (CNNs) for local features (e.g., edges, contours) and Vision Transformers (ViTs) for global shape modeling. CNNs specialize in fine-grained visual cues, while ViT divides the image into fixed-size patches and models their interdependencies via transformer encoders. The fusion of local and global streams provides a dual-scale representation that improves visual discrimination for subtle design differences.

(3) Behavioural metadata priors: Design-patent assignees tend to follow stable category preferences, clearly visible in their historical Locarno filing patterns. Yet most existing models overlook these behavioural signals. We compute the frequency distribution of Locarno codes across each assignee's past applications and embed this profile as a structured metadata vector. Leveraging such behavioural priors steers the model toward more context-aware predictions, offering pronounced benefits for applicants with highly focused portfolios or operating in long-tail categories.

(4) Three-stage fusion architecture for cross-modal coordination: To improve the quality of multimodal representations, we design a structured fusion pipeline comprising three sequential stages. The view enhancement stage captures inter-modal dependencies via a covariance matrix and refines each modality through intra-modal attention. The adaptive attention stage dynamically weighs each modality based on its relevance to the classification objective, prioritizing salient cues. The final weighted fusion stage consolidates these refined representations into a unified embedding space. This hierarchical design mitigates weak inter-modal coupling, resolves modality imbalance, and aligns heterogeneous semantics, thereby enhancing both robustness and predictive accuracy across diverse scenarios.

Benchmark experiments against all baselines—spanning traditional machine learning, single-modality, and state-of-the-art multimodal models—show that the proposed framework achieves the highest F1 score while maintaining a relatively small parameter count (197 M) and relatively low inference latency (13.9 ms/sample). Additional evaluations on USPTO and EUIPO corpora confirm strong cross-jurisdiction robustness, even when many records contain titles only. Practically, the framework can streamline design-patent examination and retrieval workflows, offering a scalable solution for intellectual-property stakeholders.

However, the study has certain limitations. During the empirical analysis, the issue of the uneven distribution of Locarno classification codes in design patents was unaddressed. Additionally, the image modality extraction process only utilized the summary drawings of design patents. Future work will focus on addressing issues such as the imbalance in sample data distribution and refining the methodology accordingly.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants 72171122.

Authors'contributions Xiaodong Xie: Responsible for research design, conducting experiments, and writing the paper.

Jie Wu: Responsible for research proposal design and providing guidance on the paper.

Mengjia Xiang: Contributed to writing the paper.

Jianting Tang: Contributed to the research proposal design.

Yongxiang Sheng: Contributed to the research proposal design and provided guidance on the paper.

Funding National Natural Science Foundation of China,72171122

Data availability The data will be made available upon reasonable request.

Declarations

Conflict of interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aristodemou L, Tietze F (2018) The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. World Pat Inf 55:37–51

Awale S, Müller-Budack E, Ewerth R (2025) Patent figure classification using large vision-language models. arXiv:2501.12751. Accessed 3 July 2025

Bakkali S, Ming Z, Coustaty M, Rusiñol M, Terrades OR (2023) VLC-DoC: Vision-language contrastive pre-training model for cross-modal document classification. Pattern Recogn 139:109419

Bakkali S, Biswas S, Ming Z, Coustaty M, Rusiñol M, Terrades OR, Lladós J (2025) GlobalDoc: A Cross-Modal Vision-Language Framework for Real-World Document Image Retrieval and Classification. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 1436–1446

Bekamiri H, Hain DS, Jurowetzki R (2024) Patentsberta: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. Technol Forecast Soc Change 206:123536

Cai J, Wang X, Guan C, Tang Y, Xu J, Zhong B, Zhu W (2022) Multimodal continual graph learning with neural architecture search. In: Proceedings of the ACM Web Conference 2022. ACM, pp 1292–1300

Caldas CH, Soibelman L (2003) Automating hierarchical document classification for construction management information systems. Automat Constr 12(4):395–406

Cohen L, Gurun UG, Kominers SD (2016) The growing problem of patent trolling. Science 352(6285):521–522

Cui L (2024) A label learning approach using competitive population optimization algorithm feature selection to improve multi-label classification algorithms. J King Saud Univ Comput Inf Sci 36(5):102083

D'hondt E, Verberne S, Koster C, et al (2013) Text representations for patent classification. Comput Linguist 39(3):755–775

Das R, Singh TD (2023) Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Comput Surv 55(13s):1–38

Dib O, Nan Z, Liu J (2024) Machine learning-based ransomware classification of Bitcoin transactions. J King Saud Univ Comput Inf Sci 36(1):101925

Ektefaie Y, Dasoulas G, Noori A et al (2023) Multimodal learning with graphs. Nat Mach Intell 5(4):340–350

Fall CJ, Törcsvári A, Benzineb K, et al (2003) Automated categorization in the international patent classification. In: ACM SIGIR Forum. ACM, pp 10–25

Fang X, Liu D, Zhou P, Hu Y (2022) Multi-modal cross-domain alignment network for video moment retrieval. IEEE Trans Multimedia 25:7517–7532

Fink TMA, Reeves M, Palma R, Farr RS (2017) Serendipity and strategy in rapid innovation. Nat Commun 8(1):2002

Haghighian Roudsari A, Afshar J, Lee W, Lee S (2022) PatentNet: multi-label classification of patent documents using deep learning based language understanding. Scientometrics 127(1):207–231

Jiang S, Hu J, Magee CL, Luo J (2022) Deep learning for technical document classification. IEEE Trans Eng Manag 71:1163–1179

Khattar D, Goud JS, Gupta M, Varma V (2019) Mvae: Multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference. ACM, pp 2915–2921

Kim YG, Suh JH, Park SC (2008) Visualization of patent analysis for emerging technology. Expert Syst Appl 34:1804–1812

Lee JS, Hsiang J (2020) Patent claim generation by fine-tuning OpenAI GPT-2. World Pat Inf 62:101983

Li S, Hu J, Cui Y, Hu J (2018) DeepPatent: patent classification with convolutional neural networks and word embedding. Scientometrics 117(2):721–744

Li J, Wang X, Lv G, Zeng Z (2023a) GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. Neurocomputing 550:126427

Li J, Bin Y, Peng L, Yang Y, Li Y, Jin H, Huang Z (2024) Focusing on relevant responses for multi-modal rumor detection. IEEE Trans Knowl Data Eng 36(11):6225–6236

Li EP, Li TR, Luo HS, Chu JL, Duan LX, Lv FM (2025) Adaptive multi-scale language reinforcement for multimodal named entity recognition. Multimodal graph fusion for patent–paper alignment. IEEE Trans Multimedia: 1–12. https://doi.org/10.1109/TMM.2025.3543105

Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW (2019) Visualbert: a simple and performant baseline for vision and language. arXiv:1908.03557. Accessed 3 July 2025

Li Y, Quan R, Zhu L, Yang Y (2023b) Efficient multimodal fusion via interactive prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp 2604–2613

Liu C, Zhu C, Xia X, Zhao J, Long H (2022) FFEDN: Feature fusion encoder decoder network for crack detection. IEEE Trans Intell Transp Syst 23(9):15546–15557

Liu Y, Wu H, Huang Z, Wang H, Ma J, Liu Q, Chen E, Tao H, Rui K (2020) Technical phrase extraction for patent mining: A multi-level approach. In: 2020 IEEE International Conference on Data Mining (ICDM). IEEE, pp 1142–1147

Luo J, Sarica S, Wood KL (2021) Guiding data-driven design ideation by knowledge distance. Knowl Based Syst 218:106873

Madaan D, Makino T, Chopra S, Cho K (2024) Jointly modeling inter- & intra-modality dependencies for multi-modal learning. Adv Neural Inf Process Syst 37:116084–116105

Miric M, Jia N, Huang KG (2023) Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. Strateg Manag J 44(2):491–519

Qi Y, Zhang Q, Lin X, Su X, Zhu J, Wang J, Li J (2025) Seeing beyond noise: Joint graph structure evaluation and denoising for multimodal recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence 39:1142–1147

Qin Z, Zhang P, Wu F, Li X (2021) Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, pp 783–792

Shalaby W, Zadrozny W (2019) Patent retrieval: a literature review. Knowl Inf Syst 61:631–660

Shomee HH, Wang Z, Ravi S, Medya S (2024) Impact: a large-scale integrated multimodal patent analysis and creation dataset for design patents. Adv Neural Inf Process Syst 37:125520–125546

Shul Y, Choi JW (2024) Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection. In: ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 8686–8690

Tan H, Bansal M (2019) LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). ACL, pp 5100–5111
Thakur N, Reimers N, Daxenberger J, Gurevych I (2021) Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, pp 296–310
Tian Y, Sun X, Yu H, Li Y, Fu K (2021) Hierarchical self-adaptation network for multimodal named entity recognition in social media. Neurocomputing 439:12–21
Trappey AJC, Hsu FC, Trappey CV, Lin CI (2006) Development of a patent document classification and search platform using a back-propagation network. Expert Syst Appl 31(4):755–765
Trappey AJ, Trappey CV, Govindarajan UH, Sun J (2019) Patent value analysis using deep learning models—The case of IoT technology mining for the manufacturing industry. IEEE Trans Eng Manag 68(5):1334–1346
Tseng YH, Lin CJ, Lin YI (2007) Text mining techniques for patent analysis. Inf Process Manag 43(5):1216–1247
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
Wang Y, Sun F, Huang W, He F, Tao D (2022) Channel exchanging networks for multimodal and multitask dense image prediction. IEEE Trans Pattern Anal Mach Intell 45(5):5481–5496
Wu CH, Ken Y, Huang T (2010) Patent classification system using a new hybrid genetic algorithm support vector machine. Appl Soft Comput 10(4):1164–1177
Wu H, Zhang L, Zhu H, Liu Q, Chen E, Xiong H (2025) Examination process modeling for intelligent patent management: A multi-aspect neural sequential approach. ACM Trans Manag Inf Syst 16(3):1–23
Wu PS, Li H, Hu LW, Ge JR, Zeng NY (2024) A local-global attention fusion framework with tensor decomposition for medical diagnosis. IEEE/CAA J Autom Sinica 11(6):1536–1538. https://doi.org/10.1109/JAS.2023.124167
Wu W, Wang X, Luo H, Wang J, Yang Y, Ouyang W (2023) Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp 6620–6630
Wu Y, Zhan P, Zhang Y, Wang L, Xu Z (2021) Multimodal fusion with co-attention networks for fake news detection. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. ACL, pp 2560–2569
Xu P, Zhu X, Clifton DA (2023a) Multimodal learning with transformers: A survey. IEEE Trans Pattern Anal Mach Intell 45(10):12113–12132
Xu S, Liu X, Ma K, Dong F, Riskhan B, Xiang S (2023b) Rumor detection on social media using hierarchically aggregated feature via graph neural networks. Appl Intell 53(3):3136–3149
Yin M, Chen W, Zhu D, Jiang J (2025) Enhancing video rumor detection through multimodal deep feature fusion with time-sync comments. Inf Process Manag 62(1):103935
Zhang H, Wu B, Yuan X et al (2024) Trustworthy graph neural networks: Aspects, methods, and trends. Proc IEEE 112(2):97–139
Zhao F, Zhang C, Geng B (2024) Deep multimodal data fusion. ACM Comput Surv 56(9):1–36