# DESIGNCLIP: Multimodal Learning with CLIP for Design Patent Understanding

**Zhu Wang** **Homaira Huda Shomee** **Sathya N. Ravi** **Sourav Medya**
Department of Computer Science, University of Illinois Chicago
{zwang260,hshome2,sathya,medya}@uic.edu

## Abstract

In the field of design patent analysis, traditional tasks such as patent classification and patent image retrieval heavily depend on the image data. However, patent images—typically consisting of sketches with abstract and structural elements of an invention—often fall short in conveying comprehensive visual context and semantic information. This inadequacy can lead to ambiguities in evaluation during prior art searches. Recent advancements in vision-language models, such as CLIP, offer promising opportunities for more reliable and accurate AI-driven patent analysis. In this work, we leverage CLIP models to develop a unified framework DESIGNCLIP for design patent applications with a large-scale dataset of U.S. design patents. To address the unique characteristics of patent data, DESIGNCLIP incorporates class-aware classification and contrastive learning, utilizing generated detailed captions for patent images and multi-views image learning. We validate the effectiveness of DESIGNCLIP across various downstream tasks, including patent classification and patent retrieval. Additionally, we explore multimodal patent retrieval, which provides potential to enhance creativity and innovation in design by offering more diverse sources of inspiration. Our experiments show that DESIGNCLIP consistently outperforms baseline and SOTA models in patent domain on all tasks. Our findings underscore the promise of multimodal approaches in advancing patent analysis. The codebase is available here: https://github.com/AI4Patents/DesignCLIP

## 1 Introduction

Patents are units for the state of the art innovation, designs and technological advancements as well as they offer legal protection for inventors' intellectual property (Moser, 2013). Patents grant the owner the authority to prevent others from manufacturing, utilizing, or distributing the patented invention without their permission[1]. Among the two popular types, *utility patents* are granted for innovations in processes, machines, manufactures, or compositions of matter, including improvements, *design patents* are awarded for new, original, and ornamental designs applied to manufactured items. While utility patents have been well studied (Fall et al., 2003; Kamateri et al., 2022; Siddharth et al., 2022; Kang et al., 2020), design patents remain relatively underexplored.

One of the most important design patent tasks is patent retrieval. This aims to determine the novelty of the patent and prevent infringements. A patent is only granted if the design significantly differs from existing ones. Most previous research (Higuchi and Yanai, 2023; Lo et al., 2024) focus primarily on image-to-image retrieval since there is a lack of informative text descriptions in design patents. Patent classification (Rademaker, 2000; Kamateri et al., 2024) is another task by patent reviewers who classify applications into various subject matters and assign design classification codes. In the US design patent system, there are 33 classes which also contain various subclasses. Automating the patent classification process can save an enormous amount of time.

Solving the above tasks would need sophisticated multimodal techniques due to several challenges. Design patent images are sketches that provide detailed design information and differ significantly from natural images (please see the examples of patent designs in Figure 2). Additionally, the class distribution in the design patent dataset is highly imbalanced, with the six most frequent classes accounting for almost half of the total data (Figure 1). Furthermore, the aim is to build one single model that can perform the patent classification as well as retrieval. *There are currently no multimodal models that are designed to address*

[1] https://www.uspto.gov/patents/basics



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10468–10490
November 4-9, 2025 

*these challenges and specifically tailored for design patents.*

**Our contributions.** In this paper, we develop a new CLIP-based framework to facilitate multimodal analyses on design patents. Our major contributions are as follows.

- **Multimodal Analysis.** We address the problem of multimodal analysis in design patents by building a comprehensive AI-based framework. The goal is to address the time-consuming patent tasks such as patent classification and patent retrieval efficiently.
- **DESIGNCLIP**: We modify CLIP (Radford et al., 2021) by incorporating the domain knowledge from design patents. It uses a class-aware learning method to balance the long-tailed distribution of classes in design patents, and is pre-trained with multiple tasks, including patent classification and multi-view image-image contrastive learning.
- **Experiments**: We conduct comprehensive experiments showing that our methods outperforms baseline models and the state-of-art patent image retrieval models on all downstream tasks and gain notable improvements on design patent representations.

## 2 Background on Patent Analysis

The patent retrieval task focuses on efficiently retrieving relevant patent documents and images based on search queries. In design patents, this task is focused on image-to-image retrieval, where the objective is to find visually similar design images that match a given image query. (Kucer et al., 2022) implement various models such as ResNet50 (He et al., 2016), and Sketchy RN50 (Sangkloy et al., 2016). Their patent-specific models are initially pre-trained on ImageNet (Deng et al., 2009) and fine-tuned on the DeepPatent dataset. Similarly, (Higuchi and Yanai, 2023; Higuchi et al., 2023) use a deep metric learning framework, utilizing cross-entropy methods like InfoNCE with ArcFace. On the other hand, (Lo et al., 2024) implement several advanced models, such as ViT (Dosovitskiy et al., 2020) and Swin (Liu et al., 2021), along with more recent multimodal models such as BLIP-2 (Li et al., 2023), and GPT-4V (Achiam et al., 2023). They introduce a novel approach by proposing a language-informed strategy for learning features from patent images. Please refer to the survey (Shomee et al., 2025) for more details. *Notably, there is currently no specific machine learning or AI-based research focused on the critical constraints (e.g., classes and multi-views) from the design patents and we address this in this paper.*

While utility patents are different from design patents, the images in utility patents are not well-studied. (Ghauri et al., 2023) classify utility patent images into distinct types of visualizations, including graphs, block circuits, flowcharts, and technical drawings. They employ the CLIP model (Radford et al., 2021), integrated with a Multi-layer Perceptron and various Convolutional Neural Networks (Krizhevsky et al., 2012) architectures, to enhance the precision of patent image classification. IMPACT dataset (Shomee et al., 2024) is a comprehensive and large-scale resource comprising over 500,000 U.S. design patents issued between 2007 and 2022. It includes 3.61 million figures accompanied by detailed captions, titles, and metadata, offering a rich multimodal dataset that integrates visual and textual information. *In this paper, we mainly focus on design patents using IMPACT dataset and address the two primary tasks associated with the design patents.* We demonstrate the design patent retrieval task in three different formats: text-to-image, image-to-text, and image-to-image. Additionally, the classification of design patent images into 33 subject matter (class) and various subclasses is a detailed and structured approach to organizing design patents. We address the classification task by categorizing patents into their class level categories using both text-based captions and images.

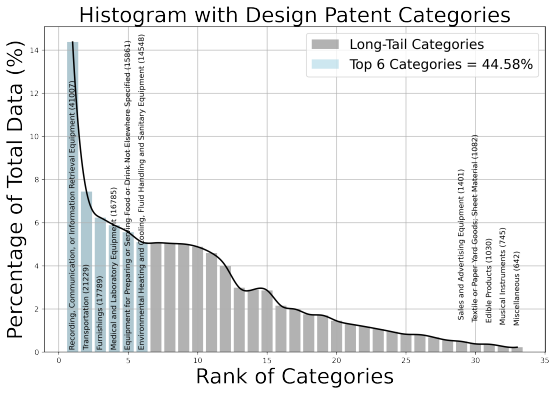



Figure 1: Design patent data category distributions in our test data. Top 6 categories consists of 44.58% of all the data which shows a long tail distribution.

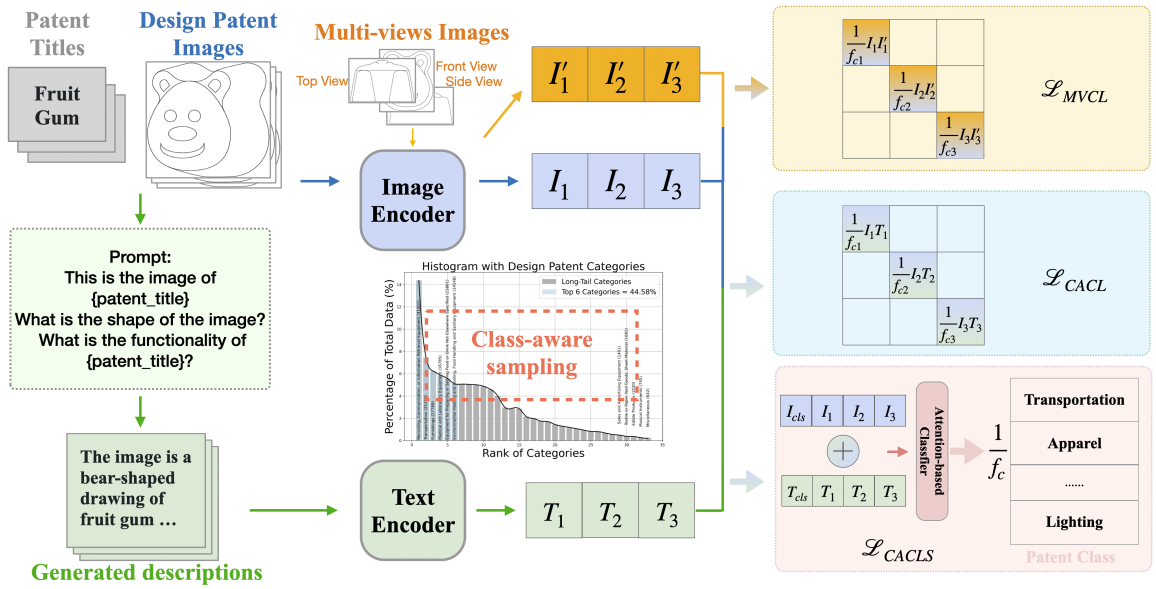



Figure 2: The framework for DESIGNCLIP. The inputs are patent images and simple text (title). First, we generate descriptions and pass them along with images to text and image encoders accordingly. Then, we pre-train CLIP-based models with adapted class-aware sampling and learning and multi-view image contrastive learning on 285,000 design patents. $\mathcal{L}_{\text{MVCL}}$, $\mathcal{L}_{\text{CACL}}$, and $\mathcal{L}_{\text{CACLS}}$ denote the multi-view contrastive loss (Eq. 5), class-aware contrastive loss (Eq. 3), and class-aware classification loss (Eq. 4) respectively.

## 3 DESIGNCLIP

In this work, we address the challenges of design patent analysis by adapting the CLIP vision-language model family. The major components of our framework, DESIGNCLIP, are as follows: *First,* considering the long tail distribution among the classes of design patents, we begin with proposing a class-aware learning method to resample construct pairs in the batch. This helps our method DESIGNCLIP to balance learning from all the classes. *Then*, we pre-train DESIGNCLIP with multiple tasks, including patent classification, image-caption contrastive loss, and multi-view image-image contrastive loss. Our architecture is illustrated in Figure 2.

**Brief Review of CLIP Framework:** CLIP (Radford et al., 2021), a multimodal model uses contrastive learning to bridge the gap between text and images. At its foundation is the principle of acquiring perceptual understanding from the guidance of natural language. During its pre-training phase, CLIP learns to identify if a text snippet and an image are matched in its dataset. The training involves a series of five ResNets and three Vision Transformers to facilitate zero-shot classification. CLIP has been popular across various fields including medical imaging (Lu et al., 2024; Müller et al., 2022; Lei et al., 2023), robotics (Shibata et al., 2024; Rana et al., 2023; Sontakke et al., 2024), biodiversity monitoring (Gong et al., 2024), e-commerce (Hendriksen et al., 2022), educational technology (Sun et al., 2024) and showed great performance.

The popular CLIP (Radford et al., 2021) is a pre-trained model to learn image-text pairs with contrastive loss. This approach enables the model to differentiate between semantically similar and dissimilar data points, enhancing its ability to capture meaningful relationships between visual and textual information. CLIP is widely applied for a variety of multimodal tasks, including zero-shot classification, and multimodal search. Specifically, given an image, CLIP can retrieve the most relevant text descriptions (e.g., captions or labels). Given image-text pairs $\{(v_i, t_i)\}_{i=1}^{N}$, the vanilla loss function in CLIP is defined as:

$$\mathcal{L}_{\text{CLIP}} = -\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_k)/\tau)}, \quad (1)$$

where $\text{sim}(\cdot)$ is the cosine similarity between the image embedding $v_i$ and the text embedding $t_i$, and $\tau$ is a temperature parameter.

### 3.1 Class-aware Learning

As illustrated in Figure 1, the class (category) distribution in the design patent dataset is highly imbalanced, among which the top six most frequent

classes constitute 44.58% of the total training data. This imbalance presents a significant challenge for the standard CLIP contrastive learning method. This long tail recognition problem tends to be biased towards the head classes and leads to suboptimal performance on the less frequent classes. In DESIGNCLIP, we address this problem by utilizing the marginal distribution in two ways: Class-aware Sampling and a modified Class-aware Contrastive Loss.

#### 3.1.1 Class-aware Sampling

While training DESIGNCLIP, it is essential for each batch to contain an adequate amount of data from all the classes, which ensures each class receives appropriate supervision signals (Wang et al., 2017; Kang et al., 2019; Zhu et al., 2022). We consider resampling categories dynamically in each batch to construct contrastive pairs. For analyzing patents in the constrastive learning framework, we use a positive sample as the patent image and its generated captions, and the negative sample as the patent image with the caption of another patent. The quality of negative image-text pairs plays a vital role in contrastive learning. Intuitively, our goal is twofold: (i) to maintain a balanced representation of all classes during the batch training, and (ii) while constructing batches, we also aim to increase the likelihood that each batch contains a more diverse set of classes, with a particular focus on sufficient representation from the tail classes. To do so, we define the probability of sampling a class $c$ as,

$$p_c = \frac{\frac{1}{f_c^\beta}}{\sum_{j=1}^{C} \frac{1}{f_j^\beta}}, \tag{2}$$

where $f_c$ is the frequency of class $c$ in the batch, $C$ is the total number of classes, and $\beta$ is a hyperparameter of rebalancing the weights of the classes.

The advantage of our definition of $p_c$ is that when an anchor from a tail class is involved, we can easily adjust the sampling process by increasing the value of $\beta$ in Eq. 2 for the head (popular) classes. This reduces the likelihood of the negative pairs mainly originating from the head classes, and thereby prevents the tail classes from being overly penalized as negatives. This adjustment helps the model to learn better representations for the tail classes and leads to improved performance in downstream tasks like patent classification and patent retrieval.

#### 3.1.2 Class-aware Contrastive Loss

In addition to Class-aware Sampling, we introduce a modified Class-aware Contrastive Loss that further mitigates the imbalance by incorporating class-dependent weighting into the CLIP contrastive loss in Eq. 1, as follows:

$$\mathcal{L}_{\text{CACL}} = -\frac{1}{f_i^\beta} \log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_k)/\tau)}, \tag{3}$$

where, $f_i$ is the frequency of class which the $i$-th sample belongs. This class-aware weighting ensures that the loss function penalizes misclassifications more heavily for tail classes, which encourages the model to allocate more capacity to learn these classes.

### 3.2 Multi-task Pre-training

Considering the properties of patent data and tasks, we train DESIGNCLIP on three tasks simultaneously, including *(1)* class-aware image-generated description contrastive learning, *(2)* class-aware classification, and *(3)* class-aware image-image (other views) contrastive learning. In task (1), we use the methods in Class-aware Learning. In this section, we introduce task (2) and task (3).

#### 3.2.1 Class-aware Classification

In this task, we aim to handle the class imbalance in addition to the Class-aware Learning. It helps to capture the subtle differences between patents which enhances the model's ability to generalize across diverse categories. Specifically, DESIGNCLIP incorporates an attention-based (Vaswani et al., 2017) classification layer with both image and text features. Let $v_i \in \mathbb{R}^{d_v}$ denote the visual features extracted from the $i$-th image by an image encoder, and $t_i \in \mathbb{R}^{d_t}$ denote the textual features extracted by a text encoder, the attention scores is computed as $\alpha_i = \text{softmax}\left(W_v v_i + W_t \mathbf{t}_i + b_a\right)$, and then the combined multimodal vector $h_i$ is calculated by $h_i = \alpha_i \odot (W'_v v_i + W'_t t_i)$, where $W$ are learnable weights and $b$ is bias. Finally, the class-aware classification loss is computed as:

$$\mathcal{L}_{\text{CACLS}} = -\frac{1}{f_i^\beta} \log \frac{\exp(w_c^\top h_i)}{\sum_{j=1}^{C} \exp(w_j^\top h_i)}, \tag{4}$$

where $f_i$ is the frequency of class in which $i$-th sample belongs, $\beta$ is a hyperparameter of rebalancing the weights of each class, and $w_c$ represents the weight vector of class $c$.

This class-aware classification approach addresses the class imbalance to ensure that DESIGNCLIP effectively pays importance to the underrepresented (tail) classes. We also add the attention mechanism to learn the most relevant features in both image and text modalities. Moreover, by applying class-aware weighting, DESIGNCLIP can be generalized across a wide variety of categories of design patents.

#### 3.2.2 Multi-view Image-image Contrastive Learning

Patents often include multiple views of the same design to provide a comprehensive understanding of the object from different angles, as shown in Figure 2 ("Multi-view Images"). To effectively learn the multi-view information, we employ a similar class-aware contrastive learning approach in Class-aware Learning. The goal is to learn consistent representations across different views of the same patent while distinguishing these from views of other patents. Note that, the number of views varies in each patent, and some contain over 10 views. Thus, we mainly focus on front, top and side views. By aligning features from different views of the same object in a shared embedding space, the model gains a better understanding of the design. Similar to Eq. 3, given image-image pairs $\{(v_i, v_i')\}_{i=1}^{M}$, the multi-view image-image contrastive learning loss is written as:

$$\mathcal{L}_{\text{MVCL}} = -\frac{1}{f_i^{\beta}} \log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{v}_i')/\tau)}{\sum_{k=1}^{M} \exp(\text{sim}(\mathbf{v}_i, \mathbf{v}_k')/\tau)}, \tag{5}$$

where $f_i$ is the frequency of class which the $i$-th sample belongs. To construct positive pairs, we randomly sample the image from the front view, side view, and top view.

Finally, our pre-train loss of DESIGNCLIP is a linear combination of three tasks as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CACLS}} + \lambda_2 \mathcal{L}_{\text{CACL}} + \lambda_3 \mathcal{L}_{\text{MVCL}} \tag{6}$$

## 4 Experiments

**Reproducibility.** We make the codebase available here: https://github.com/AI4Patents/DesignCLIP

### 4.1 Dataset

We use a total of 285,391 patents for training DESIGNCLIP, and 10,000 patents for validation from IMPACT dataset (Shomee et al., 2024). Moreover, we also generate captions following IMPACT for different views with the prompt: *This is the {patent_view} view image of {patent_title}. What is the shape of the image? What is the functionality of {patent_title}?* The "patent_view" includes top, front, and side views. In addition, we used data including 22,467 design patents from 2023 as test set, such as evaluations on zero-shot performance. Note that we use IMPACT as a baseline for patent downstream tasks. More details of the dataset are shown in the Appendix A.1.

### 4.2 Implementation details

We use an open source implementation of CLIP[2]. The backbone models are ResNet50, ResNet101, ViT-B-32 and ViT-L-14. The hyperparameters for the best performance are listed as follows: learning rate is $5e-6$, weight decay is 0.1, optimizer is AdamW for all models, $\beta$ is 1.2, $\lambda_1$ is 1, $\lambda_2$ is 0.1 and $\lambda_3$ is 0.2. The batch size is 128, except 64 for ViT-L-14. We use image size of $224 \times 224$ in all the experiments. All pre-training experiments are conducted on a cluster of 4 NVIDIA A40 GPUs, and the downstream tasks and ablation studies are on 4 NVIDIA V100 GPUs.

### 4.3 Patent downstream tasks

In this work, we mainly demonstrate DESIGNCLIP is beneficial for patent domain tasks, including patent classification and retrieval. However, there are limited studies focus on design patent, and they only work on image retrieval task. To evaluate the effectiveness of DESIGNCLIP, we consider to conduct experiments on image retrieval comparing with the state-of-the-art methods, and evaluate on design patent classification and multimodal retrieval comparing with the general CLIP which is pre-trained on natural images and IMPACT results.

#### 4.3.1 Image retrieval

Recent design patent studies mainly focus on image retrieval (IR). This process is often used for discovering new patents and assessing their novelty. These studies are validated with DeepPatent dataset (Kucer et al., 2022) containing 45,000 design patents. To verify DESIGNCLIP, we conduct experiments compare to the SOTA model (Higuchi and Yanai, 2023) on DeepPatent. We reproduce the same retrieval pipeline which used ArcFace (Deng

[2] https://github.com/mlfoundations/open_clip

Table 1: Image retrieval results (mAP) comparison between DESIGNCLIP and other patent SOTA models on DeepPatent test set. Image size is 224 × 224. The best results are in bold. *Denotes our implementation.

| Model | mAP |
|---|---|
| DeepPatent (Kucer et al., 2022) | 0.379 |
| ViT-B + ArcFace (Higuchi and Yanai, 2023) | 0.614 |
| SWIN (image) + ArcFace (Higuchi and Yanai, 2023) | 0.676 |
| SWIN (image+text) + ArcFace* | 0.684 |
| CLIP-ViT-B (image) + ArcFace* | 0.645 |
| CLIP-ViT-B (image+text) + ArcFace* | 0.671 |
| IMPACT (Shomee et al., 2024) | 0.657 |
| **DESIGNCLIP-ViT-B (image)** + ArcFace | **0.698** |
| **DESIGNCLIP-ViT-B (image+text)** + ArcFace | **0.712** |

et al., 2019) with different backbones. In our settings, we utilize DESIGNCLIP as backbones and also compare the performance of only using vision features and combining vision and textual features. More details of the implementations are shown in the Appendix (Sec. A).

Table 1 illustrates the results of image retrieval. The evaluation metric is the mean Average Precision (mAP) over all queries in the test set. DESIGNCLIP outperforms SOTA models by replacing the backbones which are pre-trained on patent data with our DESIGNCLIP models. Indeed, textual features are also beneficial to image retrieval. We show the retrieval examples in the Appendix.

#### 4.3.2 Patent classification

Patent classification is important but time-consuming for patent reviewers, and could be faster with AI-based frameworks. Thus, we consider to showcase the effectiveness of DESIGNCLIP on the classification task. The baseline model is OpenAI pre-trained CLIP and IMPACT. We use our test set (patents from 2023) to demonstrate the classification tasks, including zero-shot classification, and finetuning with a linear classifier. More details of the implementations are shown in the Appendix (Sec. A).

We demonstrate two settings of patent classification results in terms of accuracy in Table 2. We use backbones with RN101 and ViT-B to compare the results. In all settings, DESIGNCLIP performed better than CLIP and IMPACT. Notably, in finetuning settings, DESIGNCLIP outperforms CLIP by 60.1% for RN101 and by 5.4% for ViT-B, and achieves a 35.9% improvement for RN101 and a 3.6% improvement for ViT-B comparing IMPACT. Therefore, DESIGNCLIP provides better representations for patent classification.

Table 2: Accuracy (%) for patent classification: comparison between CLIP and DESIGNCLIP on 2023 test set. The best results are highlighted in bold in both settings.

| Model | Backbone | Zero-shot | Fine-tune |
|---|---|---|---|
| CLIP | RN101 | 11.91 | 22.45 |
| | ViT-B | 10.88 | 43.06 |
| IMPACT* | RN101 | 11.89 | 27.66 |
| | ViT-B | 12.39 | 43.81 |
| **DESIGNCLIP** | RN101 | 11.93 | 35.93 |
| | ViT-B | **14.70** | **45.37** |

#### 4.3.3 Multimodal retrieval

The patent retrieval task is to identify relevant patent documents and images in response to search queries. In this task, we focus on multimodal retrieval, which incorporates both text and images. This integration enhances the ability to cross-reference and verify information, thus improving the overall effectiveness and efficiency of patent searches. In addition, multimodal retrieval can enable creativity and innovation in design by providing richer and more diverse sources of inspiration. We perform experiments on zero-shot text-image (T2I) and image-text (I2T) retrieval tasks on our validation set.

We evaluate multimodal retrieval performance (Recall@K[3]) and present results in Table 3. It shows that DESIGNCLIP outperforms CLIP with all backbones on two tasks. As shown in the results, DESIGNCLIP gains significant improvements, by up to *360%* and *348%* respectively at R@5 on T2I and I2T retrievals. Note that, ViT-L obtains the best recall in all settings, which demonstrates that the larger advanced models boost the performance. Multimodal retrieval examples are provided in the Appendix.

### 4.4 Ablation studies

We perform the ablation studies to analyze the hyperparameter settings and the components of DESIGNCLIP, including the effectiveness of captions, multiple views, and pre-train tasks. Considering the limitations of computing resources, we use the patents from the recent five years for all ablation studies, including 113,887 patents in the train set and 5,000 patents in the validation set. We perform all ablation studies on the backbone of ViT-B-32.

[3]The metric R@K evaluates whether the ground truth appears within the top K results of the validation set.

Table 3: Multimodal retrieval performance comparison between CLIP, IMPACT and DESIGNCLIP on validation set. The ViT-L-14 model demonstrated superior performance over the other three backbone models tested. The highest Recall@K (%) values are highlighted in bold. *Denotes our implementations.

| Model | Backbone | Text-Image | | Image-Text | |
|---|---|---|---|---|---|
| | | R@5 | R@10 | R@5 | R@10 |
| CLIP | RN50 | 5.47 | 8.51 | 5.24 | 7.72 |
| | RN101 | 7.60 | 11.17 | 6.10 | 9.35 |
| | ViT-B | 7.49 | 10.60 | 6.90 | 10.34 |
| | ViT-L | 13.26 | 18.29 | 12.07 | 17.17 |
| IMPACT* | RN50 | 17.21 | 23.18 | 14.67 | 21.48 |
| | RN101 | 22.10 | 31.35 | 20.32 | 27.70 |
| | ViT-B | 25.60 | 34.92 | 24.88 | 35.12 |
| | ViT-L | 37.34 | 50.56 | 38.79 | 51.05 |
| DESIGNCLIP | RN50 | 25.17 | 34.50 | 23.49 | 32.70 |
| | RN101 | 26.71 | 36.51 | 25.37 | 34.84 |
| | ViT-B | 29.75 | 39.91 | 28.39 | 38.26 |
| | **ViT-L** | **42.30** | **52.80** | **40.14** | **53.98** |

#### 4.4.1 Analysis of detailed descriptions

We further analyze on the effectiveness of different texts. Specifically, we follow IMPACT dataset and classify their generated descriptions into captions only with "Patent Title", captions only with "Shape" and full captions. Based on the results shown in Table 4a, the detailed descriptions, including the shape and functions are beneficial for multimodal models to adjust to the patent domain. A notable increase of over 3% on both multimodal retrieval tasks indicates that DESIGNCLIP, which is trained with generated descriptions has, a better understanding for further assessing and inspiring on design patents.

#### 4.4.2 Impact of different views

Patent images feature views from various perspectives, and we explored how these views impact image retrieval. We illustrate different views results of image retrieval with DESIGNCLIP-ViT-B, as shown in Table 4b. Front views benefits image retrieval task with $\approx$ 1.8 % increase comparing with side views. Table 5 shows the multimodal retrieval performance across four different views with DESIGNCLIP-ViT-B. The front view yields the highest recall, while the side and top views do not produce satisfactory results. We also combine these three views which we refer as 'Multi view' but this did not improve performance either. Our experiments demonstrate that not all views contribute equally to the performance of retrieval tasks. Using only side and top views can significantly reduce the model's generalization ability. This lack of effectiveness can be attributed to the side and top views often not capturing sufficient design details. For instance, Figure 7a (Appendix) shows the design of an oven, where the top view fails to convey the design information of an oven which leads poor recall. Therefore, we believe these views may not capture the most distinguishing features of the design, leading to confusion when the model attempts to match these views with corresponding text or images. Based on these findings, we can conclude that the front view most effectively conveys the essential information of a design and aligns well with textual elements such as titles and captions.

Table 4: Ablation studies on different text and views. Backbone models is ViT-B. The best results of multimodal retrieval (Recall@K (%)) and image retrieval (mAP) are in bold.

| Text | I2T R@5 | T2I R@5 |
|---|---|---|
| Patent Title | 22.38 | 22.18 |
| Shape | 23.80 | 24.16 |
| Full captions | **25.56** | **25.90** |

(a) Text inputs

| Views | IR mAP |
|---|---|
| Side | 0.611 |
| Top | 0.605 |
| Front | **0.629** |

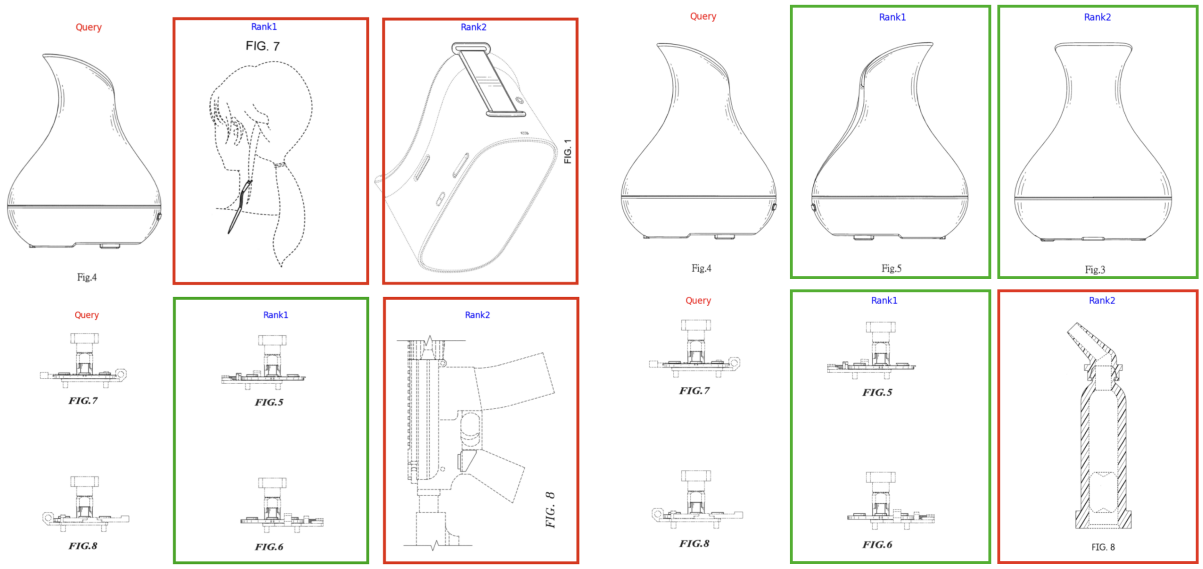
(b) Views

Table 5: Multimodal Retrieval Task Performance using multiple views. Backbone is ViT-B. Front view demonstrated higher performance over the other three views. The highest Recall@K (%) values are highlighted in bold.

| View | Text-Image | | Image-Text | |
|---|---|---|---|---|
| | R@5 | R@10 | R@5 | R@10 |
| Side view | 8.24 | 12.45 | 9.41 | 13.61 |
| Top view | 9.12 | 13.37 | 9.96 | 15.01 |
| Multi view | 8.59 | 12.57 | 9.58 | 13.82 |
| Front view | **12.05** | **17.47** | **13.36** | **18.86** |

#### 4.4.3 Effectiveness of pre-train tasks

To verify the effectiveness of our adapted multi-task pre-training, we conduct experiments on different combinations of pre-train tasks, including (1) class-aware image-generated description contrastive learning, (2) class-aware classification, and (3) class-aware image-image contrastive learning. As results shown in Table 6, pre-training on all tasks brings significant improvements for all downstream tasks in zero-shot settings. Thus, DESIGNCLIP not only enhances the model's generalization capabilities but also can be more adaptable to many patent tasks.




(a) CLIP (b) DESIGNCLIP

Figure 3: DESIGNCLIP in (b) often retrieves the accurate different views as the top results (green boxes in the first row). However, the results are not always accurate (red box in second row).

Table 6: Ablation studies on pre-train tasks of DESIGNCLIP. We evaluate classification (Accuracy (%)), multimodal retrieval (Recall@K (%)) and image retrieval (mAP) with ViT-B. The best results are in bold.

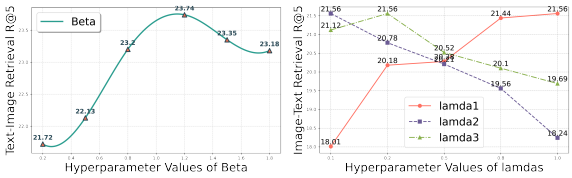
| Pre-train tasks | | | Classification | I2T | T2I | IR |
|---|---|---|---|---|---|---|
| (1) | (2) | (3) | Accuracy | R@5 | R@5 | mAP |
| ✓ | | | 3.18 | 20.94 | 21.24 | 0.649 |
| ✓ | ✓ | | 8.29 | 21.38 | 20.98 | 0.652 |
| ✓ | ✓ | ✓ | **8.64** | **21.56** | **21.71** | **0.658** |

#### 4.4.4 Impact of vision backbones

In training DESIGNCLIP, we choose 4 backbone architectures RN50, RN101, ViT-B, and ViT-L. Table 3 in the main paper shows the performance of text-image and image-text retrieval tasks for the comparisons of CLIP and our pretrained DESIGNCLIP. The results show that both ResNet50 and ResNet101 achieve comparable recall, while ViT-B performs slightly better. Notably, the Vision Transformer Large (ViT-L) model significantly outperforms the others, as evidenced by its higher Recall@5 and Recall@10 scores in both retrieval directions (Text-Image and Image-Text). The size of the parameter of the ViT-L model allows it to better handle patent data and makes it effective for patent image and text retrieval tasks.

#### 4.4.5 Hyperparameters Analysis

As results shown in Figure 4, we can conclude *(1)* increasing $\beta$ values for balancing tail classes can improve the performance, see Figure 4a, but bigger $\beta$ values may cause overfit on the tail classes which will be harmful for the overall pre-training. *(2)* Figure 4b shows that the weight assgined to $\mathcal{L}_{\text{CACL}}$ is crucial to the performance. $\mathcal{L}_{\text{CACLS}}$ ensures that the model learns to categorize images correctly, but emphasis on the classification could reduce the model's ability to generalize to new data. Similar to the results in impact of different views, while the multi-views help in learning consistent representations across different views, a smaller weight for $\mathcal{L}_{\text{MVCL}}$ indicates that image views alignment needs to be carefully considered and may diminish the benefits.

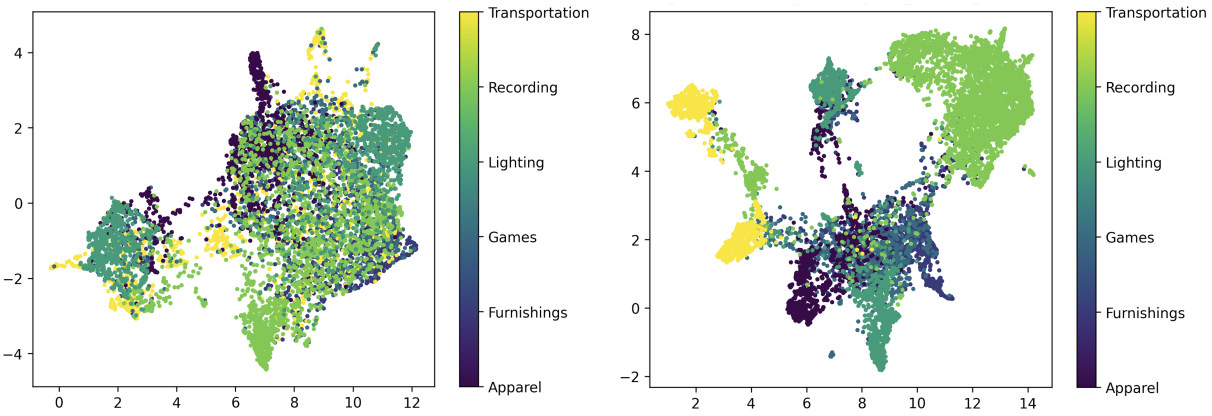



(a) T-I R@5 with different $\beta$ (b) I-T R@5 with different $\lambda$s

Figure 4: Ablation on hypermateters of $\beta$ and $\lambda$s.

### 4.5 Qualitative Analysis

We show the retrieval results of multiple views from DESIGNCLIP. Figure 3 presents the top two image retrievals. DESIGNCLIP often retrieves the different views of the query figure. Figure 9 (see the Appendix) shows that CLIP often cannot retrieve multiple views of the same query figure.




(a) CLIP (b) DESIGNCLIP

Figure 5: UMAP feature embeddings for patent images. (a) Visualization of features using CLIP models (b) Visualization of features using DESIGNCLIP. DESIGNCLIP shows well formed clusters in image features.

#### 4.5.1 U-MAP projection analysis

Figure 5 shows the learned image features for sample patents using U-MAP projection (McInnes et al., 2018). Different colors represent the clusters of the corresponding classes. We observe that DESIGNCLIP can identify clusters over the extracted image features, but CLIP is not able to classify the patent images. It shows that DESIGNCLIP might be more beneficial in the specific patent domain for many relevant downstream tasks, such as classification and retrieval.

#### 4.5.2 Class-aware classification analysis

We further analyze the impact of class-aware learning in DESIGNCLIP. On patent classification tasks, DESIGNCLIP improves by 1.22 % on the top-6 classes and 2.90 % on the long-tail classes. Among all 33 classes, DESIGNCLIP outperforms CLIP in 20 classes under fine-tuning (see Figure 6). These results highlight the importance and effectiveness of class-aware learning for improving performance on long-tail classes.

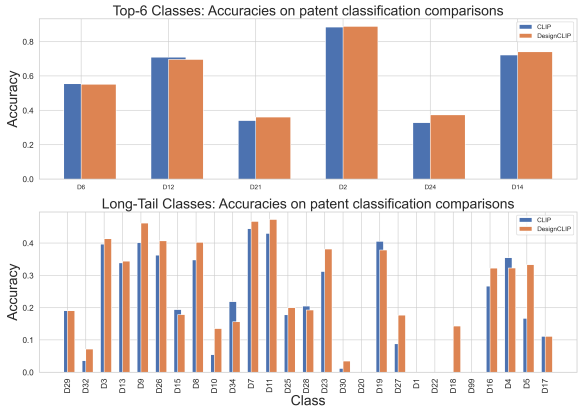



Figure 6: DESIGNCLIP improve classification accuracies for 20 classes among long-tail distributions.

## 5 Conclusions

In this paper, we introduce DESIGNCLIP to provide a domain-aware multimodal model for design patents. We first consider class-aware sampling and contrastive learning for the long-tail distribution in design patent data. In addition, multi-view image-image contrastive learning provides a comprehensive understanding of the patents from different angles, which can improve the performance in patent retrieval. Finally, we pre-train CLIP-based models on multi-tasks tailored for design patents. Our DESIGNCLIP outperforms the baseline models on patent classification, image retrieval and multimodal retrievals. These demonstrate that DESIGNCLIP can provide a better understanding and generalization ability for design patents in multimodal scenarios.

## Acknowledgments

This work was supported by in-kind contributions from University of Illinois Urbana-Champaign (UIUC) HACC Cluster and NSF ACCESS UIUC NCSA Cluster (Ref: ELE230014). The authors also acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot and the Texas Advanced Computing Center (TACC) Vista for contributing to this research result.

## 6 Ethical considerations

The ethical considerations for the multimodal analysis with DESIGNCLIP include the followings:

- **Needs for human supervision.** As with many AI tools, DESIGNCLIP could be misused for non-scientific or adversarial purposes, such as generating infringing designs. The model is intended solely for research, analysis, and balancing the use of technology with human oversight is important to maintain the quality and integrity of patent applications.
- **Legal issues.** Ethical considerations should also include to ensure that the design patents generated from DESIGNCLIP comply with legal requirements and regulations of patent laws.

## 7 Limitations

Patents often include a varying number of views, and many views are difficult to understand even for humans. Additionally, in many cases, patents have new designs on the front view but contain similar side and top views, which challenge the model to distinguish the patterns between different patents. This can lead to a poor alignment and a less effective learning. Thus, we will consider to address the challenges of incorporating multi-views as a future research direction. Moreover, while our current work primarily utilizes the CLIP model due to its feasible and strong contrastive learning framework, exploring more MLLMs, such as LLaVA (Liu et al., 2023), Qwen-VL (Bai et al., 2023), will be a potential future direction.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Caspar J Fall, Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. 2003. Automated categorization in the international patent classification. In *Acm Sigir Forum*, volume 37, pages 10–25. ACM New York, NY, USA.

Junaid Ahmed Ghauri, Eric Müller-Budack, and Ralph Ewerth. 2023. Classification of visualization types and perspectives in patents. In *TPDL*.

ZeMing Gong, Austin T Wang, Joakim Bruslund Haurum, Scott C Lowe, Graham W Taylor, and Angel X Chang. 2024. Bioscan-clip: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv preprint arXiv:2405.17537*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Mariya Hendriksen, Maurits Bleeker, Svitlana Vakulenko, Nanne Van Noord, Ernst Kuiper, and Maarten De Rijke. 2022. Extending clip for category-to-image retrieval in e-commerce. In *European Conference on Information Retrieval*, pages 289–303. Springer.

Kotaro Higuchi, Yuma Honbu, and Keiji Yanai. 2023. Patent image retrieval using cross-entropy-based metric learning. In *IW-FCV*.

Kotaro Higuchi and Keiji Yanai. 2023. Patent image retrieval using transformer-based deep metric learning. *WPI*, 74:102217.

Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. 2023. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Eleni Kamateri, Michail Salampasis, and Eduardo Perez-Molina. 2024. Will ai solve the patent classification problem? *World Patent Information*, 78:102294.

Eleni Kamateri, Vasileios Stamatis, Konstantinos Diamantaras, and Michail Salampasis. 2022. Automated single-label patent classification using ensemble classifiers. In *ICMLC*.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Dylan Myungchul Kang, Charles Cheolgi Lee, Suan Lee, and Wookey Lee. 2020. Patent prior art search using deep learning language model. In *IDEAS*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Michal Kucer, Diane Oyen, Juan Castorena, and Jian Wu. 2022. Deeppatent: Large scale patent drawing recognition and retrieval. In *WACV*.

Yiming Lei, Zilong Li, Yan Shen, Junping Zhang, and Hongming Shan. 2023. Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–412. Springer.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Hao-Cheng Lo, Jung-Mei Chu, Jieh Hsiang, and Chun-Chieh Cho. 2024. Large language model informed patent image retrieval. *arXiv preprint arXiv:2404.19360*.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, and 1 others. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Petra Moser. 2013. Patents and innovation: evidence from economic history. *Journal of economic perspectives*, 27(1):23–44.

Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2022. The role of local alignment and uniformity in image-text contrastive learning on medical images. *arXiv preprint arXiv:2211.07254*.

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19275–19284.

Charles A Rademaker. 2000. The classification of ornamental designs in the united states patent classification system. *World Patent Information*, 22(3):123–133.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Krishan Rana, Andrew Melnik, and Niko Sünderhauf. 2023. Contrastive language, action, and state pre-training for robot learning. *arXiv preprint arXiv:2304.10782*.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Kazuki Shibata, Hideki Deguchi, and Shun Taguchi. 2024. Clip feature-based randomized control using images and text for multiple tasks and robots. *Advanced Robotics*, pages 1–13.

Homaira Huda Shomee, Zhu Wang, Sathya Ravi, and Sourav Medya. 2024. Impact: a large-scale integrated multimodal patent analysis and creation dataset for design patents. *Advances in Neural Information Processing Systems*, 37:125520–125546.

Homaira Huda Shomee, Zhu Wang, Sathya N Ravi, and Sourav Medya. 2025. A survey on patent analysis: From nlp to multimodal ai. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8545–8561.

L Siddharth, Guangtong Li, and Jianxi Luo. 2022. Enhancing patent retrieval using text and knowledge graph embeddings: a technical note. *Journal of Engineering Design*, 33:670–683.

Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. 2024. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems*, 36.

Xiaoning Sun, Tao Fan, Hongxu Li, Guozhong Wang, Peien Ge, and Xiwu Shang. 2024. Clip2tf: Multimodal video-text retrieval for adolescent education. *Displays*, page 102801.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2017. Learning to model the tail. *Advances in neural information processing systems*, 30.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917.

# Appendix

## A Additional Implementation Details

We describe the implementation details of pre-training DESIGNCLIP in our paper. In addition, we adjust specific requirements for patent downstream tasks, such as image retrieval and classification, and these implementation details are provided in this section.

### A.1 Caption generation

One of the challenges with the design patents is lack of textual data. Figure 2 (see "Patent Titles") illustrates that there are only simple texts as titles in design patent data. However, detailed captions along with images are important for enhancing performance in various Natural Language Processing and multimodal tasks (Pan et al., 2023; Hu et al., 2023; Shao et al., 2023). As depicted in Figure 2 (see "Multi-views Images") , when compared to the multiple images associated with a patent, the corresponding patent text typically provides only brief and abstract descriptions, such as the patent title and view orientations. We provide detailed design patent text and image examples in 7.

Specifically, we utilized the LLaVA-1.5-13b model for caption generation, following the IMPACT (Shomee et al., 2024) paper. For each patent, we experiment with different prompts to generate suitable descriptions (captions). Initially, we attempt to create descriptions without mentioning the patent title. When queried about functionality, the responses become vague, such as 'The functionality of the image is to show the design.' Given that design patents differ from the natural images, it is challenging to generate descriptions without the title or object name. For instance, a bear-shaped design can also be an edible cookie. Therefore, we use the following prompts: *This is the image of {patent_title}. What is the shape of the image? What is the functionality of {patent_title}?* The generated descriptions provide enriched textual information, as shown in Figure 2.

### A.2 Image retrieval

We integrate DESIGNCLIP into SOTA patent image retrieval method (Higuchi and Yanai, 2023). We use pre-trained DESIGNCLIP backbones and leave other settings the same. Specifically, the dataset used in this task is DeepPatent with patents from 2018 to 2019, including a train set of 254,787 images, a validation set of 44,815 images, and a test set of 38,834 images. The loss function is Focal loss (Lin et al., 2017). The best hyperparameters are from (Higuchi and Yanai, 2023) which are listed as follows: the batch size is 256, the optimizer is AdamW, the learning rate is $1e^{-4}$ and the number of training epochs is 25.

### A.3 Patent classification

The dataset of design patents has 33 classes. For the classification task, we use data from 2023. We show design categories and number of examples of our train data in Table 7. In patent classification, we finetune a simple linear classifier with frozen CLIP (Radford et al., 2021) and DESIGNCLIP backbones (e.g., RN101 (He et al., 2016), ViT-B (Dosovitskiy et al., 2020)) for comparisons. The hyerparameters are listed as follows: the batch size is 32, the optimizer is AdamW, the learning rate is $1e^{-4}$, and the number of training epochs is 15.

## B More Qualitative Analysis

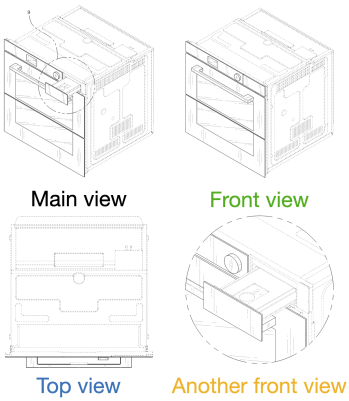
In this section, we demonstrate additional qualitative analysis. We begin with providing details of visualization of feature embeddings. Then, we show some multimodal retrieval examples.

### B.1 More details of visualization of feature embeddings.

Figure 5 in the main paper compares UMAP (McInnes et al., 2018) visualizations of feature embeddings from CLIP and DESIGNCLIP models for patent images. The patent data are from 2023, including 8,643 patents. We focus on the top most common categories: transportation, recording, lighting, games, furnishings and apparel. The CLIP model shows overlapping categories with less distinct clustering which indicates mixed feature recognition across various categories like transportation, games, and apparel. On the other hand, DESIGNCLIP shows distinct clusters for almost all the categories. The clear clusters in the DESIGNCLIP visualization show that it is very effective for accurately classifying and finding patent images.

### B.2 Additional class-aware classification analysis

We present the performance comparisons of CLIP and DESIGNCLIP with backbone of RN-101 on classification tasks. As results shown in Figure 8a, DESIGNCLIP improves significantly on the all top classes. Figure 8b shows that CLIP is not able to




(a) Example patent images with multiple views

**Title:** Oven **Patent No:** US D974110

**CLAIM**

*We claim the ornamental design for an oven, as shown and described.*

**DESCRIPTION**

FIG. **1** is a front perspective view of an oven showing our new design;
FIG. **2** is a front view thereof;
FIG. **3** is a rear view thereof;
FIG. **4** is a left-side view thereof;
FIG. **5** is a right-side view thereof;
FIG. **6** is a top view thereof;
FIG. **7** is a bottom view thereof;
FIG. **8** is another front perspective view thereof showing a state of use; and,
FIG. **9** is an enlarged view of the encircled portion in FIG. **8**.

The evenly-dashed broken lines in the figures depict portions of the oven which form no part of the claimed design. The dot-dot-dash broken lines encircling portions of the claimed design that are illustrated in enlargements form no part of the claimed design.

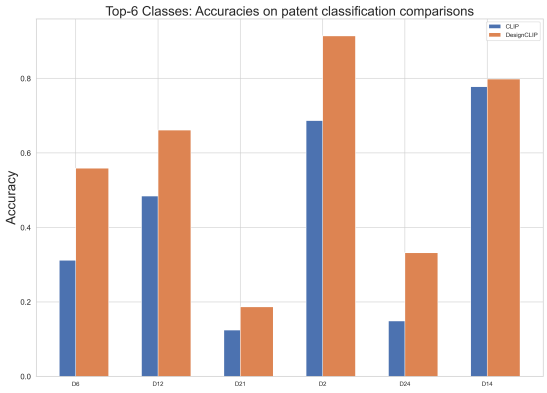
**1 claim, 9 drawing sheets**

(b) Example patent texts

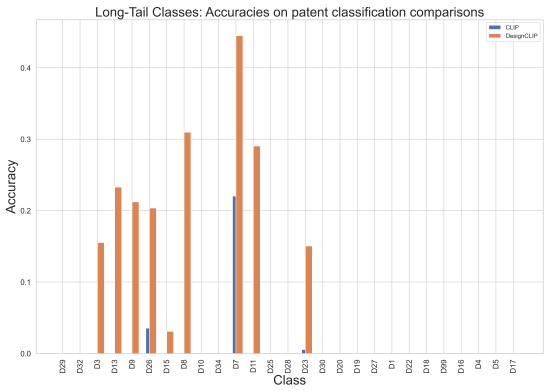
Figure 7: Example of design patent for an oven. Figure (a) shows main, front, top and enlarged front view of a single design. Figure (b) shows patent texts that includes title, Patent ID, single line claim, and description of figures.

Table 7: The table shows the list of U.S. design patent classes and number of occurrences in our classification task.

| Class | Description | Occurences |
|---|---|---|
| D1 | Edible Products | 38 |
| D2 | Apparel and Haberdashery | 930 |
| D3 | Travel Goods, Personal Belongings, and Storage or Carrying Articles | 462 |
| D4 | Brushware | 122 |
| D5 | Textile or Paper Yard Goods; Sheet Material | 20 |
| D6 | Furnishings | 1052 |
| D7 | Equipment for Preparing or Serving Food or Drink Not Elsewhere Specified | 906 |
| D8 | Tools and Hardware | 735 |
| D9 | Packages and Containers for Goods | 525 |
| D10 | Measuring, Testing or Signaling Instruments | 444 |
| D11 | Jewelry, Symbolic Insignia, and Ornaments | 369 |
| D12 | Transportation | 1261 |
| D13 | Equipment for Production, Distribution, or Transformation of Energy | 755 |
| D14 | Recording, Communication, or Information Retrieval Equipment | 1943 |
| D15 | Machines Not Elsewhere Specified | 512 |
| D16 | Photography and Optical Equipment | 359 |
| D17 | Musical Instruments | 35 |
| D18 | Printing and Office Machinery | 55 |
| D19 | Office Supplies; Artists' and Teachers' Materials | 146 |
| D20 | Sales and Advertising Equipment | 39 |
| D21 | Games, Toys and Sports Goods | 962 |
| D22 | Arms, Pyrotechnics, Hunting and Fishing Equipment | 184 |
| D23 | Environmental Heating and Cooling, Fluid Handling and Sanitary Equipment | 743 |
| D24 | Medical and Laboratory Equipment | 1044 |
| D25 | Building Units and Construction Elements | 179 |
| D26 | Lighting | 901 |
| D27 | Tobacco and Smokers' Supplies | 136 |
| D28 | Cosmetic Products and Toilet Articles | 329 |
| D29 | Equipment for Safety, Protection and Rescue | 80 |
| D30 | Animal Husbandry | 347 |
| D32 | Washing, Cleaning or Drying Machines | 221 |
| D34 | Material or Article Handling Equipment | 127 |
| D99 | Miscellaneous | 26 |




(a) Patent classification comparisons on top 6 classes




(b) Patent classification comparisons on tail classes

Figure 8: Comparisons between CLIP-RN101 and DESIGNCLIP-RN101 in classification for top-6 classes and tail classes. DESIGNCLIP improves significantly over CLIP on the all top classes.

predict on most of the tail classes, but DESIGNCLIP provides accurate predictions on 9 long-tail classes. Therefore, we believe that DESIGNCLIP can improve the performance on patent classification and class-aware learning is effective for long-tail class distribution domain data.

### B.3 Image retrieval examples

In this task, we present three examples of image-to-image retrieval using CLIP-ViT-B and DESIGNCLIP-ViT-B as backbones with ArcFace (Deng et al., 2019; Higuchi and Yanai, 2023) based on a given query image. The goal is to identify images similar to the query image. We show the top five retrieval results for both CLIP and DESIGNCLIP for qualitative analysis.

In Example 1, shown in Figure 9, CLIP fails to retrieve a single relevant image, while DESIGNCLIP correctly retrieves the top 4 out of 5 images. In Figure 10, our model achieves perfect retrieval, identifying all 5 relevant images in different views. We observe that CLIP only can retrieve images with similar shape but fails to capture patent class and different views. Therefore, DESIGNCLIP pre-trained with class-aware information and multi-views can improve performance in image retrieval tasks. However, in Figure 11, both CLIP and our model demonstrate similar performance. We leave the analysis of unsuccessful cases—such as Figure 11—for our future work.

### B.4 Multimodal retrieval examples.

We present four examples of multimodal image retrieval. For each example, based on a given query (caption) and a ground truth (GT) image, the objective is to identify similar images to the GT. The backbone is ViT-B. We show the top five retrieval results for both the CLIP model and DESIGNCLIP for case studies of their performance comparisons in matching images to textual descriptions.

- Example 1: *Text Query:* The image is a drawing of a wheel, which is a circular object with a central hub and spokes.

  *Ground truth image:* D0862341.TIF, where D0862341 is patent id.

- Example 2: *Text Query:* The image is a drawing of a control valve. The control valve is a device used to regulate the flow of fluid, such as water, steam, or gas, in a system.

  *Ground truth image:* D0858713.TIF

- Example 3: *Text Query:* The image is a truck vehicle grille. The grille serves as a protective covering for the front of the truck, covering the engine and radiator.

  *Ground truth image:* D0850330.TIF

- Example 3: *Text Query:* The image is a square-shaped vehicle floor mat, which is designed to provide comfort and protection for the vehicle floor.

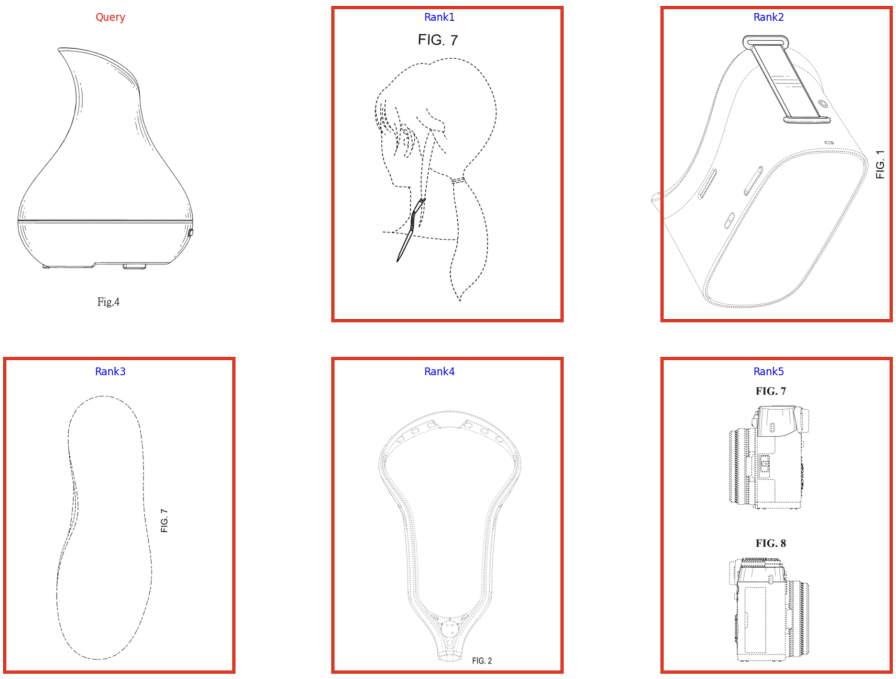
  *Ground truth image:* D0845852.TIF

**Our DESIGNCLIP model retrieves successfully the ground truth images as the top results in Figures 12, 13, and 15.** In 14, the ground truth image was retrieved as the second top result. Specifically, as shown in Figure 12, both models can retrieve wheels. However, DESIGNCLIP is pre-trained on our proposed methods tailored on

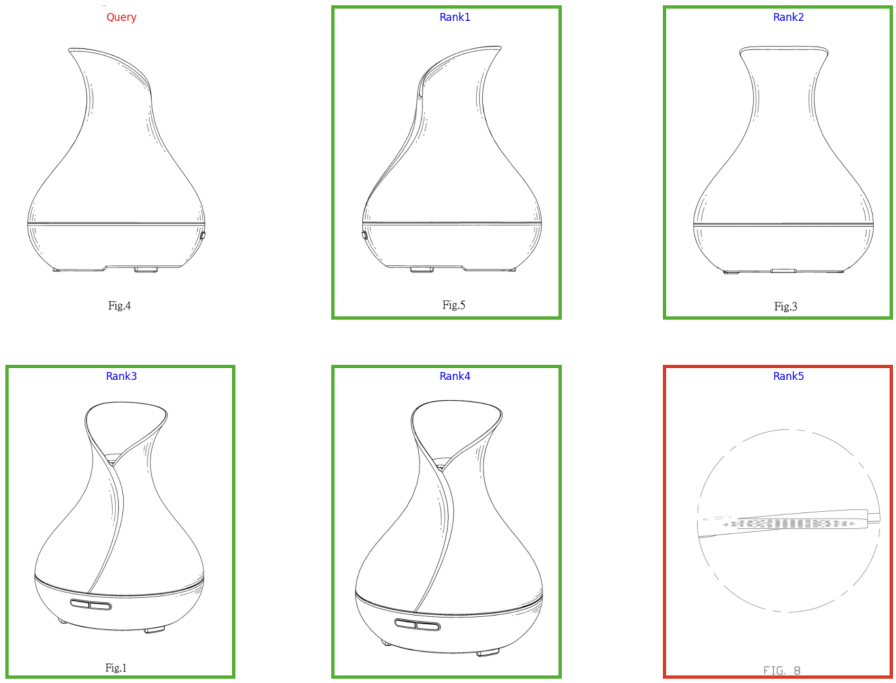
| Column | Example | Description |
|---|---|---|
| title | Electronic device with graphical user interface | The design's title |
| id | D0964399 | The patent document number serves as a means of identifying the patent. |
| claim | The ornamental design for an electronic device with graphical user interface, as shown and described. | A design patent application includes only a single claim |
| date | 20220920 | Date of the patent's publication |
| class | D14485 | U.S. design patent category. |
| no_figs | 2 | Number of figures for the design |
| sheets | 2 | The quantity of design sheets given for the figures. Some of the sheets has multiple figures |
| fig_desc | ['FIG. 1 is a front view of a display screen or portion thereof with graphical user interface showing the claimed design; and,', 'FIG. 2 is a front view of an electronic device having a display screen with the graphical user interface of FIG. 1 applied to the display screen.'] | The representation of every drawing view, including front, top, perspective, and others, is described in the figure descriptions. |
| caption | The image is square-shaped and features a graphical user interface on an electronic device. The functionality of such a device is to provide users with an interactive and visually appealing interface to access and control various features and applications. The graphical user interface makes it easier for users to navigate and interact with the device, enhancing the overall user experience. | Detailed descriptions of the design's shape and functionality |

Table 8: Description of some of the fields in the CSV file extracted from the XML files for each design patent. The description explains what each column means and includes examples from the patent ID D14485.

the design patent data, it is able to capture the details of design which demonstrates DESIGNCLIP can provide better capability for prior art search and design inspirations. Moreover, in 15, although CLIP retrieved ground truth images, other retrieved images are vehicles, not a floor mat. However, DESIGNCLIP can retrieve images that are more relevant for vehicle mats. In fact, DESIGNCLIP learns better representations for understanding design patent data in multimodal scenarios.

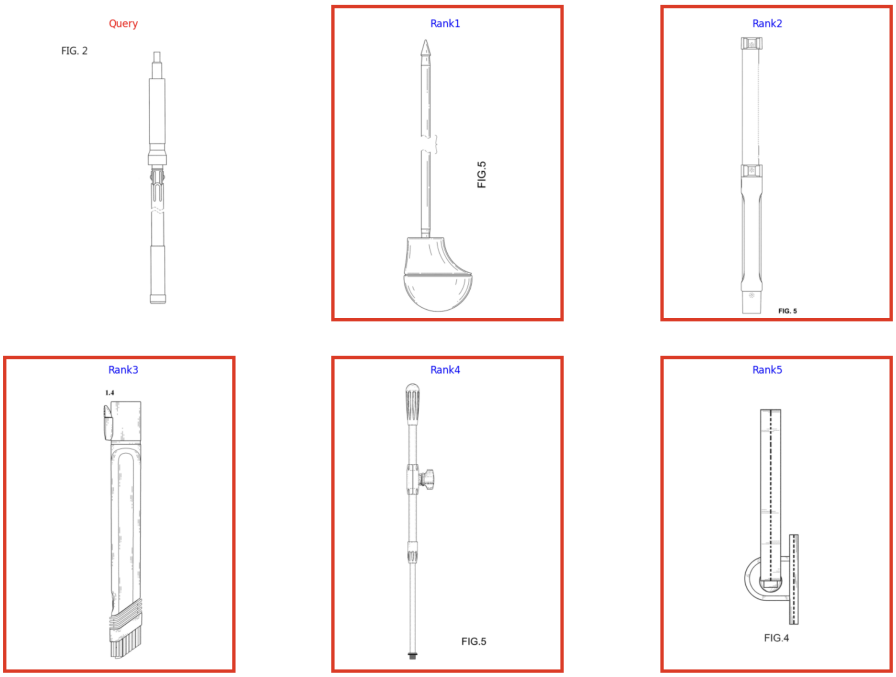



(a) Image retrieval results of CLIP-ViT-B + ArcFace

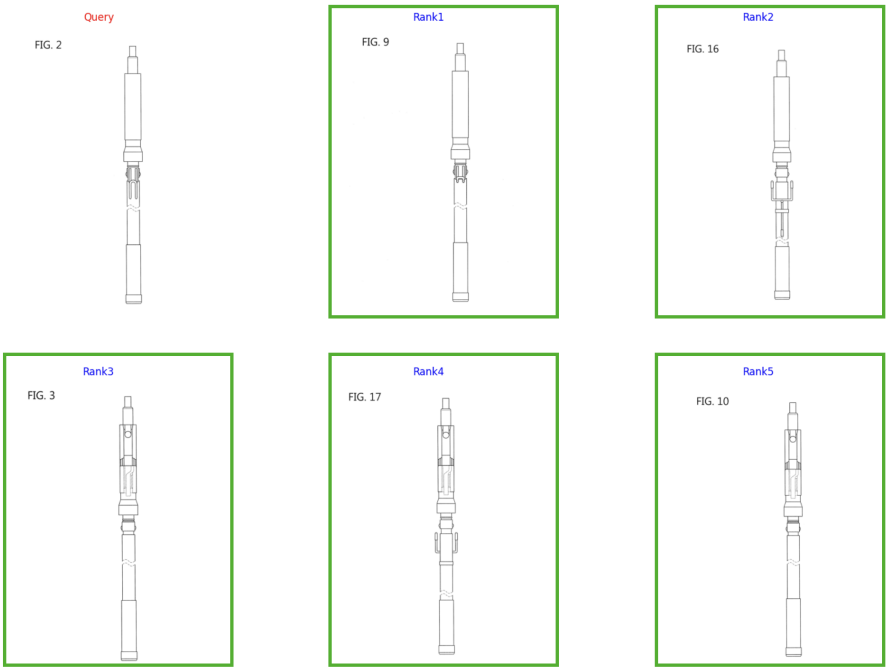



(b) Image retrieval results of DESIGNCLIP-ViT-B + ArcFace

Figure 9: Image Retrieval example 1 of patent D0815721. (a) and (b) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Green box denotes to the correct image. DESIGNCLIP can retrieve the image correctly.

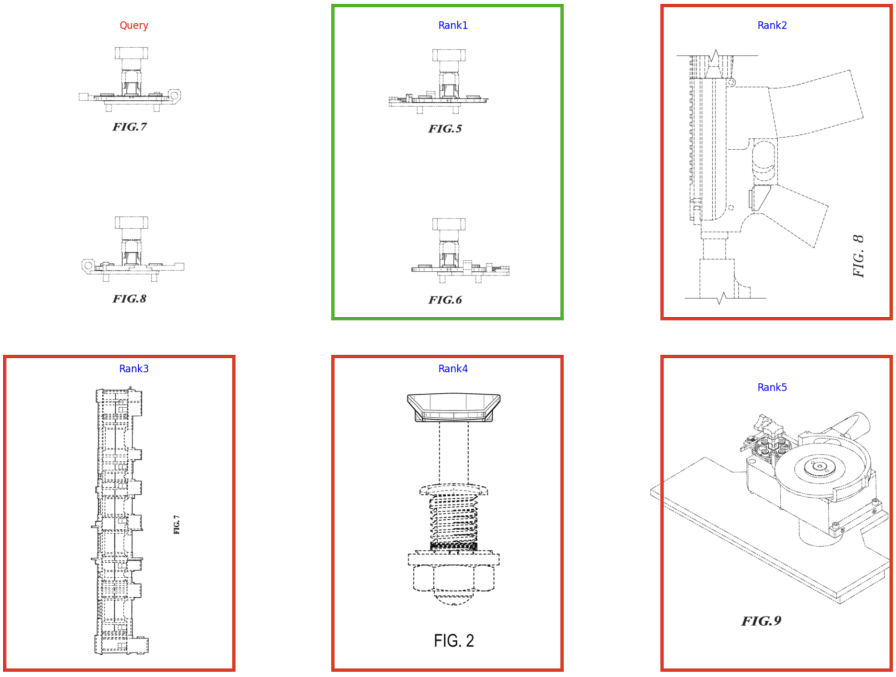



(a) Image retrieval results of CLIP-ViT-B + ArcFace

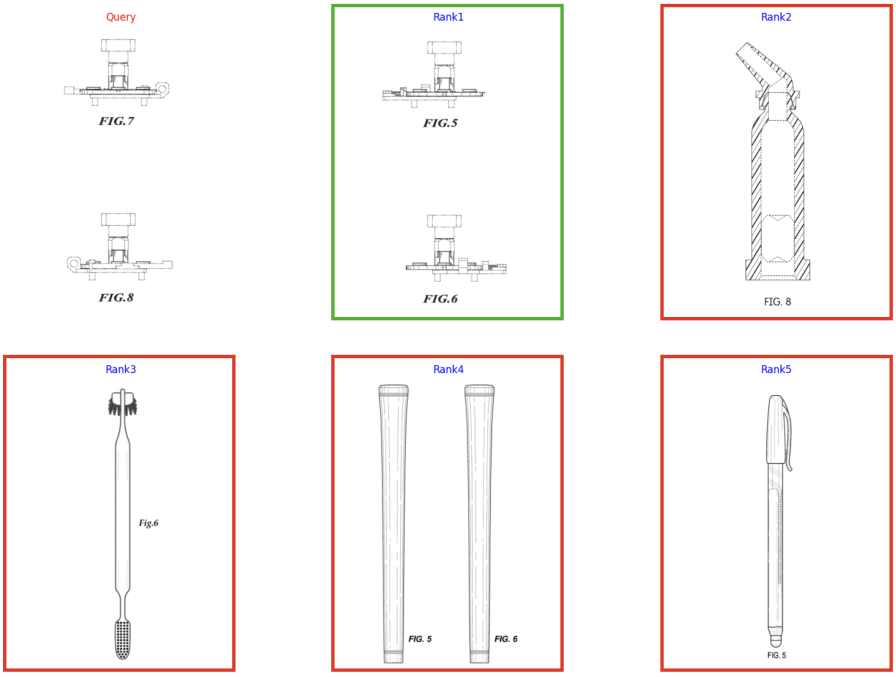



(b) Image retrieval results of DESIGNCLIP-ViT-B + ArcFace

Figure 10: Image Retrieval example 2 of patent D0817138. (a) and (b) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Green box denotes to the correct image. DESIGNCLIP can retrieve the image correctly.

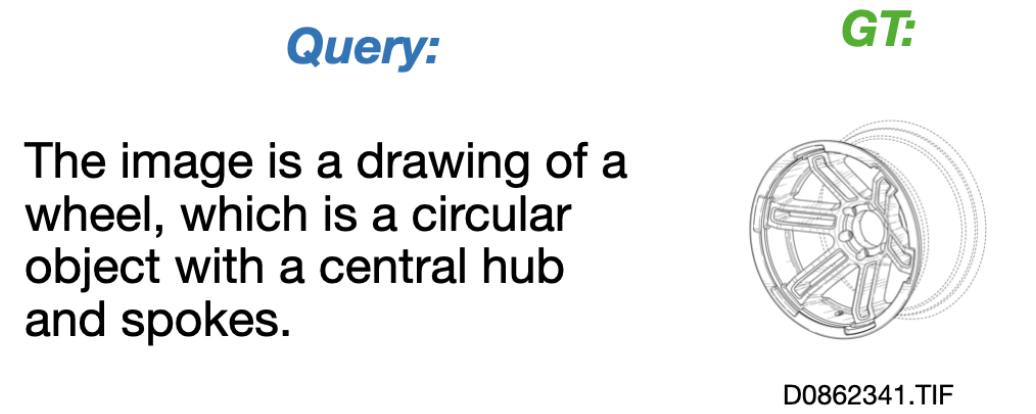



(a) Image retrieval results of CLIP-ViT-B + ArcFace

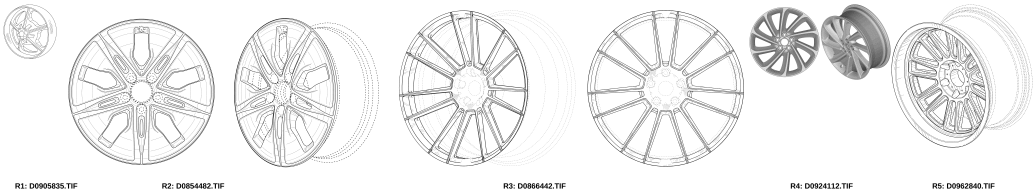



(b) Image retrieval results of DESIGNCLIP-ViT-B + ArcFace

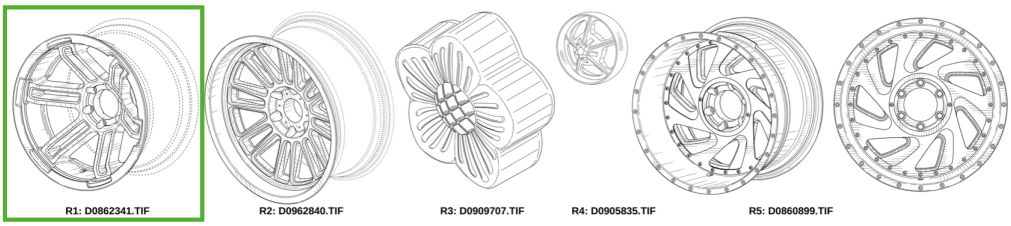
Figure 11: Image Retrieval example 3 of patent id D0827684. (a) and (b) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Green box denotes to the correct image. CLIP and DESIGNCLIP only retrieves rank 1 image correctly.




(a) Text query and ground truth image

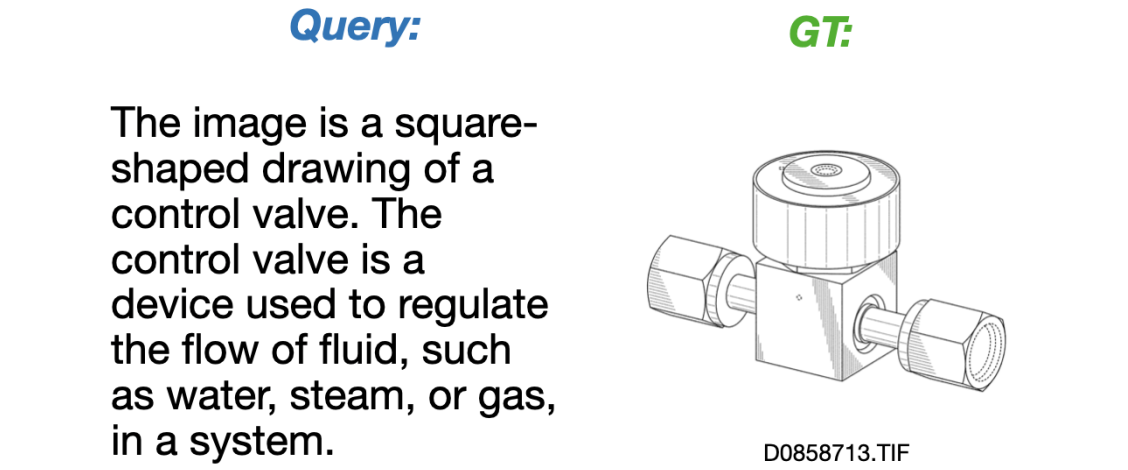



(b) Retrieval results of CLIP

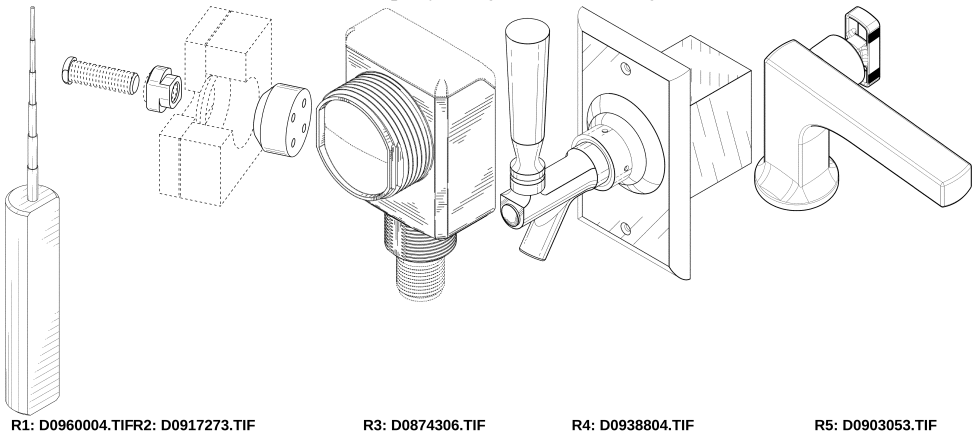



(c) Retrieval results of DESIGNCLIP

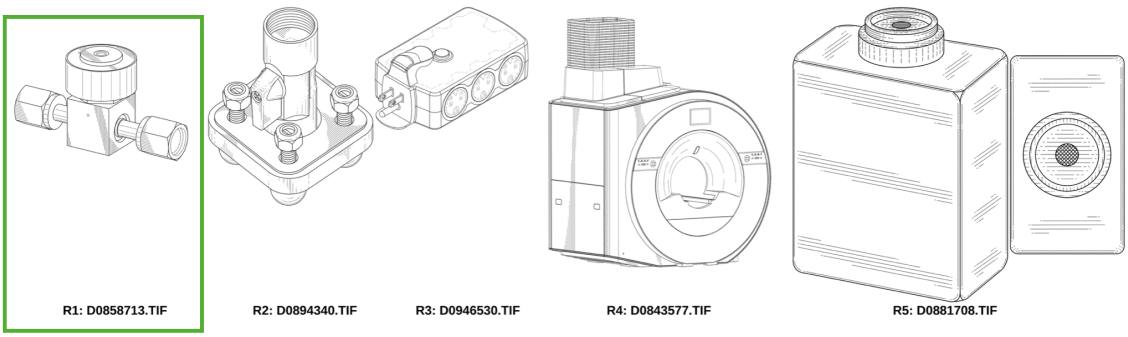
Figure 12: Text-image Retrieval example 1. Text query and ground truth image are shown in (a). (b) and (c) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. DESIGNCLIP can retrieve the image correctly.




(a) Text query and ground truth image

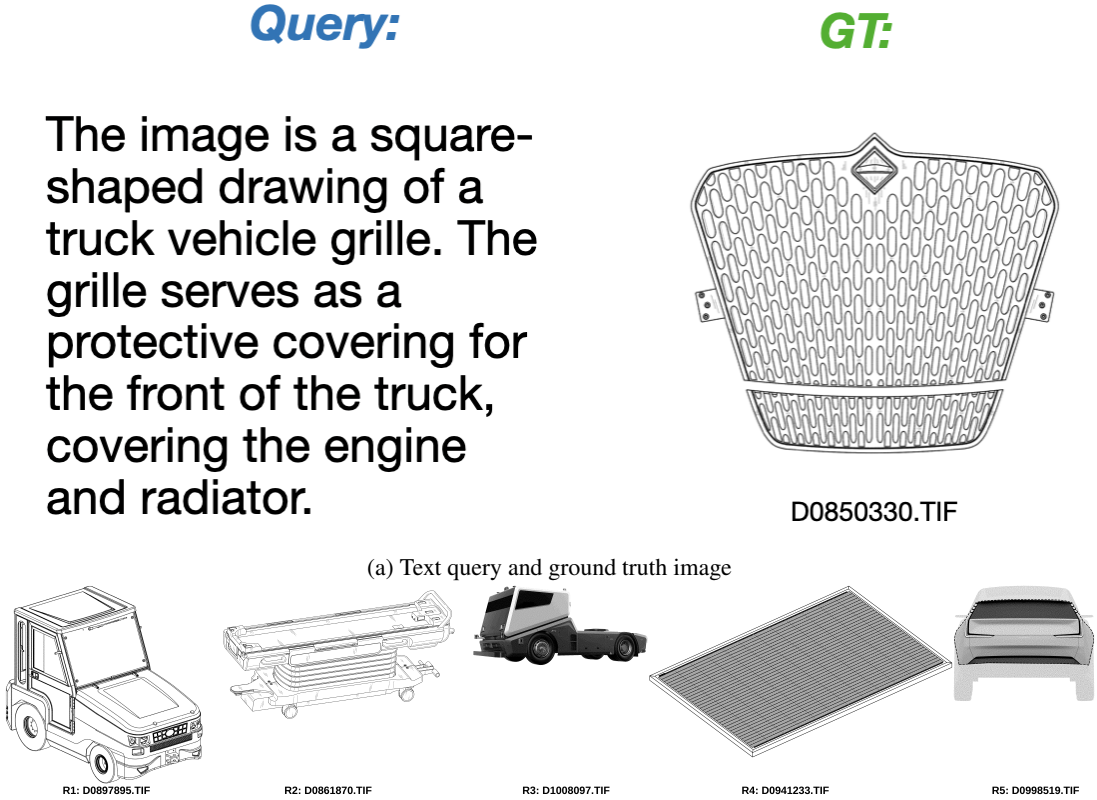



(b) Retrieval results of CLIP




(c) Retrieval results of DESIGNCLIP

Figure 13: Text-image Retrieval example 2. Text query and ground truth image are shown in (a). (b) and (c) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. DESIGNCLIP can retrieve the image correctly.

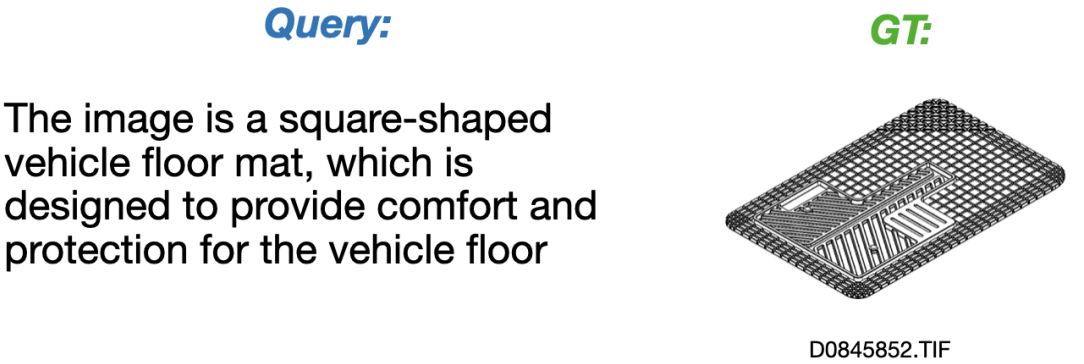



(a) Text query and ground truth image

(b) Retrieval results of CLIP

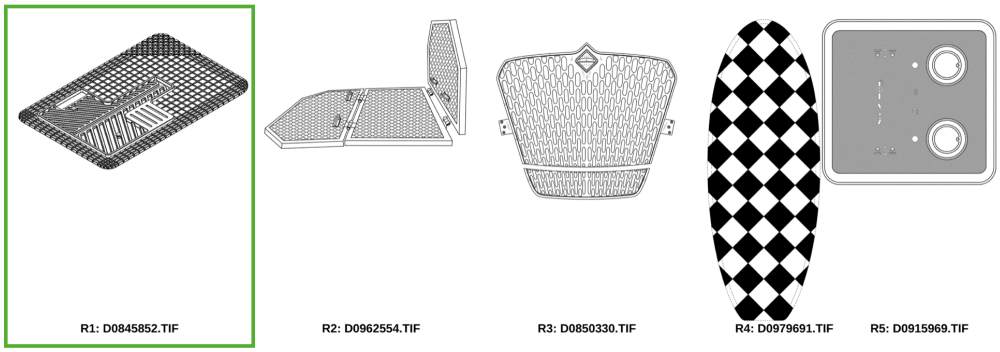
(c) Retrieval results of DESIGNCLIP

Figure 14: Text-image Retrieval example 3. Text query and ground truth image are shown in (a). (b) and (c) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. DESIGNCLIP can retrieve the image correctly.




(a) Text query and ground truth image


(b) Retrieval results of CLIP




(c) Retrieval results of DESIGNCLIP

Figure 15: Text-image Retrieval example 4. Text query and ground truth image are shown in (a). (b) and (c) are top 5 retrieval results of CLIP and DESIGNCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. Both DESIGNCLIP and CLIP can retrieval the ground truth image, but DESIGNCLIP retrieve top 1 image correctly.