A Appendix

A.1 Additional Details on Our Dataset

Number of words in a title. Design patents typically contain limited textual information. One of the fields is the title, which describes the object the design patent protects. Figure 5 shows the number of words in each title over the years. It is noteworthy that the titles average only about 3 to 3.5 words, indicating a trend towards concise and specific naming conventions in design patent filings.

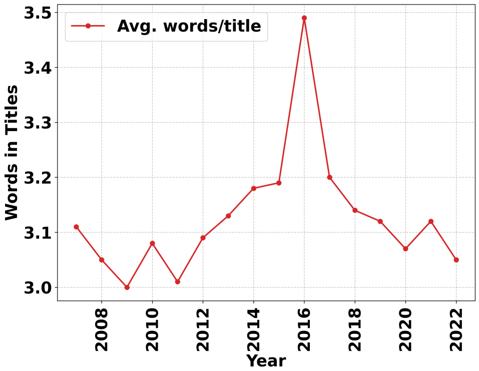


Figure 5: The distribution of the number of words in the title of the design patents over the years. This shows that the average number of words in the title of each patent is between 3 and 3.5.

Frequent objects. Figure 6 displays the top 10 objects that occur the most frequently in the data set. Display screens are the most common, with a frequency of 13,810, followed by shoes and bottles with frequencies of 4,327 and 3,718, respectively. Other frequently appearing objects include containers, mobile phones, chairs, tires, shoe uppers, electric devices, and faucets.

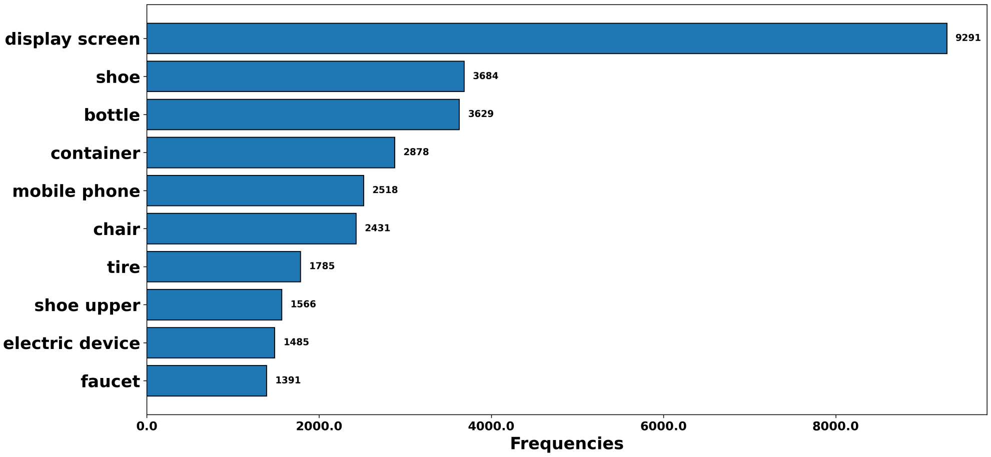


Figure 6: Distribution of the top 10 objects in the dataset by frequency. Display screens, including different types such as graphical user interfaces, animated graphical user interfaces, and transitional graphical user interfaces, are the most common objects in the dataset.

Example of the fields. In the constructed CSV files, we have 11 fields for each design patent. Table 5 shows the descriptions of the 11 fields extracted from the XML files with an example. The description defines the meaning of each column, and the examples of the fields are provided for the patent ID D0908314.

Table 5: Description of fields in the CSV file. There are 11 fields extracted from the XML files for each design patent. The description defines the meaning of each column, and the examples of the fields are provided for the patent ID D0908314.

Column	**Description**	**Example**
`title`	The title of the design is the name commonly recognized and used by the public	Garment with a side pocket
`id`	Document number of the patent starts with a 'D' followed by a series of numbers, which uniquely identifies the patent.	D0908314
`claim`	A design patent application includes only a single claim that defines the design the applicant wishes to patent, specifying the article in which the design is embodied or to which it is applied.	The ornamental design for a garment with a side pocket, as shown and described.
`date`	Publication date of the patent	20210126
`class`	U.S. design patent category under which the patent is classified	D2728, D2840
`class_search`	U.S. classification codes aiding in determining its scope and relevant prior art	['0202', 'D2728', 'D2839', 'D2829', 'D2750', 'D2839', 'D 2750', 'D2857', 'D2840', 'D2829', 'D 2840', 'D2840', 'D2840', 'D2829', 'D 2839', 'D2840', 'D21804', '293', '224153', 'D2712', 'D2720', 'D2750', 'D2831', 'D2853', 'D2865', 'D2873', 'D2874', 'D2878', 'D2840', 'D2839', 'D2857', 'D2728', 'D21801-805']
`inv_country`	Country of the inventors	US
`no_figs`	Number of figures for the design	7
`sheets`	Number of design sheets provided for the figures. Some of the sheets has multiple figure views	4
`file_names`	The filenames that contain the images of the particular design.	['USD0908314-20210126-D00000.TIF', 'USD0908314-20210126-D00001.TIF', 'USD0908314-20210126-D00002.TIF', 'USD0908314-20210126-D00003.TIF', 'USD0908314-20210126-D00004.TIF']
`fig_desc`	The Figure Descriptions specify the representation of each drawing view, such as front view, top view, perspective, and others	['FIG. 1 is a front left perspective view of the garment with a side pocket, showing my new design', 'FIG. 2 is a front view', 'FIG. 3 is a rear view', 'FIG. 4 is a left side view', 'FIG. 5 is a right side view', 'FIG. 6 is a top view thereof', 'FIG. 7 is a bottom view']
`caption`	Elaborated captions for the design which includes shape and functionality	The image is a square-shaped illustration of a garment with a side pocket. The functionality of the image is to showcase the design and features of the garment, such as the pocket, which can be useful for potential customers or designers to visualize the product and its details.

Frequent USPC class. Table 6 lists the top 10 most frequent class-subclass occurrences in the dataset. The most common class-subclass is D14-486, "Recording, Communication, or Information Retrieval Equipment," with 7,618 occurrences, specifically describing drop-down or full-screen menu types. Other notable entries include D14-485, D26-28, and D12-209, covering generated images, vehicle lamps or casings, and transportation apertures, respectively.

Table 6: Overview of the 10 Class-Subclass Occurrences. This lists the frequency of occurrences for each class-subclass, along with descriptions that specify the general category and particular functionalities or features characterized by each subclass. Note that some patents belong to multiple classes. For simplicity, we have counted only the primary class and that results in a single classification code for each patent.

Class-Subclass	**Occurrence**	**Class Description**	**Subclass Description**
D14-486	7359	Recording, Communication, or Information Retrieval Equipment	Drop down menu or full screen menu type
D14-485	5363	Recording, Communication, or Information Retrieval Equipment	Generated image
D26-28	4149	Lighting	Vehicle lamp or casing
D12-209	2700	Transportation	Aperture or simulated aperture
D2-972	2432	Apparel and Haberdashery	Vamp, toe, heel, or side panel
D12-169	2399	Transportation	Vehicle-attached front or rear type
D14-250	2320	Recording, Communication, or Information Retrieval Equipment	Cover for base or handset
D13-147	2201	Equipment for Production, Distribution, or Transformation of Energy	Linear array of identical repeating ports or contacts (i.e., in-line array)
D14-488	2022	Recording, Communication, or Information Retrieval Equipment	Visible shutter
D14-126	1998	Recording, Communication, or Information Retrieval Equipment	Receiver or monitor

Example viewpoints. We provide examples of the drawings that are available in the design patents. Figure 7 shows four viewpoints of a design object titled as *ceiling fan*. 7a and 7b represent top and bottom perspectives respectively, whereas 7c and 7d are top and bottom view, respectively.

A.2 Generated Captions Examples and Comparisons

For caption generation, we use LLaVA, GPT-4o, and Qwen-VL. Qwen-VL often lacks functional descriptions and occasionally includes Chinese words within English captions. Although GPT-4o provides high-quality captions, it comes with a higher cost. Figure 8 illustrates an example of a design patent along with the captions generated by each model.

A.3 US Design Patent Classification

The subject matter of U.S. design patents is categorized into 33 distinct classes[9]. Unlike utility patents, design patents are organized by classes and subclasses only. Table 7 lists all the categories of the design patents.

A.4 PatentCLIP

We also provide a PatentCLIP which is finetuned from OpenAI's CLIP pre-trained models with IMPACT dataset. In this section, we illustrate the implementation details and qualitative analysis using the learned feature embedding space from IMPACT.

---

[9] https://www.uspto.gov/patents/laws/examination-policy/seven-classification-design-patents

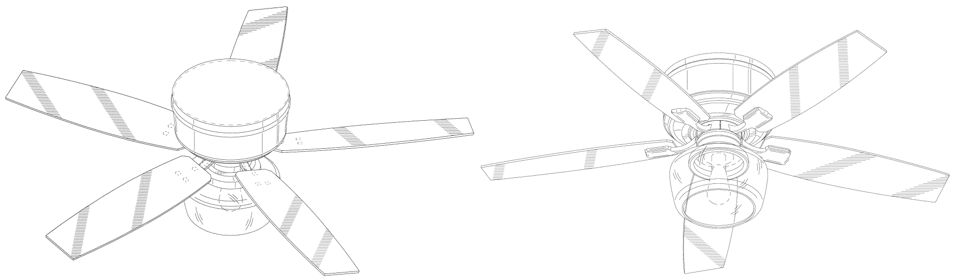


FIG. 1

(a) Top perspective

FIG. 2

(b) Bottom perspective

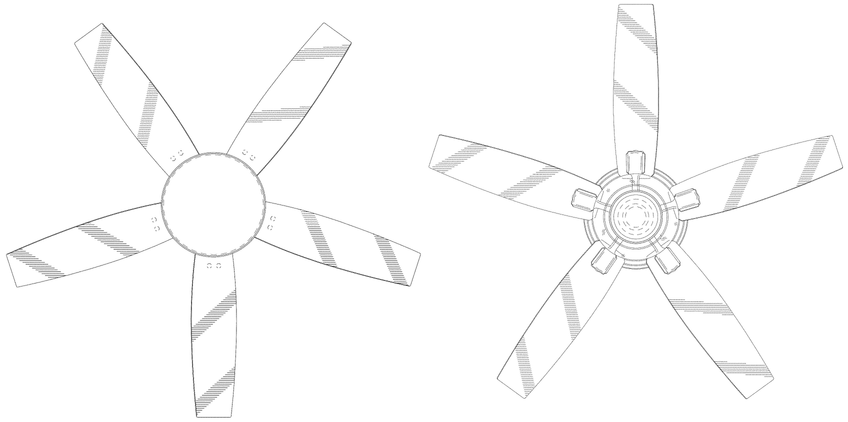


FIG. 3

(c) Top view

FIG. 4

(d) Bottom view

Figure 7: An example of image perspectives and viewpoints in a design patent. The patent is about the design of a ceiling fan.

A.4.1 Implementation details

We use an open source implementation of CLIP [10]. The models are ResNet50, ResNet101, ViT-B-32 and ViT-L-14. The hyperparameters for funetining is listed as follows: learning rate is $5e-6$, weight decay is 0.1, and optimizer is AdamW for all models. The batch size is 256 except 64 for ViT-L-14. All settings are same to image-title and image-caption pairs. All finetuning and inference are conducted on 4 NVIDIA V100 GPUs.

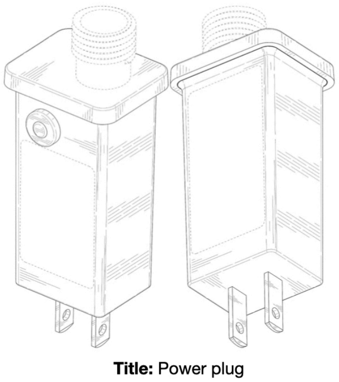
A.4.2 Qualitative analysis

To analysis the effective of finetuning CLIP on IMPACT dataset, we visualize the learned image features and text features for sample patents using U-MAP projection [37]. The sample patents are selected from the top 4 subclass in the recent five years data, including D12-209, D14-485, D14-486 and D26-28. In total, there are 5,699 patents. All the model backbone is ViT-B-32. For text features, we visualize the embeddings for the captions.

Feature embedding spaces of multiple modalities on sample IMPACT dataset are shown in Fig 9. Different colors representing the clusters of the corresponding classes. We observe that PatentCLIP can identify clusters over the extracted image features better than CLIP, see 9a, 9c and 9e. Comparing with PatentCLIP-title, PatentCLIP have the better clustering performance on the extracted text features, it can identify D14 is far to D12. Note that, CLIP also can identify text feature clusters because that the captions are generated with VLMs. However, CLIP is not able to classify the patent images. Indeed, we can see that different classes cluster clearly, and similar subclasses are often close in the embedding spaces, such as D14-485 and D14-486 (see class description in Tab 6). Therefore,

---

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



Caption (LLaVA): The image is a square-shaped drawing of two power plugs. The functionality of a power plug is to provide a connection between an electrical device and an electrical outlet, allowing the device to receive power from the outlet. Power plugs are essential for the proper functioning of various appliances and electronic devices, such as computers, and televisions. They ensure that the devices have access to the necessary electrical supply to operate efficiently and safely.

Caption (GPT-4o): The image depicts a power plug with a rectangular, box-like shape featuring a flat top and bottom. It has a cylindrical connector at the top, designed for insertion into a corresponding socket. The front side includes a circular button, likely for functionality such as power control or reset. The overall design emphasizes a compact and functional form, suitable for efficient electrical connectivity.

Caption (Qwen-VL): The image of Power plug is a rectangular shape with a round button on the upper left side.

Figure 8: Captions generated by LLaVA, GPT-4o, and Qwen-VL of a power plug design. The shapes are highlighted in yellow.

we believe that finetuning VLMs with IMPACT is beneficial in the specific patent domain for many downstream tasks, such as classification and retrieval.

A.5 Multimodal Retrieval

A.5.1 Additional results

We provide additional results on the performance of finetuned PatentCLIP-title on Text-Image and Image-Text retrievals. All the training hyper-parameters are same as the PatentCLIP model (see Sec A.4). The dataset used here is the image and title pairs in IMPACT. As results shown in Table 8, the performance pattern is similar to PatentCLIP, which is that the more advanced models and the models with more parameters perform better. Comparing the results of PatentCLIP which is finetuned on image and caption pairs in Table 3 is similar except for R@1 results of ResNet family. Others R@K results of finetuning image-captions are significantly improved than PatentCLIP-title.

A.5.2 Qualitative analysis

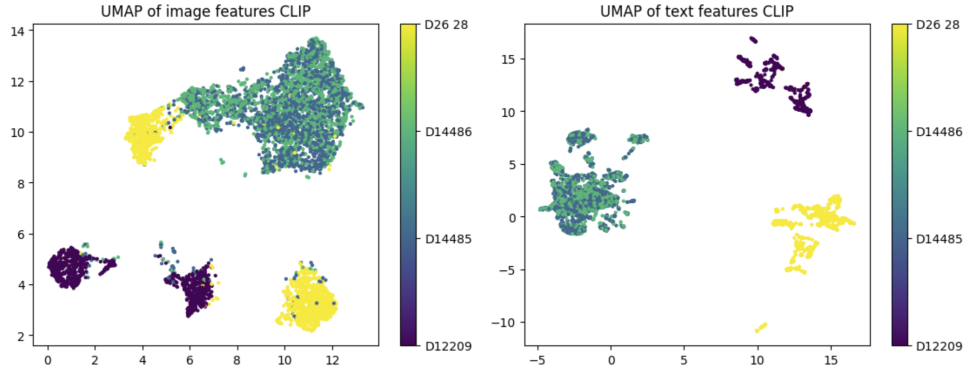
To further analyze the multiomodal retrieval results, we demonstrate three Text-image retrieval examples as follows.

- Example 1: *Text Query:* The image is a square-shaped drawing of a protective case for a game controller. *Ground truth image:* D1006114.TIF
- Example 2: *Text Query:* The image is a black and white drawing of a computer mouse, which is a device used for controlling and interacting with a computer. *Ground truth image:* D0943581.TIF
- Example 3: *Text Query:* The image is a white drawing of an FM transmitter, which is a device used to transmit audio signals through the air using frequency modulation (FM) technology. *Ground truth image:* D0985524.TIF

Figure 10 shows that PatentCLIP-title and PatentCLIP are able to retrieval game controller, but CLIP only can recall the items with square-shaped. All top 5 retrieved images of PatentCLIP are related with game controller, and the top 1 result is correct. As results shown in Fig 11, both CLIP and PatentCLIP obtained the correct images in top 5 set, but PatentCLIP produce the top 1 image correctly. Other four images and the images obtained from PatentCLIP-title are not relevant with the text query. In Figure 12, only PatentCLIP retrieve the top 1 image correctly. We see that the retrieved images by our PatentCLIP model are relevant to the text queries not only in the shape but also in terms of the function. Thus, we conclude that captions provide more information for VLMs to learn the patents. Our PatentCLIP model is also helpful for the prior art search and the design inspiration.

A.6 3D construction Examples

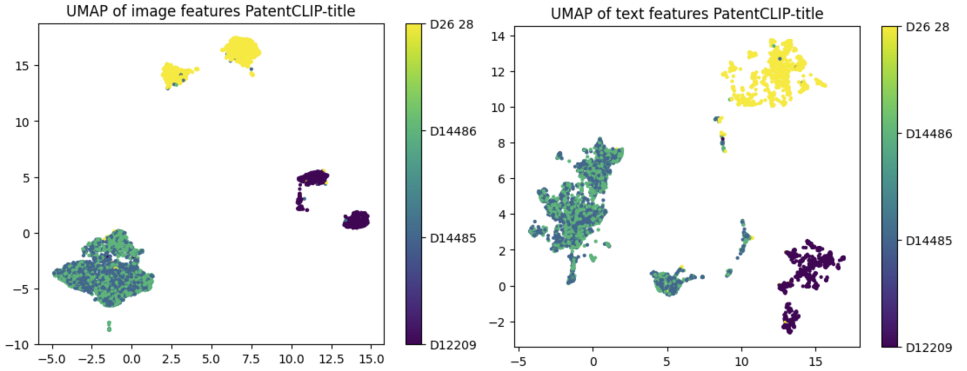
We provide two more detailed 3D constructions examples in Fig. 13. We utilize ControlNet [58] to generate 3D photos for patent images. Comparing the results of prompting with IMPACT captions


UMAP of image features CLIP

UMAP of text features CLIP

(a) Image features with CLIP

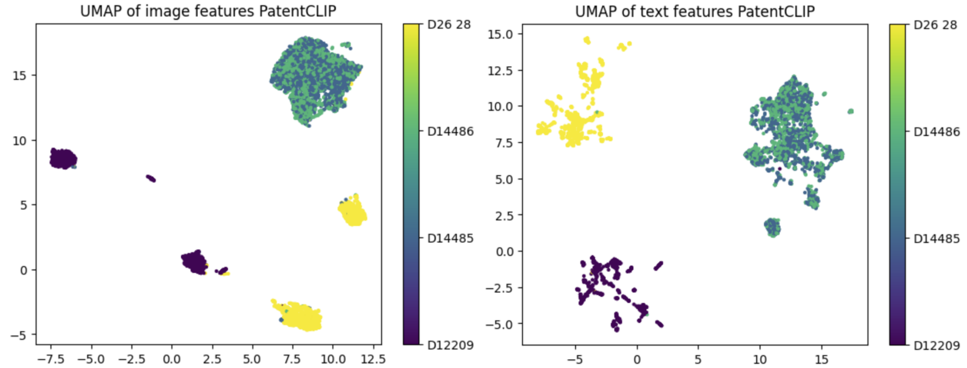
(b) Text features with CLIP


UMAP of image features PatentCLIP-title

UMAP of text features PatentCLIP-title

(c) Image features with PatentCLIP-title

(d) Text features with PatentCLIP-title


UMAP of image features PatentCLIP

UMAP of text features PatentCLIP

(e) Image features with PatentCLIP

(f) Text features with PatentCLIP

Figure 9: UMAP feature embeddings for sample patent images. (a) Visualization using CLIP models (b) Visualization using PatentCLIP finetuned on title (c) Visualization using PatentCLIP finetuned on caption. PatentCLIP shows well formed clusters in both image and text-based features.

Table 7: The table shows the list of U.S. design patent classes.

Class	**Description**
D1	Edible Products
D2	Apparel and Haberdashery
D3	Travel Goods, Personal Belongings, and Storage or Carrying Articles
D4	Brushware
D5	Textile or Paper Yard Goods; Sheet Material
D6	Furnishings
D7	Equipment for Preparing or Serving Food or Drink Not Elsewhere Specified
D8	Tools and Hardware
D9	Packages and Containers for Goods
D10	Measuring, Testing or Signaling Instruments
D11	Jewelry, Symbolic Insignia, and Ornaments
D12	Transportation
D13	Equipment for Production, Distribution, or Transformation of Energy
D14	Recording, Communication, or Information Retrieval Equipment
D15	Machines Not Elsewhere Specified
D16	Photography and Optical Equipment
D17	Musical Instruments
D18	Printing and Office Machinery
D19	Office Supplies; Artists' and Teachers' Materials
D20	Sales and Advertising Equipment
D21	Games, Toys and Sports Goods
D22	Arms, Pyrotechnics, Hunting and Fishing Equipment
D23	Environmental Heating and Cooling, Fluid Handling and Sanitary Equipment
D24	Medical and Laboratory Equipment
D25	Building Units and Construction Elements
D26	Lighting
D27	Tobacco and Smokers' Supplies
D28	Cosmetic Products and Toilet Articles
D29	Equipment for Safety, Protection and Rescue
D30	Animal Husbandry
D32	Washing, Cleaning or Drying Machines
D34	Material or Article Handling Equipment
D99	Miscellaneous

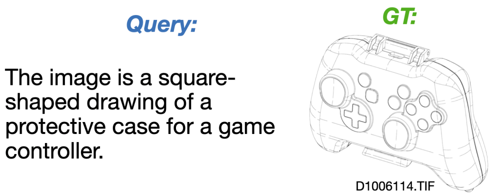
Table 8: Multimodal retrieval tasks in finetuned settings for image-text pairs used in the experiments. The best Recall@K (%) are shown in bold.

	Dataset	Backbone	**Text-Image**			**Image-Text**		
			R@1	R@5	R@10	R@1	R@5	R@10
Finetuned	Image-Title	ResNet50	5.44	14.98	21.54	5.18	14.46	20.28
		ResNet101	7.76	18.98	24.66	7.16	18.30	24.38
		ViT-B-32	9.16	22.38	29.20	8.42	22.18	28.86
		ViT-L-14	**14.88**	**33.04**	**41.42**	**13.75**	**31.98**	**39.96**

and prompting with patent title, we observe that our captions can provide more guidance for diffusion models.

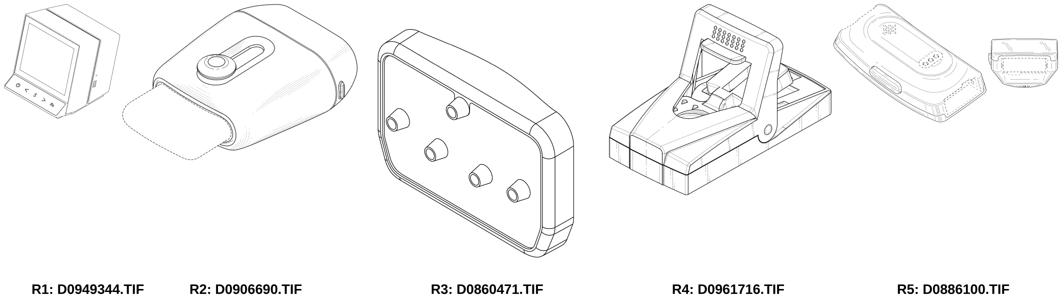
A.7 Visual Question Answering Examples

We provide two more detailed Visual question answering (VQA)examples. Based on IMPACT dataset, we design a set of questions which are relevant to design patents. Figures 14 and 15 are two examples with 3 questions and answers for each exampe patent images. We use LLaVa [35] to generate answers.

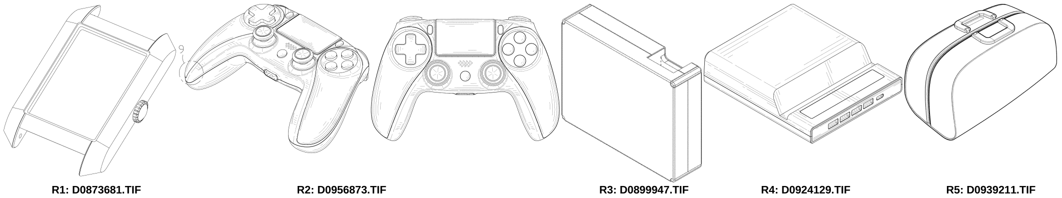


The image is a square-shaped drawing of a protective case for a game controller.

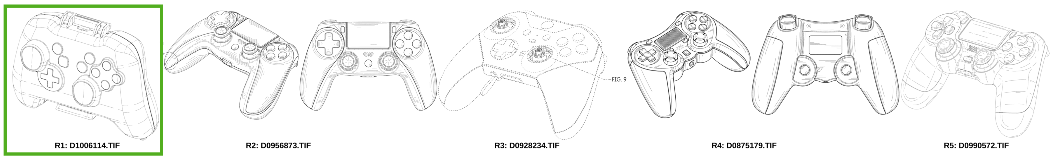
(a) Text query and ground truth image



(b) Retrieval results of CLIP



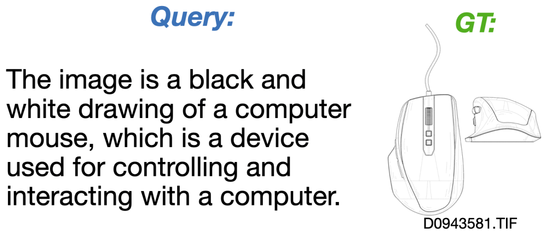
(c) Retrieval results of PatentCLIP-title



(d) Retrieval results of PatentCLIP

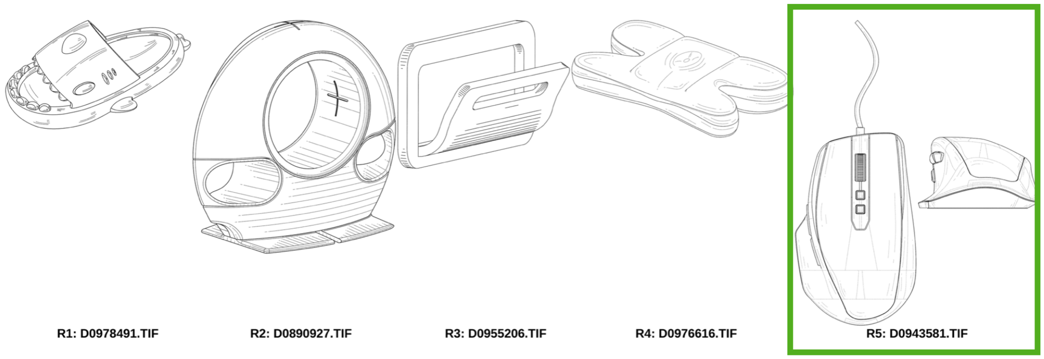
Figure 10: Text-image Retrieval example 1. Text query is shown in (a). (b), (c), and (d) are top 5 retrieval results of CLIP, PatentCLIP-title and PatentCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. In this case, only PatentCLIP retrieves correctly, which means PatentCLIP learned relevant multimodal features in the patent domain.

We can see that the answers are general text, but they are still helpful for further patent analyses. Thus, we propose a IMPACT-VQA to be a patent specific domain, which can provide more patent related information VQA systems and will be a good future direction to explore.

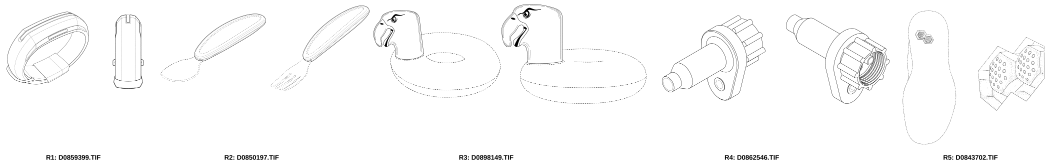


The image is a black and white drawing of a computer mouse, which is a device used for controlling and interacting with a computer.

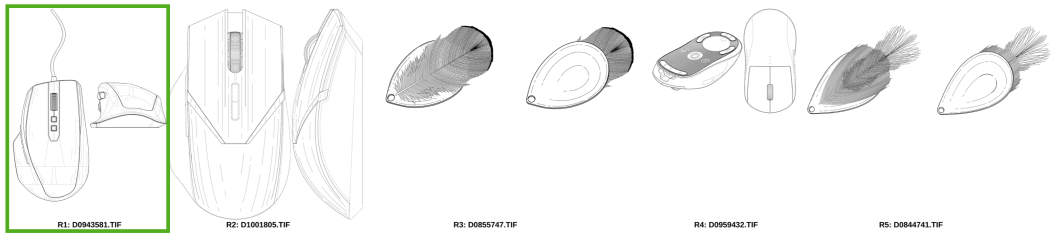
(a) Text query and ground truth image



(b) Retrieval results of CLIP

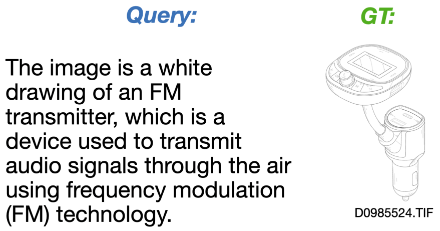


(c) Retrieval results of PatentCLIP-title



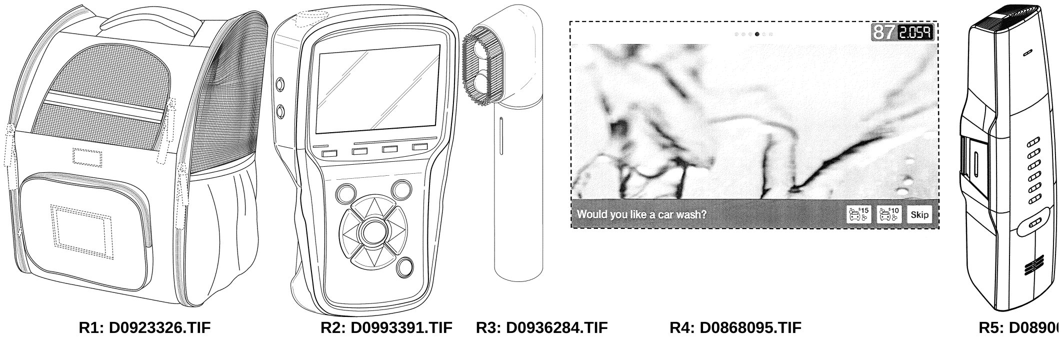
(d) Retrieval results of PatentCLIP

Figure 11: Text-image Retrieval example 2. Text query is shown in (a). (b), (c), and (d) are top 5 retrieval results of CLIP, PatentCLIP-title and PatentCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. In this case, only CLIP and PatentCLIP retrieves correctly.

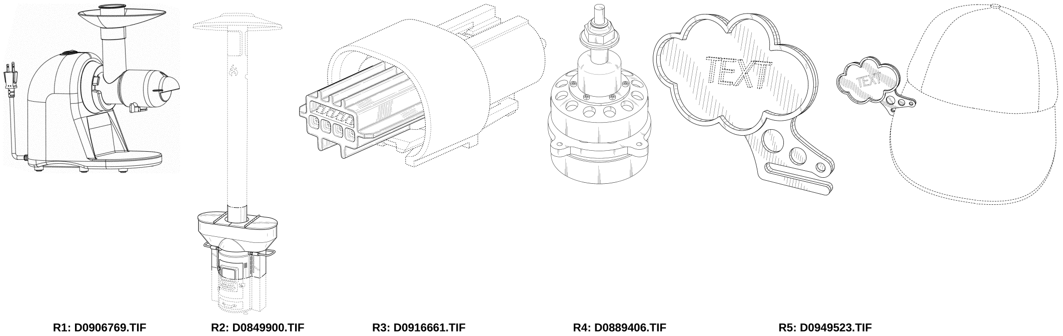


The image is a white drawing of an FM transmitter, which is a device used to transmit audio signals through the air using frequency modulation (FM) technology.

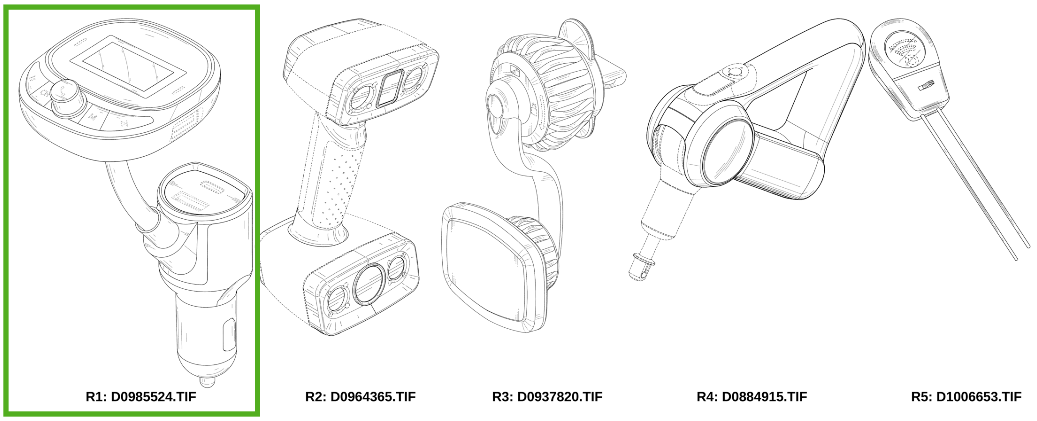
(a) Text query and ground truth image



(b) Retrieval results of CLIP

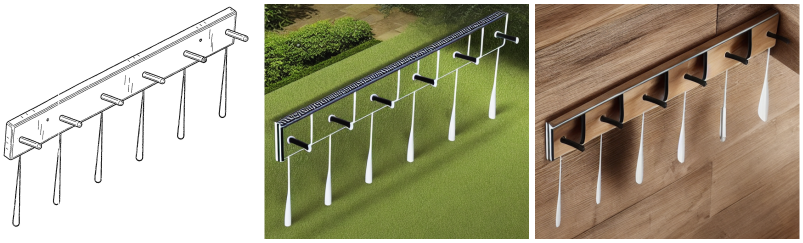


(c) Retrieval results of PatentCLIP-title

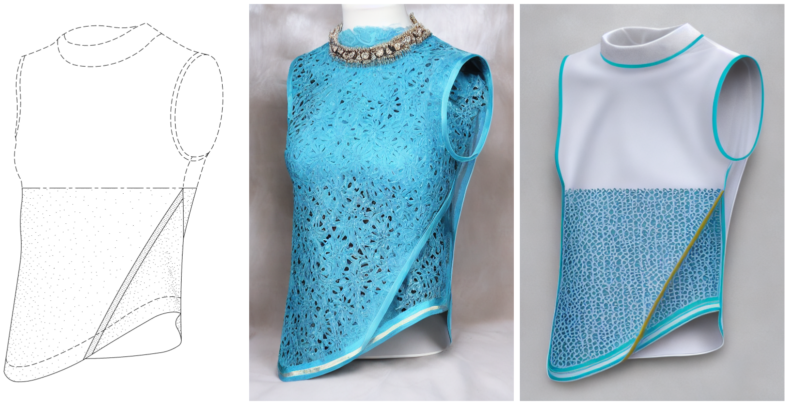


(d) Retrieval results of PatentCLIP

Figure 12: Text-image Retrieval example 3. Text query is shown in (a). (b), (c), and (d) are top 5 retrieval results of CLIP, PatentCLIP-title and PatentCLIP respectfully. Top 1-5 is from left to right. Green box denotes to the correct image. In this case, only PatentCLIP retrieves correctly.

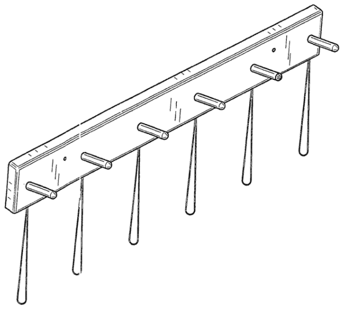


(a) Image (b) Title (c) Caption



(d) Image (e) Title (f) Caption

Figure 13: Examples of 3D reconstruction. (a) and (d) are patent images. (b) and (e) are generated images with title as text prompt. (c) and (f) are generated images with IMPACT captions.



Caption: The image is a rectangle, and it features a wooden utility rack with several hooks. The functionality of the utility rack is to provide a convenient and organized storage solution for various items, such as tools, utensils, or small equipment. The hooks allow users to hang items on the rack, keeping them off the countertops and keeping the workspace tidy and clutter-free.

Q1: what is the funtion of this item?

The item in the image is a wooden rack with multiple hooks, which is designed to hold and organize various objects, such as tools, utensils, or other items. The hooks provide a convenient and space-saving way to store and display these items, making it easy to access and keep the area tidy. The wooden rack can be used in various settings, such as a workshop, kitchen, or garage, to keep items organized and within reach.

Q2: what is the novel idea of this wooden rack with multiple hooks?

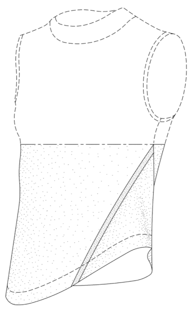
The novel idea of a wooden rack with multiple hooks is that it provides an efficient and space-saving solution for organizing and storing various items. The hooks allow for easy access to the stored items and help keep the area tidy and clutter-free. This type of rack is particularly useful in settings where space is limited, such as workshops, kitchens, or garages, where it is essential to keep tools, utensils, or other items organized and within reach. The wooden construction of the rack also adds a touch of aesthetic appeal and can blend well with the surrounding environment.

Q3: any new design suggestions for this wooden rack with multiple hooks?

Based on the image of the wooden rack with multiple hooks, here are some additional design suggestions:

1. Adjustable Hooks: The hooks on the rack could be adjustable, allowing users to customize the spacing and height of the hooks to accommodate different items and preferences.
2. Additional Storage Options: The rack could be designed with additional storage options, such as shelves or drawers, to accommodate a wider range of items and provide more flexibility in organizing the space.
3. Material Variations: The rack could be made from different materials, such as metal or plastic, to offer a more durable or weather-resistant option.
4. Customizable Design: The rack could be designed with interchangeable hooks or shelves, allowing users to customize the rack according to their specific needs and preferences. Integrated Storage: The rack could be integrated into a larger storage system, such as a wall-mounted cabinet or a freestanding unit, to create a more cohesive and functional storage solution. Multi-Purpose Design: The rack could be designed to serve multiple purposes, such as a pot rack for hanging pots and pans or a wine rack for storing wine bottles, depending on the user's needs.

Figure 14: Detailed VQA example 1, given an patent image from IMPACT, we design a few questions and use LLaVA to generate answers.



Caption:The image is a white outline of a garment, which appears to be a piece of clothing, such as a shirt or a dress. The garment has a pocket, and it is designed to provide comfort and functionality to the wearer. The pocket can be used to store small items, such as a phone or keys, while the overall design of the garment ensures that it fits well and is comfortable to wear.

Q1: what is the design of this item?

The image you've provided appears to be a technical drawing or pattern for a garment, specifically a piece of clothing that could be a top or a dress. The design features a simple silhouette with a round neckline and short sleeves. There is a horizontal seam across the chest area, which could be a design detail or a functional seam for construction purposes. The pattern includes dart details, which are used to shape the garment and give it a fitted appearance. The dotted lines likely represent the stitching lines, and the solid lines are the cut edges of the fabric.

Q2: what is the function of this cloth?

The function of this cloth is to serve as a garment, specifically a top or a dress, designed to be worn by a person. The design suggests that it is intended to provide coverage and possibly a fitted or tailored look, given the presence of darts and seams. The garment could be made from a variety of materials, such as cotton, polyester, or a blend, depending on the intended use and the desired level of comfort and durability. The exact function would depend on the specific material and the wearer's preferences and needs.

Q3: is there any similar design of this item?

Yes, there are many similar designs of this item. The design you've shown is quite basic and versatile, which means it can be easily adapted to create a wide range of garments. For example, the neckline could be changed to a V-neck, the sleeves could be lengthened or shortened, the hem could be adjusted to be longer or shorter, and the fabric could be swapped out for different textures or weights.

The simplicity of the design also means that it can be easily combined with other patterns or designs to create more complex garments. For instance, it could be used as a base for a dress with a skirt pattern attached, or it could be layered under a jacket or cardigan.

In summary, the design you've shown is quite flexible and can be adapted in numerous ways to create a variety of garments, from casual tops to more formal dresses.

Figure 15: Detailed VQA example 1, given an patent image from IMPACT, we design a few questions and use LLaVA to generate answers.